# Assigning Membership in a Fuzzy Set Analysis

JAY VERKUILEN
*University of Illinois, Urbana-Champaign*

*This article provides a largely nontechnical discussion of the acquisition of membership values in fuzzy set analyses. First the basic properties of a membership are discussed. Then the three common strategies of membership assignment—direct subjective assignment, indirect subjective assignment, and transformation—are critically examined in turn. Examples are used to illustrate the techniques. The connection with existing psychometric and statistical methods is particularly emphasized, focusing on the notion of a membership value as a random variable as a means to assess uncertainty in assignment.*

***Keywords:*** *fuzzy set analysis; direct subjective assignment; indirect subjective assignment; transformation*

## *1. INTRODUCTION*

Most, if not all, social science concepts are vague, in the sense that it is frequently very difficult to assign objects[1] to exactly defined categories. As Lazarsfeld (1972)—one of the intellectual founders of much of the modern approach to social science inquiry—cogently put it,

> All the social sciences deal with concepts that seem to have a certain vagueness. Who can precisely say what a folk society is? Who has not read many discussions as to the real meaning of public opinion? Who can, in practice, recognize an extrovert personality? There are various reasons why the social scientist's language has so many of these terms, which at first sight seem to be ill defined and even at their best "fuzzy at the fringe." (P. 1)

---

AUTHOR'S NOTE: *I would like to thank David Budescu, Gary Goertz, Carol Nickerson, Michael Smithson, and Thomas Wallsten for useful comments and discussion. It draws on work from my dissertation in political science, supervised by Gerardo Munck. Previous versions were improved by comments from James Kuklinski and two anonymous reviewers. Finally, Charles Ragin encouraged me to write it. This research was supported by a NIMH National Research Service Award (no. MH14257) to the University of Illinois. The research was conducted while the author was a predoctoral trainee in the Quantitative Methods Program at the Department of Psychology, University of Illinois, Urbana-Champaign.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 33, No. 4, May 2005 462-496
DOI: 10.1177/0049124105274498

What might the sources of this vagueness be? Again, Lazarsfeld (1972), from the same passage, offers three options:

> In some cases we can, by the nature of the concept, only observe symptoms, behind which we assume a more permanent reality. In other matters the object of investigation is so vast that we can only analyze certain aspects of it: notions like patterns of culture or Zeitgeist belong here. For still other purposes the problem itself seems to require a looser kind of formulation .... This peculiarity of the social scientist's intellectual tools has been deplored by some, considered unavoidable by others. (P. 1)

The first few sentences of the second quote point to the most important source of vagueness—namely, that generally abstract concepts are used to summarize the network of associations and implications among observable components. In general, it is very difficult to pin these concepts down to one concrete indicator. Terms such as *democracy*, *shirking*, *support for political violence*, *poverty*, *development*, or *maturity*, to name a few, are all important terms of discourse in areas of social and behavioral science, and all are vague. A person can be, for instance, poor to some degree, as could a nation be developed or unequal. But, in addition to being vague, these terms have important qualitative boundaries. We can generally recognize cases of definite shirking or working, between which there is continuous variation. Furthermore, statements such as "inequality is a necessary condition for political violence" or "participation, contestation, and representation are necessary and jointly sufficient for democracy" inherit the vagueness of the predicates, whether these statements are posed as definitions or causal statements. Thus, both measures necessary for theorizing and theories themselves are subject to this sort of qualitatively bounded vagueness.

Fuzzy sets were proposed by L. A. Zadeh in 1965. One of his motives was to propose a mathematics that could help formalize linguistic concepts subject to degree-vagueness. He was particularly interested in providing a formal language for the social and behavioral sciences, although most of the subsequent development of fuzzy set ideas took place in systems engineering, Zadeh's home discipline. Indeed, just a year before Zadeh published his seminal article, Luce (1964: 376) noted deficiencies in set theory as a mathematical language suitable for addressing questions in the social and behavioral sciences. I quote at length:

> The language of sets does not always seem adequate to formulate psychological problems. Put so baldly, the statement is almost heretical since, in practice, set theory is the accepted way to formulate mathematical problems ... and, hence applied mathematical problems.... Certainly when I think about certain psychological problems, I wish it weren't the way it is. The boundaries of many of my "sets," and the ones that my subjects ordinarily deal with, are a good deal fuzzier than those of mathematics. [... For] example, we all deal effectively with the uncertainties of everyday life in terms of extremely imprecise concepts such as "likely," "fairly likely," and so on. As theorists, we often try to cope with this sort of behavior by phrasing it in the language of probability, but I suspect that most of us do not really feel that the mathematics meshes especially well with the problem. The categories of uncertainty are not really well-defined sets and their fuzziness is not particularly well summarized by probability notions. Perhaps we can make the existing concepts work, but I doubt that we should count on it.[2]

This article proceeds under the assumption that researchers must approach vagueness in a rigorous fashion. To proceed as sciences first, we must *systematize* our background concepts (Adcock and Collier 2001). This can be quite a formidable task, and it is far too often first paid lip service but then swept under the rug. It is somewhat more contentious to say that there is little that can be done to make these so-called background concepts themselves more precise. However, I believe that most readers would probably agree that the concepts will be "essentially contested" in the sense that reasonable people will disagree about their meaning, and it is unlikely that one agreed-on "true" definition will ever be devised (Gallie 1956). It may strike some quantitatively oriented scholars that off-the-shelf statistical models will come to the rescue. This view should be avoided. For instance, Bollen and Lennox (1991) show that dramatically different measurement models are implied by different conceptualizations, and conventional wisdom such as maximization of indicator correlation may not make sense for all concepts. I will not explicitly discuss the issue of concept formation further, but it lurks behind any analysis and deserves careful attention.

Fuzzy sets are one proposed method for managing vagueness. They can be used to help make analyses, perhaps ironically, less fuzzy because vagueness is managed formally. Like all other efforts at formalization, whatever else they may buy us, they can help lay bare assumptions and force researchers to be explicit about what exactly they mean. The membership function is the fundamental quantity

necessary to use fuzzy sets. It measures the (fractional) truth value of the statement, "Object $X$ is a member of set $A$." How the membership value makes precise the notion of partial set membership will be discussed below, though this is a matter of some contention and, much like probability, there is no agreed-on interpretation of fuzziness but instead multiple interpretations. The membership assignment task is, unfortunately, far from a trivial one, a problem shared with all other areas of social science, where measurement issues are never far away.

The goal of this article is to lay out options for membership assignment. There is as yet—and quite possibly never will be—no "cookbook." There are six sections that follow. The first informally lays out the basic properties of the membership function, attempting to answer what it is that it measures. Clearly, a few pages cannot do justice to this important topic, so readers are urged to consult references. The next three sections consider strategies for assigning membership that have been used by researchers employing fuzzy sets. *Direct assignment*, by far the most common method employed, uses a judge to provide a numerical membership value based on expertise. *Indirect assignment* also uses judges to provide membership. Unlike direct assignment, the subjects do not provide membership values but instead provide some other information that is used to construct membership values via a statistical model, often one subject to an optimization process of some sort. *Assignment by transformation* uses some mixture of substantive and mathematical concerns to create a mapping that takes one or more previously existing variables into a scale of membership. By "previously existing variables," I mean to include indicators such as life expectancy, scores on a diagnostic test, gross domestic product, and so on, although transformation is often necessary for subjective indicators as well, depending on how they were elicited from judges. Each of the topical sections includes a discussion of pros and cons of the assignment methods; the last two have examples that illustrate assignment in a variety of substantive contexts. It should be noted that the assignment methods are not mutually exclusive; indeed, aspects of each type may well be a part of any one assignment. The fifth section considers the essential but difficult question of validation. The final section offers some concluding remarks.

A caveat: This article may seem to some readers to be at least a bit schizoid. It is written at a mixture of technical levels, although

I have kept mathematics to a minimum. Nevertheless, some of the *ideas* are quite technical, and it is written from the general perspective of a psychometrician, albeit one with experience in macro-level cross-national research. Unfortunately, the audience is expected to be heterogeneous, ranging from quantitative sophisticates to scholars coming from largely qualitative research traditions. Points that are everyday staples to one readership might well be exotic dishes to another. It is hoped that the interdisciplinary spirit of those interested in fuzzy set theory will prove forgiving.

For general references, Torgerson (1958) is a classic that will prove a valuable source of practical insight for all one's measurement endeavors. The prose is particularly lucid throughout, and the introductory chapters bear multiple readings. Later developments such as axiomatic measurement theory (e.g., Roberts 1979) do not appear but can be examined later by interested parties. Wallsten et al. (1986) provide psychometric foundation for membership assignment by subjects. Bilgiç and Türkşen (1997) review elicitation as seen in the fuzzy sets literature, as do Klir and Yuan (1995). From a behavioral science perspective, Smithson (1987) and Smithson and Verkuilen (forthcoming) both have extended discussions of the relationship between interpretation of the membership and assignment. Finally, Ragin (2000) discusses fuzzy set theory in a primarily macro context.

