Sociological Methods & Research
Volume 38 Number 1
August 2009 102-146
© 2009 SAGE Publications
10.1177/0049124109339371
http://smr.sagepub.com
hosted at
http://online.sagepub.com

Goodness-of-Fit Tests and Descriptive Measures in Fuzzy-Set Analysis

Scott R. Eliason
Robin Stryker
University of Arizona, Tuscon

In this article the authors develop goodness-of-fit tests for fuzzy-set analyses to formally assess the fit between empirical information and various causal hypotheses while accounting for measurement error in membership scores. These goodness-of-fit tests, and the accompanying logic, provide a sound inferential foundation for fuzzy-set methodology. The authors also develop descriptive measures to complement these tests. Examples from Stryker and Eliason (2003) and Mahoney (2003) show how goodness-of-fit tests and descriptive measures may be used to assess individual causal factors as well as conjunctions of factors. The authors show how these tools provide more information in a fuzzy-set analysis than do tests currently in use. In providing this inferential foundation, the authors also show that fuzzy-set methods (a) are no less amenable to falsificationist methods of the Neyman-Pearson type than are standard statistical techniques and (b) may be usefully applied in either an exploratory/inductive or a confirmatory/deductive research design.

Keywords: *Fuzzy-set; goodness-of-fit; causal inference, necessity; sufficiency*

Fuzzy-set theory and methods were developed in part to address perceived deficiencies in probability theory when dealing with specific, largely linguistic and semantic, types of uncertainty in empirical information (Zadeh 1995). They now have become an increasingly important and powerful methodological lens in social science research. As with the increased visibility of Bayesian statistical methods in the social sciences (e.g., Western and Jackman 1994; Western 1998, 2001), the past decade or so has witnessed a noticeable, though relatively modest, increase in the use of fuzzy-set theories and methodologies. These include theoretical considerations in understanding the self as a fuzzy-set system of social roles (Montgomery 1998, 2000) and in understanding decision-making processes (Arfi 2005, 2006), measurement considerations in the study of sterilization

(Rindfus and Liao 1988), methodological considerations in grade-of-membership techniques (Manton et al. 1987; Manton et al. 1992), and myriad other research settings (e.g., Mahoney 2003; Stryker and Eliason 2003, 2004; Goertz and Mahoney 2004; Stryker, Eliason, and Tranby 2008; Ragin and Pennings 2005).[1] Perhaps most notable in sociology is Charles Ragin's (2000) recent harnessing of fuzzy-set logic for use in qualitative-comparative research in assessing necessary and sufficient causal relations. This is a major advance given (a) that many statistical models typically used by social scientists are not well suited to assessing hypotheses involving necessary and sufficient causal relations and (b) the rather ubiquitous nature of such hypotheses and explanations in the social sciences. See, for example, Goertz's (2003:76-94) list of 150 hypotheses in the social science literature that involve necessary conditions.[2]

Development of fuzzy-set theory and methods, however, dates back to the 1930s when Max Black (1937) initially laid the theory's logical foundation. It was not until the 1960s that the theory of fuzzy sets significantly evolved in the work of engineer Lotfi Zadeh (1965). From these roots, concepts and applications of fuzzy-set logic developed largely in research on expert systems and artificial intelligence, helping to solve problems requiring the identification of an object as belonging to some class or set of objects (e.g., McNeill and Freiberger 1993; Laviolette et al. 1995). Fuzzy-set methods and accompanying computer algorithms are now commonly used in diverse areas including optical character recognition (such as that used in pen-based handheld and tablet computers), so-called smart devices (such as intelligent household appliances), and identification recognition (implemented in airport security).

Newly developed methods are often met with much debate, critique, and intense scrutiny—as they should be in any science. Consistent with

Authors' Note: The methods discussed in this article were presented at the Robert Schuman Center for Advanced Studies, European University Institute, Florence; the Department of Social and Political Studies, University of Milan; the Department of Sociology, University of Stockholm; the Max Planck Institute for the Study of Societies, Cologne; the University of Iowa Theory Workshop; the Minnesota Population Center, University of Minnesota; and the annual meetings of Research Committee 19 of the International Sociological Association, Toronto (2003) and Paris (2004). We thank Joakim Palme, Sheldon Stryker, and all the participants in these seminars and workshops for their constructive feedback. We also thank the Robert Schuman Center, which supported this research while Stryker was a Jean Monnet Fellow and Eliason was a visiting fellow. Please address correspondence to Scott Eliason, Social Sciences Building, 1145 E. South Campus Drive, Tucson, AZ 85721; e-mail: seliason@email.arizona.edu. An online appendix is available at http://smr.sagepub.com/supplemental.

this, debates about the status of fuzzy-set theory relative to probability theory appear in full bloom (e.g., Almond 1995; Kandel, Martins, and Pacheco 1995; Laviolette et al. 1995; Zadeh 1995). At the same time, scholars are further establishing the theoretical foundations underlying the idea of fuzzy random variables in statistics as well as the use of fuzzy-set techniques and algorithms in engineering and computer science (e.g., Puri and Ralescu 1985; Klament, Puri, and Ralescu 1986; Stojakovic and Stojakovic 1996). Thus, it appears that fuzzy-set techniques in many areas of engineering and science, including the social sciences, are here to stay.

In this article we build on Ragin's (2000) methods for assessing the empirical relation between a hypothesized cause and outcome based on fuzzy-set logic. We extend Ragin's fuzzy-set methodology by (a) formally accounting for measurement error; (b) constructing descriptive measures of the distance and consistency between an observed fuzzy-set graph and specific causal and null hypotheses; and (c) constructing goodness-of-fit F tests to assess the fit between some fuzzy-set graph and causal necessity, sufficiency, and necessity and sufficiency hypotheses. In extending fuzzy-set methods in this way—and especially in developing the goodness-of-fit tests—we seek to place fuzzy-set methodology on a firm inferential foundation. In so doing, we show how assertions that general qualitative comparative analysis (QCA) methodology—of which fuzzy-set methods are part—is severely limited and cannot "employ a probabilistic perspective" or "deal with data errors" are in fact misguided (Lieberson 1994:1225; see also Lieberson 1991; Sobel 1995). To the contrary, the extensions developed in this article show that fuzzy-set techniques are no more or less amenable to falsificationist methods and the hypothesis-testing framework of the Neyman-Pearson type than are standard statistical modeling and testing techniques commonly used in sociology.[3] In general, we show how these goodness-of-fit tests and descriptive measures may be usefully applied in the context of case-oriented research, for exploratory inductive or confirmatory deductive purposes.

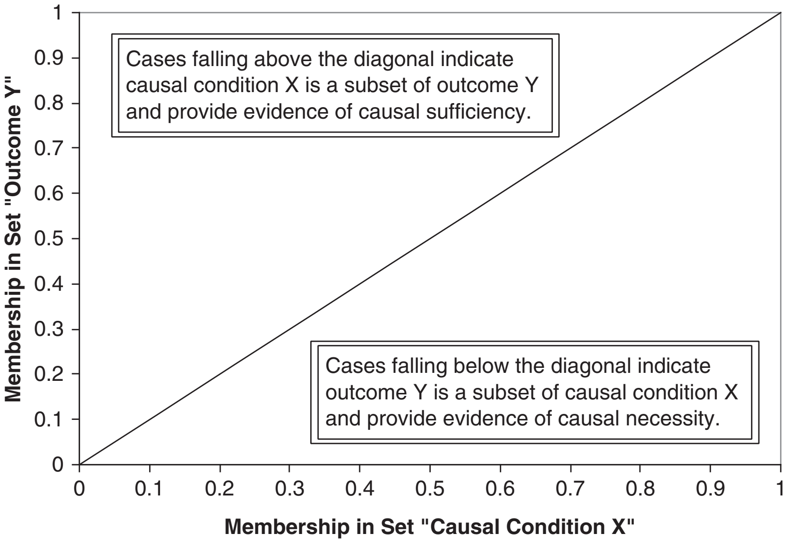
Assessing the Empirical Content of a Fuzzy-Set Graph: Preliminaries

Ragin's (2000) fuzzy-set methodology builds on the subset principle in a manner similar to that found in previous case-oriented QCA (e.g., Ragin 1987).[4] However, compared to the "crisp set" logic of QCA, in which the researcher assesses the relation between the presence or absence of some hypothesized cause and the presence or absence of some outcome, fuzzy-set logic is based on the degree to which a case belongs to a class defined on some causal condition (or conjunction of conditions) and a class defined on some outcome. Although some argue for a strong distinction in the logic of fuzzy sets between this "degree of membership" and the "probability of membership" (e.g., Zadeh 1995), membership scores in fuzzy-set methods typically take on values in the range similar to probabilities: A minimum score of 0 indicates the minimum degree, and a maximum score of 1 indicates the maximum degree, of a case belonging to some set. Scores between 0 and 1 then indicate the relative degree of belonging to the set.

Ragin (2000) shows how causal relations in a fuzzy-set analysis may be understood by graphing a biplot of membership scores in the cause against those in the outcome. Ragin's approach makes use of the subset principle applied to the membership scores (instead of the presence or absence of some causal condition and outcome as in QCA). The idea is elegant and simple, yet very powerful. By comparing membership to the cause and to the outcome for all cases under study, a researcher can establish if one may be considered a subset of the other. If so, then depending on the subset pattern observed in the biplot, a researcher may claim evidence in the data for a causally necessary, causally sufficient, or causally necessary and sufficient relationship.

Figure 1 gives the patterns expected in the graph under different causal relations reflected in the data. If the data are fully consistent with a causally necessary, but not sufficient, relationship, then membership in the outcome will be less than membership in the causal condition for all cases. This, in turn, indicates the outcome is in general a subset of the cause. Under a causally necessary relationship, then, all cases (points) will be distributed *below* the main diagonal. Although not fully consistent with strict adherence to fuzzy-set logic, it may be useful to think of this probabilistically so that we consider the membership scores (i.e., the degree of belonging to some set) as indexing the likelihood of observing the causal condition or the outcome.[5] For a causally necessary, but not sufficient, relationship, the likelihood of observing the cause must be as high as or higher than the likelihood of observing the outcome. Under this scenario, for any one randomly selected case in the data, we may observe the cause while not observing the outcome, but we are unlikely to observe the outcome while not observing the cause. Mapping these likelihoods (i.e., fuzzy-set membership scores) in an *xy* scatterplot produces the lower diagonal pattern as shown in Figure 1.

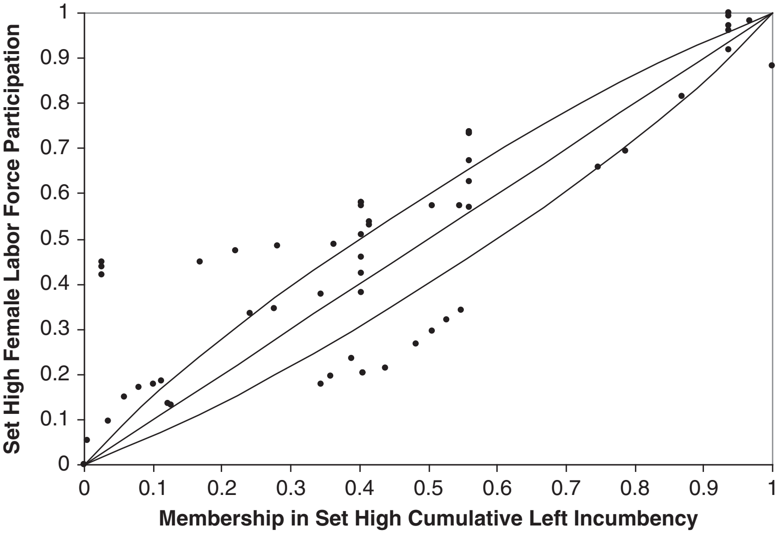
Figure 1
Fuzzy-Set Graph Showing Relation of Cases to Diagonal and the Causal Arguments Supported



If the data are fully consistent with a causally sufficient, but not necessary, relationship, then membership in the causal condition will be less than membership in the outcome for all cases. This, in turn, indicates the causal condition is in general a subset of the outcome. Under a causally sufficient relationship, then, all cases (points) will be distributed *above* the main diagonal, as shown in Figure 1. Using the probabilistic way of looking at a sufficient, but not necessary, causal relation, the likelihood of observing the outcome must be as high as or higher than the likelihood of observing the cause. For any one randomly selected case in the data, we may observe the outcome while not observing the cause, but we are unlikely to observe the cause while not observing the outcome.

