For Comment. Do Note Cite or Quote Without Permission

A Theory of Measurement

Nancy L. Cartwright[1] and Norman M. Bradburn

London School of Economics/UCSD NORC/University of Chicago

Abstract

This paper discusses basic issues about the nature of measurement for social science concepts. It considers some advantages and disadvantages of standardization of constructs and operations in social science, including the relationships between measures and theory.

For science and policy purposes we are concerned with quantities or categories rather than material objects. What does it mean to measure 'how much' of a quantity or which category to assign to something? Measurement isn't just assigning values or numbers; it is assigning values in a systematic and grounded way. This involves applying some well-grounded metric to the thing and expressing the quantity of the thing in terms of that metric. Measurement implies three requirements: 1) We have to have a characterization of the quantity or category, that is we have to be able to identify its boundaries and know what belongs to it and what does not (*characterization*).; 2) we have to have a metrical system that appropriately represents the quantity or category (*representation*); and 3) we have to have rules for applying the metrical system to produce measurement results (*procedures*). Common metrics depend on the characterization, representation and procedures being well specified and meshing in a consistent form so that the measurements can be used cumulatively.

We distinguish between concepts that refer to a single quantity or category that can be precisely defined and those that refer to things that are loosely related but for which the boundaries of the concept are not clear (*Ballung* concepts). The use of concepts for different purposes often leads to changes in characterization, representation and/or procedures that disrupt the possibility of common metrics. This is especially likely for *Ballung* concepts.

Many policy-related social science concepts are value-laden or at least exactly how they are defined and measured will have many value-relevant consequences. Often we lack a firm scientific basis for important choices in how they should be defined and measured. The varying purposes for which they are used and the varying values assigned to the consequences make common metrics very difficult in these cases, if not impossible.

[1] Nancy Cartwright's work on this paper was sponsored by grants from the Templeton Foundation, the British Academy, the UK AHRC and the UK LSE ESRC Centre for Climate Change, Economics and Policy.

A Theory of Measurement

Nancy L. Cartwright[2] and Norman M. Bradburn

London School of Economics NORC/University of Chicago

In this paper we will discuss basic issues about the nature of measurement for social science concepts. We will consider some advantages and disadvantages of standardization of constructs and operations in social science, including the relationships between measures and theory.

I. *What does 'measuring' mean?*

For science and policy purposes we are concerned with quantities or categories rather than material objects. What does it mean to measure 'how much' of a quantity or which category to assign to something? Measurement isn't just assigning values or numbers; it is assigning values in a systematic and grounded way. This involves applying some well-grounded metric to the thing and expressing the quantity of the thing in terms of that metric. This in turn implies that:

1. We have a characterization of the quantity or category.
2. We have a metrical system that appropriately represents the quantity or category.
3. We have rules for applying the metrical system to produce the measurement results.

So to begin, we need to: 1) identify the quantity or category to be measured (characterization); 2) define the properties of the measurement scheme (representation); and 3) formulate rules for applying the scheme (procedures). The reasons we undertake a measurement project – what we want to use the measurement results for – will play into all of these steps. As we shall see, this is especially the case in measuring what we call *Ballung* concepts: Since many important concepts in social science sort things into categories based on a variety of often imprecise and relatively loosely interconnected characteristics, it is especially important in treating these to be clear about the job the categorization is to do and the kinds of inferences to be made. But this is important not only for Ballung concepts but no matter what kind of concept is at stake. The 3 steps must be responsive to the role the concept will play and to the reasons the category or quantity is worth measuring in the first place. Although 1-3 are listed as separate steps to help analyze measurement processes, what happens under each step may influence other steps. We may, for example, come to re-characterize a category on the basis of results derived relative to one or another metrical representation of it. This may be the case with characterizations of economic freedom, as we illustrate below. Or we may pick a metric system because the procedural rules for applying it are well-defined or the staff knows these methods better or they are easier to implement, etc.

All of these steps-- characterization, representation and procedures-- need explication. For an adequate measurement, these three must line up properly together: The representation of the category

[2] Nancy Cartwright's work on this paper was sponsored by grants from the Templeton Foundation, the British Academy, the UK AHRC and the UK LSE ESRC Centre for Climate Change, Economics and Policy.

measured must be appropriate to the central features taken to characterize it; equally, the procedures adopted to carry out the measurement must be appropriate to the formal representation adopted, and we should have good reason to expect that within acceptable bounds of accuracy the values assigned indicate the underlying values we are aiming to measure in a systematic way. Showing that a measure satisfies these requirements can be done more or less formally, more or less explicitly and more or less well. But it should not be neglected. Note George Bohrnstedt's concerns that after item trimming within an IRT approach it is important to ensure that the final items measure the concept originally intended.

II. *The three steps*

II.1 *Characterization: What concept is being measured?*

The characterization of concepts in natural and social science may be more or less precise. The more precise the characterization, the more likely it is to be defined in terms of its measurement (operational definition). But most concepts in social sciences and particularly in the policy realm are rather loosely defined at the start. Further the definitions may vary depending on the use to which the concept is being put. For example, "disability" may mean different things depending on whether we are talking about particular individuals, about a policy goal, a variable in a psychological theory, or a characteristic of a group of individuals. Concepts used in political discussions are often used in quite loose senses.

Many social science concepts seem to refer to specific qualitative or quantitative features that things might have, or to sets or functions of these. Other concepts sort things into categories based on a loose set of criteria where the members of the same category do not share any specific set of features but rather have what Wittgenstein referred to as 'family resemblance'. The number of women refugees in a province, the average age of school leavers, the NAIRU (the non-accelerating inflation rate of unemployment), or even just the unemployment rate (though see below) seem to be in the first category. Concepts with relatively clear evaluative implications, such as human welfare, human rights, quality of PhD programs or number of people in poverty, generally fall in the second category. The distinction here is not one between natural features on the one hand and ones socially constructed or dependent on social relations on the other. "Being a stepmother" depends entirely on what social relations obtain but can be a specific feature possessed by a woman. "Race" is a paradigm of a concept generally taken to be socially constructed but has been, in many societies, meticulously defined to allow labels to be applied unambiguously so that all the members of the same category do resemble each other in precisely the ways laid out in the definition (even though these may not have any other importance).

