Modeling Strategies for Measuring Phenomena In- and Outside the Laboratory

Marcel Boumans, University of Amsterdam

Introduction

Measurement is the assignment of numerals to a property of a phenomenon – 'measurand' – according to a rule with the aim of generating reliable information about this phenomenon. The central measurement problem is the design of rules so that the information is a reliable as possible. To arrive at reliable numbers for a measurand, the rules have to meet specific requirements. The nature of these requirements depends on the nature of the measurand *and* on the circumstances in which the measurements will be made.

The dominant measurement theory of today is the Representational Theory of Measurement (RTM) (Krantz e.a. 1971, 1989, 1990).[1] The core of this theory is that measurement is a process of assigning numbers to attributes or characteristics of the empirical world in such a way that the relevant qualitative empirical relations among these attributes or characteristics are reflected in the numbers themselves as well as in important properties of the number system.

The origins of RTM can be traced in Maxwell's method of using formal analogies. A first glimpse of it appeared in Maxwell's article 'On Faraday's lines of force' (1965). In discussing his method of using analogies, the 'representational view' is made *en passant*: "Thus all the mathematical sciences are founded on relations between physical laws and laws of numbers, so that the aim of exact science is to reduce the problems of nature to the determination of quantities by operations with numbers" (p. 156). Helmholtz took up Maxwell's view and continued to think in this direction. Usually Helmholtz 1887 is taken as the starting point of the development of the representational theory. The development since Helmholtz's seminal paper is described by Michell (1993) and Savage and Ehrlich (1992).

In the formal representational theory, measurement is defined set-theoretically as:

> Given a set of empirical relations $\mathbf{R} = \{R_1, \ldots, R_n\}$ on a set of extra-mathematical entities **Y** and a set of numerical relations $\mathbf{P} = \{P_1, \ldots, P_n\}$ on the set of numbers **N** (in general a subset of the set of real numbers), a function ϕ from **Y** into **N** takes each R_i into P_i, $i = 1, \ldots, n$, provided that the elements $Y_1, Y_2, \ldots$ in **Y** stand in relation R_i if and only if the corresponding numbers $\phi(Y_1), \phi(Y_2), \ldots$ stand in relation P_i.

In other words, measurement is conceived of as establishing homomorphisms from empirical relational structures $\Psi = \langle \mathbf{Y}, \mathbf{R} \rangle$ into numerical relational structures $N = \langle \mathbf{N}, \mathbf{P} \rangle$. We say then that the ordered triple $\langle \Psi, N, \phi \rangle$ is a *scale*.

[1] See for an early account Suppes and Zinnes 1963.

A numerical relational structure representing an empirical relational structure is also called a model, therefore the RTM is sometimes called the Model Theory of Measurement.

This theory is developed in the context of experimental psychology, and the theory accounts properly only for measurements taken in laboratories where the environment is always neutralized (so no noise), and where information is not provided by an instrument (but often the outcome of e.g. throwing a dice, or turning a card). In general, the RTM therefore provides no account for measurement procedures, devices and methods, no account for errors and noise, and so fails to cover questions about the reliability of measurement outside the laboratory (Boumans 2005).

This paper will show that a modelling strategy that accounts for measurement outside the laboratory, where one cannot base measurements on a single simple law, will have to drop the requirement that the model is a homomorphic mapping of the empirical relational structure. The models used for measurement will be simulacra, that is, "having merely the form or appearance of a certain thing, without possessing its substance or proper qualities" (OED 1933). This definition is also used by Cartwright (1983) to denote what models are, stressing the 'anti-realist' aspect of models. She could have used the term 'simulation', but probably did not because it refers to the assumption of false appearances for the sake of deception. But today the term is employed without this connotation of deception: "the assumption of the appearance of something without having its reality" (Dawson 1962: 1-2). Or to put it briefly: The aim of the paper is to show that *outside the laboratory, models used for measurement aims are simulations*.

The reliability of measurement

To make comparisons between strategies in- and outside the laboratory, the scope of the strategies is strongly simplified to a common aim of finding a 'true' value of a system variable, denoted by x.[2] The reliability of a measurement result can be characterized by three features: 'invariance', 'accuracy' and 'precision'. 'Invariance' refers to the stability of the relationship between measurand, measuring system and environment. 'Accuracy' is defined as the closeness of the agreement between the result of a measurement and a true value of the measurand, and 'precision' is defined as the closeness of the agreement between quantity values obtained by replicate measurements of a quantity, under specified conditions.