Finally, if the data are fully consistent with a causally necessary and sufficient relationship, then membership in the causal condition will be equal to membership in the outcome for all cases. Under a causally necessary and sufficient relationship, then, all cases (points) will be distributed

Figure 2
Fuzzy-Set Graph Showing Relation Between Membership in Sets High Cumulative Left Cabinet Incumbency and High Female Labor Force Participation



Source: Stryker and Eliason (2003) data for France, Belgium, Germany, Italy, Denmark, and Britain 1977-1994.

along the main diagonal. From a probabilistic viewpoint, this would indicate that the likelihood of observing the cause and outcome jointly for any one randomly selected case will be exactly equal; knowing the probability of observing the cause tells us precisely the probability of observing the outcome.

Figure 2 provides an empirical example from Stryker and Eliason's (2003) fuzzy-set analysis of the relation between cumulative governance by left political parties and the supply of and demand for female labor, and how these in turn influence the feminization of left political support in France, Belgium, Germany, Italy, Denmark, and Britain from 1977 to 1994.[6] As part of their analysis, Stryker and Eliason posit a causal chain to explain how cumulative governance by left political parties in these countries feeds back over time to help create the feminization of left support. An important part of the hypothesized causal chain posits that countries with strong left governance traditions create high female labor force participation by expanding female-typed jobs in the civilian public sector as well as such programs as public day care and maternity leave that free women to enter the labor market.

Ignoring for the moment the intervening mechanisms, Figure 2 gives a fuzzy-set plot of membership in the hypothesized causal condition high cumulative left cabinet incumbency by membership in the outcome high female labor force participation for these data. Of the 60 country-period cases in the Stryker/Eliason data, 15 (25 percent) are inconsistent with a causal sufficiency hypothesis and the remaining 45 (75 percent) are inconsistent with a causal necessity hypothesis. Strictly speaking, all cases are inconsistent with the causal necessity and sufficiency hypothesis.

Thus, visual inspection of the empirical content in this fuzzy-set graph provides at best an ambiguous picture of the causal relationships and does little to help us decide for or against any causal hypothesis for these data. The graph may, however, suggest a simple linear correlation between fuzzy-set scores on high cumulative left cabinet incumbency and those on high female labor force participation. While assessing this association may be useful in some settings, it does little to address hypotheses regarding causal necessity and sufficiency. In the next section, we describe our goodness-of-fit tests for these causal hypotheses. We show how these tests help the researcher overcome the ambiguity suggested by a visual inspection of the empirical content in Figure 2. Once we have guided the reader through the necessary technical validation of our approach, we show that applying our goodness-of-fit F tests, along with our descriptive measures of distance and consistency, reveals especially strong empirical support in these data for a relationship of causal sufficiency between high cumulative left governance and high female labor force participation.

Goodness-of-Fit Tests for Causal Hypotheses

While there have been significant developments over the past few years in descriptive measures for fuzzy-set analysis (e.g., Goertz 2006; Ragin 2006), there has been little movement toward providing it a sound inferential framework.[7] One notable exception is Ragin's (2000:111-15) z test based on the proportion of cases consistent with some causal argument. This approach has intuitive appeal, but it also has some consequential limitations. First, while the proportion of cases consistent with some causal argument is an important indicator of how well the argument compares to the data, the proportion itself does not take into account the distance of each case from that expected under some causal argument. Thus, some cases may be inconsistent with a causal hypothesis (e.g., causal necessity and sufficiency) but still close to that expected under the causal argument (e.g., the diagonal on the graph). Other cases may be inconsistent with a causal hypothesis and relatively distant from that expected under the causal argument. Any test based on the proportion alone will not be sensitive to this distinction and instead will treat all cases inconsistent with some causal argument as the same.

A second limitation lies in the inferential infrastructure underlying the z test on the proportion. Specifically, the construction of the standard error for the test is not consistent with how many studies using fuzzy-set analysis are designed. To see this, recall that there are often two primary sources of error in most empirical analyses, sampling error and measurement error. As used in the denominator of the proportionate z test, the standard error reflects the variability in the (function of the) proportion under repeated random sampling from some population. That is, the standard error for the test reflects sampling error, and its very mathematical form is based on the theory of random sampling. Rarely do researchers using fuzzy-set analysis presume that their data constitute a random sample. Instead, samples are often constructed using various forms of strategic sampling principles rather than those used to derive random samples, or the data instead constitute the entire population of interest to the research question. So while the proportion itself is somewhat useful (even though it does not include information on the distance as described previously), the standard error used to obtain the proportionate z test is less useful because it reflects the sampling error in that proportion, and not measurement error in the fuzzy-set scores.[8]

Finally, and perhaps most importantly, both the proportion and the proportionate z test are not informative on the causal necessity and sufficiency hypothesis. That is, while cases may lie on one side or the other of the diagonal, once we consider measurement error, cases close to the diagonal may constitute evidence for a causally necessary and sufficient relationship. The proportion and accompanying z test leave the researcher blind to this information. As a result, and as we show in our second example later, researchers relying on the proportion may miss an important part of the causal story supported by the data.[9]

To help address these limitations and to provide a more informative inferential framework for assessing the empirical content in a fuzzy-set graph, we adopt a goodness-of-fit strategy. This strategy is based on comparing the observed distance of cases in a fuzzy-set graph from some causal hypothesis with the distance that would be expected given the truth of the causal hypothesis while accounting for measurement error. To illustrate, first consider what the graph should look like were it completely consistent with an argument of causal necessity and sufficiency. Temporarily ignoring consideration for measurement error, under complete consistency with causal necessity and sufficiency, all cases should line up along the diagonal. Any deviation away from the diagonal will count as evidence against causal necessity and sufficiency. Measuring the distance of the points away from the diagonal gives us the observed distance of the graph from the argument of causal necessity and sufficiency.

Next, consider what the graph would look like were it completely consistent with a null relationship between a hypothesized cause and some outcome. Under a null relationship, the distribution of fuzzy-set membership scores on the outcome is unaffected by membership scores on the hypothesized cause. That is, the distribution of scores on the outcome should be the same everywhere we look relative to scores on the hypothesized cause.[10] If that pattern were observed in the graph, we would then conclude that the distance of the graph from that expected under a null relation is zero. Any deviation of cases away from that expected under a null association provides evidence against a null association. Measuring that distance tells us the observed distance of the graph from the null association argument.

In between these two extremes are considerations for hypotheses of causal sufficiency and causal necessity, separately. Again temporarily ignoring errors in fuzzy-set membership measurement, for a graph to be completely consistent with a causal necessity hypothesis, all cases must fall on or below the main diagonal. If that pattern were observed in the graph, we would conclude that the distance of the graph from that expected under the hypothesis of causal necessity is zero. Any deviation of cases away from this pattern (or, equivalently, any cases that fall above the main diagonal) provides evidence against the hypothesis of causal necessity. Measuring that distance tells us the observed distance of the graph from the hypothesis of causal necessity.

Finally, the minimum requirement for a graph to be completely consistent with the hypothesis of causal sufficiency is that all the cases must fall on or above the main diagonal. If that pattern were observed in the graph, we would conclude that the distance of the graph from that expected under the hypothesis of causal sufficiency is zero. Any deviation of the cases away from this pattern (or, equivalently, any cases that fall below the main diagonal) provides evidence against the hypothesis of causal sufficiency. Measuring that distance tells us the observed distance of the graph from the hypothesis of causal sufficiency.

Constructing the Goodness-of-Fit Tests

This logic provides the foundation for a precise measure of the distance of the graph from a null association, from causal sufficiency, from causal necessity, and from causal necessity and sufficiency combined. It also provides the foundation for goodness-of-fit tests of each causal hypothesis.[11] To see this, let x_i and y_i be the fuzzy-set membership scores for the hypothesized cause and outcome, respectively, for case i.[12] To ensure that distributional properties hold for the goodness-of-fit tests constructed below regardless of sample size, we transform the fuzzy-set membership scores into their respective standardized normal scores, $z_{x(i)} = \Phi^{-1}\{x_i\}$ and $z_{y(i)} = \Phi^{-1}\{y_i\}$, where $\Phi^{-1}\{\cdot\}$ is the inverse cumulative distribution function of the standard unit normal distribution.[13] Finally, let d_i be an indicator (dummy) variable coded 1 when $y_i > x_i$ and 0 when $y_i \leq x_i$.

For an xy biplot with $N(x_i,\ y_i)$ pairs, the accumulated squared Euclidean distance of the normalized fuzzy-set membership scores from that expected under each argument may now be defined.[14]

Squared distance from a null association:

$$D_{null} = \sum_{i=1}^{N} \left(z_{y(i)} - E\left\{Z_{y(i)}|\text{null XY association}\right\}\right)^2,$$

Squared distance from causal necessity: $D_{nec} = \sum_{i=1}^{N} d_i\left(z_{y(i)} - z_{x(i)}\right)^2$,

Squared distance from causal sufficiency: $D_{suf} = \sum_{i=1}^{N} (1 - d_i)\left(z_{y(i)} - z_{x(i)}\right)^2$,

Squared distance from causal necessity and sufficiency:

$$D_{(nec\&suf)} = \sum_{i=1}^{N} \left(z_{y(i)} - z_{x(i)}\right)^2 = \sum_{i=1}^{N} d_i\left(z_{y(i)} - z_{x(i)}\right)^2 + \sum_{i=1}^{N} (1 - d_i)\left(z_{y(i)} - z_{x(i)}\right)^2 = D_{nec} + D_{suf},$$

where $E\left\{Z_{y(i)}|\text{null XY association}\right\}$ is the expected value of the standardized outcome membership score for case i given a null association between the hypothesized cause and the outcome.[15]

With $Z_{y(i)}$ and $Z_{x(i)}$ normally distributed by definition, a null association implies independence of $Z_{y(i)}$ and $Z_{x(i)}$ and thus $E\left\{Z_{y(i)}|Z_{y(i)} \otimes Z_{x(i)}\right\} = E\left\{Z_{y(i)}\right\} = \bar{Z}_y$, where $Z_{y(i)} \otimes Z_{x(i)}$ indicates independence and $\bar{Z}_y$ gives the sample mean of $Z_{y(i)}$. Thus, substituting $\bar{Z}_y$ for $E\left\{Z_{y(i)}|\text{null XY association}\right\}$ gives the minimum-distance expected value of $Z_{y(i)}$ under the null association. Similarly, $z_{x(i)}$ gives the minimum-distance expected value of $Z_{y(i)}$ under the argument of causal necessity for cases with $y_i > x_i$; $z_{x(i)}$ also gives the minimum-distance expected value of $Z_{y(i)}$ under the argument of causal sufficiency, but for cases with $y_i \leq x_i$. Finally, for all cases in the graph, $z_{x(i)}$ gives the minimum-distance expected value of $Z_{y(i)}$ under causal necessity and sufficiency combined.