## *2. THE MEMBERSHIP FUNCTION*

The key idea in a fuzzy set is twofold: A fuzzy set has (1) *qualitative boundaries* like a ordinary (crisp or classical) set with (2) *continuous variation* between these two poles. Many terms of discourse seem to work this way, and one of the main points of fuzzy set theory is to provide a faithful translation of theoretical statements into a formal language. For instance, the classical paradox known as The Sorities considers the meaning of the term *heap of sand*. (A variant uses *bald* with the base set *men's heads*.) When does a heap cease to be a heap? Nearly everyone would agree that the contents of a wheelbarrow of sand in the middle of one's living room would constitute a heap. Nearly everyone would agree that a few grains of sand in the middle of the same floor would not be a heap. The problem lies in finding

a firm cutting point between heap and not-heap. Any boundary we would pick would ultimately be arbitrary since adding or removing a few grains of sand would not constitute a *qualitative* change to the "heap-ness" of the sand . . . until only a few grains were left. Another popular example is the term *young*. It is difficult to find a qualitatively satisfying boundary between words such as *young* and *adult* or *adult* and *old*, even after a context has been specified precisely, despite the fact that in a given context, it is possible to identify cases that are clearly one or the other.[3]

Fuzzy sets dodge the problem of finding a clear-cut boundary by proposing a number, the membership function, that indexes the degree to which the object in question is in the set. It is customary (although not strictly speaking necessary) for this number to range from 0 (full nonmembership) to 1 (full membership). Formally, it is a function for an attribute $A$ over some space of objects $x \in \Xi$ (which may or may not be numerical) mapping to the closed-unit interval, [0, 1] (or some subset of it):

$$m_A(x) : \Xi \rightarrow [0, 1].^4$$

It is an index of "set-hood" that measures the degree to which an object with property $x$ is a member of a particular defined set $A$. It measures the fractional truth value of the proposition "$x$ is an element of $A$." If membership values are restricted to {0, 1}, the set is not fuzzy at all but instead referred to as *crisp* (i.e., it is an ordinary set).

Because it measures *subjective meaning*, a membership value is typically latent, in the wide sense discussed in Bollen (2002); the more restrictive definition based on the principle of local independence generally does not apply. It may also be subject to individual differences even within the same contexts, something that has been found empirically in efforts to scale memberships for probability terms (Wallsten et al. 1986). It is also unidimensional, being only one number. Of course, many concepts cannot be captured by only one dimension, and so multiple, disaggregated sets may well be necessary to accommodate the concept fully. Finally, the boundaries of the set are crucial (more on this below).

It is important to specify as clearly as possible what a membership is *not*, which in turn will help say what it is.

1. *A membership is not simply a quantitative variable of the interval level. Rather, its measurement level is complex and does not fit easily in the standard classification of scale types*. The fact that the membership function is continuous is important, but what makes a fuzzy set different from an interval variable is the fact that the endpoints matter as more than simply being nuisances as they are in regression based on the normal distribution applied to Likert-type items or sum scores, for instance. A fuzzy set measures continuous variation between qualitative poles, so the membership has a meaningful zero (no membership) and top (full membership). By convention, this is scaled into [0, 1]. The neutral point (membership = .5) is also typically considered important. In between these reference points, a fuzzy set might well be *ordinal*, having the not very well-known level of measurement termed "ordinal with natural zero" by Torgerson (1958: 16). That is, many fuzzy sets will not, strictly speaking, have the properties of a ratio scale but will still have meaningful zeros and tops. By contrast, an interval scale has no meaningful origin and thus can be subject to arbitrary positive linear transformation, which would destroy the endpoints. In sum, a membership value does not sit easily in the usual nominal-ordinal-interval-ratio-absolute classification because the reference points are so important. A simple way to think about it in familiar terms, an ordinary set is like a dummy variable, which takes on a value of 1 if the dummy variable condition is true and 0 otherwise. A membership generalizes this to include shades of gray in between and is thus a generalization of a dichotomy, not simply a standard interval scale variable.
2. *A membership is not a probability*. Despite being normalized to the unit interval, a fuzzy set is not a probability. They measure different things. A probability gives the mass of a particular event in a normalized space, while a membership is a generalized truth value. The most important property of probability is additivity (i.e., the fact that the sum over all events in the space sums to 1). There is no such restriction on memberships. The sum of memberships has an interpretation (fuzzy cardinality, which measures the size of the fuzzy set), but it could equal any nonnegative number up to the number of objects in the set. I distinguish the two because probability theory and fuzzy set theory can be useful adjuncts, and they should not be confused, although they have been. Statistical procedures are useful to construct membership functions and to construct analyses based on fuzzy sets (e.g., tests of necessity discussed elsewhere). Because it is not a probability, it does not share the interpretation in terms of bets that can

help clarify probability. An axiomatic basis exists for fuzzy sets, but it does not (yet) have the intuitive appeal and near-universal acceptance of the Kolmogorov axioms in probability. There is a deep connection between interpretation and assignment. See Smithson and Verkuilen (forthcoming, chap. 2) for more details. A special issue of *Fuzzy Sets and Sets and Systems* (Hisdal 1988) discussed the interpretation problem in great detail. See Klir and Yuan (1995) or Singpurwalla and Booker (2004) for more discussion of the relationship between fuzziness and probability.

Let me provide two examples to help fix the discussion above. Consider the standard pass-fail binary test item, which is conventionally represented as 0 = fail and 1 = pass. If we allow for one degree of "partial credit," it might make sense to score this as .5 for half right and half wrong—the exact numerical value depends on the scoring procedure, which is itself a rule to assign the response to a numerical scale. Other intermediate values are possible, representing different grades of correctness. The qualitative boundaries matter here since the student might be able to solve the item completely, to some degree, or not at all. Despite the gradation, an individual student's partial credit is not a probability. Instead, it represents the degree to which the student possesses the knowledge required to answer the question. Without new information, the student will probably get the same parts right and wrong if given the item again. Second, consider the notion of a developmental process such as physical maturation. We can agree when a child is physically immature or mature but recognize there are intermediate states between these two points where some aspects of maturation are present but others are not and want to talk about maturation as something that can be measured. Again, a child's value will be the same if measured in roughly the same period of time, up to measurement error, so whatever the concept maturation represents, it does not represent a binary gamble in an individual. (Whether it can be fruitfully treated as a gamble across children is a different matter.)

To quantify measurement error, it is reasonable to treat the membership function as a random variable, though there is not universal agreement on this point. For statistical purposes, the membership function is viewed as a random variable $M \in \{0, 1\} \cup (0, 1) = [0, 1]$. Because of its peculiar nature of continuous variation between two

qualitative endpoints, it is often of mixed type with a density of the form

$$\begin{aligned} f(m) = p_0\Delta(0) + p_1\Delta(1) \qquad & \text{endpoints: } \{0, 1\} \\ + (1 - p_0 - p_1)g(m), \qquad & \text{interior points: } (0, 1) \end{aligned}$$

where $g(m)$ is some continuous density in the unit interval, $\Delta(m)$ is Dirac's delta (i.e., the "spotting" function that has unit mass at a point $m$), and $p_0$, $p_1$ are the probabilities of being on the boundary. (The fuzzy set literature has been sloppy on this point at times, so the reader should be warned.) The beta density is a particularly convenient two-parameter family that includes many different specific cases of bimodality and unimodality with varying degrees of skew for memberships that are strictly continuous. A mixture distribution can be constructed to handle the endpoints or bimodality on the interior of the unit interval (Gupta and Nadarajah 2004). Indeed, the measures of shirking used in Brehm and Gates (1993) could be easily interpreted as direct assignment fuzzy sets, and they use a beta-dependent variable for regression models.

## *3. MEMBERSHIP BY DIRECT ASSIGNMENT*

In direct assignment, a judge provides a numeric or linguistic membership value "out of her head" after considering the objects and relevant evidence. Direct assignment of some sort is, of course, common across the social and behavioral sciences—for example, in magnitude scaling used in psychophysics or Likert scales, which are routinely given to subjects. One of the biggest selling points for direct assignment is low cost: For $k$ objects to be assigned, only $k$ values provided by the judge are necessary (or a multiple if replications are desired).

*I will emphasize right out that there is nothing inherently wrong with direct subjective assignment, although there are better or worse ways of doing it.* In many circumstances, particularly in more macroscale areas such as sociology, political science, or economic history, the likely error in subjective assessments is less than those found in seemingly objective indicators, which may have substantial bias. For

instance, it is well known that official statistics are often quite "soft," perhaps representing the story that the government wants to put out, not what is really happening on the ground, or perhaps simply representing reporting biases in the indicators that are not the results of conscious manipulation (e.g., the well-known underreporting of rape and domestic violence). A subjective scale might well be a better reflection of what is actually going on. In the case where meaning differs across contexts, some kind of expert adjustment might well be necessary for comparisons across units even with objective indicators (Przeworski and Teune 1970). Finally, it might well be that nothing else will do (e.g., in a historical study where hard data are simply unavailable). Direct assignments are also highly useful to check assignments by other means for consistency, particularly in establishing the reference points of full membership, nonmembership, and the neutral point or establishing the general shape of a curve to be used as a means for assigning a numerical variable to a membership scale.

