Volume 23, No. 3, September 2016

The Journal of Economic Methodology

Journal of Economic Methodology

ISSN: 1350-178X (Print) 1469-9427 (Online) Journal homepage: http://www.tandfonline.com/loi/rjec20

Suppes's outlines of an empirical measurement theory

Marcel Boumans

To cite this article: Marcel Boumans (2016) Suppes's outlines of an empirical measurement theory, Journal of Economic Methodology, 23:3, 305-315, DOI: 10.1080/1350178X.2016.1189124

To link to this article: http://dx.doi.org/10.1080/1350178X.2016.1189124

Published online: 09 Jun 2016.

Journal of Economic Methodology, 2016
Vol. 23, No. 3, 305–315, http://dx.doi.org/10.1080/1350178X.2016.1189124

Suppes's outlines of an empirical measurement theory

Marcel Boumans*

School of Economics, Utrecht University, Utrecht, The Netherlands

(*Received 2 October 2015; accepted 21 December 2015*)

According to Suppes, measurement theory, like any scientific theory, should consist of two parts, a set-theoretical defined structure and the empirical interpretation of that structure. An empirical interpretation means the specification – 'coordinating definitions' – of a 'hierarchy of models' between the theory and the experimental results. But in the case of measurement theory, he defined the relationship between numerical structure and the empirical structure specifically in terms of homomorphism. This is rather a highly restrictive relation between models, and therefore he never succeeded in giving his measurement theory empirical content. This paper discusses what an empirical measurement theory will look like if we would use less restrictive 'coordinating definitions' to specify the relationships between the various models.

Keywords: measurement theory; model of data; correspondence rule; Patrick Suppes

Introduction

One of Patrick Suppes's most influential contributions is his theory of measurement, which later evolved into what is now the most dominant theory of measurement: the representational theory of measurement. This latter theory was canonized in the three-volume survey *Foundations of Measurement*, edited by Krantz, Luce, Suppes, and Tversky (1971/1989/1990). These volumes present measurement theory as a highly formalistic axiomatic theory. The first version of this axiomatic theory of measurement is the 'Basic Measurement Theory' by Suppes and Zinnes (1963).

This highly abstract theory of measurement is criticized for its abstractness, that is, its lack of giving an account of actual practical measurement. It does not account for measurement procedures, devices, and methods; and it applies only to error-free data, in the sense that it says nothing about handling the response variability in real data. A recent discussion of the representational theory of measurement aptly characterizes it as 'a library of mathematical theorems […] useful for investigating problems of concept formation' (Heilmann, 2015, p. 788).

Suppes's contributions to the development of this abstract account of measurement is in sharp contrast with his accounts of experiments in psychology, which have a much more practical focus; they are about procedures, devices, methods, errors, and variability. It is as if Suppes was never able to connect these two kinds of studies. This paper is a proposal of how this potential connection could be conceived, based on expanding Suppes's account of experimental practices into the practices of measurement.

According to Suppes (2002), a scientific theory should consist of two parts, a set-theoretical defined structure and the empirical interpretation of that structure.[1] The reason

*Email: m.j.boumans@uu.nl

for having set-theoretical structures play a central role is that 'such structures provided the right settings for investigating problems of representation and invariance in any systematic part of past or present science' (p. xiii). Suppes argues that these 'right settings' cannot be provided by any of the logical languages,

> the artificial-language treatment of problems of evidence are inadequate. [It gives] a much too simplified account of the extraordinarily complex and technically involved practical problems of assessing evidence in the empirical sciences. (Suppes, 2002, p. 2)

In addition to his interest in theory structure, Suppes's also emphasized empirical details: particularly those of the experiments in psychology, some of which he had conducted himself. He hoped to extend his work on set-theoretical structures to include an account of set-theoretical representations of data, 'as a necessary, but still desirable, abstraction of the complicated activity of conducting experiments' (p. xiv), but he never managed to complete the project . In his 2002 book, however, section 8.6 'exemplifies' what he had in mind.