Often, however, we may not wish to treat cases close to the main diagonal as constituting sharp evidence against a causal relationship, even though those cases may be inconsistent with some causal hypothesis. This may be due to imprecision in the information used to code membership scores, variability in procedures used to measure degree of membership, or other measurement and coding considerations. Therefore, to assess the fit of the data to that expected under some causal hypothesis, we compare the observed distances to expected distances under the assumption that a specific causal hypothesis is true, up to a specified degree of measurement error.[16]

To see this, assume now that fuzzy-set membership scores y_i and x_i are measured with error. Because membership scores are bound between 0 and 1, any reasonable measurement error in the observed membership scores should be consistent with these bounds. Although more complex relationships may be used, here we assume an additive error in the standardized membership scores. That is, $z_{y(i)} = z^t_{y(i)} + \varepsilon_i$ and $z_{x(i)} = z^t_{x(i)} + \eta_i$, where $z^t_{y(i)}$ and $z^t_{x(i)}$ are the standardized membership scores measured without error and ε_i and η_i are the errors in measurement. Additive errors in the standardized scores provide the simplest approach for the researcher while ensuring that the bounds are respected for the observed fuzzy-set scores y_i and x_i measured with error as well as the fuzzy-set scores y^t_i and x^t_i measured without error.[17]

Considering first the combined causal necessity and sufficiency hypothesis, rewrite the observed distance in the normal scores $D_{(nec\&suf)}$ as

$$D_{(nec\&suf)} = \sum_{i=1}^{N} \left(z_{y(i)} - z_{x(i)}\right)^2 = \sum_{i=1}^{N} \left[\left(z^t_{y(i)} + \varepsilon_i\right) - \left(z^t_{x(i)} + \eta_i\right)\right]^2$$
$$= \sum_{i=1}^{N} \left[\left(z^t_{y(i)} - z^t_{x(i)}\right) + (\varepsilon_i - \eta_i)\right]^2.$$

If the causal necessity and sufficiency hypothesis is true, then $\left(z^t_{y(i)} - z^t_{x(i)}\right) = 0$ for all i. Thus, the expected distance under the truth of the causal necessity and sufficiency hypothesis is given by

$$E\left\{D_{(nec\&suf)} | \text{causal necessity \& sufficiency}\right\} = \sum_{i=1}^{N} (\varepsilon_i - \eta_i)^2.$$

Given that $z_{x(i)}$ and $z_{y(i)}$ derive from standard unit normal distributions by definition, the difference $(z_{y(i)} - z_{x(i)})$ is necessarily normally distributed (Puri and Ralescu 1985; Stuart, Ord, and Arnold 1999).[18] With errors independent of the expected (y_i, x_i) relationship under the truth of the causal necessity and sufficiency hypothesis, the ratio[19]

$$F^{*}_{(nec\&suf)} = \frac{D_{(nec\&suf)}\big/N}{E\big\{D_{(nec\&suf)}|\text{causal necessity \& sufficiency}\big\}\big/N}$$

will be distributed as an F random variable on (N, N) degrees of freedom if the causal necessity and sufficiency hypothesis is indeed true.[20]

Comparing the calculated $F^{*}_{(nec\&suf)}$ to the reference F distribution on (N, N) degrees of freedom provides a goodness-of-fit test between the causal necessity and sufficiency hypothesis and the information in the fuzzy-set biplot.[21] Given a Type I error rate for the test (typically .05), if the calculated $F^{*}_{(nec\&suf)}$ is larger than the critical F value on (N, N) degrees of freedom, then the test provides evidence for a lack of fit between the causal necessity and sufficiency hypothesis and the data. Conversely, if the calculated $F^{*}_{(nec\&suf)}$ is smaller than or equal to the critical F value, then the test provides evidence for goodness of fit between the causal necessity and sufficiency hypothesis and the data.[22]

Importantly, this goodness-of-fit F test is appropriate regardless of whether the data constitute some sample (random or not) from a population or the entire population. The test measures the goodness of fit between the causal necessity and sufficiency hypothesis and the data at hand relative to some acceptable degree of measurement error in the fuzzy-set membership scores. Additionally, our goodness-of-fit F test does not require any assumptions about the functional relationship—beyond that expected under the subset principle—between fuzzy-set membership scores on the outcome and fuzzy-set membership scores on the hypothesized cause (or conjunction of causes). We do not assume that any type of statistical model (linear or otherwise) fits the fuzzy-set data. Nor do we assume that any type of statistical model fits the empirical information on which the fuzzy-set scores are based.

This logic and method for assessing data fit (or lack of fit) to a hypothesis of causal necessity and sufficiency carry forward to assessing separately the causal necessity and causal sufficiency hypotheses. The numerators for the separate tests for causal necessity and causal sufficiency are easily obtained by making use of the fact that the total observed sums of squared distances from causal necessity and sufficiency can be separated into two independent components: (a) the sums of squared distance from causal necessity plus (b) the sums of squared distance from causal sufficiency. That is, $D_{(nec\&suf)} = D_{nec} + D_{suf}$, as shown previously.

To obtain the denominators for each test statistic, we need to find the expected sums of squared distances under the truth of each causal hypothesis. Unfortunately, unique expected sums of squared distances under the truth of either the causal necessity or the causal sufficiency hypothesis do not exist. This is because we cannot know under either hypothesis the precise values of the normalized fuzzy-set scores $z^t_{y(i)}$ and $z^t_{x(i)}$ measured without error. Instead, all we can know is that if the causal necessity hypothesis is true, then $z^t_{y(i)} \leq z^t_{x(i)}$, and if the causal sufficiency hypothesis is true, then $z^t_{y(i)} \geq z^t_{x(i)}$, for all cases i.

Nevertheless, by assuming that the mean values for the errors are equal, $E\{\varepsilon_i\} = E\{\eta_i\}$, we can obtain unique minimum expected sums of squared distances under the truth of either the causal necessity or the causal sufficiency hypothesis. While alternative assumptions may be reasonably considered, and goodness-of-fit tests derived under those alternative assumptions, the equal-mean-values assumption provides a conservative test of each hypothesis. This in turn increases our confidence in the truth of some causal hypothesis when this test indicates a good fit for some specific set of data.

To see this more completely, assuming $E\{\varepsilon_i\} = E\{\eta_i\}$ gives

$$E\left\{\sum_{i=1}^{N}\left[\left(z^t_{y(i)} - z^t_{x(i)}\right) + (\varepsilon_i - \eta_i)\right]^2\right\} = E\left\{\sum_{i=1}^{N}\left[\left(z^t_{y(i)} - z^t_{x(i)}\right)^2 + (\varepsilon_i - \eta_i)^2\right]\right\}.$$

It follows from the result on the right-hand side of this equality that for all possible values of the measurement errors (i.e., the ε_i and η_i), and for all possible combinations of the fuzzy-set membership scores measured without error (i.e., the $z^t_{y(i)}$ and $z^t_{x(i)}$), the minimum expected sums of squared distances under the truth of either the causal necessity or the causal sufficiency hypothesis is obtained when $z^t_{y(i)} = z^t_{x(i)}$ for all cases. The minimum expected distances under the truth of either hypothesis is given by

$$\min E\{D_{nec}|\text{causal necessity}\} = \min E\{D_{suf}|\text{causal sufficiency}\} = \sum_{i=1}^{N}(\varepsilon_i - \eta_i)^2.$$

These in turn provide the following test statistics for the causal necessity hypothesis,

$$F^*_{nec} = \frac{D_{nec}/N_{nec}}{\min E\{D_{nec}|\text{causal necessity}\}/N},$$

and the causal sufficiency hypothesis,

$$F^*_{suf} = \frac{D_{suf} \big/ N_{suf}}{\min E\big\{D_{suf}|\text{causal sufficiency}\big\}\big/N},$$

where N_{nec} is the number of cases such that fuzzy-set membership scores on the hypothesized causal factor are a subset of the membership scores on the outcome (i.e., $y_i > x_i$) and where N_{suf} is the number of cases such that fuzzy-set membership scores on the outcome are a subset of the membership scores on the hypothesized causal factor (i.e., $y_i < x_i$). If the causal necessity hypothesis holds, F^*_{nec} will be distributed as an F random variable on (N_{nec}, N) degrees of freedom. Similarly, if the causal sufficiency hypothesis holds, F^*_{suf} will be distributed as an F random variable on (N_{suf}, N) degrees of freedom. As we illustrate below, this in turn provides the basis from which to test the causal necessity hypothesis and the causal sufficiency hypothesis.

To calculate any of the above test statistics, the expected distances under the truth of each causal argument (the denominators for each test) are required. This in turn requires specification of the degree of certainty (or uncertainty) that a researcher has in the constructed fuzzy-set membership scores. Currently there are no published guidelines in this regard. Nevertheless, it appears reasonable that researchers will most often be most certain of the likelihood of membership when coding extreme membership scores around 0 and 1, and least certain of the likelihood of membership in the set as we move away from these extremes and toward the midpoint membership score 0.5.

There are a number of ways to capture this type of uncertainty in measuring degree of membership. One very useful, yet simple, approach is to assume some maximum measurement error at the midpoint (0.5) on the fuzzy-set membership scores, assume some minimum measurement error on the endpoints (0 and 1), and allow the measurement error to diminish smoothly from the maximum to the minimum for membership scores moving away from the midpoint and toward the two endpoints. This is readily achieved by assuming a small constant error in the standardized normal scores, $z_{x(i)}$ and $z_{y(i)}$, such that the above property holds for some chosen minimum and maximum. An important property of this measurement error in the standardized scores is that it ensures that measurement errors are independent of the expected (y_i, x_i) relationship under the truth of the causal hypotheses and that $E\{\varepsilon_i\} = E\{\eta_i\}$.[23] Both properties are required for the goodness-of-fit F tests to have the expected distributions under the truth of some causal hypothesis.

More complicated approaches certainly may be used. However, doing so necessarily introduces into the goodness-of-fit statistics artifacts due

Table 1
Goodness-of-Fit Statistics and Relative Consistency Measures for Relationship Between Membership in Sets High Cumulative Left Cabinet Incumbency and High Female Labor Force Participation

Hypothesis	*SD*	*df*	*MSD*	*F*	*p*	*R* (%)
Null association	42.04	59	0.71	—	—	26.34
Necessity	11.55	42	0.27	1.07	.41	79.76
Sufficiency	3.48	16	0.22	0.85	.63	93.89
Necessity and sufficiency	15.03	60	0.25	0.98	.54	73.66

solely to these more complex functions. Because of this, we adopt a simple strategy based on the above logic. That is, we allow for a maximum degree of uncertainty of 0.1 in our fuzzy-set scores at the midpoint and a minimum degree of uncertainty of 0.0 in our fuzzy-set scores at the endpoints, with the measurement error moving smoothly from a maximum of 0.1 at membership score of 0.5 to a minimum of 0 at the extreme membership scores of 0 and 1.

Goodness-of-Fit Tests for the Stryker/Eliason Data

Table 1 gives the goodness-of-fit statistics for the relationship shown in Figure 2, the plot of membership in the set High Cumulative Left Cabinet Incumbency by membership in the set High Female Labor Force Participation. The first column in Table 1 gives the three causal hypotheses to be tested along with the null association hypothesis. The second column gives the squared distance of the data from the null association, causal necessity (D_{nec}), causal sufficiency (D_{suf}), and causal necessity and sufficiency ($D_{(nec+suf)} = D_{nec} + D_{suf}$). The third column gives the accompanying degrees of freedom for each of these distances. The fourth column provides the mean squared distances, which are the squared distances divided by their respective degrees of freedom. The fifth and sixth columns give the goodness-of-fit F statistics and the corresponding p values. (The last column provides descriptive measures to be discussed in the next section of the article, providing additional information on which to base substantive interpretations.)

Using a maximum midpoint measurement error of 0.1 and a Type I error rate of .05,[24] each causal hypothesis fits these data, as indicated by p values larger than .05 (the Type I error rate) for the tests on each causal hypothesis. That is, given the degree of measurement error specified, each causal hypothesis—causal necessity, causal sufficiency, and causal necessity and sufficiency—fits these data at the .05 level. Thus, from the goodness-of-fit tests given in Table 1, we conclude that these data are consistent with each causal hypothesis.

If we were to end our analysis here, rather than with visual inspection of the biplot in Figure 2, we would conclude that our data for Britain, France, Denmark, Italy, Germany, and Belgium from 1977 to 1990 support not just the hypothesis that high cumulative left governance is causally sufficient for High Female Labor Force Participation. Allowing for measurement error and a Type 1 error rate of .05, the data for this set of countries also support the hypotheses that high cumulative left governance is causally necessary for high female labor force participation and that high cumulative left governance is causally necessary and sufficient for high cumulative left cabinet incumbency. The next section takes us further into our empirical detective work.