Having said that, I will note its deficiencies. There are five main problems with direct assignment:

1. The first is one is interpretation. Simply put, interpretation of directly scaled numbers is difficult since it is rare that something concrete underlies the number, although some methods for assignment based on combining more concrete variables and/or the use of careful, systematic, and public coding rules can alleviate it. Still, interpreting what a membership value means can be difficult, particularly for the sorts of abstract sets that are used in the social sciences that are not based on an underlying quantitative variable with meaningful units. Interpretation of memberships is a contentious issue, but it is particularly difficult to interpret the meaning of numbers when they are not related to more easily interpreted variables.
2. The second main problem is that direct assignment may be too hard for the judges to do reliably, particularly for very abstract concepts such as economic development, democracy, or physical maturity. This relates intimately to the first problem, of course. If this is the case, it is usually preferable to break the concepts down into components, subjectively assign membership for the components—preferably according to explicit rules—and then reconnect the components according to a model. For instance, the Human Development Index (HDI; discussed more fully below) breaks development into three components and connects them using an average to form a composite top-level

index because the index creators felt that a compensatory model was appropriate. Other aggregators are possible. Indeed, fuzzy set theory provides a great number of them that accommodate disjunctive ("or," which represents a situation of redundancy among components), compensatory ("average," which represents trade-offs among components), and conjunctive ("and," which represents a lack of substitution among components) models naturally (Zimmerman 1993). While the indicators used in the HDI are objective variables, there is no in-principle reason a model could not be used with subjective indicators (although see validation below). Alternatively, an indirect scaling procedure that involves a simpler cognitive task for judges makes sense. Cross-modal matching, which involves relationships between a pair of direct measurements, is a possibility (Baird and Noma 1978).

3. The most relevant criticism of direct assignment with regards to other methods of assignment is that it is frequently biased. Of course, the purpose of direct assignment is to tap into a judge's expertise, which in a sense *is* bias. There are, of course, different sources of bias, and many are nuisances, not expertise. Many are common judgment biases discussed in detail in sources such as Poulton (1989), who summarizes more than a century of experience by psychophysicists in eliciting numerical responses from human subjects. As Baird (1997) notes, no method of elicitation of subjective numerical estimates seems to be free of known and potentially important biases. Huber, Ariely, and Fischer (2002) show this in the context of utility measurement in a principal-agent task, so even careful instructions given to disinterested judges do not protect against judgment biases. Careful instructions provide control over the *content* of the subjects' biases, however, and make them predictable to the investigator. This comes with the cost of forcing the subjects to respond according to the preexisting scheme.

   For instance, one prominent bias relevant to membership elicitation noted by Thole, Zimmerman, and Zysno (1979) is the *endpoint bias*. Subjects will systematically bias membership values away from the interior of the membership interval toward the endpoints compared to membership assigned via an indirect procedure (paired comparison) that is considered more reliable, with the bias larger for points closer to the neutral point $m = .5$. They suggest applying the arcsine square root transformation,

$$m' = \frac{2}{\pi} \arcsin \sqrt{m}.$$

It is often used as a variance-stabilizing transformation in the analysis of proportions, as it moves assigned membership values $m$ away from the endpoints, giving corrected values $m'$. Any other contrast diffuser would also work. (Naturally, this sort of transformation should not be applied pro forma.) Other studies demonstrate substantial method bias from numerical elicitation methods in direct measurement tasks. Chameau and Santamarina (1987) have a discussion specifically related to the elicitation of memberships. On a related point, generally there is no easy way to test whether axioms of measurement such as unidimensionality, weak ordering, and continuity are met since they are met by assertion.

4. Most direct scaling methods do not generate uncertainty estimates that would allow users to put error bars on assigned scores. This is unfortunate because all measurements have uncertainty attached to them, and ignoring measurement error has potentially serious consequences. It is simply honest science to make the level of precision of scores public, and this is not done frequently enough in practice. A very simple procedure is to elicit a range of possible values from the judge (e.g., low, medium, and high values of membership for each object). Hesketh et al. (1988) use a variation on the semantic differential to generate exactly this sort of information from judges. There are also direct scaling procedures such as the "staircase method" that generate uncertainty estimates as a by-product of the elicitation process (Cornsweet 1962; Tversky and Koehler 1994). There are ways to generate uncertainty estimates by simulation, so direct assignments by a single judge can be handled in some fashion (Smithson and Verkuilen forthcoming, chap. 2, provide an example).
5. Combining the results of direct assignments by multiple judges is, of course, useful but also more difficult than one might think. Wallsten et al. (1986) found substantial individual differences among subjects in their experiments assigning membership values for linguistic probability words such as *likely* or *unlikely*. They note that the individual differences are strong enough that they do not recommend averaging across values to generate a composite membership function (which is common practice in fuzzy set applications) since the standard deviations would be quite wide. More important, the differences probably reflect systematic differences in meaning across subjects. For empirical scientists, individual differences might be a blessing in disguise, however. If several judges disagree widely about an object, it is at minimum a sign of a matter for dispute and thus further study.

I do not consider an example here to conserve space. Other articles in this special issue make use of direct assignment, and I refer readers to them. Ragin (2000) also provides several examples.

## *4. MEMBERSHIP BY AN INDIRECT SCALING MODEL*

Indirect scaling elicits responses of some kind about the objects to be scaled from judges (broadly speaking) and then applies a model to the judgments to generate scale values. I focus on models based on paired comparison because the technique is common in fuzzy set contexts, but it should be noted that there are other scaling models. For instance, Healy and Goldstein (1976) use an variant of multiple correspondence analysis/optimal scaling to score the developmental process of physical maturation in a manner consistent with a membership function. Manton, Woodbury, and Tolley (1994) use maximum likelihood methods in their Grade of Membership program, which does a fuzzy set version of latent class analysis.

One reason to focus on paired comparison is that it represents a "gold standard" for membership assignment, as shown by Wallsten et al. (1986), where ratio scale properties are demonstrated for their procedure through a test of measurement axioms.[5] Thus, it represents a useful check on direct assignment, even if direct assignment is most commonly done in real practice.

Before diving into the example, I consider the pros and cons of indirect scaling. In many ways, the debate between direct and indirect scaling is as old as scientific psychology. The founders of psychophysics (the study of the subjective perception of physical variables), such as G. T. Fechner in the nineteenth century, used indirect scaling to determine empirical laws for subjective perception. In the 1920s, it was noted by L. L. Thurstone that it is possible to scale an abstract continuum representing utility or other mental constructs that do not have direct parallels in the physical world. Generally, these researchers also felt that subjects could not give meaningful answers directly but instead used indirect scaling procedures of various kinds (e.g., the law of comparative judgment). In the 1950s, S. S. Stevens showed that it is often possible for subjects to provide

meaningful—or at least reasonably consistent—numerical responses to abstract continua common to social sciences.[6] Baird and Noma (1978) have an excellent review of the history with extensive citations. Economics had parallel developments in utility theory.

The early psychophysicists adopted indirect scaling because they felt that subjects (often themselves) would not be able to give reasonable responses. The reason in the modern day (given that direct scaling has been established as a workable alternative) is that types of indirect scaling were devised to address the flaws listed for direct scaling. Point 2 above notes that direct scaling may be too hard for subjects to manage reliably. Indirect scaling often substitutes more, cognitively easier tasks for fewer, harder ones in a direct scaling. Point 3 notes that the axioms of measurement for a given scale type are typically *asserted* in direct scaling but can often be *tested* in indirect scaling. It is obvious that assuming less and testing more is generally desirable, so indirect scaling provides a means to that end. Point 4 notes that uncertainty estimates are not typically provided in direct scaling. Most indirect scaling procedures generate error estimates as a by-product of the model-fitting process (e.g., maximum likelihood), which gives error estimates from the information matrix. Where they do not, it is possible to use resampling or permutation.

Of course, the downside of most indirect scaling methods is high cost in terms of both data gathering and model formulation. As mentioned above, direct scaling is dramatically cheaper if there are many objects, although of course, this cheapness comes at the cost of making more and/or stronger assumptions. It should be noted that indirect scaling is most beneficial precisely where it is most needed—when the objects in question are confusable by reasonable judges, to some degree, but there is still an underlying gradient to be found. Many scaling procedures break down in the presence of "perfect" data because they depend on the presence of variation. Thus, indirect and direct scaling can certainly be complementary. For instance, an indirect scaling technique could be used to establish a "ruler" over important anchoring objects, with the more inexpensive direct assignment being used to fill in other objects relative to that ruler. It is possible, indeed desirable, to use one method to check the other.