Nor is it the distinction between 'observable' and unobservable', which at any rate is a tricky matter. Some things are observable 'directly' and without the aid of instruments, but even some of these are not observable with precision: age, for instance. Others can be inferred only indirectly, from their more readily observable effects. In the case of instruments, such as a thermometer where the effect is the height of the column of mercury, their readings are supposed to be linked with the values of the measured quantity by the structure of the instrument. In other cases readings from the measuring

device and the instrument are expected to be correlated in natural conditions. In either case background assumptions, or *theory,* are required to underwrite the supposed link between the unobservable quantity and the more readily observable measurement outcomes. This is a version of the problem normally labeled 'the theory-ladenness of observation', which besets the natural and social sciences equally. It is probably the case, however, that social science measurement procedures depend far less on instruments that force a correlation and depend far more on carefully selected naturally occurring correlations, as for instance in the IRT measure that Bohrnstedt discusses. The observable/unobservable distinction is at any rate not the one we are looking for at this point.

Sometimes the distinction we want to highlight takes the labels 'realism' versus 'nominalism' about concepts:[3] Is there a *real* quantity or feature referred to by the concept or is it just a name we use for various purposes? These labels can be misleading however. Family resemblance can certainly be based in reality, and it can be a perfectly objective fact that individuals clump into categories according to these resemblances. A good example is plant taxonomy. It is partly because the labeling by networks of family resemblances works so well that biologists continue to hope for some single feature or fixed set of features that can ground the similarities and differences, such as genetic structure, evolutionary history, morphological structure, or ecological role.[4]

Otto Neurath (socialist, sociologist, one of the founding members of the Vienna Circle, and spearhead of the unity of science movement of the 1930's) maintained that most concepts used in daily life are of the second type. He called them *Ballungen* ('congestions'),[5] as in the German "Ballungsgebiet' for a congested urban area with ill-defined edges: There is a lot packed into it; there is often no central core without which one doesn't merit the label; different clusterings of features from the congestion (*Ballung)* can matter for different uses; whether a feature counts as in or outside the concept, and how far, is context and use dependent We employ Neurath's word here because other words more commonly in use, like 'umbrella concept', generally already have interpretations, and different ones for different scholars and different fields.

Neurath was influenced in his discussions of *Ballung* concepts by Max Weber,[6] who argued that the study of society could almost certainly not become a proper science. Weber supposed that the hallmark of proper science is the use of precise and unambiguous concepts that figure in exact relations with one another. Physics, he believed, has the latitude to pick and choose the concepts it studies in an effort to find precise concepts that figure in exact laws. The study of society, he maintained, has no such latitude since it is supposed to help us understand and manage the concepts we are concerned with, few of which have the right character to participate in exact science. Scientific definitions must attempt to regulate the divergence of meaning. But any concept familiar to general society, such as

[3] Cf. Chang and Cartwright 2008.

[4] For further discussion, see Dupre (1996).

[5] See Cartwright et al 1996 for discussion and references.

[6] Cf. Weber in Weber (1949).

'disability' or 'poverty' or 'functional literacy', is bound to have a multifaceted meaning. So the measurement of such a concept with any precision is likely to sacrifice or alter aspects of the meaning.

In the course of constructing a measurement of a concept it is important to be clear what kind of concept it is. Is it supposed to refer to a single quantity (or function of a set of quantities) on the one hand or is it a *Ballung* concept on the other, or somewhere in between? But this is only a first step. Understanding just what the concept is matters essentially to how it is to be measured. Explicit definition is the most straightforward way to go, but this is seldom possible in either the natural or social sciences. When there is a well articulated body of knowledge already accepted, implicit definition via the role the concept plays with respect to others in a system of claims or axioms is the next tightest way to characterize a concept. This is the method that Bohrnstedt refers to in introducing Northrop's 'concepts of postulation'. Generally we are not in a position to do this in the social sciences, and often not in the natural sciences. After all, part of the point of measuring a concept is to find out enough about it and its relations to other concepts to characterize it more fully. Usually we need to start with some rough defeasible characteristics of the concept and through a gradual back-and-forth process refine the concept itself simultaneously with refining our claims about its relations to other concepts and with the methods taken to produce accurate and precise measurements of it. It should also be noticed that the fact that we often start with rough open-ended characterizations does not imply that the concept in view is a *Ballung* concept since this is the historical trajectory of many natural science concepts that refer to precise quantities.

When our understanding of the concept we are trying to measure is weak and our knowledge of what other features might serve as good indicators of it is weak as well, we sometimes resort to a kind of unhelpful explicit definition: operational definition. We point to a set of relatively well-articulated measurement procedures and define it in terms of those procedures: The concept is whatever it is that these procedures assign values to. IQ is the canonical example: 'IQ', some maintain, 'is just what IQ tests measure'. This form of characterization makes knowledge accumulation difficult since by definition no other procedure can be used to measure the same quantity. We also often gain confidence in our measurement results by noting that different procedures for measuring the same quantity yield roughly equivalent results, which is impossible when quantities are defined operationally.