This penultimate section of his 2002 book (the very last section is an epilog) discusses Suppes's most recent experimental work on brain-wave representations of words and sentences. He suggests that most of the activities related to these kinds of experiments are actually attempts to 'clean up' the data:

> As in many areas of science, so with EEG recordings, statistical and experimental methods for removing artifacts and other anomalies in data constitute a large subject with a complicated literature. […] I am happy to end with this one example of a typical method of 'cleaning up' data.[2] (Suppes, 2002, p. 465)

In addition to favoring a set-theoretical account of structure over the syntactic account of theories, he also favored an approach in which the correspondence rules between theory and data were defined in terms of models instead of empirical interpretations of the syntactical terms. Correspondence should be defined – 'coordinating definitions' – in terms of a 'hierarchy of models' between theory and experimental results. This hierarchy of models consists of various levels of models, with the top level being a model of the theory and the bottom level an empirical model. The reason for using a hierarchy of models instead of direct empirical interpretations of theories is that the correspondence between theory and data is 'much more complicated,' in part because 'the model of the experimental results is of a relatively different logical type from that of any model of the theory' (p. 7).

In case of measurement, the empirical model is:

> an abstraction from most of the empirical details of the actual empirical process of measurement. The function of the empirical model is to organize in a systematical way the *results* of the measurement procedures used. (Suppes, 2002, p. 4)

But as an experimental practitioner he knew that these abstractions do not account sufficiently for the various practices of experimentation, so he hoped to supplement these abstractions of the experimental results with abstractions of the procedures:

> It would be desirable also to develop models of the experimental *procedures*, not just the results. A really detailed move in this direction would necessarily use psychological and related psychological concepts to describe what experimental scientists actually do in their laboratories. This important foundational topic is not developed here and it has little systematic development in the literature of the philosophy of science. (Suppes, 2002, p. 7)[3]

In contrast with the acknowledgment of multiple levels of models between theory and data in experimental practice, in his theory of measurement Suppes distinguishes only two levels: a numerical and an empirical level. Moreover, the correspondence between the numerical structure and the empirical structure is defined solely in terms of a homomorphism. This narrow definition of correspondence is in contrast with the more liberal coordinating definitions determining the hierarchy of models for experiments. These experimental models could be different in nature for each correspondence between two consecutive levels in the hierarchy, and thus include other types of correspondence than homomorphism alone. This narrow definition of correspondence in measurement is probably the main reason that Suppes's theory of measurement never has become a theory that accounts for practice of measurement, and remained primarily a mathematical theory.

This paper will explore the framework Suppes set out for experimental practices: as he developed for the experimental results and also as he hoped that someone would develop for procedures, see last quotation. The aim of this paper is to see how Suppes's framework might be applied to measurement practice, with the goal of ending up with an empirical measurement theory. To do so, I will use my own measurement account that does not require a homomorphic relationship between the numerical and empirical level, because the model structures are modular instead of relational.[4]

Models of data

Before I discuss Suppes's hierarchy of models in more detail, it is useful to clarify what Suppes meant by the terms 'model' and 'theory.' Unfortunately his two textbooks *Introduction to Logic* (1957) and *Axiomatic Set Theory* (1960a) do not provide clear unique definitions of models and theories. The *Axiomatic Set Theory* does not discuss models and theories at all, and the *Introduction to Logic* gives three different definitions of a 'model for a theory': one used in logic, one in mathematical economics, and one in empirical science:

> Logic: 'when a theory is axiomatized by defining a set-theoretical predicate, by a model for the theory we mean simply an entity which satisfies the predicate' (1957, p. 253).
>
> Mathematical economics: 'the model for a theory is the set of all models for the theory in the logicians' sense. What the logicians call a model is labeled a structure' (p. 253).
>
> Empirical science: a model is 'an exact mathematical theory', and a theory is a set of 'non-mathematical, relatively inexact statements about the fundamental ideas of a given domain in science' (p. 254).

Suppes's habit of listing the definitions of models used in various disciplines rather than providing a single definition of a model was continued in his 1960 article on the 'Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Science.'[5] In that case the disciplines were mathematical logic, physics, economics, psychology, and mathematical statistics, but he also made his preferences more explicit. He considered Tarski's definition as 'a fundamental concept' in all the above disciplines: 'I would assert that the meaning of the concept of model is the same in mathematics and the empirical sciences' (1960b, p. 289). Tarski defined a model of a theory T as 'a possible realization in which all valid sentences of a theory T are satisfied' (Tarski quoted in Suppes, 1960b, p. 287). A theory is thus a linguistic entity consisting of a set of sentences and models are non-linguistic entities in which the theory is satisfied (1960b, p. 290).[6]

It is striking that in these early papers, models were exclusively defined in relation to a theory, evidenced by calling them 'model of a theory,' or 'model for a theory.' It is, however, in the same period that Suppes started to think about models in relation to data, which he called 'models of data.'