Before doing so, however, it may be useful to compare fuzzy-set causal claims with the type of causal statements often made in the context of estimating counterfactual treatment effects (e.g., Imbens and Rubin 1997; Heckman, Lalonde, and Smith 1999; Winship and Morgan 1999; Morgan and Winship 2007). In the latter case, researchers often pinpoint some average causal or treatment effect. These very specific statements about average treatment effects are made with reference either to the general population (as in the average treatment effect), to those receiving treatment (as in the average treatment effect on the treated), to those who comply with some instrument selecting subjects into various treatment and control groups (as in the local average treatment effect and the complier average causal effect), or to myriad other subpopulations of interest.[25]

In fuzzy-set analysis, statements referencing some specific point estimate of some specific (typically average) treatment effect on the outcome due to the cause are generally not possible.[26] Thus, for this example, we cannot say that some specific degree of high cumulative left cabinet incumbency will give rise, on average or otherwise, to a specific degree of high female labor force participation. Instead, the goodness-of-fit F tests reveal that these fuzzy-set data are consistent with the subset relations underlying the assertion of the three causal relations—necessity, sufficiency, and necessity and sufficiency—between the factor High Cumulative Left Cabinet Incumbency and the outcome of High Female Labor Force Participation. In this case, there is no evidence in these data to reject the subset relations underlying the three causal claims.

Descriptive Measures for Relative Distance and Consistency

The inferential framework outlined in the prior section to assess data fit to each causal argument does not apply to assessing a null association. Thus, to compare the distances from each causal hypothesis to the distance from the null association hypothesis, we also provide information about the *relative consistencies* of the three causal hypotheses and that of the null association with the empirical information in the graph. These descriptive measures tell us if the graph is more consistent with a null association, a causally sufficient relation, a causally necessary relation, or a causally necessary and sufficient relation. We first define proportional distances relative to all causal and null hypotheses.

Relative distance from a null association: $D^*_{null} = \frac{D_{null}}{\left(D_{null} + D_{nec} + D_{suf}\right)}$,
Relative distance from causal necessity: $D^*_{nec} = \frac{D_{nec}}{\left(D_{null} + D_{nec} + D_{suf}\right)}$,
Relative distance from causal sufficiency: $D^*_{suf} = \frac{D_{suf}}{\left(D_{null} + D_{nec} + D_{suf}\right)}$,
Relative distance from causal necessity and sufficiency:
$D^*_{(nec+suf)} = \frac{D_{nec} + D_{suf}}{\left(D_{null} + D_{nec} + D_{suf}\right)} = D^*_{nec} + D^*_{suf}$.

Relative proportional consistencies for all causal and null hypotheses are then calculated as

Relative consistency with a null association: $R_{null} = 1 - D^*_{null}$,
Relative consistency with causal necessity: $R_{nec} = 1 - D^*_{nec}$,
Relative consistency with causal sufficiency: $R_{suf} = 1 - D^*_{suf}$,
Relative consistency with causal necessity and sufficiency:
$R_{(nec+suf)} = 1 - D^*_{(nec+suf)}$.

These descriptive measures give the relative closeness of the information in the graph to some hypothesis.[27] More precisely, R gives the proportion of information in the graph consistent with a specific hypothesis. When R equals the maximum value of 1, the corresponding D^* equals the minimum of 0, indicating that the relative distance of the information in the graph from the corresponding hypothesis is 0. Thus, an $R = 1$ for some hypothesis means that the information in the graph is completely consistent with that hypothesis. When R equals the minimum value of 0, the corresponding D^* equals a maximum of 1, indicating that the total dispersion of empirical information in the graph is away from, or inconsistent with, the corresponding hypothesis.

Another way to think of this is that when $R = 0$ for some specific argument, one of the other Rs must be equal to 1. An $R = 0$ for some specific hypothesis means that there is no empirical information in the graph to support that hypothesis. Between these two extremes R may be interpreted as the proportion of information, relative to the set of hypotheses considered (null association, causal necessity, causal sufficiency, and causal necessity and sufficiency combined), in the graph consistent with the corresponding hypothesis. For example, R_{nec} can be interpreted as "the data are $100(R_{nec})\%$ consistent with a causal necessity hypothesis." Similar interpretations obtain for the other relative consistency measures.

Returning to our example concerning left political governance and female labor force participation, recall that the goodness-of-fit tests indicated that the data are consistent with each causal hypothesis. There is no evidence in these data to reject either the causal necessity, the causal sufficiency, or the causal necessity and sufficiency hypothesis. We thus turn to our descriptive measures to determine which causal hypothesis, if any, is most consistent with these data.

The last column in Table 1 gives these relative consistency measures for the null association and the three causal hypotheses. With an $R_{null} = .2634$, these data are only 26.34 percent consistent with the null association hypothesis. By contrast, these data are 73.66 percent consistent with the causal necessity and sufficiency hypothesis. Comparing the two hypotheses, these data are $.7366/.2634 = 2.80$ times more consistent with the causal necessity and sufficiency hypothesis than with the null association hypothesis. Similarly, these data are 79.76 percent consistent with the causal necessity hypothesis, which is $.7976/.2634 = 3.03$ times more consistent with the data than is the null association hypothesis. However, the data provide the strongest support for the causal sufficiency hypotheses, where we see a 93.89 percent consistency level. This in turn indicates that the causal sufficiency hypothesis is $.9389/.2634 = 3.57$ times more consistent than is the null association hypothesis with the relationship these data exhibit between high cumulative left governance and high female labor force participation.

In short, once we combine our goodness-of-fit tests with our newly developed technique for using relative distances to assess the relative consistency of the data with the null association and various causal hypotheses, we would conclude that our data shows especially strong support for the idea that a strong history of left political governance is causally sufficient for high female labor force participation.

What then of other factors plausibly associated with high female labor force participation? In particular, what of measures tapping more directly into the demand for and supply of female labor? The next section takes us into this terrain by developing and illustrating the use of goodness-of-fit tests for higher order conjunctions.

Testing Higher Order Conjunctions

The goodness-of-fit tests described previously may be applied to single factors or conditions and any order of conjunction of factors or conditions. To obtain goodness-of-fit tests for conjunctions, replace the fuzzy-set membership score x_i in the previous discussion with the minimum of a set of scores corresponding to those factors or conditions in the conjunction. In addition to testing the goodness-of-fit of a specific conjunction to some causal argument, it is informative to compare that conjunctural goodness of fit to the goodness of fit of each factor or condition making up the conjunction. Doing so reveals whether the conjunction of factors provides a significant improvement in fit to some causal hypothesis over and above what each factor provides separately. Given that a conjunction constitutes a more restrictive subset of each factor making up the conjunction, comparing the goodness of fit of the conjunction with that of each factor also provides a test for the degree of generality of the causal argument, assuming that the conjunction and also each factor tested separately indeed fit the causal hypothesis being tested.[28]

Constructing Tests for Conjunctions

To begin, assume that we wish to assess whether the conjunction $(x_1 \cap x_2)$ provides a better fit to a causal necessity and sufficiency hypothesis than does x_1 alone.[29] This is equivalent to asking whether we can generalize the causal necessity and sufficiency statement from the conjunction $(x_1 \cap x_2)$ to the less restrictive causal necessity and sufficiency statement involving x_1 alone. Using notation developed in the previous section, the observed distance of the conjunction $(x_1 \cap x_2)$ from the distance expected under causal necessity and sufficiency can be given by (dropping the [*nec* and *suf*] subscript to simplify the expression)

$$D\{x_1 \cap x_2\} = \sum_{i=1}^{N} \left(z_{y(i)} - z_{\min\{x_1 \cap x_2\}(i)}\right)^2,$$

where $z_{\min\{x_1 \cap x_2\}(i)} = \Phi^{-1}\{\min\{x_{1i}, x_{2i}\}\} = \min\left\{\Phi^{-1}\{x_{1i}\}, \Phi^{-1}\{x_{2i}\}\right\}$. Finally, let $\Delta\{x_1 \cap x_2\}$ and $\Delta\{x_1\}$ be the true distances—that is, the distances measured without error—of the conjunction $(x_1 \cap x_2)$ and single factor x_1, respectively, from the causal necessity and sufficiency hypothesis.

Our question can now be given in the form of null and alternative hypotheses,

$$\begin{aligned} H_0 &: \Delta\{x_1\} \leq \Delta\{x_1 \cap x_2\} \\ H_a &: \Delta\{x_1\} > \Delta\{x_1 \cap x_2\}, \end{aligned}$$

where the null hypothesis reflects a better fit to the data for the single factor x_1 and the alternative hypothesis for the conjunction $(x_1 \cap x_2)$. Under the truth of the null hypothesis, the ratio $F^* = D\{x_1\}/D\{x_1 \cap x_2\}$ will be distributed as an F random variable on (N, N) degrees of freedom. If the test indicates that we cannot reject the null hypothesis, then we conclude that the conjunction $(x_1 \cap x_2)$ does not provide a better fit to a causal necessity and sufficiency hypothesis than does x_1 alone. As well, we can generalize the causal necessity and sufficiency statement from the conjunction of factors $(x_1 \cap x_2)$ to the less restrictive statement involving factor x_1 alone. If, on the other hand, the test indicates a rejection of the null in favor of the alternative hypothesis, then we would conclude that the conjunction of factors $(x_1 \cap x_2)$ does provide a better fit to a causal necessity and sufficiency hypothesis than does factor x_1 alone. Accordingly, we could not generalize the causal necessity and sufficiency statement from the conjunction $(x_1 \cap x_2)$ to the less restrictive statement involving x_1 alone.

This test readily generalizes to higher order conjunctions. Assume that we wish to test the conjunction of J factors, $(\bigcap_{j=1}^{J} x_j)$, against a lower order conjunction. Without loss of generality, assume that the lower order conjunction is given by dropping the last factor, $(\bigcap_{j=1}^{J-1} x_j)$. In this case, the null and alternative hypotheses are given by

$$\begin{aligned} H_0 &: \Delta\left\{\bigcap_{j=1}^{J-1} x_j\right\} \leq \Delta\left\{\bigcap_{j=1}^{J} x_j\right\} \\ H_a &: \Delta\left\{\bigcap_{j=1}^{J-1} x_j\right\} > \Delta\left\{\bigcap_{j=1}^{J} x_j\right\}. \end{aligned}$$

Under the truth of the null hypothesis, the ratio $F^* = D\{\bigcap_{j=1}^{J-1} x_j\}/D\{\bigcap_{j=1}^{J} x_j\}$ will be distributed as an F random variable on (N, N) degrees of freedom.

If the test indicates that we cannot reject the null hypothesis, then we would conclude that the conjunction $(\bigcap_{j=1}^{J} x_j)$ does not provide a better fit to a causal necessity and sufficiency hypothesis than does the lower order conjunction $(\bigcap_{j=1}^{J-1} x_j)$ and that we can generalize the causal necessity and sufficiency statement from $(\bigcap_{j=1}^{J} x_j)$ to the less restrictive statement on $(\bigcap_{j=1}^{J-1} x_j)$. If, on the other hand, the test indicates a rejection of the null in favor of the alternative hypothesis, then we would conclude that the conjunction $(\bigcap_{j=1}^{J} x_j)$ does provide a better fit to the causal necessity and sufficiency hypothesis than does the lower order conjunction $(\bigcap_{j=1}^{J-1} x_j)$ and that we cannot generalize the causal necessity and sufficiency statement from the conjunction $(\bigcap_{j=1}^{J} x_j)$ to the less restrictive statement on $(\bigcap_{j=1}^{J-1} x_j)$.