*Example 1*. This example uses a paired-comparison procedure that can be implemented using any statistical package with a logistic regression program (e.g., SAS, Systat, Stata, or SPSS). It is based on the Bradley-Terry-Luce (henceforth BTL) model for the two-alternative forced-choice experiments. It is simpler than the procedure from Wallsten et al. (1986) but similar in spirit. It is very similar to Case V of Thurstone's law of comparative judgment, differing only in some relatively minor distributional assumptions (Baird and Noma 1978). In the paired-comparison setup, a subject answers the following question for the $k(k-1)/2$ pairs that can be formed of $k$ objects: Which of the two objects you are presented possess the attribute more?

The big advantage of the BTL paired comparison is that the decision that subjects need to make is simple, and the way they compare objects is more predictable than in a direct procedure. It is a binary, yes/no decision rather than a direct number. Assignment by paired comparison tends to be a lot more consistent than direct assignment. Furthermore, the scaling model provides useful diagnostic information about the measurement axioms that direct assignment asserts. For instance, it is too much to expect that a subject will be perfectly consistent across all pairs as would be required by the weak ordering implied by the numerical structure of a membership function. The scaling model uses an explicit loss function (here maximum likelihood for the logistic distribution) to quantify the quality of the solution of the scaling model. It also generates standard errors for the scale values, which provides further useful information.

The big disadvantage of the paired-comparison procedure is that it is tedious because each choice made by the subject gives relatively little information. The number of necessary comparisons, $k(k-1)/2$, increases rapidly as $k$ increases: For 5 objects, the number of pairs is 10; for 10 objects, it is 45; and for 20 objects, it is 190. Furthermore, replication puts an even larger demand on subjects. Böckenholt (2001) notes that the hierarchical modeling framework provides some substantial advantages in terms of estimation and economy of design (other incomplete designs exist), as well as a means to handle individual differences among raters. Since this is often a problem in practice (i.e., "simple scalability" will typically fail), the BTL model is probably too simple for a real-world problem, but it is a useful example.

In a fuzzy set context, it is necessary for the investigators to decide (somehow) what objects are full members and nonmembers to anchor the scale. These objects should be included in the scaling task for subjects for two reasons. First, they provide anchors for the subjects and thus ensure that the assignments are valid. Second, a researcher's understanding may differ substantially from subjects', and the scaling task provides an opportunity to find this out! This example uses paired comparison to generate a fuzzy set "prestigious medical occupations" over 10 different occupations in the following set: {general practitioner, specialist, surgeon, nurse, orderly, emergency medical technician (EMT), janitor, health educator, lab tech, admissions clerk}.[7] The BTL model uses the paired-comparison frequencies to estimate the probability that object $i$ will be chosen over object $j$. As developed in Luce (1959), this probability is modeled as

$$p_{ij} = v_i/(v_i + v_j),$$

where $v_k$ are the utilities of the $k$th object; note particularly that if the utilities are equal, $p_{ij} = .5$, which naturally represents the state of indifference. The BTL model assumes that the pairwise probability of choice depends only on the two utilities. The choice of $v_k = \exp(u_i)$ has a number of desirable properties. It leads to a logistic regression since

$$\begin{aligned} p_{ij} &= \exp(u_i)/(\exp(u_i) + \exp(u_j)) \\ &= 1/(1 + \exp(u_j)), \end{aligned}$$

which is the cumulative distribution function of a standard logistic distribution. The design matrix is of the form shown in Table 1 (note the absence of an intercept), and the dependent variable = 1 if object $i$ is preferred to object $j$ and 0 otherwise. Readers should consult sources (e.g., Böckenholt 2001; Agresti 2002) for details on fitting the simple BTL model.

Parameter estimates from the BTL model, along with the rescaled membership values, are shown in Table 2. White-corrected standard errors are also included (not all packages will use White-correction). The solution fits very well, with McFadden's pseudo-$R^2 = (\ln L_{\text{initial}} - \ln L_{\text{final}})/\ln L_{\text{initial}} = .59$. Membership values have been generated by rescaling the BTL scores into the unit interval using

$$m_i = (u_i - \min\{u_i\})/\text{range}\{u_i\}$$

**TABLE 1: Sample Bradley-Terry-Luce (BTL) Model Design Matrix for Four Objects**

| Object 1 | Object 2 | Object 3 | Object 4 |
|---|---|---|---|
| 1 | –1 | 0 | 0 |
| 1 | 0 | –1 | 0 |
| 1 | 0 | 0 | –1 |
| 0 | 1 | –1 | 0 |
| 0 | 1 | 0 | –1 |
| 0 | 0 | 1 | –1 |

**TABLE 2: Bradley-Terry-Luce (BTL) Estimates With Rescaled Membership Values**

| Occupation | $u_i$ | $SE(u_i)$ | → | $m_i$ | $SE(m_i)$ |
|---|---|---|---|---|---|
| Janitor | –3.29 | 0.43 | → | 0.00 | 0.06 |
| Orderly | –2.56 | 0.39 | → | 0.10 | 0.05 |
| Admissions clerk | –2.02 | 0.37 | → | 0.17 | 0.05 |
| EMT | –1.26 | 0.34 | → | 0.28 | 0.05 |
| Lab tech[a] | 0 | — | → | 0.45 | — |
| Health educator | 0.97 | 0.33 | → | 0.58 | 0.04 |
| Nurse | 1.53 | 0.34 | → | 0.65 | 0.05 |
| Specialist | 2.54 | 0.37 | → | 0.79 | 0.05 |
| General practitioner | 2.59 | 0.38 | → | 0.80 | 0.05 |
| Surgeon | 4.08 | 0.47 | → | 1.00 | 0.06 |

[a] Value constrained to equal 0 to identify model.

for each object; the standard errors are transformed by dividing by the range. In essence, the interval scale generated by the BTL model from the paired comparisons is "promoted" to ratio by this rescaling, but since the objects that are assigned membership values 0 (janitor) and 1 (surgeon) seem to be sensible given the domain of objects under consideration, this assumption does not seem too bad. It may not be justified in other circumstances. For instance, someone else's understanding may have all MDs as the anchor for full membership, which would merge the three occupations with highest status and stretch out the rest of the scale. Intermediate values are scored by the model, which uses the pattern of preferences among objects to interpolate between the two extremes. Figure 1 shows the membership values with approximate error bars (±.10) attached. Objects relatively close to each other on the membership scale are probably not

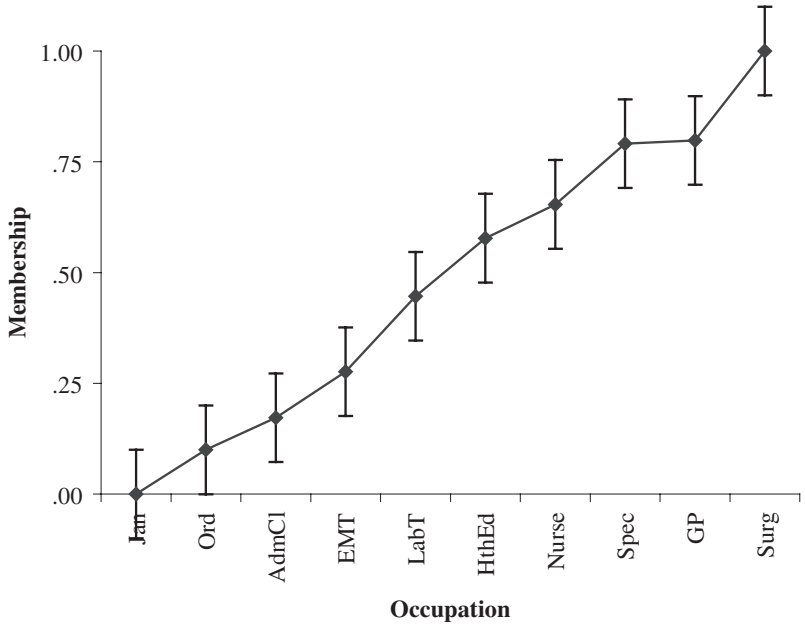


**Figure 1: Occupations Example Membership Values With Error Bars**

distinguishable; caution should be taken in drawing any conclusions depending on knife-edged distinctions among them.

## *5. MEMBERSHIP BY TRANSFORMATION*

The basic idea in transformation is to take a numerical variable and map it into membership values with a theoretically motivated transformation. In a sense, this strategy epitomizes what Torgerson (1958) calls *measurement by fiat*. The term *fiat* has a connotation of arbitrariness, although as he notes, measurement by fiat is frequently required if inquiry is to go forward in the absence of fundamental measurement (Torgerson 1958: 21-5). As has been seen in Example 1, transformation is often necessary given data gathered by a subjective scaling procedure. I will illustrate primarily by using examples as the number of possible transformations is literally endless. Indeed, the numerous

ways of proceeding is precisely the main flaw of measurement by fiat since choices—which may or may not have important downstream consequences—will often be motivated by nothing more than convenience. Before going to the examples, however, I will discuss some general issues.