The difference between invariance, accuracy and precision can be illustrated by an analogy of measurement with rifle shooting, where the bull's eye represents the (in practice unknown) true value x. A group of shots is precise when the shots lie close together. A group of shots is accurate when it has its mean in the bull's eye. When during the shooting the alignment of rifle and target remains stable, even in a turbulent environment, this is a matter of invariance.

To explore these three requirements and to show how different strategies deal with them, a more formal, though simplified, framework will be developed. It is assumed

[2] 'True value' is an idealized concept and is by nature indeterminate.

that x is not directly measurable. In general, the value of x is inferred from a set of available observations y_i (i = 1, …, n), which inevitably involve noise ε_i:

$$y_i = F(x) + \varepsilon_i. \quad (1)$$

This equation will be referred to as the observation equation. The term for this kind of measurement is 'indirect measurement'. We have 'derived measurement' if $y = F(x)$ is an empirical law.

To clarify the requirements of invariance, accuracy and precision when control is possible and when it is not, it is useful to rewrite Eq. (1) as a relationship between the observations y, the target variable x, and background conditions B:

$$y = f(x, B) = f(x, 0) + \varepsilon. \quad (2)$$

In other words, it is assumed here that noise is (only) caused by disturbing background influences. The observed quantity y can only provide information about the system variable, x, when this variable x does influence the behavior of y. In general, however, it will be the case that not only x will influence y, but that there will be also many other influences, B, too. To express more explicitly how x and other possible factors (B) influence the behavior of the observed quantities, the relationship is transformed into the following equation:

$$\Delta y = \Delta f(x, B) = f_x \Delta x + f_B \Delta B \quad (3)$$

where f_x and f_B are functions of x, y, and B that denote how much y will change proportionally due to changes in x and B, respectively.

To achieve reliable measurement results, the following problems have to be dealt with:

Inside the laboratory
Taking care that the observations are as informative as possible about the measurand, or in other words, are as accurate and precise as possible, we have to reduce the influences of the other background factors B. In a laboratory, where we can control the environment, this can be achieved by imposing *ceteris paribus* conditions: $\Delta B = 0$. By noise reduction, both requirements of accuracy and precision are met simultaneously:

$$\Delta y_{CP} = f_x \Delta x \quad (4)$$

f_x is the element of Eq. (3) that expresses the relation between the observed quantity y and the measurand x. This element should be, as much as possible, an invariant function – that is, it has to remain stable or unchanged for, and to be independent of, two kinds of changes: variations over a wide range of the system variable, Δx, and variations over a wide range of background conditions, ΔB (Woodward 2000).

In the laboratory, the stability and shape of f_x can be investigated because of the possibility of creating a *ceteris paribus* environment and the possibility of controlling the measurand:

$$f_x = \frac{\Delta y_{CP}}{\Delta x} \tag{5}$$

If the ratio of the variation of y_{CP} and the variation of x appears to be a stable function, the correlation $y = F(x)$ is an invariant relationship (a law) of which the shape can be inferred and subsequently used for the measurements of x. This relation $y = F(x)$ then has become the measurement equation, also called the measurement formula.

Outside the laboratory
Outside the laboratory, where observations are 'passive', the assessment of invariance is much more complicated. To discover invariant observation relations to be used for measurement, one has no *ceteris paribus* environments at one's disposal, or at least they are very rare. One has, instead, to look out for *ceteris neglectis* environments. These are environments were disturbing influences are negligible, that is, where $f_B \Delta B \approx 0$. If in these circumstances

$$f_x = \frac{\Delta y_{CN}}{\Delta x}, \tag{6}$$

the ratio between the variation of the passive observations y_{CN} and the variation of x appears to be a stable function, the observation relationship could be used for measurement purposes.