Testing Conjunctions Against Single Factors and Testing Higher Order Conjunctions Against Lower Order Conjunctions in the Stryker/Eliason Data

Returning to our welfare state and gendered labor markets example, here we bring back into consideration some of the additional factors proposed by Stryker and Eliason (2003) to have an impact on female labor force participation. Along with a strong tradition of left governance (given by membership in the set High Cumulative Left Cabinet Incumbency), these include an expanded civilian public sector (High Civilian Public Sector Size), maternity leave support (High Maternity Leave Support), and support for public day care for young children (High Support for Public Day Care Ages 0-2) and older children (High Support for Public Day Care Ages 3 to School Age).[30]

All of these factors are presumed to shape either the demand for or supply of female labor. Where available, affordable day care and maternity leave should facilitate female labor supply, and an expanded civilian public sector generally increases demand for female labor through the sex typing of such public sector jobs (for additional explanation, see Eliason, Stryker, and Tranby 2008). For maternity leave and day care factors, the term "support" refers to government support, not to public opinion support. Fuzzy-set membership scores are derived from a combination of public expenditures, the proportion of people taking advantage of some program, and in the case of maternity leave, the duration and wage replacement rate for the leave. To create more compelling illustrations and reveal stronger relations than we would otherwise see, we exclude the British data from the rest of our methodologically illustrative examples.[31]

Table 2 gives goodness-of-fit statistics for a select set of conjunctions of these five factors as well as for each factor separately. (See the online appendix for the full set of conjunctions.) Table 2 also provides the test statistics comparing lower order conjunctions against each higher order conjunction as described previously. Once again we assume a maximum midpoint measurement error of 0.1 and a Type I error rate of .05 for each test.[32]

There are a number of strategies that may be used to assess the information in Table 2. One useful approach is to begin with the highest order conjunction that fits the causal sufficiency hypothesis. We would then test whether lower order conjunctions nested within the higher order conjunction provide an equally good fit, indicating empirically unnecessary (or superfluous) factors included in the higher order conjunction. For example, suppose a conjunction of High Cumulative Left Cabinet Incumbency, High Civilian Public Sector Size, High Maternity Leave Support, High Support for Public Day Care Ages 0-2, and High Support for Public Day Care Ages 3 to School Age is sufficient to produce high female labor force participation. We then might wish to test whether any one of these factors included in this five-way conjunction is unnecessary to the goodness of fit of the data to the causal sufficiency hypothesis. That is, we may wish to test whether one or more four-way conjunctions nested within this five-way conjunction provide equally good fits to the causal sufficiency hypothesis.

This would be useful to know not just because it gives us a more parsimonious explanation, but also potentially for more pragmatic reasons. That is, if some combination of day care and public sector size alone is sufficient to produce high female labor force participation in our data, then countries that lack legacies of high cumulative left governance but nonetheless have built high public sector size and high day care support could be more confident of their capacity to produce high female labor force participation absent high cumulative left cabinet incumbency.

As well, from a goodness-of-fit standpoint, if we shed unnecessary factors from our original conjunction of five explanatory factors—High Cumulative Left Cabinet Incumbency, High Civilian Public Sector Size, High Maternity Leave Support, High Support for Public Day Care Ages 0-2, and High Support for Public Day Care Ages 3 to School Age—we may be able to identify (a conjunction of) explanatory factors that turn out to be both sufficient and necessary for producing high female labor force participation. For example, perhaps when we examine the conjunction of

Table 2
Select Set of Goodness-of-Fit and Test Statistics for Fuzzy-Set Relations

Factors and Conjunctions	Hypothesis	*SD*	*df*	*MSD*	*F*	*p*
Factors	Null association	38.57	49	0.77	—	—
High Cumulative Left Cabinet Incumbency (A)						
	Necessity	10.27	31	0.33	1.29	.21
	Sufficiency	3.48	19	0.18	0.71	.79
	Necessity and sufficiency	13.75	50	0.28	1.07	.40
High Civilian Public Sector Size (B)						
	Necessity	22.39	38	0.59	2.29	.00
	Sufficiency	0.44	12	0.04	0.14	1.00
	Necessity and sufficiency	22.83	50	0.46	1.78	.02
High Maternity Leave Support (C)						
	Necessity	1.41	7	0.20	0.78	.60
	Sufficiency	28.76	43	0.67	2.60	.00
	Necessity and sufficiency	30.16	50	0.60	2.35	.00
High Support for Public Day Care Ages 0-2 (D)						
	Necessity	26.73	40	0.67	2.60	.00
	Sufficiency	0.96	10	0.10	0.38	.95
	Necessity and sufficiency	27.69	50	0.55	2.16	.00
High Support for Public Day Care Ages 3 to School Age (E)						
	Necessity	3.85	30	0.13	0.50	.98
	Sufficiency	4.07	20	0.20	0.79	.71
	Necessity and sufficiency	7.93	50	0.16	0.62	.95
Select two-way conjunctions						
AC	Necessity	11.63	34	0.34	1.33	.18
	Sufficiency	3.34	16	0.21	0.81	.66
	Necessity and sufficiency	14.97	50	0.30	1.17	.29
	A vs. AC	—	—	—	0.92	.62
	C vs. AC	—	—	—	2.02	.01
AE	Necessity	12.64	37	0.34	1.33	.17
	Sufficiency	0.52	13	0.04	0.16	1.00
	Necessity and sufficiency	13.16	50	0.26	1.02	.47
	A vs. AE	—	—	—	1.05	.44
	E vs. AE	—	—	—	0.60	.96
CE	Necessity	4.89	34	0.14	0.56	.96
	Sufficiency	3.54	16	0.22	0.86	.61
	Necessity and sufficiency	8.44	50	0.17	0.66	.93
	C vs. CE	—	—	—	3.58	.00
	E vs. CE	—	—	—	0.94	.59

(continued)

Table 2 (continued)

Factors and Conjunctions	Hypothesis	*SD*	*df*	*MSD*	*F*	*p*
Select three-way conjunction						
ACE	Necessity	13.65	39	0.35	1.36	.15
	Sufficiency	0.43	11	0.04	0.15	1.00
	Necessity and sufficiency	14.08	50	0.28	1.10	.37
	AC vs. ACE	—	—	—	1.06	.41
	AE vs. ACE	—	—	—	0.93	.59
	CE vs. ACE	—	—	—	0.60	.96
Select four-way conjunctions						
ABCE	Necessity	31.85	45	0.71	2.76	.00
	Sufficiency	0.24	5	0.05	0.19	.96
	Necessity and sufficiency	32.09	50	0.64	2.50	.00
	ABC vs. ABCE	—	—	—	0.98	.53
	ABE vs. ABCE	—	—	—	0.97	.54
	ACE vs. ABCE	—	—	—	0.44	1.00
	BCE vs. ABCE	—	—	—	0.77	.82
ACDE	Necessity	33.27	45	0.74	2.88	.00
	Sufficiency	0.30	5	0.06	0.23	.95
	Necessity and sufficiency	33.57	50	0.67	2.62	.00
	ACD vs. ACDE	—	—	—	1.00	.50
	ACE vs. ACDE	—	—	—	0.42	1.00
	ADE vs. ACDE	—	—	—	0.97	.54
	CDE vs. ACDE	—	—	—	0.84	.73
Five-way conjunction						
ABCDE	Necessity	35.02	45	0.78	3.03	.00
	Sufficiency	0.24	5	0.05	0.19	.96
	Necessity and sufficiency	35.26	50	0.71	2.75	.00
	ABCD vs. ABCDE	—	—	—	1.00	.50
	ABCE vs. ABCDE	—	—	—	0.91	.63
	ABDE vs. ABCDE	—	—	—	0.97	.54
	ACDE vs. ABCDE	—	—	—	0.95	.57
	BCDE vs. ABCDE	—	—	—	0.85	.72

Note: Outcome is membership in the set High Female Labor Force Participation. See the online appendix for the complete set.
Source: Stryker and Eliason (2003) data for France, Belgium, Germany, Italy, and Denmark, 1977-1994.

High Civilian Public Sector Size, High Maternity Leave Support, High Support for Public Day Care Ages 0-2, and High Support for Public Day Care Ages 3 to School Age, we discover that these four factors conjoined are not just sufficient but also necessary for High Female Labor Force Participation. Thus, we would have found that the remaining factor—High

Cumulative Left Cabinet Incumbency (the potential causal factor with which we began in Figure 1)—was unnecessary for the fit to the causal sufficiency hypothesis and that shedding that factor revealed that the remaining explanatory conditions combined fit the causal sufficiency and necessity hypothesis.

In short, our general approach assumes that the researcher is interested in obtaining the lowest order, most general sufficient conjunctural (or possibly single-factor) condition for the outcome (in this case, female labor force participation) that fits the data at the prescribed Type I error rate. Thus, the researcher begins with the highest order causal conjunctions (in our example, the five-way causal conjunction) and tests first the next highest order causal conjunctions (in our example, all the four-way causal conjunctions). If all tests indicate that lower order conjunctions fit the data for sufficiency as well as the parent higher order conjunction (tests with p values greater than .05), the researcher repeats the procedure on the next set of lower order conjunctions (in our example, all the three-way causal conjunctions).

This strategy enables us to find the lowest order, most general sufficient conjunctural or single-factor condition producing the outcome. At the same time, it ensures that if there is a conjunction of factors or a factor that is necessary and sufficient to produce the outcome, it will be found. One way to look at this approach then, is that at each step, the researcher sheds unnecessary (or redundant) conjunctions and conditions from consideration until the most general sufficient (and possibly necessary) condition from those under consideration is obtained that fits the data.[33]

Another important point that should be made here has to do with the nature of goodness-of-fit tests in general. As is often the case when using goodness-of-fit techniques in other contexts (e.g., with log-linear models for categorical data analysis), we may find that the data are consistent with multiple, and perhaps contradictory, hypotheses at the prescribed Type I error rate. If the empirical tests could not be made more stringent to adjudicate among competing hypotheses, we would be left with the task of obtaining more informative data. Here, however, we show how to make our goodness-of-fit tests more exacting to more finely adjudicate between competing hypotheses given a specific set of data.

Comparing five-way, four-way, and three-way conjunctions. Goodness-of-fit tests given in Table 2 on the highest order five-way conjunction of potential explanatory factors indicate a strong fit with the hypothesis that for these data, a conjunction of High Cumulative Left Cabinet Incumbency, High Civilian Public Sector Size, High Maternity Leave Support, High Support for Public Day Care Ages 0-2, and High Support for Public Day Care Ages 3 to School Age is sufficient to produce high female labor force participation ($F = 0.19$, $p = .96$, $R = 99.67$ percent).[34] All tests of the four-way conjunctions compared to the parent five-way conjunction indicate no evidence in these data to suggest that the more restrictive five-way conjunction provides a better fit to a causal necessity and sufficiency hypothesis than do each of the four-way conjunctions. Similarly, while each four-way conjunction provides a strong fit to the causal sufficiency hypothesis for these data (see the online appendix for the full set), all tests of the lower order three-way conjunctions indicate no evidence that the more restrictive four-way conjunctions provide better fit to the causal necessity and sufficiency hypothesis than do the three-way conjunctions.

Comparing three-way to two-way conjunctions. This same pattern of fit for a causal sufficiency hypothesis holds for the three-way conjunctions when compared to the sets of lower order two-way conjunctions. Additionally, the three-way conjunction given in Table 2, involving High Cumulative Left Cabinet Incumbency × High Support for Maternity Leave × High Support for Public Day Care Ages 3 to School Age (ACE in Table 2), fits the causal necessity and sufficiency hypothesis for these data ($F = 1.10$, $p = .37$) and is 2.74 times more consistent with these data than with the null association ($R = 73.26$ percent). Were we to complete the analysis at this point, we would conclude that strong government support for maternity leave and publicly provided day care for older children combined with a strong tradition of left political governance is necessary and sufficient to produce high female labor force participation. However, while this hypothesis is consistent with the data, a more thorough analysis testing this three-way conjunction against the set of nested two-way conjunctions indicates that the three-way conjunction is unnecessarily complex.

All two-way conjunctions nested in the three-way conjunction involving High Cumulative Left Cabinet Incumbency, High Support for Maternity Leave, and High Support for Public Day Care Ages 3 to School Age (ACE in Table 2) provide a good fit to the causal necessity and sufficiency hypothesis (AC: $F = 1.17$, $p = .29$; AE: $F = 1.02$, $p = .47$; CE: $F = 0.66$, $p = .93$). That is, the combination of High Cumulative Left Cabinet Incumbency and High Support for Maternity Leave fits the necessary and sufficient causal hypothesis (AC: $F = 1.17$, $p = .29$), the combination of High Cumulative Left Cabinet Incumbency and High Support for Public Day Care Ages 3 to School Age fits the necessary and sufficient causal hypothesis (AE: $F = 1.02$, $p = .47$), and the combination of High

Support for Maternity Leave and High Support for Public Day Care Ages 3 to School Age fits the necessary and sufficient causal hypothesis (CE: $F = 0.66$, $p = .93$).