There is a third type of concept to be noted as well, what might be called 'concepts of pure understanding'. Since at least Aristotle onwards, thinkers have noted that the sciences make important use of concepts that have no referents in reality. These concepts are supposed to be useful for understanding, for representation (often mathematical representation, as in Aristotle's reminder that lines have no breadth so don't exist in reality despite the usefulness of geometry as a representational device), or for organization, but not for literal description. Consider Carl Menger,[7] who was one of the three thinkers generally credited with creating the notion of marginal utility and who, in the late 19th century 'Methodenstreit' (battle of the methods) in social science, was a great proponent that theory should be both formal and general . Menger argued that theoretical economics could be a proper

[7] Menger 1985 {1883].

science, which in his mind involved having exact and exceptionless laws, but at the cost of using not 'real types' but rather concepts that do not pick out features that exist in 'full empirical reality'. Despite its lack of immediate connection with the empirical world, theoretical economics is nevertheless what leads to genuine understanding of the empirical world. That's why we use the term 'concepts of the understanding'. Probably Weber's ideal types can be classed in this way, as well as Goethe's 'ur' objects (e.g. ur-pflanze, or basic plant types) and possibly much that goes under the title 'idealization' in modern natural and social science.

We don't use the term 'idealization' for the general category, however, because this term suggests a theory about how these concepts are to be understood: that they are limiting cases of things that do exist, as in the frictionless plane, which is ideal in Galileo's rolling ball experiments for seeing the 'pure' unimpeded effect of gravity. This is just one theory among many. Another is that these concepts are like Plato's forms. They are what matter for understanding; the concepts that describe empirical reality are like the fuzzy shadows on the walls that people living in a cave observe of the real objects located in the bright sun outside. Something like this was probably Menger's own view, and maybe Einstein's, about the mathematical concepts of general relativity. Another theory, that Pierre Duhem defends,[8] is that these mathematical concepts are mere organizing devices. It is not clear which if any of the accounts on offer is correct; and of course there is no reason to assume that one account covers all the cases that might roughly be grouped together under the heading 'concepts of the understanding'. We mention them to flesh out our categorization of types of concepts that need to be treated differently when it comes to issues of measurement. But we would like to dismiss them from further consideration here since if a concept is not supposed to apply to real things, a theory of what counts as measuring their values in real things is going to be at best complicated, controversial and highly dependent on the account presupposed of what these concepts are for; and at worse, otiose and a waste of time.

No matter which of the three types of concepts we are considering is at stake, in characterizing the concept we are usually pulled in two different directions. Whether we stay as faithful to our concepts as possible, or make them as fit for purpose as possible, we are likely to proliferate concepts and their measures in order to get ones that do well what we need. But purpose-built concepts make the accumulation of knowledge difficult so we are often reasonably pulled to rely on concepts that do not quite fit and to hope that the ones we use are close enough for results established in one situation to be reasonably accurate in others. It takes a serious case-by-case scientific decision to determine when this use can be expected to work well enough and when not and what if any the losses will be.

The importance of characterization cannot be overemphasized. Our measurement procedures should measure the very concepts we aim to measure (or something 'near enough' for the purposes at hand), not something different. But that cannot in any way be assured until the concept has been refined and delimited well enough to allow us to show that the representation and the procedures we propose to adopt are suited to it. Consider the definitions of poverty by Smith and Townsend described

[8] Duhem 1962 [1906].

in George Mitchell's paper. These may well have been great advances, but they are far from a basis for devising measurements. It is not just the well-known issue of relative versus absolute concepts that produces divergent measures, but as we discuss in more detail below,[9] a vast array of questions must be settled about the basic poverty concept we aim for in the course of devising a measurement for it, including, for example, the one that Mitchell stresses – choice of units to be counted: Are we looking for the concept 'number of people in poverty' or 'number of households in poverty'? If we want to turn the *Ballung* concept of poverty into a precise unambiguous one, poverty concepts are bound to proliferate.[10]

II.2 *Representation*

Although the distinction between these two kinds of concepts is not a sharp one, it is useful to keep it in mind in thinking about representation and measurement. For instance, representing a concept by a linear function with perhaps an interval scale, as in length or a ratio scale, as in loudness, would be inappropriate for a *Ballung* concept, which could be better represented by a table of indicators or an index. This is illustrated in the thinking behind the EU measure of social exclusion.[11] The recommendation there is for a 3-tiered measure. The first tier contains 7 or 8 leading factors that can be taken to pick out important aspects of social exclusion across the European community, such as inequality and extent of poverty. These will not be very comprehensive; they are restricted in number because getting any kind of a rough view of the meaning of a larger array is difficult for non-experts. The second tier makes up for this deficiency by including a far larger number of factors that are thought to make the representation more comprehensive. Finally the third tier allows for the introduction of society-specific features: features that might make for social exclusion given the way life is conducted in one society, but not in another, or that have special value in one European society but not another.

Conversely, it is misleading to represent concepts of the first type by sets of indicators or indices. This is not to say that we may not be forced to measure a concept of the first kind in a host of indirect ways, none of which suffices to zero in on it sufficiently reliably or precisely, in which case good practice would be to report the array of results. But to represent such a concept in a theoretical structure in indirect ways loses the possibility of laying out any fairly exact relations it figures in. It blurs the line between what is vague in the social world and what we are uncertain of, and blurs it in an unhelpful way.

Representation of concepts of the first type is usually in some metrical system. These systems have some underlying mathematical structure. Stevens (1951) enumerated four levels of measurement (representation). Measures may simply enumerate instances of a concept, e.g. proportion of a

[9] See also Atkinson 1998.

[10] See Efstathiou 2009 for an interesting account, *found science*, on the analogy of found art, of how 'everyday' Ballung concepts are transformed into proper scientific concepts, with the use of 'race' in medicine as a primary illustration.

[11] Atkinson 2002.

population that is male; they may order instances of the concept, e.g. rank colleges on the quality of their students; they may order instances of the concept on a scale with equal intervals, e.g. GDP, or they may order instances of the concept on a scale with equal ratios and a true zero point, e.g. perceptual scales of loudness.

Bohrnstedt gives an excellent account of the current state of thinking about the scientific measurement requirements for concepts that refer to specific features, the requirements for representing them mathematically and the necessary development of procedures to meet the requirements of the models. These concepts are represented as a single factor or a multi-dimensional set of factors and require rigorous procedures to produce data that meet the standards of the measurement model. He correctly notes that the difficulty in meeting these requirements necessitates more attention of the theoretical clarification of the concepts and to the kinds of observations or items that make up the procedural operations that produce the measurements. This kind of measurement can produce the invariance of results that one expects of science.