The problem, however, is that it is not possible to identify the reason for a disturbing background influence being negligible. We cannot distinguish, 'identify', whether its 'potential influence' is very small, that is when $f_B \approx 0$, or whether the factual variation of this factor in the set of observations under consideration is too small, $\Delta B \approx 0$. In the first case, it is justified to ignore this background influence, but this is not true for the latter case. The variation of B is determined by other relationships within the environment. In some cases, a virtually dormant factor may become active because of changes in the environment elsewhere. Each found empirical relationship is a representation of a specific data set. So, for each data set it is not clear whether potential influences are truly negligible or only dormant.

This problem, called the problem of passive observation (Haavelmo 1944), can be dealt with by the strategy of comprehensiveness and it works as follows (see Sutton 2000): when a relationship appears to be inaccurate, this is an indication that a non-negligible potential background factor has been omitted in the model. As long as the resulting relationship is inaccurate, potential relevant factors should be added to the model. The expectation is such that this strategy will result in the fulfillment of two requirements:

(1) the resulting model captures a complete list of factors that exert large and systematic influences;
(2) all remaining influences can be treated as a sufficiently small noise component.

The problem of passive observations is solved by accumulation of data sets: the expectation is that we converge bit by bit to a closer approximation of the complete model, as all the most important background factors reveal their influence. In other

words, the strategy aims at modeling not only the measurand but also by modeling its relevant environment as complete as possible.

As a result, outside the laboratory, where we cannot control the environment, accuracy and precision have to be dealt with by using models as 'virtual laboratories', representing *ceteris neglectis* measuring system, in which reliability is not materialized (as in a laboratory) but achieved by setting the parameters carefully (Morgan 2003). To measure *x*, a model, denoted by *M*, has to be specified, for which the observations y_i function as input and $\hat{x}$, the estimation of *x*, functions as output:

$$\hat{x} = M[y_i; \alpha] \tag{7}$$

where α denotes the parameters of the model. Substitution of the observation equation (1) into model *M* (Eq. 5) shows what should be modeled (assuming that *M* is a linear operator):

$$\hat{x} = M[F(x) + \varepsilon; \alpha] = M_x[x; \alpha] + M_B[\varepsilon; \alpha] \tag{8}$$

A necessary condition for $\hat{x}$ to be a measurement of *x* is that model *M* must be a representation of the observation equation (1), in the sense that it must specify how the observations are related to the measurand. Therefore we first need a representation of the measurand, M_x. This specification should be completed with a specification of the error term, that is, a representation of the environment of the measurand, M_B. As a result, outside the laboratory, accuracy and precision has to be dealt with in two different ways. To see this, we split the measurement error $\hat{\varepsilon}$ in two parts:

$$\hat{\varepsilon} = \hat{x} - x = M_B[\varepsilon; \alpha] + (M_x[x; \alpha] - x) \tag{9}$$

To explore how this measurement error is dealt with, it may be helpful to compare this with the 'mean-squared error' of an estimator as defined in statistics:

$$E[\hat{\varepsilon}^2] = E[(\hat{x} - x)^2] = Var\hat{x} + (E\hat{x} - x)^2 \tag{10}$$

The first term of the right-hand side of Eq. (10) is a measure of precision and the second term is called the bias of the estimator. If we expand Eq. (10) further, we have:

$$E[\hat{\varepsilon}^2] = Var(M_B[\varepsilon; \alpha]) + (M_x[x; \alpha] - x)^2 \tag{11}$$

Comparing expression (9) with expression (11), one can see that the error term $M_B[\varepsilon; \alpha]$ is reduced, as much as possible, by reducing the spread of errors, that is by aiming at precision. The second error term $(M_x[x; \alpha] - x)$ is reduced by finding an as accurate as possible representation of *x*.

To obtain a reliable measurement result with an immaterial mathematical model, the model's *parameters* have to be adjusted in such a way that both precision and accuracy are maximized. So, tuning, that is separating signal *x* and noise $\hat{\varepsilon}$, is done by adjusting the parameter values α. The parameters should be adjusted such that

simultaneously $M_B[\varepsilon; \alpha]$ and $(M_x[x; \alpha] - x)$ are reduced. In some fields this is called 'filtering', in actuary 'graduation'.