Transformation often makes use of statistical data gathered for other reasons. Thus, it is (potentially) very cheap. The fact that objective indicators such as gross domestic product (GDP), psychological test scores, and so on can be used is also a strength. These variables are often have much more nuance than simple subjective indicators that might have, say, five or seven points. They are usually much more generally interpretable in terms of variables that we understand (or think we understand). On the other hand, existing variables may not be very good, but the notion that they are "objective" as opposed to "subjective" has a seductive ring. Carefully researched and documented subjective indicators are often much better data than objective indicators, particularly when data generation is out of the hands of the investigators, as it is in the case of official statistics or other secondary sources. At minimum, uncertainty estimates in the original variables should be propagated through the transformations to give uncertainty bounds for the transformed membership. Simple methods from mathematical statistics exist to do this—for example, the delta method for approximating the variance of a nonlinear transformation $g(x)$ of random variable $X$, $\text{var}(g(x)) \approx \text{var}(x)|g'(x)|^2$.

Transformation is often necessary in direct or indirect subjective elicitation. The natural range of output of many scaling or elicitation procedures is not the unit interval. For instance, one way to avoid the endpoint bias mentioned above under direct scaling is to ask subjects to provide magnitude ratings, which have a response range $(0, \infty)$. These scores would then have to be translated into the unit interval to be used in a fuzzy set–based analysis, and there are many possible transformations that would do the job. Furthermore, the investigator will need to make decisions about issues such as subnormality—that is, whether the fuzzy set actually contains objects with memberships 0 or 1 (in direct scaling assignment, subnormality turns out to be common). Transformation, then, will almost certainly be needed to establish the boundaries of the fuzzy set.

Of course, the issue of boundaries is in its own right potentially contentious. It is well known that ceiling or floor effects cause trouble in conventional data analysis, and many of the transformations that are applied to generate a membership function could be seen as throwing away variation by introducing an active ceiling and/or floor. For instance, the Human Development Index, discussed more fully below in Example 2, applies a linear filter to two variables (GDP and life expectancy). Any variation outside the boundaries of the filter is simply chopped off and mapped to full membership or nonmembership. Of course, the main point of a fuzzy set analysis is to maximize theoretical fidelity in the formalization of verbal theory. To the extent that conceptual boundaries are often a part of a theory, they should be represented even at the cost of variation.

However, it is essential that the connection of variable to membership be argued explicitly, which means that the concept that the membership is measuring must be understood clearly so that an appropriate mathematical transformation can be chosen. For instance, essentially monotonic "more (or less) is better" concepts, such as "young" or "old," are different from one such as "middle aged" since, as arrayed along the axis age, young monotonically decreases and old monotonically increases, while middle age is unimodal, representing an ideal point "just right" structure around, say, 50. This is illustrated in Figure 2. Klir and Yuan (1995) discuss this point (and the issue of normalizing the fuzzy set). It is often implicit in the theory underlying a membership assignment that there is diminishing returns in membership near the endpoints. Many transformations used to represent "more is better" type concepts (e.g., life expectancy as a numerical stand-in for health development) will tend to be sigmoid shaped, representing the squeezing action of the endpoints. Similarly, for a "just right"–type concept, a bell-shaped curve in which membership falls off from the ideal point first slowly, then fast, then slowly again, will often be called for; it should be noted that this may not be symmetric. Direct scaling is often a useful way to get an idea of the kind of transformation needed.

As I said above, the list of possible transformations, even ones that match the necessary qualitative properties, is endless. Nevertheless, we can make an important distinction between purely a priori/theoretical transformations and data-based ones. In a purely

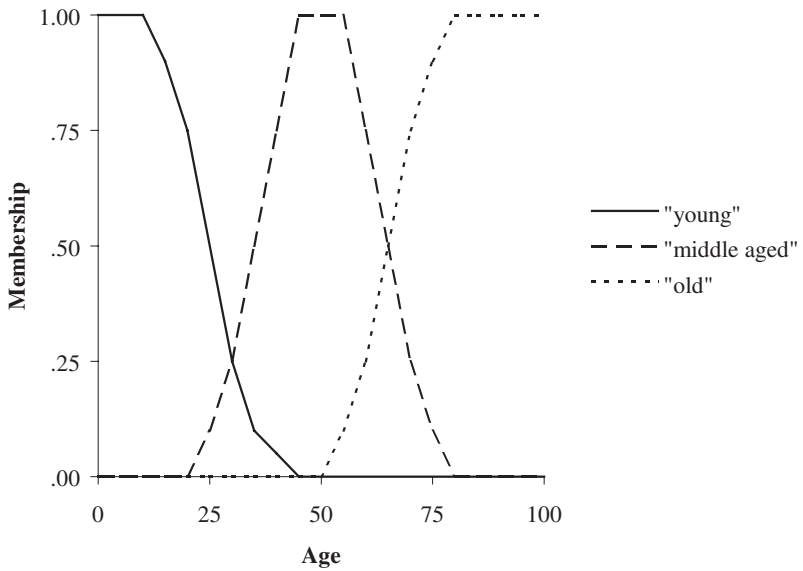


**Figure 2: "Young," "Middle Aged," and "Old"**

a priori/theoretical transformation, the entire transformation is specified in advance and does not depend on the data at all. The HDI transformations are like this. In a data-based one, the membership values depend on the relative distribution of data in some fashion. For instance, the membership values assigned by the use of the BTL model in Example 1 are data dependent. Another data-dependent strategy uses the cumulative distribution function ( cdf ), $F(x)$, of the variable in question, for instance, assigning $m(x) = F(x)$. Of course, most cdfs are sigmoid shaped and all weakly monotonically increasing, so a cdf is appropriate for a more-is-better concept.

Two examples will be considered here. The first is the United Nations Development Program (UNDP) Human Development Index, or HDI, which illustrates a macro-level problem (United Nations Development Program 2004). It is not an well-advertised fact, but fuzzy set theory underlies the construction of the HDI; indeed, fuzzy set theory has been important in the literature on the measurement of poverty. The second considers some data generated by

Muller (1972) on subjects' attitude toward and willingness to engage in political violence. It illustrates the use of survey data in a fuzzy set analysis.

*Example 2*. The purpose of the HDI is to provide a more valid, conceptually rich measure of human development. In the study of development, it is typical for a single indicator such as gross domestic product per capita or energy expenditure per capita to be used as a proxy for development. The authors of the HDI wanted to recognize that wealth is only one aspect of human development alongside others, getting a complete picture. The normative theory behind the HDI draws heavily on work by Sen (1999) and others and is discussed in detail in the methodological appendix of the annually issued Human Development Report (United Nations Development Program 2004; Fukuda-Parr and Kumar 2003). Cerioli and Zani (1990) provided the basic logic as it relates to fuzzy sets, and Qizilbash (2003) is a recent, critical review. A number of variations of the index in use focus on different aspects of development or focus on different cases, but we will only discuss the basic HDI here. The basic index was designed to be applied to all countries in the world, and thus it does a good job discriminating among low, medium, and high levels of development but not necessarily within a given range.

The index authors disaggregated the top-level concept, development, into three components: economic, health, and education. To combine these components, each needed to be put on a common scale. The unit interval was chosen. In addition, the authors felt that certain lower and upper "goalposts" represented important key points on the continuum of development. A country with a value above the upper goalpost could be considered fully developed on that component. Conversely, a country with a value below a lower goalpost could be considered fully undeveloped on that component. Variation between the goalposts was important, but outside, it was not. The basic strategy was to use a linear filter to assign membership. Table 3 shows the components, the indicators chosen to measure them, the goalposts, and the equation (a linear filter) used to assign membership for each component. The top-level index is a simple average of the three components. Of course, other choices of aggregator might make sense (e.g., a conjunctive one based on the minimum), which would represent a "weakest link"–type conception of development,

**TABLE 3: Human Development Index (HDI) Example Component Membership Assignments**

| Component | Indicator | Goalposts | Membership Between Goalposts |
|---|---|---|---|
| Economic | Log gross domestic product (GDP) per capita ($PPP) | ($100, $40,000) | $\text{econ} = \ln(\frac{GDPpc}{40{,}000})$ |
| Health | Life expectancy at birth | (25 years, 85 years) | $\text{health} = \frac{LE-25}{85-25}$ |
| Education | Adult literacy rate and gross enrollment | (0 percent, 100 percent) | $\text{educ} = \frac{2}{3}AL + \frac{1}{3}GE$ |

where deficiencies in one area of development could not be made up for by proficiencies in others. While all reasonable transformations will be strongly monotonically related, the relationship of the differently aggregated HDI values to other variables may well differ quite a bit.

One glaring flaw of the HDI that has become apparent in practice is that it lacks an uncertainty estimate.[8] Does anyone *really* believe that the life expectancy at birth is known without error in many sub-Saharan countries, for instance? Even in the developed world, it is known that economic statistics are not entirely accurate due to the presence of a black and gray economy often making up a nontrivial amount of economic activity. No uncertainty estimates are provided with the HDI scores even though three decimal places are reported. This is a major shortcoming.