Comparing two-way conjunctions to single factors. We next consider each individual factor nested within each of these two-way conjunctions. Here, we have evidence to suggest that the conjunction of High Cumulative Left Cabinet Incumbency and High Support for Maternity Leave provides a better fit to the causal necessity and sufficiency hypothesis than does High Support for Maternity Leave alone (C vs. AC: $F = 2.02$, $p = .01$). We also have evidence to suggest that the conjunction of High Support for Maternity Leave and High Support for Public Day Care Ages 3 to School Age provides a better fit to the causal necessity and sufficiency hypothesis than does High Support for Maternity Leave alone (C vs. CE: $F = 3.58$, $p = .00$).

Combined, these results suggest that high support for maternity leave by itself does not influence female labor force participation in the same way, nor to the same extent, as it does when maternity leaves are offered in conjunction with a strong tradition of left governance or in conjunction with publicly provided day care for older children.[35] Furthermore, tests given in Table 2 show that High Cumulative Left Cabinet Incumbency and High Support for Public Day Care Ages 3 to School Age, both of which condition the effect of maternity leave on female labor force participation, operate separately of one another. This is indicated because the conjunction of High Cumulative Left Cabinet Incumbency and High Support for Public Day Care Ages 3 to School Age does not fit the causal necessity and sufficiency hypothesis better than each factor alone (A vs. AE: $F = 1.05$, $p = .44$; E vs. AE: $F = 0.60$, $p = .96$).

Assessing single factors. Finally, we examine fit statistics for each factor separately, but in the context of the knowledge gained from the analysis of the conjunctions. Keeping the focus on maternity leave, these data are consistent with a causally necessary, but not sufficient, relationship between High Support for Maternity Leave and High Female Labor Force Participation ($F = 0.78$, $p = .60$). As shown in the previous assessment of conjunctions, this further indicates no support in these data for the idea that maternity leave operates on its own in producing high female labor force participation.

On the other hand, there is evidence in these data to suggest that High Civilian Public Sector Size ($F = 0.14$, $p = 1.00$) and High Support for Public Day Care Ages 0-2 ($F = 0.38$, $p = .95$) are (separately) sufficient to produce high female labor force participation. Combined with analysis of the conjunctions (see the previous discussion, Table 2, and the online appendix), there is no evidence in these data to suggest that causal hypotheses involving any conjunction containing either or both of these factors provide a better fit to the data than each factor separately. Thus, these results suggest that for these data, High Civilian Public Sector Size and High Support for Public Day Care Ages 0-2 provide alternative routes to high female labor force participation.

Finally, we cannot reject the hypothesis that High Cumulative Left Cabinet Incumbency is necessary and sufficient for high female labor force participation for these data ($F = 1.07$, $p = .40$, $R = 73.72$ percent). We also cannot reject the hypothesis that High Support for Public Day Care Ages 3 to School Age is necessary and sufficient for high female labor force participation for these data ($F = 0.62$, $p = .95$, $R = 82.95$ percent).

Adjudicating from among competing (and perhaps contradictory) hypotheses that fit the data. As mentioned previously, goodness-of-fit techniques may at times indicate the data are consistent with multiple, and perhaps contradictory, hypotheses. This is common in nearly all methods that make use of goodness-of-fit techniques (such as log-linear analysis), techniques that are well established in many branches of science and are nearly a century old (e.g., see Stuart et al. 1999, chap. 25). In our case, there are a number of hypotheses that fit the data—or, equivalently, cannot be rejected by the data—that are logically inconsistent. For example, it is not logically possible for high support for maternity leave to be necessary, and at the same time high support for public day care for younger children to be sufficient, for high female labor force participation. Similarly, it is not logically possible for high cumulative left cabinet incumbency to be necessary and sufficient, and at the same time high support for day care for older children to also be necessary and sufficient, for high female labor force participation

Here we demonstrate one strategy enabling our goodness-of-fit method to more precisely adjudicate between two competing hypotheses, comparing the necessary and sufficient hypotheses involving left governance and day care for older children. To do this, we assess the sensitivity of each hypothesis to ever tighter degrees of measurement error, decrementing that error by 0.01 units from the maximum (midpoint) error of 0.1. The hypothesis that continues to fit the data under tighter degrees of measurement error indicates the hypothesis more strongly supported by the data. A maximum measurement error of 0.09 (=0.1-0.01) reveals no distinction between these two hypotheses. However, a maximum measurement error of 0.08 (= 0.1-0.02) leads to a rejection of the hypothesis that High Cumulative Left Cabinet Incumbency is necessary and sufficient for high female labor force participation. Thus, we find stronger support in these data for the hypothesis that High Support for Public Day Care Ages 3 to School Age is necessary and sufficient for high female labor force participation.

More generally, the previous analysis reveals the following patterns in these data. First, there is no evidence to suggest that five-way or four-way conjunctions are necessary in explaining high female labor force participation in these data. The only three-way conjunction worth considering is that involving High Cumulative Left Cabinet Incumbency, High Support for Maternity Leave, and High Support for Public Day Care Ages 3 to School Age. While the evidence for this three-way conjunction in these data is weak, it appears at least suggestive for subsequent analysis on data involving more countries and/or time periods. At the least, it suggests that researchers examine this three-way conjunction in subsequent analysis to identify and interpret effects of maternity leave provision on female labor force participation.

Second, the separate effects on female labor force participation due to a strong tradition of left governance and high public provision of day care for older children may well suggest a process characterized by a causal chain rather than one characterized by high-order conjunctions or contexts involving these two factors. While the conjunction does seem to matter for maternity leave, these two factors appear to operate separately of one another and also separately in providing a context in which maternity leave will have an impact. Interpreted against existing research on the welfare state, these findings suggest a causal chain in which a strong tradition of left governance promotes public day care for older children, which in turn promotes female labor force participation. A strong tradition of left governance may also promote strong government support of maternity leave. However, it appears that these maternity leave programs have a measurable impact on female labor force participation only in that context, that is, a state/market context as created from a strong tradition of left governance. See Stryker and Eliason (2003) and Stryker, Eliason, and Tranby (2008) for further elaborations.

An Additional Example: Social Underdevelopment in Latin America

As part of his in-depth analysis of the lasting impact that Spanish colonialism has had on Latin American development, Mahoney (2003:52) explored a set of "intervening processes" through which colonialism may have shaped "long-run development." Using fuzzy-set methods, he examined whether economic and social development or underdevelopment might be causally related to the presence or absence of a dense indigenous population, labor-intensive estates, mineral exports, tropical export agriculture, and a strong liberal political faction and/or a strong conservative political faction.

Here, we reexamine some of Mahoney's (2003, tables 6 and 7) results using his published data. To simplify, we focus only on the social underdevelopment outcome, measured based on country-level literacy and life expectancy rate data. As Mahoney (2003:78) points out, dense indigenous populations in Latin America "may have been associated with an exclusionary political elite that was unwilling to invest resources in social development, such as broad based education initiatives or improvements in rural sanitation." Similarly, labor-intensive estates may have inhibited social development because such estates are associated with high levels of economic inequality and with patronage relations that impede investment in health care and education for broad segments of society (Mahoney 2003). Political factionalism too may have played a role. With colonial elites typically divided into factions that are more liberal versus more conservative, and liberals promoting not only markets and church-state separation but also an important state role in development, a strong liberal faction may have facilitated economic and social development.

Using Ragin's proportionate *z* test (discussed previously), Mahoney's (2003) analysis revealed that a dense indigenous population is "usually necessary" for underdevelopment to be present, while the conjunction of a dense indigenous population with (a) the absence of significant labor-intensive estates and (b) the absence of strong liberal factions are "usually sufficient" for social underdevelopment to be present. Mahoney comments that while these results may be "difficult to interpret, they do suggest that under certain circumstances the presence of a dense indigenous population may have been usually necessary *and* sufficient for social underdevelopment" (2003:85). No further empirical analysis is offered to support this conjecture.

In fact, the proportionate test is not well suited to uncovering empirical support for a causally necessary and sufficient hypothesis, should that support be present in the data. In contrast, our goodness-of-fit tests enable improved empirical detective work.

Table 3 gives select goodness-of-fit statistics and tests on conjunctions for the outcome "socially underdeveloped country" and the set of factors

Table 3
Select Goodness-of-Fit and Test Statistics for Mahoney (2003) Data

Factors and Conjunctions	Hypothesis	*SD*	*df*	*MSD*	*F*	*p*
Factors	Null association	19.74	14	1.32	—	—
Dense Indigenous Population (A)						
	Necessity	1.45	1	1.45	5.66	.03
	Sufficiency	4.32	6	0.72	2.80	.05
	Necessity and sufficiency	5.77	15	0.38	1.50	.22
Absence of Labor-Intensive Estates (B)						
	Necessity	31.64	8	3.96	15.41	.00
	Sufficiency	28.59	4	7.15	27.84	.00
	Necessity and sufficiency	60.23	15	4.02	15.64	.00
Mineral or Tropical Exports (C)						
	Necessity	5.04	4	1.26	4.90	.01
	Sufficiency	23.26	9	2.58	10.06	.00
	Necessity and sufficiency	28.29	15	1.89	7.35	.00
Absence of Strong Liberals (D)						
	Necessity	20.09	9	2.23	8.69	.00
	Sufficiency	1.65	2	0.82	3.20	.07
	Necessity and sufficiency	21.73	15	1.45	5.64	.00
Strong Conservatives (E)						
	Necessity	9.58	5	1.92	7.46	.00
	Sufficiency	6.65	6	1.11	4.32	.01
	Necessity and sufficiency	16.22	15	1.08	4.21	.00
Select two-way conjunctions						
AB	Necessity	32.62	8	4.08	15.88	.00
	Sufficiency	0.48	1	0.48	1.86	.19
	Necessity and sufficiency	33.09	15	2.21	8.59	.00
	A vs. AB	—	—	—	0.17	1.00
	B vs. AB	—	—	—	1.82	.13
AC	Necessity	5.04	4	1.26	4.90	.01
	Sufficiency	2.86	5	0.57	2.23	.10
	Necessity and sufficiency	7.90	15	0.53	2.05	.09
	A vs. AC	—	—	—	0.73	.73
	C vs. AC	—	—	—	3.58	.01
AD	Necessity	20.09	9	2.23	8.69	.00
	Sufficiency	0.00	0	—	—	—
	Necessity and sufficiency	20.09	15	1.34	5.22	.00
	A vs. AD	—	—	—	0.29	.99
	D vs. AD	—	—	—	1.08	.44

(continued)

Table 3 (continued)

Factors and Conjunctions	Hypothesis	*SD*	*df*	*MSD*	*F*	*p*
AE	Necessity	9.58	5	1.92	7.46	.00
	Sufficiency	2.86	5	0.57	2.23	.10
	Necessity and sufficiency	12.44	15	0.83	3.23	.01
	A vs. AE	—	—	—	0.46	.93
	E vs. AE	—	—	—	1.30	.31

Note: Outcome is membership in the set Socially Underdeveloped Country. See the online appendix for the complete set.

Dense Indigenous Population, Absence of Labor-Intensive Estates, Mineral or Tropical Exports, Absence of Strong Liberals, and Strong Conservatives. The full set of fit statistics and tests on the conjunctions are given in the online appendix. For definitions and measures of causal and outcome factors, see Mahoney (2003).

We are particularly interested in comparing Mahoney's (2003) results on the proportionate test described previously against the results suggested in Table 3 using our goodness-of-fit tests. In comparing the results, three things are important to keep in mind. First, while our goodness-of-fit tests take into account the distance of each case from that expected under some causal hypothesis, the proportionate test is insensitive to that distance. Second, while our goodness-of-fit tests are based on measurement error, the proportionate test is based on sampling error. Third, the language and logic buttressing the proportionate test rest with rather arbitrary conventions linking arbitrarily derived sharp cutoffs in the proportion of cases consistent with some hypothesis (e.g., .65) to rather arbitrary terms (e.g., "usually") to describe the causal relations. In contrast, the language and logic buttressing our goodness-of-fit tests rest on the firm foundation of Neyman-Pearson hypothesis testing. Nearly 100 years of research from all areas of science rely on this foundation in assessing whether empirical information is consistent with some hypothesis (Stuart et al. 1999). Differences between our results and subsequent substantive conclusions and those of Mahoney stem from these distinctions.