Modeling and tuning in this way, however, does not yet solve the problem of invariance. To see what this problem entails outside the laboratory, we will first have to go back to the laboratory. According to Cartwright (1999), a law is invariant because it is the product of a nomological machine. A nomological machine is "a fixed (enough) arrangement of components, or factors, with stable (enough) capacities that in the right sort of stable (enough) environment will, with repeated operation, give rise to the kind of regular behaviour that we represent in our scientific laws" (p. 50). So, this machine will only produce invariant relationships in a stable, that is, *ceteris paribus* environment (a laboratory).

This idea of nomological machine shows that a measuring instrument must function as a nomological machine to fulfil its task. Measurement with an instrument is derived measurement, which uses (at least) one invariant relation (a law) between the instrument's readings and the measurand. The instrument must be designed and constructed in such a way that the invariance of the measurement relation is guaranteed.

Calibration

As I have discussed above, nomological machines can also work successfully in a *ceteris neglectis* environment, so also outside the laboratory. A necessary requirement for a nomological machine to function as measuring instrument is that it should be stable (enough). For a *ceteris paribus* nomological machine the stability is guaranteed by its environment. In the case of a *ceteris negelectis* nomological machine the stability must be a feature of the machine itself. The problem, however, is to find *ceteris neglectis* nomological machines that are stable. These are the natural systems that can be used as a (natural) measurement system, on the condition that they are stable. The evaluation of whether a natural nomological machine is stable, that is the issue of invariance, can only be done at the level of the numerical representation (model) of this natural machine. The modelling strategy of comprehensiveness does not necessarily lead to the identification of a representation of an invariant machine. One cannot definitely decide whether one has such a machine even when one thinks to have found one. Any so far neglected potential factor can appear suddenly to be active and to influence the measurement's accuracy negatively.

In many field sciences, the solution to this problem is calibration. Calibration is the establishment of the relationship between values indicated by a measuring instrument and the corresponding values realized by standards. In the laboratory, a standard is an instrument or a constructed signal chosen as reference: under specific determined conditions it performs in a specific determinate way. Because one can control the conditions in a laboratory, calibration there is only a technical problem. Outside the laboratory, the idea of a standard is that it is often based upon naturally occurring phenomena when these possess the required degree of stability. A standard, in this context, is a representation of the stable properties of a phenomenon – stable facts. So, to apply the calibration strategy outside the laboratory, one needs stable facts as a reference. In natural science, obvious candidates are the universal constants, if available for the relevant phenomenon (the list of universal physical constants is

however small). In other fields, calibration is achieved by the involvement of (other) models to define or reveal standards or stable facts.

In experimental science, calibration is one of the epistemological strategies used to distinguish between a valid observation and an artefact created by the instrument (Franklin 1997). Franklin (1997, 31) defines calibration as "the use of a surrogate signal to standardize an instrument. If an apparatus reproduces known phenomena, then we legitimately strengthen our belief that the apparatus is working properly and that the experimental results reproduced with that apparatus are reliable". This kind of calibration is to establish the relationship between the values of quantities indicated by the instrument in one specific dimension and the corresponding standard values in the same dimension, to acquire reliability of the values indicated by the instrument in other dimensions. But one should be warned, this kind of calibration does not guarantee a correct result; though its successful performance does argue for the validity of the result.

Franklin defines calibration in relation to instruments. But his definition can also be applied to the calibration of models. In the above framework this entails the following steps. A 'surrogate' input signal y_0 is supposed to (re)produce output signal x_0. These input and output data are used to calibrate the model, that is, to set the parameters α such that:

$$M[y_0;\alpha] = x_0 \tag{12}$$

These parameters are denoted by α_0.

A result of this calibration strategy is that invariance has become an external feature of the model, instead of an internal feature that one or more of the empirical relations the model represents should be invariant. For a specific input it should always have a specific output referring to a stable fact about the measurand.

What we have seen above is that the assessment of models as measuring instruments outside the laboratory is not based on the evaluation of a homomorphic correspondence between the empirical relational structure and the numerical relational structure. The assessment of these models is more like what is called *validation* in systems engineering. Validity of a model is seen as 'usefulness with respect to some purpose'. Barlas (1996) notes that for an exploration of the notion validation it is crucial to make a distinction between white-box models and black-box models. In black-box models, what only matters is the output behavior of the model. The model is assessed to be valid if its output matches the 'real' output within some specified range of accuracy, without any questioning of the validity of the individual relationships that exists in the model. White-box models, on the contrary, are statements as to how real systems actually operate in some aspects. Generating an accurate output behavior is not sufficient for model validity; the validity of the internal structure of the model is crucial too. A white-box model must not only reproduce the behavior of a real system, but also explain how the behavior is generated.