*Example 3*. The second example of transformation uses survey data from an article by the late Edward N. Muller (1972) on the relationship between subjects' approval of political violence and their intention to engage in political violence. The data have been extensively analyzed by Smithson (1987), who in turn got them from Hildebrand, Laing, and Rosenthal (1977), but I will consider issues that previous authors did not. With this example, I hope to illustrate some of the choices that need to be made when using discrete ordinal data—very common in practice—to construct fuzzy sets.

There are two variables in Figure 3, approval for political violence (henceforth APV) and intention to engage in political violence (henceforth IPV). They are Guttman scales formed from survey data taken

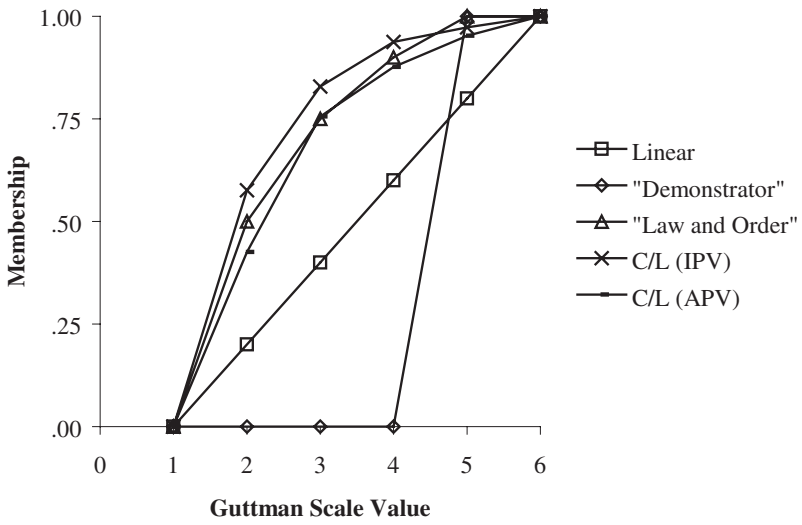


**Figure 3: IPV × APV From Muller (1972)**

NOTE: APV = approval for political violence; IPV = intention to engage in political violence.

in Iowa in the early 1970s. I will describe them more fully below.[9] While Muller (1972) notes that the Guttman scale did not fit perfectly, it is certainly the case that these items contain much more content than the typical survey instrument. In particular, as will be discussed below, the items that make up the Guttman scale give us substantial advantage in deciding how to assign membership in a way that has real meaning, particularly at the endpoints of the membership scale.

It is useful to take a look at the data before going any further. First note that in general, this population does not approve of political violence, as most of the cases are concentrated in the lower ends of the two variables. In his article, Muller (1972) hypothesized that APV was a necessary condition for IPV, which is certainly a reasonable hypothesis stating that the relevant attitude proceeds willingness to engage in a given behavior. In set-theoretic terms, APV $\supseteq$ IPV. As discussed in Smithson and Verkuilen (forthcoming), this in effect states that the membership value for APV puts a ceiling on that of IPV or, equivalently, that $m_{\text{IPV}} = m_{\text{APV}\cap\text{IPV}}$. Simple examination of the data table shows this to be the case. The frequencies in the cells

where the scores on APV equal those on IPV are underlined, and it is quite clear that most cases lie on or below the diagonal (482 of 499, or 97 percent, with 249, or 51 percent, strictly below). However, to provide a formal test of necessity—or even the informal eyeball test—we must be able to assert with confidence at minimum that APV and IPV are on the same scale. Otherwise, the statements just listed are simply fiction because the scales would not be calibrated and cannot meaningfully be compared directly. This is in contrast with ordinary regression, where the slope has units that convert between units of the independent and dependent variables. (But see Smithson and Verkuilen [forthcoming, chap. 5], who discuss an essentially purely ordinal method for assessing whether necessity holds based on conditional quantiles that does not depend on scale comparability.)

Here is where the richness of the data underlying the scale comes into play. To construct the scale, subjects were asked whether they approved of the following acts:

(a) Protest marches or meetings permitted by authorities
(b) Disobeying an unjust law
(c) Engage in sit-ins or takeovers to disrupt the government
(d) Rioting, fighting the police, or destroying property
(e) Armed insurrection against the authorities

Then they were asked whether they would engage in such acts, might engage in such acts, or not. It seems on the surface that these items should be ordered in terms of "difficulty." It is easier to agree with (c) than (e), for instance. The Guttman scaling technique tests to see if these items can in fact be ordered, such that if one "passes" a higher numbered item, all lower numbered items will also be passed.[10] Except for some difficulties with (a) and (b), this was indeed found to be true. Someone receiving a 1 on either scale disapproved of all items, someone receiving a 2 approved of (a) but not the rest, and so forth. Thus, the numbers here have substantially more meaning than is the case with typical rating scales, where we have good reason to believe that particular numbers might mean different things to different subjects. Furthermore, the scale values are interpretable in terms of concrete acts or beliefs. Having a scale value of 4 *means something*—namely, that a person approves of sit-ins and other efforts to disrupt government, as well as lesser acts, but does not approve of

| IPV \ APV | 1 | 2 | 3 | 4 | 5 | 6 | Total |
|---|---|---|---|---|---|---|---|
| **6** | 0 | 0 | 0 | 0 | 2 | 5 | 7 |
| **5** | 0 | 0 | 1 | 1 | 3 | 4 | 9 |
| **4** | 0 | 0 | 3 | 13 | 8 | 4 | 28 |
| **3** | 0 | 5 | 38 | 8 | 11 | 2 | 65 |
| **2** | 5 | 75 | 45 | 17 | 4 | 2 | 148 |
| **1** | 97 | 89 | 43 | 9 | 2 | 2 | 242 |
| **Total** | 102 | 169 | 131 | 48 | 30 | 19 | 499 |

**Figure 4: Five Possible Membership Assignments for Muller (1972) Data**

rioting or insurrection. I also think it is fair to say that the two scales are comparable in that the subjects were asked about the same acts, with sufficient anchoring information provided by the items.

In general, it is best to use as weak a measurement assumption as one can get away with. To answer whether APV $\supseteq$ IPV, we only need ordinal information and thus could stop at the cross-table given above. How would we assign membership if we wanted to create a ratio scale of membership, which might be the case if linguistic hedges were to be used? I believe that it is clear from the content of the items that each scale value 1 should have membership 0, and scale value 6 should have membership 1. Thus, we have at minimum an ordinal scale with meaningful endpoints. Intermediate values are substantially more difficult and depend on one's notion of how different particular acts relate to the underlying concept. Furthermore, the content of the items seems to demand that $m_{\text{APV}}(x) \leq m_{\text{IPV}}(x)$ since these are "more is better" concepts. Smithson (1987) simply assigned scale value 1 to have membership 0, 2 to 0.2, . . . , 5 to 0.8, and 6 to 1. That is, he set $m(x) = (x-1)/5$ for each variable, a linear assignment, and then tested containment based on that assignment.

However, it might be that other understandings of the concepts will dictate different transformations. Figure 4 illustrates five possibilities,

**TABLE 4: Scaled Euclidean Distances for Five Membership Assignments**

| | Linear | Demonstrator | Law | CL(IPV) | CL(APV) |
|---|---|---|---|---|---|
| Linear | — | | | | |
| Demonstrator | .29 | — | | | |
| Law | .12 | .60 | — | | |
| CL(IPV) | .16 | .68 | .00 | — | |
| CL(APV) | .10 | .58 | .00 | .01 | — |

NOTE: APV = approval for political violence; IPV = intention to engage in political violence; CL = Chelli-Lemmi assignment.

each of which has scale value 1 with membership 0 and scale value 6 with membership 1. One is linear. It would be a common default, though not necessarily a sensible one. "Demonstrator," by contrast, assigns 0 membership to scale values 1 through 4 and 1 to scale values 5 and 6. In effect, this assignment states that any protest activities up to but not including rioting are not at all political *violence*. Note that "Demonstrator" does not give a fuzzy assignment at all. By contrast, "Law and Order" reflects one possible membership function for a "law and order" person since any deviation from scale value 1 increases membership dramatically. The next two assignments are based on the data. I use the proposal of Cheli and Lemmi (1995), which transforms the empirical cdf of the given variable to generate a membership. The assignment is given by

$$m(x) = \max\left(0, \frac{\hat{F}(x) - \hat{F}(x_0)}{1 - \hat{F}(x_0)}\right),$$

where $x_0$ is a cutoff value for minimum membership; here, $x_0 = 1$. Note that both data-dependent assignments are very similar to the a priori "Law and Order" assignment and distinct from "Demonstrator." Table 4 shows scaled Euclidean distances between each membership assignment. Of course, norm-based procedures such as multidimensional scaling could be used to study differences in a larger problem.

Unfortunately, without a solid theoretical justification, we have no really good criteria to decide on a particular transformation, which is clearly a dilemma a researcher would have to face in practice. Many social science theories do not claim more than monotonicity and so provide little basis to preferring one transformation over another. One

way to dodge this issue is to make sure that the conclusions drawn do not depend on the particular assignment but instead are invariant under monotonic transformation. In other words, we would want to use an ordinal comparison method such as the conditional quantile method found in Smithson and Verkuilen (forthcoming), which does not depend on the scale values. However, if a numerically based procedure is desired, the assignment may well matter. In this case, a sensitivity analysis is definitely in order to show that the conclusions hold up over varying assignments.