Suppes's account of hierarchy of models was introduced for the first time in his 1960 article on the meaning and uses of models. His reason for introducing this idea of hierarchy of models was the 'radical' difference between the 'logical type' of models used in theory and those used in experiment: 'The maddeningly diverse and complex experience which constitutes an experiment is not the entity which is directly compared with a model of a theory' (p. 297). To make a comparison between theory and experiment possible 'drastic assumptions of all sorts are made in reducing the experimental experience […] to a simple entity ready for comparison' (p. 297). A plurality of models between these two levels could reduce the need for drastic assumptions.

A more detailed discussion of the hierarchy of models appeared in his 'Models of Data' (1962). He argued that this paper was written to overcome the 'sins of philosophers of science […] to overly simplify the structure of science' (p. 260) by representing scientific theories as logical calculi and then to 'go on to say that a theory is given empirical meaning by providing interpretations or coordinating definitions for some of the primitive or defined terms of the calculus' (p. 260). Instead of this overly simplistic view of how theories are related to data, Suppes argued that 'a whole hierarchy of models stands between the model of the basic theory and the complete experimental experience' (p. 260). A model at one level is given empirical meaning by a specifically defined connection with the model at a lower level. Because the models at each level are of a different 'logical type,' the connections between them will be also of different types.

According to Suppes a systematic account of these connections should be formal, which for him meant set-theoretical. He did not make clear why he took this position ('a general defense of this conclusion cannot be made here', p. 261), but it seems to contradict his more liberal principle acknowledging the difference between logical types of models, and, as will be shown below, it also prevented him from providing an account that would connect all the various levels down the hierarchy.

The lowest level of the hierarchy, however, could not be modeled. This lowest level is that pertaining to 'noises, lighting, odors, phases of the moon,' see Table 1, 'here is placed every intuitive consideration of experimental design that involves no formal statistics' (p. 258). In contrast to this lowest level, the level just above, the level of experimental design, can be formalized, which makes the relationship between the level of experimental design and the level above it (models of data), explicit. This was considered to be impossible for the lowest level because of 'the seemingly endless number

Table 1. Hierarchy of theories, models, and problems.

Theory of	Typical problems
Linear response models	Estimation of θ, goodness of fit to models of data
Models of experiment	Number of trials, choice of experimental parameters
Models of data	Homogeneity, stationarity, fit of experimental parameters
Experimental design	Left–right randomization, assignment of subjects
Ceteris paribus conditions	Noises, lighting, odors, phases of the moon

Source: Suppes (1962, p. 259).

of unstated *ceteris paribus* conditions' (p. 259). In other words, this lowest layer of dealing with the *ceteris paribus* conditions cannot be covered by any model because of the infinite number of conditions one has to account for. Therefore it cannot be connected to the level of experimental design above it.

The level of *ceteris paribus* conditions aims at reducing sources of errors: to mute loud noises, to fresh the air from bad odors, or to reorganize the schedule for observations. These attempts to reduce sources of errors are what I would like to call cleaning activities, since they reduce errors or even remove the sources of them. According to Suppes, these activities, unlike experimental design, cannot be accounted for by any model or theory, and hence cannot be connected to the higher levels.

Notwithstanding that this hierarchy of models account was too restricted to allow for a systematic account of the basic – often most time-consuming – research activities, such as cleaning the environment before any experiment can be run, it is more liberal than the standard view on the relationship between theories and data. In the first place it allows for other types of correspondence rules; and secondly, ignoring the problems of correspondence, the level of the cleaning activities was explicit in the hierarchy. This was exceptional at the time and still is today.

Theory of measurement

During the period when Suppes was developing his account of models of data, he, together with Scott (Scott & Suppes, 1958), was also working on a theory of measurement. Actually his model account is closely related to his theory of measurement, because both were based on Alfred Tarski's theory of models: 'The main point of the present paper is to show how foundational analyses of measurement may be grounded in the general theory of models' (p. 113). The core idea of such a theory was 'to lay bare the structure of a collection of empirical relations which may be used to measure the characteristic of empirical phenomena corresponding to the concept' (p. 113), and therefore the main goal was 'to construct relations which have an exact and reasonable numerical interpretation and yet also have a technically practical empirical interpretation' (p. 113).