Barlas (1996) discusses three stages of model validation: direct structural tests, structure-oriented behavior tests and behavior pattern tests. For white models, all three stages are equally important, for black box models only the last stage matters.

Direct structure tests assess the validity of the model structure, by direct comparisons with knowledge about the real system structure. This involves taking each relation individually and comparing it with available knowledge about the real system. Barlas emphasizes that for these kinds of tests no simulation is involved. The second category, the structure-oriented behavior tests, assesses the validity of the structure indirectly, by applying certain behavior tests on model-generated behavior patterns. These tests involve simulations, and can be applied to the entire model, as well as to isolated sub-models of it. Barlas emphasizes the special importance of structure-oriented behavior tests: these are strong behavior tests that can provide information on potential structure flaws. The information, however, provided by these tests does not give any direct access to the structure, in contrast to the direct structure tests.

The most interesting structure-oriented behavior test that Barlas lists is the Turing test. This test was originally described by Turing (1950) as an 'imitation game' to investigate the question "Can machines think?" Today, a Turing test is generally described as follows: Reports based on output of the quantitative model and on measurements of the real system are presented to a team of experts. When they are not able to distinguish between the model output and the system output, the model is said to be valid.

The enormous advantage of Turing's approach to artificial intelligence is that it freed scientists from building replicas of the human mind to achieve machine thinking that meets the standard of human intelligence. In the same way, this kind of testing frees field scientists to build detailed, quantitatively accurate replicas of the actual nomological machine. Turing testing legitimizes to work with simpler models on the condition that it provides equally good answers as the more comprehensive models.

The interesting feature of a Turing test is that validates a model along the same dimensions as the models has been calibrated, or in other words, it tests a model on whether it is calibrated properly. To have confidence that a computer is intelligent, it should give known answers to familiar questions. This induces trust that the machine will also give proper answers to question for which we do not have yet an answer and for which we actually had built the machine. Likewise, the measurements of the real target system that are used by the experts to assess a model are usually not the large sets of available data on this system, but a smaller set of stable facts well-known to the experts.

Gray-box models

Though Barlas emphasizes that structure-oriented behavior tests are designed to evaluate the validity of the model structure, his usage of the notion of structure for this category of tests needs some further qualification. The way in which he describes and discusses these tests shows that his notion of structure is not limited to accurate descriptions of the individual relations of the target systems; it also includes other kinds of arrangements, namely the assemblages of subsystems.

To trust the results of a simulation for measurement purposes, the models that are run should pass the Turing test but need not to be accurate representations of the relevant economic systems. To picture the architecture of these models passing the structure-oriented behavior tests and behavior pattern tests, let us first label them as gray-box models – in line with the labeling of the other two types of models.

In addition to the fact that structure-oriented behavior test legitimize simpler models of the nomological machines than the white-box models, they also provide heuristics of how to simplify these too complex models, namely by partitioning.

Let us therefore recapitulate why calibration is an appropriate strategy for determining the model's parameter set α. Therefore, I will use Woodward's (1989) distinction between data and phenomena. According to Woodward, phenomena are relatively stable and general features of the world and therefore suited as objects of explanation. Data, that is, the observations playing the role of evidence for claims about phenomena, on the other hand, involve observational mistakes, are idiosyncratic, and reflect the operation of many different causal factors (see Eqs. (1)-(3)). Phenomena are more 'widespread' and less idiosyncratic, less closely tied to the details of a particular nomological machine. Because of the idiosyncrasy of the observations they are not appropriate to determine invariance. Facts about phenomena, however, do have the required non-idiosyncratic stability, and are therefore particularly apt to calibrate a measuring instrument outside the laboratory. Calibration does not take all observations into account but only picks out those stable facts the instrument should reproduce.

Models are not calibrated on the level of the structural relations, in particular when these models are too comprehensive and detailed. For these kinds models it is not clear how the reproduction of stable facts relates back to individual structural relations, it may even be not possible to do so. In other words, in complex models it may not be possible to relate specific output characteristics to specific individual relations. Output characteristics are more the result of the *interplay* of these relations. It therefore more appropriate to locate the *sub*model which is responsible for specific behavior characteristics than to try to find an individual relation.