II.2.1 *Getting the representation right*

For proper measurement our scientific representation of a concept must reflect and be warranted by our idea of what kind of feature or category we are measuring and what kinds of characteristics it is taken to have. Similarly the procedures employed for assigning values to the concept must be grounded in our assumptions about what the feature or category we are measuring is like; and they must line up properly with the formal representation we adopt for it. For instance, a scale of 1 -10 should be treated as only a pure ordering if 9 units ≠ 3 times the amount of the quantity possessed by individuals with 3 units.

If in one accepts the distinction between concepts that refer to specific features of the thing being studied and *Ballung* concepts that involve bundles of features whose boundaries are not very well specified, then one needs to consider the implications of the difference for how the connections between characterization and representation of a features are warranted at both stages. We begin with concepts of the first kind.

II.2.1.1 *Concepts that pick out specific features or sets of features*

In this paper we discuss the general theory of the nature of measurement, especially in the social science. Sometimes the term *theory of measurement* is used more narrowly to refer to concerns about the connection between our assumptions about the feature and our formal representation of it.[12] Although these concerns apply equally in the natural and social sciences, social scientists are more attentive to its demands. The first task within measurement theory in this sense is to provide a mathematical representation of the targeted concept so that it can be integrated into a scientific context with an existing set of concepts. The second task is to provide a *representation theorem* to show that this representation is adequate. A representation theorem first provides a set of characteristics

[12] See Suppes 1998 for an accessible introduction.

taken to be true of the targeted concept, and then proves that the concept as represented has those characteristics.

Consider economic freedom. We talk loosely of economic versus political freedom, of negative versus positive freedoms, and the like. Can economic freedom be represented more exactly in the framework of, say, social choice theory? The simplest idea is a pure cardinality measure that identifies the degree of economic freedom agents have with the number of options available to them. Is this a good representation? Suppose we agree that economic freedom has some basic features: for example, if one set contains every option that a second contains and more, the first offers more economic freedom that the second, or if the same option is added to two sets of options deemed to provide equal economic freedom, the two expanded sets will offer equal economic freedom as well. In a good exemplar of measurement theory at work, Pattanaik and Xu[13] provide axioms describing three such features, then prove that an ordering among sets of options satisfies those axioms just in case it orders the sets according to their size. Their proof is an example of a *representation* theorem, a theorem that shows that the scientific representation – here a cardinality measure on the choice set – is just what is required to capture the characteristics supposed in the axioms to be true of the feature to be represented. Later writers provide more nuanced accounts of the characteristics of the notion "freedom of choice" that they are trying to represent.[14] In each case measurement theory requires that the definitions be defended by a representation theorem.

Perhaps a representation theorem seems an overly formal demand. But it is important for there to be serious and explicit consideration of the characteristics and structure of the concept we are trying to measure. Consider cardinality measures. Suppose we adopt a cardinality measure to incorporate a concept of freedom of choice into social choice theory. We may think that a choice among three kinds of ice cream offers as much freedom of choice as a choice among three kinds of cake. But are we happy to agree that adding a new kind of cake as a fourth choice to each set leaves them equivalent? Many feel that a choice among 3 kinds of cake and a choice among 3 kinds of ice cream offer about the same amount of economic freedom, but 3 kinds of ice cream plus a cake offers more freedom of choice than does 4 kinds of cake. If so then the characteristics of the options matter, not just how many there are. So we had better not use a mere counting measure (i.e., a cardinality measure) to represent the concept we are looking for. Or poverty. Are we really concerned with an all-or-nothing concept, like above/below the poverty line, or more/less than $2/day? If so, a cardinality measure will do. But we may feel that transfers from the poorer poor that bring the richest poor above the line should not count as alleviating poverty. In that case we need a concept that gives more weight to those further below the poverty line. Sometimes of course we might for expediency reasons adopt a concept and a concomitant representation that is not exactly the one we want. But we should be clear when we are doing this and what the implications will be in various contexts of use.

[13] Pattanaik and Xu 1990.

[14] Cf Bavetta and Guala 2003 and references therein.

Another place where the need for representation theorems looms large is when different indices for the same quantity are agglomerated into a single number, usually by some kind of weighting procedure. There is of course a great deal of pressure to do this since a total ordering of the units measured will then be possible, whereas with sets of indices, usually at best only a partial ordering is possible. But in this case there should be good arguments both that the weightings, and the methods for choosing the weightings, are appropriate to the concept to be measured and that the final representation does not imply features that the concept does not have, as the sheer cardinality measure for freedom of economic choice does.

II.2.1.2. *Ballung concepts: sorting things into categories*

With *Ballung* concepts there are two obvious strategies for representation. One is to represent them with a table or vector of features laying out the dimensions along which the family resemblances in question lie, as in the EU measure for social exclusion discussed above. The other is to shed much of the original meaning and zero in on some more precisely definable feature from the congestion that constitutes the concept. An example are the various attempts to measure the quality of PhD programs in American universities "Quality of PhD programs" is a quintessential *Ballung* concept. It has intuitive meaning but has indefinite boundaries. Each of the past efforts has used a different method of measuring quality so that results across time are not comparable. The most recent attempt to measure it (National Research Council, 2009) has proceeded empirically by surveying a large sample of faculty in many fields to ascertain their views on twenty objective characteristics of good PhD programs in their field. This exercise resulted in a table of indicators of quality with varying degrees of support by the faculty raters. An overall measure of quality was then constructed by computing a weighted average of the indicators using weights derived from the faculty survey.