To put it in set-theoretical terms, one has to define two relational systems $\mathrm{A} = < A, R_1, \ldots, R_n>$, and $\mathrm{B} = < B, S_1, \ldots, S_n>$, where A is a non-empty set of qualitative empirical data, $R_1, \ldots, R_n$ are relations on A, B is the set of all real numbers, and $S_1, \ldots, S_n$ are numerical relations such that B is a homomorphic image of A. B is a homomorphic image of A if there is a function f from A onto B such that, for each $i = 1, \ldots, n$ and for each sequence $< a_1, \ldots, a_{m_i} >$ of elements of A, $R_i(a_1, \ldots ., a_{m_i})$ if and only if $S_i(f(a_1), \ldots ., f(a_{m_i}))$. In other words, a 'reasonable numerical interpretation' of an empirical relational system is a numerical relational system that is homomorphic to this empirical relational system (see also Suppes & Zinnes, 1963, pp. 5, 6).

Forty years later, Suppes (1998) published this theory of measurement in a more transparent but fairly condensed way as an entry in the *Encyclopedia of Philosophy.* Of interest here is that this later account has an additional section on 'Variability, Thresholds and Errors,' which examines the kind of problems that one encounters in the empirical practice of measurement.

Variability in the quantity measured, as Suppes explains, can have different sources. One source can be variability in the empirical properties of the object being measured. The height of a person for example varies on a diurnal basis. Another source of variability lies in the procedures of measurement being used, and this kind of variability is usually attributed to measurement error. Suppes distinguished various kinds of errors: instrumental errors due to imperfections of the measuring instrument, personal errors due to the response characteristics of the observer, systematic errors due to circumstances ‘that are themselves subject to observation and measurement’ (p. 248), random errors due to variability in the conditions surrounding the observations, and computational errors.

Although Suppes explicitly mentions these sources of variability, he also admits that ‘it is not possible here to examine in detail how the foundational investigations of measurement procedures have been able to deal with such problems of errors’ (1998, p. 248). However, he gave no such account elsewhere.

A model of *ceteris paribus* conditions

The requirement for connecting the lowest level of *ceteris paribus* conditions to the level above it is the existence of a model representing this level. At this level ‘is placed every intuitive consideration of experimental design that involves no formal statistics. Control of loud noises, bad odors, wrong times of day or season go here’ (Suppes, 1962, p. 258). The level of *ceteris paribus* conditions is the level of controlling variability, that is, of reducing sources of errors.

Although this level is not accounted for within Suppes’s theory of measurement, nor in the more general representational theory of measurement (Krantz et al., 1971, 1989, 1990), it is accounted for by the current metrological theory of measurement. This metrological measurement theory is mainly a practice-based account.[7]

The target of modeling the measurement process in metrology is the measurement function f: In most cases, a measurand Y is not measured directly, but is determined from N other quantities $X_1, X_2, \ldots, X_N$ through a functional relationship f: $Y = f(X_1, X_2, \ldots, X_N)$, where $X_1, X_2, \ldots, X_N$ are called the input quantities and Y the output quantity (JCGM 100 2008, p. 8). If data indicate that f does not model the measurement to the degree imposed by the required accuracy of the measurement result, additional input quantities must be included in f to reduce this inaccuracy (see JCGM 100 2008, p. 9).

The problem with this modeling strategy, however, is that accuracy of measurement does not provide a straightforward way to validate a measurement model. This is because accuracy is defined with respect to the true value of the measurand: ‘closeness of agreement between a measured quantity value and a true quantity value of a measurand’ (JCGM 200 2012; p. 21). But a true value would only be obtained by ‘a perfect measurement,’ which is ‘only an idealized concept’ (JCGM 100 2008, p. 50); therefore, ‘true values are by nature indeterminate’ (p. 32). This indeterminateness is because there are potentially an infinite number of conditions that can influence the measurand.[8]

> The first step in making a measurement is to specify the measurand – the quantity to be measured; the measurand cannot be specified by a value but only by a description of a quantity. However, in principle, a measurand cannot be *completely* described without an infinite amount of information. (JCGM 100 2008, p. 49)

Regarding this incomplete knowledge of the measurand, current metrology generally acknowledges that measurement should be expressed in terms of uncertainty:

> it is not possible to state how well the essentially unique true value of the measurand is known, but only how well it is believed to be known. Measurement uncertainty can therefore be described as a measure of how well one believes one knows the essentially unique

> true value of the measurand. This uncertainty reflects the incomplete knowledge of the measurand. (JCGM 104 2009, p. 3)

Thus, instead of evaluating measurement results in terms of errors, it is now preferred to assess them in terms of uncertainty.