The basis of this modeling strategy to deal with complexity is von Neumann's (1963) General and Logical Theory of Automata. According to von Neumann, the problem of complexity consists of two parts. The first part is partitioning into elements:

> The natural systems are of enormous complexity, and it is clearly necessary to subdivide the problem that they represent into several parts. One method of subdivision, which is particularly significant in the present context, is this: The [natural systems] can be viewed as made up of parts which to a certain extent are independent, elementary units. (von Neumann 1963, 289).

The second part consists of understanding how these elements are organized into a whole, and how the functioning of the whole is expressed in terms of these elements. The first part of the problem could be removed by the "process of axiomatization":

> *The Axiomatic Procedure*. Axiomatizing the behavior of the elements means this; we assume that the elements have certain well-defined, outside,

> functional characteristics; that is, they are to be treated as ‘black boxes”. They are viewed as automatisms, the inner structure of which need not be disclosed, but which are assumed to react to certain unambiguously defined stimuli, by certain unambiguously defined responses. (von Neumann 1963, 289)

The general approach is of partitioning into elementary units, which can be treated as black boxes that are calibrated. It is remarkable that these elements are labeled as ‘axioms’, indicating that they have some fundamental status.

In current systems engineering, this calibrated black box is called a module: a self-contained component with a standard interface to other components within a system (White 1999, 475). The great advantage of modular design is that it simplifies final assembly because there are fewer modules than subcomponents and because standard interfaces typically are designed for ease of fit (“plug and play”). Each module can be tested prior to assembly. Different measuring systems can be realized by different combinations of standard components. While each module has to be calibrated individually, the overall test for an assemblage of these modules is a kind of Turing test.

As a result, a gray-box model is a specific assemblage of calibrated black-box models such that it functions as an instrument to measure a specific property of a phenomenon, and which is validated by a kind of Turing test.

Conclusions

The Representational Theory of Measurement conceives measurement as establishing homomorphisms from empirical relational structures into numerical relation structures, called models. This theory is rooted in the work of James Clark Maxwell, in particular his ideas about the use of analogies; “that partial similarity between the laws of one science and those of another which makes each of them illustrate the other” (Maxwell 1965: 156). In other words, to the extent that two physical systems obey laws with the same mathematical form, the behaviour of one system can be understood by studying the behaviour of the other, better known. Moreover, this can be done without formulating any hypothesis about the real nature of the system under investigation.

Heinrich Hertz recognized the value of the concept of formal analogy in trying to understand the essential features of the natural world. For Hertz, representations of phenomena, models, could only be understood in the sense of Maxwell’s analogies. “In order to determine beforehand the course of the natural motion of a material system, it is sufficient to have a model of that system. The model may be much simpler than the system whose motion it represents” (p. 176), but the model was only to be considered as a representation of a system under investigation if the consequences of (that is, the inferences from) a representation of that system are the representation of the consequences of that system. This requirement, however, would allow for many different models meeting this requirement. Hertz, therefore, formulated three additional requirements: First, a representation should be ‘(logically) permissible’, that is, it should not contradict the principles of logic. Second, permissible representations should be ‘correct’, that is, the relations of the representation must not contradict the system relations. Third, of two correct and permissible representations of the same system, one should choose the most

'appropriate'. A representation is more appropriate when it is more distinct, that, when it contains more of the essential relations of the system; and when it is simpler, that is, when it contains a smaller number of superfluous or empty relations. Hertz explicitly noted that empty relations cannot be altogether avoided: "They enter into the images because they are simply images, - images produced by our mind and necessarily affected by the characteristics of its mode of portrayal" (Hertz 1956: 2).