One additional troublesome aspect of these data is that they are discrete. This is, of course, common in social science. The trouble comes because some of the necessity tests proposed are sensitive to boundary cases. For instance, Ragin (2000) suggests that, for a test of necessity, all cases with membership 0 on the causal/including set should be excluded.[11] If many cases are mapped to the boundary, a substantial proportion might be excluded from analysis. The status of observations with tied membership scores is also potentially a matter for concern. With continuous data, boundary cases and ties are unlikely to occur, and thus few cases will be excluded.[12] However, with discrete data, it is likely that many cases will be either on the boundary or tied. For instance, after transforming the Muller data, this could be a substantial number of cases. "Demonstrator" applied to APV, for instance, would end up excluding 450 cases from analysis by assigning them to 0 membership. Many methods of ordinal analysis also exclude ties. At present, I do not believe there is a satisfactory solution to the problem of ties aside from care. If the data are coarse-grained, it might be better to use methods specifically designed for discrete data, such as those in Hildebrand et al. (1977).

## *6. VALIDATION*

All measures need to be validated in the sense that it is incumbent on the investigator to demonstrate that the concept being measured is indeed measured. The topic of validation is very broad and not amenable to a simple treatment. Nevertheless, I would be seriously remiss not to remark on it. As with any very broad concept, validation is multifaceted. I focus on two aspects, *internal validation*, primarily

focusing on measurement axioms, and *parallel validation* of multiple measures of a given fuzzy set. The process of assignment has implications for what sorts of validation are possible.

In internal validation, the researcher is concerned with the degree to which a set of desirable axioms is satisfied by a given measure. For instance, a metric space satisfies three axioms for all triples of objects in the space $x$, $y$, $z$:

1. Identity: $d(x, x) = 0$;
2. Symmetry: $d(x, y) = d(y, x)$;
3. Triangle inequality: $d(x, y) \leq d(x, z) + d(y, z)$.

To test these axioms, it is necessary to consider different pieces of information. To test *all* the axioms of a metric space, it is necessary to have all $k^2$ comparisons between pairs of objects, including self-comparisons. Testing symmetry requires all non-self-comparisons to see if they are consistent. Testing the triangle inequality only requires the lower triangle of the comparison matrix. If it is desirable to test all the axioms, it is necessary to gather sufficient data. Data gathered from direct numerical ratings of objects (as opposed to pairs of objects) do not generally allow a test of the three metric space axioms at all. Other axioms such as cancellation and double cancellation from conjoint measurement will have similar requirements to be testable. Generally testing these axioms is costly since more data will be necessary, often a lot more than if one uses cheaper methods and simply avoids testing axioms altogether. There is no free lunch.

In parallel validation, the researcher has $j$ membership assignments to the same set $A$, arrived at by different procedures. The purpose is to determine whether they are parallel in the sense that they all measure the same thing, up to noise. This is exactly the traditional psychometric concern, but how this agreement is assessed typically needs to be altered to suit the situation. In the context of numerical membership assignment, Pearson correlation is too weak a measure because it ignores location and scale shifts that are important components of the agreement between different membership assignments. Thus, it makes sense to use a stronger standard, such as Lin's (1989) coefficient of agreement, which shrinks the usual Pearson correlation by an "accuracy" factor that ranges from 0 to 1, depending on the degree of difference in location and scale between the variables in question.

A graphic procedure such as a scatter plot matrix combined with numerical estimators is the best approach. The reference line $y = x$ shows deviation from perfect agreement, and this is in fact exactly what Lin's coefficient measures. Congruence coefficients are also reasonable.

One set of specifically fuzzy set axioms that can be subject to test is whether the operators of fuzzy set theory are empirically valid. For instance, $m_{A\cap B} \equiv \min(m_A, m_B)$ and $m_{A\cup B} \equiv \max(m_A, m_B)$, assuming the usual min-max norms, and $m_{\sim A} \equiv 1 - m_A$, assuming the usual complement. It is, of course, possible to ask subjects to rate objects in the sets $A$ and $B$ separately and then to have them rate the union, intersection, and complement directly. If the min and max norms are valid, the derived memberships should agree with the directly rated stores, up to noise. It is not unusual for this consistency test to fail, particularly for union, so the reader should be warned that fuzzy set theory is not a model for the way people actually *do* think about many common categories. Linguistic hedges have also been studied; it seems that Zadeh's (1965) power transformations are often inadequate models for the way linguistic hedges seem to work in natural language and thus should be viewed with at least a grain of salt. Linguistic hedges as implemented by power transformations also require higher levels of measurement (ratio) than researchers might be comfortable with for many data, whereas the max-min norms require really only ordinal information as well as calibration of scale (and there are ways around this, too). Smithson (1987, chap. 2) has a survey of the cognitive psychological literature on this point. Despite the fact that fuzzy set theory does not adequately model vagueness in natural language, it is still possible to use it as a modeling framework for scientific statements, provided the investigator takes the time to establish validity.

## *7. CONCLUSION*

I hope it is clear that the acquisition of membership values in a fuzzy set analysis presents many of the same challenges common to all measurement tasks faced by social and behavioral scientists. While the task is not easy and requires careful thought on the part of

the investigator, it can be a worthwhile one. Even if a conventional analysis may *seem* easier, it is often the case that little solid attention is paid to making a valid test of the theoretical propositions at hand. Most models used in a traditional analysis are models for conditional means, but theoretical predictions often say little about conditional means. This is particularly true for substantive theory framed in terms of logical propositions, where many-to-one relationships, networks of implications, and qualitative boundaries are common.

One temptation that should be avoided is to view fuzzy set–based analyses as alternatives to the tools of measurement that have been developed by statisticians, psychometricians, and others in general. Models such as the necessary condition/fuzzy inclusion model are alternatives to specific statistical models such as regression in circumstances where a linear additive model does not faithfully translate the theoretical propositions to be tested. Empirical traces of a necessary condition are common, and it is usual to transform them away as nuisance in a traditional analysis so these may make sense as alternates to the *systematic component* of statistical models.

But a given statistical model has a measurement error component as well as a systematic one. The toolbox of statistical techniques is essential for dealing with measurement error. Since it is unlikely that measurement error will be small enough to be ignored in most practical situations, *some* method is necessary to address them. Attempts at purely fuzzy set–based error theory have been made in the past (e.g., "Level 2 fuzzy sets"), which attempt to quantify the uncertainty in membership assignments using fuzzy numbers to quantify the uncertainty, but so far, none has led to useful technology (Smithson 1987). In my view, statistical methods are the best we have at the moment and for the foreseeable future. Fuzzy set methods are best viewed as an additional item in the old kit bag we already own, not as a replacement for it.

Despite my words of caution, I believe that fuzzy set theory provides a useful and tractable way to address relationships that are too often ignored in traditional analysis. As a mathematical language that is closer to that of verbal theory, it provides a useful bridge between worlds that are far too often separated. Without careful attention to the problem of assignment, however, this opportunity is likely to be missed.

## NOTES

1. *Object* represents the objects of inquiry, be they human or animal subjects, experiment, survey respondents, examinees, organizations, countries, or commodities. It is usual in various research traditions to have specialized terms. *Objects* is the most general I can find. When *subjects* or *judges* is used, it is solely in relation to a process of assignment that depends on subjective judgments.

2. It should be noted that Luce (1995, 1997) has been critical of fuzzy set theory as it has been applied in psychology, primarily because fuzzy sets do not formalize crucial aspects of his concern—namely, vagueness in the universe of objects itself. I do not feel that Luce's criticisms are unsound. (But even if one disagrees with Luce, it is wise to take his point of view seriously.) Set theory itself is less than a century old, and fuzzy set theory—which itself is simply a generalization of ordinary set theory that presupposes it even while relaxing one aspect—is even younger, so it is to be expected that both fail in places. Furthermore, much of the development of fuzzy set theory has been in the hands of engineers, and so it is not surprising that their concerns and those of empirical scientists who want to make use of it differ, a point discussed extensively by Zimmerman (1993).

3. Interactive Web demos can be found on the Web via a bit of searching and give useful hands-on experience. I found this page by Dr. Richard Morris at Leeds University: www.scs.leeds.ac.uk/pfaf/rich_home.html. Unfortunately, with the Web being what it is, this demo may not exist for future readers.

4. It is customary in the fuzzy set literature to use $\mu$ to represent membership. Because $\mu$ is already "taken" in statistics to represent the population mean, to avoid pointless confusion, I use $m$ instead.

5. I should note that despite showing that their procedure generates a ratio scale, the top value of membership—and thus the unit of change—was established "by fiat" since the value $m = 1$ was chosen for the object with maximum membership.