This uncertainty approach has consequences for the way in which measurement models are built. Models should be built ‘to express what is learned about the measurand’ (JCGM 104 2009, p. 3). Uncertainty, defined as the ‘non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used’ (JCGM 100 2008, p. 25), reflects ‘the lack of knowledge of the value of the measurand’ (JCGM 100 2008, p. 5), and consists of several components, that is, sources of uncertainty.

In metrology the following sources of uncertainty are identified:

(a) incomplete definition of the measurand;
(b) imperfect realization of the definition of the measurand;
(c) the sample measured may not represent the defined measurand;
(d) inadequate knowledge of the effects of environmental conditions on the measurement or imperfect measurement of environmental conditions;
(e) personal bias in reading analogue instruments;
(f) finite instrument resolution or discrimination threshold;
(g) inexact values of measurement standards and reference materials;
(h) inexact values of constants and other parameters obtained from external sources and used in the data-reduction algorithm;
(i) approximation and assumptions incorporated in the measurement method and procedure;
(j) variations in repeated observations of the measurand under apparently identical conditions. (JCGM 100 2008, p. 6)

These sources are not necessarily independent, and an unrecognized causal factor will contribute to measurement error. It is also acknowledged that ‘blunders in recording or analyzing data can introduce a significant unknown error in the result of a measurement’ (JCGM 100 2008, p. 8), but such blunders are not supposed to be accounted for by the measurement model.

To evaluate uncertainty of measurement results, in metrology the recommendation is to use two different ways of evaluating uncertainty components, a Type A evaluation and a Type B evaluation:

> Type A evaluation is the ‘method of evaluation of uncertainty by the statistical analysis of series of observations’. (JCGM 100 2008, p. 3)

> Type B evaluation is the ‘method of evaluation of uncertainty by means other than the statistical analysis of series of observations’. (JCGM 100 2008, p. 3)

Type A evaluation can be objectively established as soon as a metric is chosen, since it is a quantitative concept. Type B evaluation, however, is not based on a series of observations. It is considered to be a ‘scientific judgement’ based on professional skill ‘that can be learned with practice’ (JCGM 100 2008, p. 12) depending on qualitative and subjective knowledge of the measurand and ‘experience with or general knowledge of the behavior and properties of relevant materials and instruments’ (p. 11).

This distinction between Type A and Type B evaluations implies two different stages of modeling, a Type A stage and a Type B stage. A Type A stage exploits the measurement conditions under which the observations are obtained: 'If all of the quantities on which the result of a measurement depends are varied, its uncertainty can be evaluated by statistical means' (JCGM 100 2008, p. 7). A Type B stage depends on 'skilled judgement' and external sources, such as quantities associated with calibrated measurement standards, certified reference materials, and reference data obtained from handbooks, which may be used as an additional pool of information about whether the model is complete. Combined, both stages lead to the following strategy of modeling:

> Because the mathematical model may be incomplete, all relevant quantities should be varied to the fullest practicable extent so that the evaluation of uncertainty can be based as much as possible on observed data. Whenever feasible, the use of empirical models of the measurement founded on long-term quantitative data, and the use of check standards and control charts that can indicate if a measurement is under statistical control, should be part of the effort to obtain reliable evaluations of uncertainty. The mathematical model should always be revised when the observed data, including the result of independent determination of the same measurand, demonstrate that the model is incomplete. (JCGM 100 2008, p. 7)

To arrive at a model of the 'ceteris paribus conditions,' both types of uncertainty evaluations have to be accounted for. Modeling Type A evaluations is no more problematic than any other kind of statistical modeling. The crucial problem is how to model the judgments based on 'other means.'[9]

The basic idea of modeling type B evaluations can be briefly summarized as follows: When modeling the measurement process, one should include every potential input quantity, X_i, suggested by theory, experience, and general knowledge, regardless of whether there are (enough) observations to assume its potential influence. Subsequently the validity of this encompassing model should be tested. The model may still be incomplete, but the tests will tell whether a significant input quantity is still missing or whether the input quantities not included in the model are negligible. To deal with input quantities that are not measurable or for which there are not enough observations for a Type A evaluation, the proposal is to use a gray-box modeling approach instead of a white-box modeling approach.