Previous efforts at measurement of quality had relied heavily on subjective measures of the reputation of faculty in the programs. Because of criticism of the past studies' use of subjective measures, the new effort tried to avoid using any subjective indicators. But when the results of the objective measures produced ratings of programs that did not reflect widely held views about the quality of many programs, an additional measure based on a statistical model that incorporated a subjective reputational measure was constructed. The difficulty in adequately characterizing the concept of "quality" led to measurement problems. The final definitions of the boundaries of the concept necessary to get a measure that could be represented by an ordering of the programs left many with the feeling that the final results did not fully reflect the intuitive meaning of the concept.

Which is the better strategy will depend on the purposes for which the measure is being constructed. This is always an issue, even with concepts that are supposed to pick out a single feature in the natural or social world: Exactly which feature among many that may be closely related are we trying to produce a measure for? The answer will generally depend on why we want information about that feature and what we are going to do with that information. It matters even more for *Ballung* concepts that generally serve a variety of different purposes to begin with. In constructing measurements, it may be useful to think of purposes in two broad categories: first, understanding where we stand, and second, prediction and explanation.

Sometimes our chief concern is to get a very good picture of how a society or a group stands with respect to a social concept of concern: How much social exclusion or poverty or disability is there and what is it like? And it's *that* concept we are interested in, with all its many aspects and implications, not some 'made-up' substitute. In that case representation with a vector or table of indicators is most appropriate. It gives the most accurate answer to the question we ask.

This comes with drawbacks, of course. Vectors and tables do not make for easy comparison, either across time or across groups. For example, *Healthy People, 2010* (http://www/healthypeople.gov) has 10 Leading Health Indicators, 28 focus areas, and 467 "science based objectives", progress over which is to be tracked over the decade. Comparison is not impossible, though. For instance high rankings on all indicators orders a group above one with low rankings on all; and there may be further reasonable ways to rank groups where differences on most indicators are large or 'enough' indicators go in the same direction. Nevertheless generally at best we can expect only a partial ordering. But that is not a problem with the measure. It is often, as Weber urged, the problem with the concept itself that we are interested in: There simply is no fact of the matter about which, among a large group of European countries with mixed results on different indicators, has more social exclusion.

Sometimes we try to 'solve' the problem by, say, taking a weighted average. This can allow for comparisons. But for many cases this will simply be the second strategy in disguise. It will be a device for constructing a new, more manageable concept rather than for informing us about the concept of interest. And it raises its own questions, as we mentioned in discussing methods of representation: How are the weights selected? Of what interest is this new concept? What purposes are served by measuring it? It may be that the new concept is useful for scientific theorizing, for prediction and for explanation. In this case it is probably most useful to think of it no longer as a *Ballung* concept but as one that refers to a feature properly possessed by the individual or group, since playing a role in a network of predictive principles is one of the chief grounds on which we judge concepts to pick out 'real' features.

The second strategy is appropriate for purposes of prediction and explanation, which rely on principles involving concepts that are precise and unambiguous. This strategy has its own comparability problems. Any precise characterization that brings a *Ballung* concept into a scientific context properly necessarily leaves behind a good bit of its meaning and implications. Moreover the new concept needs to be tailored to fit the scientific field in which it is to be used and the purposes which it is to serve. But this will lead to a proliferation of different concepts, each fit for a different discipline and purpose. These different concepts may share the same name but they do not pick out the same thing, so results established for one cannot be transferred to the others.

Consider, for example, the controversial variable 'race', and its introduction into different medical contexts. But the point, as Efstathiou (2009) points out, in precisely this: What variation 'race' picks out for epidemiologists and what variation it picks out for geneticists is different in type. Epidemiologists mostly care for variation in common health outcomes and risk factors for particular diseases and ask whether 'race' picks out this type of variation successfully. On the other hand, geneticists care about markers in the genomes of individuals and so what they examine is whether 'race' picks out any interesting patterns on the level of molecular variation (which, in the case of the U.S.

population, it seems 'race' does). This is precisely a case where we get a loaded, *Ballung* concept, 'race', to work in the service of different scientific research projects. But the concept named 'race' changes significantly in meaning in this process. 'Race' in the epidemiologists' speak would be defined in terms of disease risks, while 'race' would map onto sets of genetic polymorphisms so far as geneticists see it. Efstathiou calls these new, substitute concepts "founded" concepts on the analogy of found art: objects that exist in the common world around us but become art by being transformed in appropriate ways as they are brought into artistic settings; found art objects are recognizably related to what they were in common life but, as found art, they function in different, more limited ways and should not be mistaken for the objects they were originally.[15]

III. *Measurement: the procedures for carrying it out*

We think of measurement in science as a set of procedures that apply the metrical system to the observation of some phenomena in the real world. In setting up measurement procedures effort should be made to ensure that the procedures are both *accurate* and *precise*. In common parlance "precision" is often confused with "accuracy." Here is one way to regiment the use of the words that we recommend: *Accuracy* is about whether measurement results agree with the true values or locate individuals in the correct category; *precision* indicates how specific a measurement result is.

Where a genuine quantity is measured, the observations are often done with an instrument that is calibrated to the metrical system that represents the quantity, such as a ruler or thermometer or by a simple counter. The observations can be transformed into various metrical systems by algorithms, such as converting feet into meters or Fahrenheit into Celsius scales. In many cases these instruments do not look at the quantity directly but rely on some pre-established connection between the quantity to be measured and another more directly observable quantity. For instance, with a mercury thermometer we measure the temperature of a body by recording the height of a column of mercury that we take to be at the same temperature as the body.

This gives rise to what is sometimes called 'the problem of nomic measurement' [Hasok Chang (2004: 59)]. To be confident that the mercury thermometer measures temperature accurately, we must be confident that mercury expands uniformly with temperature. But to establish this empirical regularity we need an independent and accurate method of measuring temperature. This problem of justification is common to all measurement methods based on empirical laws. There are several obvious ways to try to avoid this problem. First, determine the values of the quantity we want to measure, like temperature, by another measurement method; this only postpones the problem, since now that other method needs to be justified. Second, derive the empirical law from a more general theory. This is not straightforward, either, since the theory relied on must be empirically justified, which can be especially difficult in the social sciences where few theories are accepted uncontroversially.