In short, the three requirements that a representation of a system should fulfill are: (1) logical consistency; (2) 'correctness', that there is correspondence between the relations of the representation and those of the system; (3) 'appropriateness', that it contains the essential characteristics of the system as simply as possible. Hertz considered the last requirement as most problematic to meet:

> We cannot decide without ambiguity whether an image is appropriate or not; as to this differences of opinion may arise. One image may be more suitable for one purpose, another for another; only be gradually testing many images can we finally succeed in obtaining the most appropriate. (Hertz 1956: 3)

These requirements can be used to compare and characterize measurement strategies in- and outside the laboratory. Models used for measurement purposes inside the laboratory satisfy the correctness requirement and so are white-box models. To achieve accurate measurements outside the laboratory one needs to take account of the environment. This has two problematic consequences: a white-box modeling strategy, reflecting the complexity of the environment due to its correctness requirement, will readily lead to immensely large models. These models are representations of nomological machines outside the laboratory. But outside the laboratory there is no guarantee that the machine remains stable, which means that representations of it may not be accurate anymore for a new set of observations, with the subsequent consequence that its measurement results then would not be accurate. To arrange invariance, the models should be calibrated, that is, bringing them into accordance with stable facts about the measurand. A model strategy that allows for simplification such that it still remains accurate is gray-box modeling. This kind of model satisfies the appropriateness requirement, but drops the correctness requirement.

Models for measurements outside the laboratory are not homomorphic mappings, but simulacra validated by a kind of Turing test.

References

Barlas, Y. (1996) Formal aspects of model validity and validation in system dynamics, *System Dynamics Review* 12.3: 183-210.

Boumans, M. (2005). Measurement outside the laboratory, *Philosophy of Science* 72, 850-863.

Cartwright, N. (1983) *How the Laws of Physics Lie*. Oxford: Clarendon Press.

Cartwright, N. (1999) *The Dappled World. A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

Dawson, R.E. (1962) Simulation in the social science, in *Simulation in Social Science: Readings*, ed. H. Guetzkow. Englewood Cliffs, NJ: Prentice Hall.

Franklin, A. (1997) Calibration, *Perspectives on Science* 5: 31-80.

Haavelmo, T. (1944) The probability approach in econometrics, supplement to *Econometrica* 12.
Helmholtz, H. von (1887). Zählen und Messen, erkenntnistheoretisch betrachtet, in *Philosophische Aufsätze Eduard Zeller gewidmet*. Leipzig: Fuess.
Hertz, H. (1956) *The Principles of Mechanics Presented in a New Form*. New York: Dover.
Krantz, D.H., R.D. Luce, P. Suppes and A. Tversky (1971, 1989, 1990) *Foundations of Measurement*. 3 Vols. New York: Academic Press.
Maxwell, J.C. (1965) On Faraday's lines of force, in *The Scientific Papers of James Clerk Maxwell*, ed. W.D. Niven, Vol. I, 155-229. New York: Dover.
Michell, J. (1993) The origins of the representational theory of measurement: Helmholtz, Hölder, and Russell, *Studies in History and Philosophy of Science* 24.2, 185-206.
Morgan, M.S. (2003) 'Experiments without material intervention: Model experiments, virtual experiments, and virtually experiments', in *The Philosophy of Scientific Experimentation*, ed. H. Radder, 216-235. Pittsburgh: University of Pittsburgh Press.
OED (1933) *Oxford English Dictionary*. Oxford: Clarendon Press.
Savage, C.W. and P. Ehrlich (1992) 'A brief introduction to measurement theory and to the essays', in *Philosophical and Foundational Issues in Measurement Theory*, eds. C.W. Savage and P. Ehrlich, 1-14. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Suppes, P. and J.L. Zinnes (1963) 'Basic measurement theory', in *Handbook of Mathematical Psychology*, eds. R.D. Luce, R.R. Bush, and E. Galanter, 1-76. New York, London and Sydney: Wiley.
Sutton, J. (2000) *Marshall's Tendencies. What Can Economists Know?* Cambridge: MIT.
Turing, A.M. (1950) Computing machinery and intelligence, *Mind* 59: 433-460.
von Neumann, J. (1963) The general and logical theory of automata, in *John von Neumann. Collected Works*, Vol. 5., ed. A.H. Taub. Oxford: Pergamon Press.
White, K.P. (1999) System design, in *Handbook of Systems Engineering and Management*, eds. A.P. Sage and W.B. Rouse, 455-481. New York: Wiley.
Woodward, J. (1989) Data and phenomena, *Synthese* 79: 393-472.
Woodward, J. (2000) Explanation and invariance in the special sciences, *The British Journal for the Philosophy of Science* 51: 197-254.