6. Of course, so long as humans have been trading with money, quantitative comparisons of abstract value have been made since the price system in effect provides a direct ratio scale. Scientists, it seems, often take a long time to catch up with ordinary folks.

7. There are 810 cases in the example, 18 sets of 45 pairs. The data are simulated from my own scaling of one complete set of pairs with noise added. The data file and Systat code to run the reported logistic regression are available upon request.

8. The author recalls news coverage in the mid-1990s that considered the horse race between countries in the top of the scale for who was in the top slot on the Human Development Index (HDI). These decisions often came down to values in the third decimal place, which is clearly asking for a ridiculous level of precision from the data.

9. In general, it is better in the days of modern computers to use an item response model than the older Guttman scaling criteria. Item response theory (IRT) models are probabilistic generalizations of the deterministic Guttman scale. Since the original data are not available, it is impossible to fit a better model. This does not affect the utility of the Muller data for expository purposes.

10. The cumulative scale structure is one of the really compelling notions in social science because it starts with the notion that social reality is fundamentally categorical and sees if ordering can be found based on a particularly logical criterion, that of nested subsets. For $k$ sets, there is a cumulative structure if it is true that $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_k$ (Roberts 1979). "Passing" $A_1$ indicates that all previous items are passed and so on down the sequence, which provides a means for defining an order among the items in question based only on (crisp) set membership. Smithson (1987) illustrates two fuzzy set generalizations of Guttman scaling.

Polytomous IRT is a probabilistic generalization of the Guttman scale for polytomous items, as mentioned above.

11. There is some debate about this, but it is clear that all cases with 0 membership in both sets should be excluded since it is always the case that a statement that proceeds from a false premise has formal truth value 1, even though it clearly has no actual truth.

12. Of course, as noted above, most membership functions will be of mixed type (i.e., both discrete and continuous), with nonzero probability of having value 0 or 1, so ties are definitely a problem for boundary cases.

## REFERENCES


Adcock, Robert and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95 (3): 529-46.

Agresti, Alan. 2002. *Categorical Data Analysis*. 2nd ed. New York: John Wiley.

Baird, John C. 1997. *Sensation and Judgment: Complementarity Theory of Psychophysics*. Mahwah, NJ: Lawrence Erlbaum.

Baird, John C. and Elliott Noma. 1978. *Fundamentals of Scaling and Psychophysics*. New York: John Wiley.

Bilgiç, Taner and I. Burhan Türkşen. 1997. "Measurement of Membership Functions: Theoretical and Empirical Work." Unpublished manuscript, University of Toronto.

Böckenholt, Ulf. 2001. "Hierarchical Paired Comparisons." *Psychological Methods* 6 (1): 49-66.

Bollen, Kenneth A. 2002. "Latent Variables in Psychology and Social Sciences." *Annual Review of Psychology* 53: 605-34.

Bollen, Kenneth A. and Richard L. Lennox. 1991. "Conventional Wisdom on Measurement: A Structural Equation Perspective." *Psychological Bulletin* 110 (2): 305-14.

Brehm, John and Scott Gates. 1993. "Donut Shops and Speed Traps: Evaluating Models of Supervision of Police Behavior." *American Journal of Political Science* 37 (2): 555-81.

Cerioli, Andrea and Sergio Zani. 1990. "A Fuzzy Approach to the Measurement of Poverty." In *Income and Wealth Distribution, Inequality and Poverty*, edited by Camilio Dagum and Michele Zenga. Berlin: Springer-Verlag.

Chameau, Jean-Lou and Juan Carlos Santamarina. 1987. "Membership Functions I: Comparing Methods of Measurement." *International Journal of Approximate Reasoning* 1 (3): 287-301.

Cheli, B. and A. Lemmi. 1995. "A 'Totally' Fuzzy and Relative Approach to the Measurement of Poverty." *Economic Notes* 94 (1): 115-34.

Cornsweet, T. N. 1962. "The Staircase Method in Psychophysics." *American Journal of Psychology* 75 (3): 485-91.

Fukuda-Parr, Sakiko and A. K. Shiva Kumar. 2003. *Readings in Human Development: Concepts, Measures, and Policies for a Development Paradigm*. Oxford, UK: Oxford University Press.

Gallie, W. B. 1956. "Essentially Contested Concepts." *Proceedings of Aristotelian Society* 51: 167-98.

Gupta, Arjun K. and Saralees Nadarajah, eds. 2004. *Handbook of Beta Distribution and Its Applications*. New York: Marcel Dekker.

Healy, M. J. R. and Harvey Goldstein. 1976. "An Approach to the Scaling of Categorized Attributes." *Biometrika* 63 (2): 219-29.

Hesketh, B., R. G. Pryor, M. Gleitzman, and T. Hesketh. 1988. "Practical Applications and Psychometric Evaluation of a Computerised Fuzzy Graphic Rating Scale." Pp. 425-54 in *Fuzzy Sets in Psychology*, edited by T. Zetenyi. Amsterdam: North-Holland.

Hildebrand, David K., James D. Laing, and Howard Rosenthal. 1977. *Prediction Analysis of Cross Classifications*. New York: John Wiley.

Hisdal, Ellen, ed. 1988. "Interpretations of Grades of Membership" [special issue]. *Fuzzy Sets and Sets and Systems* 25(3).

Huber, Joel, Dan Ariely, and Gregory Fischer. 2002. "Expressing Preferences in a Principal-Agent Task: A Comparison of Choice, Rating, and Matching." *Organizational Behavior and Human Decision Processes* 87(1): 66-90.

Klir, George A. and Bo Yuan. 1995. *Fuzzy Sets and Fuzzy Logic*. New York: Academic Press.

Lazarsfeld, Paul F. 1972. *Qualitative Methods*. New York: Free Press.

Lin, Lawrence I-Keui. 1989. "A Concordance Correlation Coefficient to Evaluate Reproducibility." *Biometrics* 45 (1): 255-68.

Luce, R. Duncan. 1959. *Individual Choice Behavior*. New York: John Wiley.

———. 1964. "The Mathematics Used in Mathematical Psychology." *American Mathematical Monthly* 71 (4): 364-78.

———. 1995. "Four Tensions Concerning Mathematical Modeling in Psychology." *Annual Review of Psychology* 46:1-26.

———. 1997. "Several Unresolved Conceptual Problems in Mathematical Psychology." *Journal of Mathematical Psychology* 41(1): 79-87.

Manton, Kenneth G., Max A. Woodbury, and H. Dennis Tolley. 1994. *Statistical Applications Using Fuzzy Sets*. New York: John Wiley.

Muller, Edward N. 1972. "A Test of a Partial Theory of Potential for Political Violence." *American Political Science Review* 66 (3): 928-59.

Przeworski, Adam and Henry J. Teune. 1970. *Logic of Comparative Social Inquiry*. New York: John Wiley.

Qizilbash, Mozzafar. 2003. "Vague Language and Precise Measurement: The Case of Poverty." *Journal of Economic Methodology* 10 (1): 41-58.

Ragin, Charles C. 2000. *Fuzzy-Set Social Science*. Chicago: University of Chicago Press.

Roberts, Fred S. 1979. *Measurement Theory*. Reading, MA: Addison-Wesley.

Sen, Amartya. 1999. *Development as Freedom*. New York: Knopf.

Singpurwalla, Nozer D. and Jane M. Booker. 2004. "Membership Functions and Probability Measures of Fuzzy Sets." *Journal of the American Statistical Association* 99 (467): 867-77.

Smithson, Michael J. 1987. *Fuzzy Set Analysis for Behavioral and Social Sciences*. New York: Springer.

Smithson, Michael J. and Jay Verkuilen. Forthcoming. *Fuzzy Set Theory*. Thousand Oaks, CA: Sage.

Torgerson, Warren S. 1958. *Theory and Methods of Scaling*. New York: John Wiley.

Tversky, Amos and D. J. Koehler. 1994. "Support Theory: A Nonextensional Representation of Subjective Probability." *Psychological Review* 101 (3): 547-67.

United Nations Development Program. 2004. *Human Development Report 2004*. Retrieved from http://hdr.undp.org/reports/global/2004/

Wallsten, Thomas S., David V. Budescu, Amnon Rapoport, Rami Zwick, and Barbara H. Forsyth. 1986. "Measuring the Vague Meanings of Probability Terms." *Journal of Experimental Psychology–General* 115 (3): 348-65.

Zadeh, Lotfi A. 1965. "Fuzzy Sets." *Information and Control* 8: 338-53.

Zimmerman, H.-J. 1993. *Fuzzy Set Theory and Its Applications*. 2nd ed. Boston: Kluwer-Nijhoff.

*Jay Verkuilen is a Ph.D. candidate in the Department of Psychology, Quantitative Division, University of Illinois, Urbana-Champaign. He received a Ph.D. in political science from UIUC in 2002. Publications include a forthcoming monograph on fuzzy sets in the social sciences with Michael Smithson, work in behavioral game theory, and measurement in comparative democracy research. Current research considers applications of multilevel models for clustered data to scaling and psychophysical models of time perception in intertemporal choice.*