[15] See Efstathiou 2009, chapters 1 and 3 and references therein for discussion of both 'found science' and of these uses of 'race' in medicine.

A mild version of operationalism can also be seen as an attempt to circumvent the regress. If empirical concepts are defined by well-specified measurement operations, observational data can be fixed without reference to theories and be made secure, while theoretical concepts and laws fluctuate and develop. This tactic of course brings with it all the well-known problems of narrowness, comparability of results from different methods, and the danger that we are no longer talking about what we started out to study.

Another tactic is to look for operations that depend on relatively uncontentious empirical principles or ones that should give near-enough the same results across the range of competing empirical principles that are deemed plausible. Whether these are available or not depends on the circumstance. Herbert Feigl[16] argued that our most basic measurement operations are grounded in middle-level regularities that seem to have a remarkable degree of stability, such as Archimedes's law of the lever and Snell's law of refraction. Again, finding these seems especially problematic in the social sciences.

For psychological constructs that are messy and in principle unobservable, Campbell and Fiske (1959] advocated a multi-trait, multi method approach to validation. Only when the construct can be measured by several different methods and with different representations, can it be used as a scientific construct. Hoyle's discussion of the state of research on self-regulation illustrates how a concept that gives rise to much research lacks basic demonstration of its characterization.

Coherence along a variety of dimensions seems to provide the best solution in practice in both the natural and the social sciences: Does the quantity as measured by the proposed method behave as it is expected to? Do the results cohere with those of other reasonably defended methods? Do the empirical principles needed to support the method cohere with other reasonably justified empirical principles and theories? In typical cases in the natural sciences, settling on measurement procedures is part of a back-and-forth process in which definitions of concepts, measurement procedures, and empirical principles become mutually adjusted. Another common strategy is to provide a vector or table of results from different methods, with perhaps some attempt to describe the spread of results statistically. It is important to keep in mind, however, that in this case there is a very different reason for using vectors or tables of results than with *Ballung* concepts. The difference in reason can have important consequences for how we use the information thus presented and for how we proceed to develop our science and our measurement procedures since in the first instance we suppose that there is a single true value to be ascertained and in that latter, not.

There are many models in the social sciences that are used to combine observations into a measure of a concept. Some of these assume similarity of observations, such as responses to Likert type questions; some combine responses to different types of questions by looking at patterns of correlation, such as factor analytic or latent trait models, some create indices, such as those of socio-economic status or poverty, based on theoretical relations or as described above, tables of indicators. These models may have well established levels of measurement or, as is often the case, assume an equal-

[1616] Feigl 1970

interval level of measurement for computational ease. Even if the properties of the system are clear, the actual measuring instruments may be precise or imprecise, may be definite or probabilistic, and may be objective or subjective.

The procedures for measuring a concept should provide values of the appropriate kind for the concept in the systems/individuals measured. What is appropriate depends on how the concept is characterized and on the kind of representation chosen for it. As we have stressed, characterization, representation and procedures must work together consistently.

Often the procedures for measuring a concept do so only indirectly. We infer the value of the concept from the results of measurement on some quantities that can be observed more readily. We measure, as already mentioned, the temperature of a fluid by observing the height of a column of mercury in a mercury thermometer; we measure the charge of a particle by the deflection of its trajectory in an electromagnetic field; and we measure the strength of an attitude by answers to questions in a survey. Sometimes the more immediately observed quantity will be a cause of the targeted concept, sometimes an effect, sometimes the two are correlated for some other reason. What matters is that the two quantities be linked by reliable regularities. Laying out and defending these regularities is one of the central tasks in designing a measurement.

III.1. *Getting the procedures right*

Not only must the metric system chosen to represent a concept be matched to the concept but the procedures to be carried out to assign values to it must also be appropriately matched with the representation adopted for it. Mere counting may be adequate if we are only representing "poverty" as a dichotomous category, but it is a poor procedure for assigning a value to the amount of poverty for a concept of poverty that is sensitive to degree. Measurement systems for subjective phenomena, that is, those for which there are in principle no appeals to consensus of external observation, are particularly difficult because there is no direct way of knowing that the subjective judgments are using the same scales or at least some transformations of the same scale. For subjective phenomena measurement typically starts with observations that depend on responses from individuals to (more or less) common stimuli. The responses are then recorded in some fashion and represented in some metrical system with (more or less) well-defined properties. Sometimes these observations may have a one-to-one relation between the response and the metric, as in the case of the "just noticeable difference" (jnd) measures of sensation; others by a frequency distribution of answers to a survey question. Often, however, the observations are combined in some (more or less) well specified way and put forth as a measure of a complex subjective phenomenon such as an "attitude".

The procedures used to measure a social or economic concept may end up producing measures of something that does not correspond to the way the concept is meant to be understood. For example, the concept of *unemployment* appears to be an easily understood concept. Although being employed is relatively unambiguous, what it means to be unemployed is not so clear. The procedures used to measure *unemployment* in official statistics consist of a set of questions regarding employment status and whether a person is looking for work. The unemployment rate, an important policy measure for

economic policy, is calculated by expressing the number of those not working as a proportion of those who are in the labor force. The concept of *labor force,* however, turns out to be a difficult concept to measure in part because it is itself not clearly delineated and in part for procedural reasons and practical problems with implementation.

The measurement problem is that not everyone who is not currently employed is considered to be part of the labor force, that is, they are not considered to be unemployed. Those who choose not to work or for some reason are not able to work are not considered part of the labor force. Whether or not you are part of the labor force is determined by asking if you are "looking for work." But "looking for work" is a subjective judgment and not a particularly reliable indicator. For example, when the unemployment rate was calculated for youth on the basis of the first wave of the National Longitudinal Survey of Youth (NLSY '78) the rate was higher than the youth unemployment rate calculated from the Current Population Survey (CPS), the Labor Department's survey that is used for the official unemployment statistics. After much investigation of possible factors causing the discrepancy, it was determined that the reason was that the NLSY interviewed only the youth themselves, while the CPS depended for many of its reports of youth work status on parents' reports of youth's employment status and, more importantly, whether the youths were looking for work or not. More youth were reporting that they were looking for work, hence were defined as part of the labor force, than did adults who were reporting for them. Apparently parents have a higher threshold for what it means to look for work than do teenagers.