The relationship between white-, gray-, and black-box modeling is as follows. A white-box model is a set of causal-descriptive statements of how some aspect of a real system actually operates. Testing this kind of model involves taking each relationship individually and comparing it with observations of the real system. As will be shown below, a Type B evaluation does not require this kind of model. For Type B evaluations the model can be a less demanding gray-box model. A gray-box model is a modular designed model, where each of the modules are black boxes. Testing this kind of model does not require having observations for each individual relationship.

To clarify this distinction between white-box, gray-box, and black-box models and the different kinds of testing they require, Barlas's (1996) distinction between three stages of model validation is useful. These three stages are (1) direct structure tests, (2) structure-oriented behavior tests, and (3) behavior pattern tests. *Direct structure tests* assess the validity of the model structure, by direct comparison with knowledge about the real system structure. This involves taking each relationship individually and comparing it with available knowledge about the real system. *Structure-oriented behavior tests* assess the validity of the structure indirectly, by applying certain behavior tests on model-generated behavior patterns. These tests involve simulation, and can be applied to the entire model, as well as to isolated sub-models of it. ‘These are “strong” behavior tests that can help the modeler uncover potential structural flaws’ (Barlas, 1996, p. 191). *Behavior pattern tests* do not evaluate the validity of the model structure, either directly or indirectly, but measure how accurately the model can reproduce the major behavior patterns exhibited by the real system.

For white-box models all three stages are equally important, while for black-box models it is only the last stage of behavior pattern tests that matters. Barlas (1996) does not refer to gray-box models. Although Barlas emphasizes that structure-oriented behavior tests are designed to evaluate the validity of the model structure, his usage of the notion of structure with respect to these tests allows for a notion of structure that is not limited to realistic descriptions of real systems; it also includes other kinds of arrangements like modular organizations. Structure-oriented behavior tests are also adequate for the validation of modular-designed models and for these models the term structure refers to the way the modules are assembled.

A module is a self-contained component (to be treated as a black box) with a standard interface to other components within a system. I call these modular-designed models gray-box models and they should pass the structure-oriented behavior tests and the behavior pattern tests.

This concept of a gray-box model and the way it should be validated is useful for outlining how to account for the lowest level of Suppes’s hierarchy of models. The first step is to acknowledge that the model of the *ceteris paribus* conditions does not need to be a complete representation of the relational system of these conditions, that is, a white-box model. It is not required that the *ceteris paribus* model has to capture detailed statistical knowledge about the complete set of the input quantities and the relations between them. Notwithstanding these weaker requirements on knowledge of these conditions and available observations, strong validation test – structure-oriented behavior tests – exist that are able to identify and even to estimate the magnitude of the uncertainty of neglected, ignored, or unknown influence quantities. As a consequence, the model of the *ceteris paribus* conditions can be a validated gray-box model, which does not require that an infinite number of conditions be accounted for, and moreover, the involvement of ‘intuitive considerations that involve no formal statistics’ can nevertheless be validated by structure-oriented behavior tests.

Conclusions: connecting the bottom level of cleaning up activities

According to Suppes, a theory of empirical research practices, whether of measurement or of experiment, should adequately account for the complex and technically involved practical problems of assessing evidence. A major part of the activities involved in such research practices are attempts to clean up the data, that is, treatments of errors and their sources. Such a theory should focus more on procedures than on empirical results.

In connecting evidence with theory, Suppes preferred a set-theoretical interpretation of structure because this would allow for a richer account of the correspondence between theory and data than a logical calculus. This account proposes a hierarchy of models as a layered connection between theory and data. The great benefit of such a hierarchy of models is that it acknowledges that models on different levels can be of different logical types. Another consequence is that the correspondences between consecutive models can also be of different types, determined by the types of models that are connected.

While Suppes had developed this challenging framework for a theory of experimental practices, particularly in his account of hierarchies of models, his theory of measurement lacks these features and suffers from too drastic simplifications. One of these crucial simplifications is that he restricted the kind of correspondence between an empirical and a numerical structure exclusively to homomorphisms.