Tests on lower order conjunctions against higher order parent conjunctions indicate that for these data, we need not consider higher than two-way conjunctions. (The full set of test statistics is given in the online appendix.) Table 3 presents the two-way conjunctions informing the comparison of results involving the condition Dense Indigenous Population.

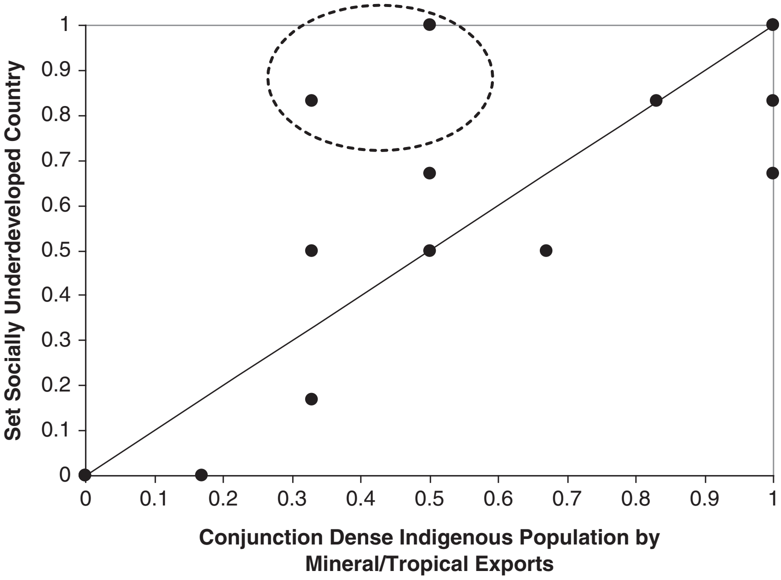
Our goodness-of-fit tests suggest that the conjunction of a Dense Indigenous Population with the Absence of Labor-Intensive Estates fits the causal sufficiency hypothesis for these data ($F = 1.86$, $p = .19$). This is consistent with Mahoney's (2003) findings.

However, whereas Mahoney suggested that the effects of a dense indigenous population and an absence of labor estates may operate in combination, when we test each factor— Dense Indigenous Population and Absence of Labor-Intensive Estates—against the conjunction of the two factors, we find that the conjunction does not provide a better fit than does each factor separately (A vs. AB: $F = 0.17$, $p = 1.00$; B vs. AB: $F = 1.82$, $p = .13$). Thus, density of the indigenous population and labor-intensive estates operate separately in inhibiting social development in Latin America. Based on our analyses, the same conclusion can be drawn with respect to the density of the indigenous population and the absence of strong liberal factions or the presence of strong conservative factions. That is, neither of the two-way conjunctions combining Dense Indigenous Population with the presence of Strong Conservatives or with the Absence of Strong Liberals provides a better fit to the data than does each factor alone.

Comparing our results to Mahoney's (2003) so far highlights differences between the goodness-of-fit and proportionate tests and differences in what is revealed to the researcher. An even more telling difference between our results and Mahoney's is revealed when we examine the conjunction of a dense indigenous population with the presence of mineral or tropical exports. In Mahoney's analysis, this conjunction does not pass the proportionate test for causal sufficiency. In contrast, our analysis shows that the conjunction of a dense indigenous population with the presence of mineral or tropical experts is sufficient for social underdevelopment.

Figure 3 shows the conjunction of a dense indigenous population and mineral or tropical exports plotted against social underdevelopment. Putting aside for a moment that the proportionate test is based on sampling error, notice that any test based on the proportion alone will treat each case below the diagonal as providing equal weight against the causal sufficiency hypothesis. Of cases below the diagonal, those closer to the diagonal—closer to that expected under the hypothesis of causal sufficiency—weigh just as much against the hypothesis of causal sufficiency as those further from the diagonal—further from that expected under the hypothesis of causal sufficiency. Whereas the proportionate test is insensitive to these distances, the goodness-of-fit F tests take these distances, along with the number of cases inconsistent with some causal hypothesis, into account.[36] Thus, although there are 5 cases out of 15 (33.3 percent)

Figure 3
Fuzzy-Set Graph Showing Relation Between Membership in Set Socially Underdeveloped Country and the Conjunction of Membership in Sets Dense Indigenous Population and Presence of Mineral or Tropical Exports



Note: Of the 15 cases, there are 2 cases plotted at the (1, 1) coordinates and 2 cases plotted at the (0, 0) coordinates. Thus, only 13 plot marks of the 15 cases are visible.
Source: Data from Mahoney (2003).

below the diagonal, the mean squared distance of 0.57 is entirely consistent with the hypothesis of causal sufficiency given the specified degree of measurement error and a Type I error rate of .05.

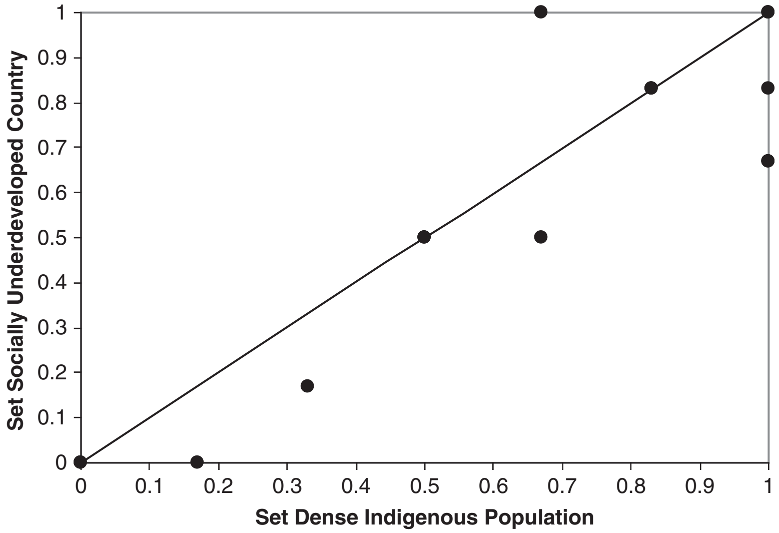
Now examine the cases above the diagonal. These cases are inconsistent with the causal necessity hypothesis. There is 1 fewer case inconsistent with the causal necessity hypothesis (4 out of 15 cases, 26.6 percent) than was so for the causal sufficiency hypothesis (5 out of 15 cases, 33.3 percent). This suggests that the proportionate test would indicate that the data are a better fit to the hypothesis of causal necessity than to the hypothesis of causal sufficiency. But when we examine the mean squared distance of the 4 cases from the distance expected under causal necessity, our mean squared distance of 1.26 is too large given the specified degree of measurement error and a Type I error rate of .05. If we were to quit at this point, we would reject the hypothesis of causal necessity for these data. However, we still need to examine the goodness-of-fit statistics for the hypothesis of causal necessity and sufficiency combined.

When we examine these last goodness-of-fit statistics, we find that the combined weight of all 15 cases suggests that we cannot reject the hypothesis of causal necessity and sufficiency for these data ($F = 2.05$, $p = .09$). Why should this be so, when we just observed an apparent lack of fit for the hypothesis of causal necessity itself? The answer lies with the amount of empirical content—the number of cases in the data—taken into account when testing for causal necessity, causal sufficiency, and causal necessity and sufficiency combined. Whereas tests for causal necessity and sufficiency take into account only the number and distance of cases above and below the main diagonal, respectively, the test for causal necessity and sufficiency combined takes into account the combined empirical content from all 15 cases.

In short, the conclusion to be drawn from comparing the test for causal necessity alone to that for causal necessity and sufficiency reflects well the reality—and ambiguity—of these data. That is, while the 2 outlying cases circled in Figure 3 weigh heavily against the causal necessity hypothesis alone, when considered against the combined weight of all the data, these 2 cases are not sufficiently distant nor are they sufficient in number to outweigh the other 13 cases in testing the combined causal necessity and sufficiency hypothesis.

There is one additional set of conclusions to be drawn from our goodness-of-fit statistics for two-way conjunctions. That is, we have evidence in these data to suggest that the conjunction of the presence of Mineral or Tropical Exports with a Dense Indigenous Population provides a better fit to the causal necessity and sufficiency hypothesis than does the presence of Mineral or Tropical Exports alone (C vs. AC: $F = 3.58$, $p = .01$). However, the same is not true when testing this conjunction against the presence of a Dense Indigenous Population alone. That is, the conjunction of Mineral or Tropical Exports with a Dense Indigenous Population does not provide a better fit to the causal necessity and sufficiency hypothesis for these data than does the more general condition involving the presence of a Dense Indigenous Population alone (A vs. AC: $F = 0.73$, $p = .73$). In sum, for these data, the presence of a dense indigenous population is strongly related to social underdevelopment independently and regardless of the context of other potential causal factors considered.

Figure 4
Fuzzy-Set Graph Showing Relation Between Membership in Sets Socially Underdeveloped Country and Dense Indigenous Population



Note: Of the 15 cases, there are 2 cases at the (1, 1) coordinates, 2 cases at the (.83, .83) coordinates, 2 cases at the (.5, .5) coordinates, 2 cases at the (0, 0) coordinates, and 2 at the (1, .67) coordinates. Thus, only 10 plot marks of the 15 cases are visible.
Source: Data from Mahoney (2003).

Figure 4 shows the biplot for this relationship. Although caution is warranted with any visual inspection, it appears that the information in Figure 4 provides some support for a causally necessary and sufficient relation between the presence of a dense indigenous population and social underdevelopment in Latin American countries. In fact, an examination of the fit statistics for the main factors suggests that the causal necessity and sufficiency hypothesis involving the presence of a Dense Indigenous Population fits these data well ($F = 1.50$, $p = .22$). Moreover, the relative consistency measure of 77.38 percent indicates further that these data are 3.42 times more consistent with this hypothesis than with the hypothesis of no association. This in turn provides the solid empirical support that Mahoney (2003) himself could not provide for his conjecture about the causally necessary and sufficient role that density of the indigenous population plays in shaping social underdevelopment in Latin America.

In sum, once armed with tools provided by the goodness-of-fit F tests and tests for conjunctions demonstrated in this article, we are able to provide a more nuanced and precise fuzzy-set empirical analysis of possible causal factors shaping social underdevelopment in Latin America than was Mahoney (2003) with the tools he had available. Above all, we are able to assess empirically the fit of the data to a relationship of causal necessity and sufficiency combined. Because the proportionate z test cannot assess the fit of the data to this type of relationship, Mahoney was constrained to leave as speculative a substantively important hypothesis that can, in fact, be supported empirically by applying fuzzy-set methodology to Mahoney's data. This highlights perhaps the most significant strength of our goodness-of-fit approach relative to the prior proportionate test approach.

Conclusions

Fuzzy-set ideas and methods have already provided researchers with powerful tools to explore and better understand social phenomena. In this article, we helped to strengthen this young tradition by buttressing and extending in significant ways Ragin's (2000) use of fuzzy-set methods to assess causal hypotheses through controlled comparison of cases. We did so by (a) formally accounting for measurement error in fuzzy-set scores, (b) providing precise measures of the distance and consistency of the information in a fuzzy-set graph from specific causal arguments and an argument of no association, and (c) constructing goodness-of-fit F tests and tests on conjunctions. Perhaps most important, our goodness-of-fit tests and the logic underlying those testing procedures established a firm inferential foundation for fuzzy-set methodology. This foundation is no different from that supporting countless empirical research studies in sociology—and in science in general—based on Neyman-Pearson-type hypothesis testing. Our foundations and extensions of fuzzy-set methods also clearly indicate that these methods are not susceptible to key criticisms previously aimed at QCA and related methodologies. Thus, suggestions that general QCA methodology—of which fuzzy-set methods are part—cannot usefully apply a probabilistic approach or deal with errors in data are misguided (Lieberson 1994:1225; see also Lieberson 1991; Sobel 1995).