Also, it is believed that many people who have looked for work for a number of weeks and failed to find any cease looking for work. These so-called "discouraged workers" are then dropped from the labor force and not counted as unemployed, although they may still call themselves unemployed and would be so categorized by most people. Thus in periods of high unemployment, the official unemployment rate is often described as underestimating the actual unemployment rate. To say this is to admit that the reported measure does not correspond to our common understanding of the concept. For this reason, many economists think that it is better to look at the employment rate rather than the unemployment rate. The former is less ambiguous than the latter.

Another way in which measures may not correspond to the intended concept is that the procedures used to measure the concept combine measures that have different underlying relations with other factors that are considered at least partially causal. For example, there is renewed interest in measures of happiness. When one looks at the actual questions used, however, they are often questions about life satisfaction rather than about happiness. The assumption here seems to be that "satisfaction" and "happiness" are more or less synonymous terms. While it is true that answers to the questions about life satisfaction and happiness are positively correlated, the questions are related to other variables, such as age, in opposite directions. Questions that use the term "satisfaction" are positively correlated with age, while those that use the term "happiness" are negatively correlated with age. Younger people are happier, but less satisfied with their lives as a whole (Bradburn, 1968). While it is interesting to speculate why this difference should exist, it does not seem right to treat them as if they were synonymous terms.

Even when there are good measurement properties, different operations may be used for different purposes. For example, there is considerable interest in the medical world about measures of Quality of Life (QOL) used as outcome measures to evaluate the effectiveness of different treatments. As mentioned in the Bohrnstedt paper, one such set of measures with good IRT properties is the Patient Report Outcomes Information System (PROMIS) being developed by NIH (http://www.nihpromis.org). This effort arose out of concern for the large number of items that were being used by various researchers to measure aspects of medical QOL. Many of these measures had poor or unknown measurement properties. As part of a larger NIH effort to improve measurement, the PROMIS project, through elaborate review processes, has categorized domains of interest, revised items and submitted them to IRT scaling procedures to produce measures that have the desirable equal interval properties described by Bohrnstedt. The scales have been standardized on large general populations and some specialized clinical populations. So far scales for 11 domains have been developed. Many more are in the process of being constructed and validated.

Although each individual domain, such as depression, anxiety, pain, mobility, etc., may be well measured, differing domains may be used depending on the purpose to which the QOL measures are put. Thus those who are studying outcomes of treatments for neurological diseases may use a different set of domains than those studying outcomes of cancer treatments. The goal of improved QOL may be common to the different studies as desired outcomes, but the domains, hence the measures used to represent QOL, may be quite different depending on the purpose of the treatments. Consequently, it would be impossible to compare QOLs across different kinds of treatments unless it were clear that the domains making up the QOL indices were the same, even though each of the domains had been measured by the same set of items which were all well scaled,.

Interest in QOL measures has also been growing in other areas. Perhaps the most notable is the report of the Stiglitz Commission (http://www.stiglitz-sen-fitoussi.fr) which recommends developing measures of well-being that will be routinely reported by state statistical offices. The Commission offers a set of indicators, as seems most appropriate in treating a concept taken from ordinary life with multiple, loosely-connected meanings and different boundaries in different contexts and for different purposes. They recognize that the choice of indicators is a value judgment rather than a technical exercise. They believe, however, that there is a sufficient consensus on domains that are constituent parts of assessments of quality of life that they could come up with a non-controversial list. They suggest that quality of life depends on people's education and health, their everyday activities including jobs and housing, their participation in political and social activities as well as the environment in which they live, and factors that shape their personal and economic security. As they note: "The challenge in all these fields is to improve upon what has already been achieved, to identify gaps in available information, and to invest in statistical capacity in areas (such as time-use) where available indicators remain deficient." (Recommendation 6).

The report also notes that there will be a demand for a single summary measure of quality of life. Assuming that there were a consensus on the components of QOL and that good measures of the various components had been developed, there remains the vexatious issue of how to combine the measure into a single index. The problem here is one of weighting the various component measures.

How does one decide on what weight to give to each component? Does one use the same weighting scheme for all the different uses one might make of the index? If different weighting schemes produce differential results for different groups, how does one reconcile the results? These very difficult issues involve value and political judgments that may threaten the professional reputation of the statistical agencies producing the measures and undermine the integrity of the process. These are just the kinds of reasons that lead social scientists to aim for 'value-free' measures.

IV. *Value-laden concepts and measures*

Measures in social science are often not value-free despite our best efforts and our most ardent hopes. Frequently, they make sense as measures only in relation to certain values or uses to which the results will be put, both epistemic and practical uses. This may be obvious in a case like the human development index, which includes life expectancy, level of education, and GDP. Should it include a measure of political freedom as well? That presumably depends on whether political freedom is accepted as a constituent of human flourishing.

The intrusion of values or purposes may be less expected elsewhere, but it seems exceedingly difficult to avoid. Consider the Boskin Commission proposals in the US for revising the consumer price index (CPI). One proposal argued that the price for many goods was overestimated because it was based on samples from retail stores, whereas the goods tend to be much cheaper in outlet stores, which were not properly represented when prices are sampled. But, many argue,[17] adjusting the CPI in this way will disadvantage the elderly, those without cars, and other groups who have poor access to outlet stores, which are generally far from town centers.