According to Suppes's theory of measurement, the key requirement of measurement is to find a homomorphism that maps the relations between the relevant features of the measurand into a numerical model. This model is a representation of the empirical relational structure. The implicit consequence of the homomorphism requirement is that for the measurement to be reliable, the model needs to be as complete as possible. Completeness means in this case that the model encompasses all possible influences that may affect the measurand. Because the *ceteris paribus* conditions cannot be covered completely by any white-box model (because of the potentially infinite number of conditions one has to account for), Suppes assumed that this level could not be captured by a model at all, and that only 'intuitive considerations' could play a role. The argument in this paper, however, suggests that with specific validation tests – structure-oriented behavior tests – combined with a specific model design – gray box – a model of *ceteris paribus* conditions is feasible. The consequence of this is that the measurement model does not have to be a homomorphism of the structural relations describing the measurand.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes

[1] Although this book is published nearly at the end of Suppes's life, it represents his rather invariant ideas as he had developed over many years: 'I began this book as a young man. Well, at least I think of under 40 as being young, certainly now. I finished it in my tenth year of retirement, at the age of 80' (p. xv).

[2] An electroencephalogram (EEG) is a test that records the electrical activity of a brain.

[3] These 'psychological and related psychological concepts' were not further explicated, but I assume that these are related to the subjective judgements that have to be made while setting up and running an experiment. As such they will be discussed in later sections.

[4] This account can be found in Chapter 2 of Boumans (2015).

[5] A habit which he actually also continued in his later work.

[6] In his later work, Suppes would not anymore attempt to give such a definition of a 'theory.' For example, in his 2002 book in a section titled 'What Is a Scientific Theory,' he answers this question by stating that 'scientific theories cannot be defined simply or directly in terms of other nonphysical, abstract objects' (p. 2).

[7] Metrology is a field within instrument and control engineering involved with measurement and is the shared view on measurement of the international metrological organizations. This shared view can be found in the publications of the Joint Committee for Guides in Metrology. These publications are used here to outline this metrological measurement theory.

[8] A quantity is very generally defined as 'property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and a reference' (JCGM 200 2012, p. 2). This definition of quantity is more general than the traditional definition of quantity where it is a property of an object.

[9] A more detailed outline of this kind of modeling have been presented in Boumans (2013, 2015). The next paragraphs of the section are based on excerpts from these two publications.

References

Barlas, Y. (1996). Formal aspects of model validity and validation in system dynamics. *System Dynamics Review, 12*, 183–210.

Boumans, M. (2013). Model-based Type B uncertainty evaluations of measurement towards more objective evaluation strategies. *Measurement, 46*, 3775–3777.

Boumans, M. (2015). *Science outside the laboratory. Measurement in field science and economics*. New York, NY: Oxford University Press.

Heilmann, C. (2015). A new interpretation of the representational theory of measurement. *Philosophy of Science, 82*, 787–797.

Joint Committee for Guides in Metrology 100 2008 (2008). *Evaluation of measurement data: Guide to the expression of uncertainty in measurement*. JCGM

JCGM 104 2009 (2009). *Evaluation of measurement data: An introduction to the 'guide to the expression of uncertainty in measurement' and related documents*. JCGM

JCGM 200 2012 (2012). *International vocabulary of metrology: Basic and general concepts and associated terms* (3rd ed.). JCGM.

Krantz, D. H., Luce, R.D., Suppes, P., & Tversky, A. (1971/1989/1990). *Foundations of measurement* (Vols. 3). New York, NY: Academic Press.

Scott, D., & Suppes, P. (1958). Foundational aspects of theories of measurement. *The Journal of Symbolic Logic, 23*, 113–128.

Suppes, P. (1957). *Introduction to logic*. New York, NY: Van Nostrand Reinhold Company.

Suppes, P. (1960a). *Axiomatic set theory*. Princeton, NJ: D. Van Nostrand Company.

Suppes, P. (1960b). A comparison of the meaning and uses of models in mathematics and the empirical sciences. *Synthese,* Proceedings of the Colloquium: The Concept and the Role of the Model in Mathematics and Natural and Social Sciences*,12*, 287–301.

Suppes, P. (1962). Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology, and philosophy of science: Proceedings of the 1960 international congress* (pp. 252–261). Stanford, CA: Stanford University Press.

Suppes, P. (1998). Theory of measurement. In E. Craig (Ed.), *Routledge encyclopedia of philosophy* (pp. 243–249). London: Routledge.

Suppes, P. (2002). *Representation and invariance of scientific structures*. Stanford, CA: CSLI Publications.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*, Vol. 1 (pp. 1–76). New York, NY: Wiley.