A stock response to these problems urges that decisions involving value-laden choices in the construction of a measure be given to users of the measure – policy-makers of all sorts who will use the measure in their deliberations. This has major drawbacks. First it leads to a proliferation of measures which become difficult to understand and keep track of; we also get the same problems of theory-testing and comparison discussed already with respect to universal versus purpose-built measures. Second, it is an extremely difficult strategy to execute. Consider poverty measures. After the National Academies report criticizing the official poverty measure (See paper by R. Michael in this session) the Census Bureau started publishing a set of alternative poverty measures with the expectation that one new version would finally replace the official measure which is universally viewed as a bad measure. This has not happened, however, because any change in the official measure would show that some groups were better off than had been the case with the old measure and some groups would be worse off. No one up to now has been willing to take the political heat for bringing about such a change. There is a recent proposal to publish in 2011 an additional measure recommended by a National Academy panel 15 years ago that may be used to track the effectiveness of policies to mitigate poverty (U.S. Dept. of Commerce, 2010). For some policy purposes a concept may be defined in absolute rather than relative terms. In the U.S., in contrast to Europeans countries, poverty is defined in absolute terms so that in principle it is possible to talk about the elimination of poverty.

[17] Cf. Reiss 2007.

But suppose that a legislative body or the populace were willing and able to think about whether the measure should be relative, and, if relative, relative to what. there are still a host of further questions that need answering in developing a measure of poverty: Should we set the poverty line at two-thirds of the median income? Should it be mean or median? Should we consider income or wealth? Should we count households or individuals? How should we weight individuals in a household? These decisions both affect different groups in different ways and also can dramatically change the assessment of how much poverty there is and the poverty-rankings among different regions. To understand the impact of those decisions requires much thought and more economic and social knowledge than even experts have, let alone those who want to use the information. Here again is a problem that makes designing measures in the social sciences far more difficult than in the natural sciences.

V. Implications for the possibility of common metrics

We noted at the beginning of this paper that measurement implies three requirements: 1) We have to have a characterization of the quantity or category, that is we have to be able to identify its boundaries and know what belongs to it and what does not.; 2) we have to have a metrical system that appropriately represents the quantity or category; and 3) we have to have rules for applying the metrical system to produce measurement results. Common metrics depend on these three requirements being met in a consistent form so that the measurements can be used cumulatively. Only if the characterization, representation and procedures are well specified are common metrics possible.

We distinguished between concepts that refer to a single quantity or category that can be precisely defined and those that refer to things that are loosely related but for which the boundaries of the concept are not clear (*Ballung* concepts). The use of concepts for different purposes often leads to changes in definition, representation and/or procedures that disrupt the possibility of common metrics. This is especially likely for *Ballung* concepts.

Many policy-related social science concepts are value-laden or at least exactly how they are defined and measured will have many value-relevant consequences and often we lack a firm scientific basis for important choices in how they should be defined and measured. The varying purposes for which they are used and the varying values assigned to the consequences make common metrics very difficult in theses cases, if not impossible.

References

Atkinson, A. B., (1998). *Poverty in Europe.* Oxford: Blackwell.

Atkinson,A.B., Cantillon, Bea, Marlier, Eric and Nolan, Brian.(2002) *Social Indicators: The EU and Social Inclusion.* Oxford: Oxford University Press.

Bavetta, Sebastiano and Guala, Francesco, (2003), 'Autonomy Freedom and Deliberation', *Journal of Theoretical Politics*, Vol. 15, No. 4, 423-443

Bradburn, N. (1969). *The structure of psychological well-being.* Chicago: Aldine Publishing Co.

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

Cartwright, Nancy and Jordi Cat, Lola Fleck, Thomas E Uebel, (1996) *Otto Neurath: Philosophy between Science and Politics*, New York: Cambridge University Press, Ideas in Context Series No. 38.

Chang, Hasok (2004) *Inventing Temperature: Measurement and Scientific Progress*, New York: Oxford University Press.

Chang, Hasok and Nancy Cartwright.(2008). 'Measurement', *The Routledge Companion to Philosophy of Science*, Stathis Psillos and Martin Curd (eds.), Routledge: London and New York, pp. 367-375.

Duhem, Pierre (1962 [1906]) *The Aim and Structure of Physical Theory*, New York: Atheneum.

Dupré, J. (1996), *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*, Harvard University Press.

Efstathiou Sophia, (2009), *The Use of 'Race' as a Variable in Biomedical Research*, PhD Thesis, UC San Diego.

Feigl, Herbert (1970) "The 'Orthodox' View of Theories: Remarks in Defense as well as Critique," in Michael Radner and Stephen Winokur (eds) *Analyses of Theories and Methods of Physics and Psychology*, Minneapolis: University of Minnesota Press, pp. 3–16.

Menger, Carl. (1985) *Investigations into the Method of the Social Sciences with Special Reference to Economics.* New York: New York University Press. Original publication 1883, trans. Francis J. Nock, University of Illinois, Urbana, 1963

National Research Council, Committee to Assess Doctorate Programs. (2009) *A Guide to the Methodology of the National Research Council Assessment of Doctorate Programs*, unpublished.

Pattanaik, Prasanta and Xu, Yongsheng (1990) "On Ranking Opportunity Sets in Terms of Freedom of Choice," *Louvain Economic Review* 56: 383–90.

Reiss, Julian. (2007). *Error in Economics: The Methodology of Evidence-Based Economics*, London: Routledge

Stevens, S. S. (1951) "Mathematics, measurement, and psychophysics." In S.S. Stevens (ed.) *Handbook of Experimental Psychology* New York: Wiley.

Suppes, Patrick (1998) "Measurement, Theory of," in E. Craig (ed.) *The Routledge Encyclopedia of Philosophy*,London:Routledge; available: http://www.rep.routledge.com/article/Q066?ssid=388873355&n=8.

U.S. Department of Commerce, Economics and Statistics Administration. (2010)**. "**Census Bureau to Develop Supplemental Poverty Measure." News release, March 2, 2010.

Weber,Max.(1949) *"Objectivity,"* in E. A. Shils and H. A. Finch (*eds. and trans.) The Methodology of Social Sciences.* Glencoe, Ill.: Free Press.