Aside from placing fuzzy-set methodology on a firmer inferential footing, our goodness-of-fit tests provide more information to the researcher on the fit of causal hypotheses to the data than does the proportionate test currently in use. This is in part because our tests make explicit use of the distance of each case from that expected under some causal hypothesis and in part because our framework provides a test for the causal necessity and sufficiency hypothesis. That the proportionate test leaves the researcher blind to some relations suggested by the data is highlighted by our example building on Mahoney's (2003) results, comparing them with the results suggested by our goodness-of-fit methodology. Additionally, our descriptive measures provide a comparison between the distance of the data from each causal hypothesis and from the hypothesis of no association. Rarely do those using the proportionate z test compare patterns in the data to that expected under a null relationship. This too leaves the researcher blind to some patterns in the data, as some data may be more consistent with the null relationship hypothesis than with any of the causal hypotheses.

Finally, we provide a clear testing procedure with which to assess higher order conjunctions against lower order conjunctions and single conditions or factors. These tests are invaluable in that they allow the researcher to assess the degree of causal complexity supported by the data. In addition, researchers can use these tests to assess whether data are more consistent with a conjunctural causal relationship, with one involving a causal chain including sequential nonconjunctural (independent) conditions, or with a combination of the two. In using the goodness-of-fit tests and tests on conjunctions to distinguish among these types of causal relations, researchers have a powerful tool to help assess empirical social processes.

Notes

[1] See the many useful contributions listed on the Comparative Methods for the Advancement of Systematic Cross-Case Analysis and Small-N Studies Web site (http://www.compasss.org) and those found in the special issue of *Sociological Methods & Research* (Ragin and Pennings 2005) devoted to fuzzy-set methodology.

[2] For valuable insights into the methodology of necessary conditions, see Braumoeller and Goertz (2000, 2003), Ragin (2003), and other useful contributions found in the edited volume by Goertz and Starr (2003).

[3] See Stuart, Ord, and Arnold's (1999) useful discussion of the Neyman-Pearson theory of hypothesis testing (particularly chapters 20, 21, 22, and 26) and its relation to other methodologies (e.g., Bayes).

4. This lineage extends at least back to Mill's ([1843] 1967) work on causal logic. Although rarely brought to the front of these discussions, Mill's causal logic is distinct from the manipulative counterfactual account of cause (Holland 1986; Marini and Singer 1988; Sobel 1995, 1996). Rather than enter into what is often an endless, fruitless debate on which of these logics is more flawed, we instead assume that researchers are already informed of the advantages and limitations in both. We likewise assume that Ragin's (2000) reformulation of Mill's causal logic is reasonably suited to the research question of interest. See Sobel (1995, 1996) and Holland (1986) for informative treatments and Marini and Singer (1988) for a helpful introduction.

5. For strict adherence to the logic, the empirical content of a case relates not to the probability that some cause or outcome may be observed, but rather to the degree to which a case belongs to the set defined by the causal condition and the set defined by the outcome, thus providing evidence of the cause and of the outcome, respectively. For some purposes, this distinction is crucial. However, here it does little if any damage to consider the "likelihood of observing" and the "degree of belonging to a set" as nearly interchangeable. Indeed, some have argued that membership scores are in fact conditional probabilities, the probability that a randomly chosen observer (researcher) will claim that a specific observation (case) belongs to a specific set (causal condition or outcome). See Loginov (1966) for an articulation of this view of membership scores; see Zadeh (1995) for an opposing view.

6. For details, see Stryker and Eliason (2003); for subsequent extensions, including combining fuzzy-set methods with estimation of complier average causal effects, see Eliason, Stryker, and Tranby (2008).

7. At first glance, it may appear that what we develop here is similar to Ragin's (2006) measures of consistency. However, while our approach develops an inferential foundation for assessing goodness of fit, Ragin's measure of consistency is instead a purely descriptive measure indicating the degree to which the data are consistent with either necessity or sufficiency. While this is a useful descriptive measure, it fails as an inferential tool in that there is no known expected distribution for Ragin's consistency ratios, assuming necessity or sufficiency holds for a specific collection of cases. Instead, there is a single value—namely, 1—that the data should conform to were they completely consistent with necessity or sufficiency (depending, of course, on the measure being calculated). Beyond that, these measures provide only a descriptive account of the proportion of the fuzzy scores consistent with some causal argument and cannot be used in an inferential manner as with our goodness-of-fit tests.

8. Ragin (2000:223-26) does address measurement error through an adjustment factor on the rule governing consistency with some causal argument. Using this adjustment could very well affect the proportion of cases considered inconsistent with some causal argument, giving what may be considered a measurement-adjusted proportion. Nevertheless, the standard error used to obtain the z test on the adjusted proportion still does not reflect measurement error. Rather, it reflects the sampling error in the newly adjusted proportion.

9. Another drawback worth mentioning is not so much in the proportion itself, but rather in using arbitrary descriptors attached to arbitrary values on the proportion to convey the strength of the causal relationship. This is similar to the ill-advised practice of attaching arbitrary descriptors to the r^2 in some regression models. For example, it is unclear why a proportion of .65 refers to the label "usually" when describing the causal relation (e.g., Ragin 2000; Mahoney 2003). In this case, both the proportion and the label are arbitrary. We might just as well claim that a proportion of .64 (or .643, should we desire more precision) constitutes "usually" or that we prefer to use "typically" to refer to a proportion of .65 instead of

"usually." Moreover, given that fuzzy-set scores reflect degree of membership in the set given by the causal condition or the outcome, and given that to assess a causal relation the subset principle is applied to those membership scores, the proportion of cases consistent with some causal hypothesis in a fuzzy-set graph does not reflect the proportion of times that the causal relation holds in the data. So even if all researchers agreed on labeling conventions, attaching descriptors such as "*X* sufficient" or "*X* necessary" when *Y* percent of the cases are consistent with the corresponding causal hypothesis conveys misleading information about the causal relation as suggested in the fuzzy-set analysis. Interestingly, fuzzy-set researchers have developed a detailed calculus around what they call fuzzy probabilities—words such as "usually" that reflect a subjective probability or range of probabilities (e.g., Zadeh 1995). To use such fuzzy probabilities as "usually" in a rather arbitrary way thus appears inconsistent with the theoretical and logical foundations of fuzzy-set methodology as developed in other areas of science.

[10] This logic is used in linear statistical models to establish the so-called null model. But it also is used in nonlinear statistical models to do the same. It is used in other nonlinear systems analyses as well. In short, there is nothing inherently linear about the logic. Fuzzy-set analysis, qualitative comparative analysis, and the causal analysis of sufficiency and necessity in general may be considered a very general nonparametric model of a general nonlinear system.

[11] While these goodness-of-fit tests do not assess coverage as described by Ragin (2006), our test for necessity and sufficiency can be considered a test for Goertz's (2006) notion of relevance. Specifically, any goodness-of-fit test for necessity and sufficiency is necessarily a test for a highly relevant necessary condition. See Goertz for details.

[12] While Ragin (2000) and others (e.g., Mahoney 2003) suggest a coding strategy for fuzzy-set membership scores based on an ordinal semantic structure, the membership scores themselves derive from the continuous range [0, 1]. Foundational fuzzy-set theories from which fuzzy-set membership scores derive consider these scores continuous on the range [0, 1] (e.g., Klement, Puri, and Ralescu 1986).

[13] Importantly, that $z_{x(i)}$ and $z_{y(i)}$ derive from a standard unit normal distribution with zero mean and unit variance is not an untestable assumption but is true by definition of the transformation on the function $\Phi^{-1}\{\cdot\}$. One could, of course, induce some other mean and/or variance on either set of normalized fuzzy-set scores. Or for that matter, one could impose an entirely different distributional form (e.g., log-normal or gamma distribution). However, we would still need to transform these nonstandardized, perhaps nonnormal, scores into their standardized normal form to calculate the goodness-of-fit *F* tests to be discussed later. This is a transformation of a fuzzy random variable whose resulting distributional properties are known by definition of the transformation itself. It is not some untestable assumption such as one may encounter in the case of, for example, properties assumed for some error term for some statistical model that are required, in turn, for properties of an estimator to hold. Moreover, no assumptions need be made on the fuzzy-set scores x_i or y_i, except that they fall in the range [0,1]. This is always the case for the type of fuzzy-set analysis discussed here. See DeGroot (1986, chap. 3) and Stuart and Ord (1987, sec. 6.27) for general introductions to transformations on random variables. See Puri and Ralescu (1985) for useful discussion, derivations, and theorems on normally distributed fuzzy random variables.

[14] Other distances or loss functions—for example, one based on absolute deviations—may be of interest. However, the Euclidean distance provides the actual unit distance between any two (normalized) points in an *xy* biplot, such as those used in fuzzy-set analysis. Thus the

Euclidean distance is a natural choice to measure the collective distance of the points in the plot from that expected under some argument. We use and report the squared Euclidean distances for the normalized scores, as these provide the component information for the relative distance measures we use and the basis for the goodness-of-fit *F* tests.

15. Note that we follow standard conventions where uppercase letters refer to random variables and lowercase letters refer to observations on a random variable.

16. While fuzzy-set membership scores themselves can be thought of as reflecting measurement error, that error is with respect to membership of a case in some set and not with respect to the scores themselves. Instead, our goodness-of-fit tests can be thought of as capturing the degree of measurement variability in the scores themselves due to disagreements and other perturbations in the research community about coding cases.

17. We are grateful to one reviewer for pointing out that measurement error itself can be considered from a fuzzy-set theoretic standpoint. This appears to be a fruitful avenue for future research.

18. Proof of Theorem 4.1 given in Puri and Ralescu (1985) shows that the difference of a normal fuzzy random variable and its expected value is normally distributed, insofar as the expected value in the domain of support of the normal fuzzy random variable is bounded (i.e., not infinite). So treating $z_{x(i)}$ as the expected value of a normal fuzzy random variable $z_{y(i)}$ also results in $(z_{y(i)} - z_{x(i)})$ being normally distributed.

19. We would like to point out that it was after reading the first draft of our attempt to provide a sound inferential infrastructure for fuzzy-set analysis that Sheldon Stryker encouraged us to more fully develop the goodness-of-fit tests described here. Because of this, we warmly dedicate this *F* statistic to him and suggest it be referred to as Sheldon's *F* statistic in future work (rather than Stryker's *F* statistic, as one of the authors shares his last name).

20. Importantly, this independence assumption is not an "independence of observations" assumption. Rather, it is an assumption about the independence of (a) the function in the numerator based on the expected relationship from (b) the denominator that is a function of the errors. Simply put, the numerator and denominator need to be independent for the ratio to have the expected *F* distribution under the null hypothesis. See Stuart and Ord (1987) on this property for the central *F* ratio, which is the ratio of two independent χ^2 random variables.

21. See Weisberg (1985:89-91) for similar derivations for the goodness-of-fit *F* test in the context of a regression model with repeated measures.

22. It is interesting to note here that a highly relevant necessary condition, as suggested in Goertz (2006), is one that is in fact necessary and sufficient. From this, a test of a necessary and sufficient condition is necessarily a test of a highly relevant necessary condition as defined by Goertz. Therefore, developing the inferential apparatus as we do here for necessary and sufficient conditions in fuzzy-set analyses at the same time does so for highly relevant necessary conditions as well.

23. Note that the expected values of the errors need not be zero.

24. A Type I error rate often refers to the rate of incorrectly rejecting the null hypothesis over repeated samples from some population. However, it may also refer to the rate of incorrectly rejecting the null hypothesis when the data actually do derive from, or exhibit, the relationship posited by the null hypothesis allowing for a specified degree of measurement error (or, equivalently, precision in measurement instrumentation). This latter interpretation of the Type I error rate, often associated with experimental designs in the physical sciences, is consistent with our goodness-of-fit test.

[25] See Morgan and Winship (2007) for an extensive and insightful treatment of this work.

[26] However, see Eliason et al. (2008) for methods combining the two causal traditions.

[27] Importantly, the descriptive measure R requires no assumption about normality, linearity, or any equation underlying the relationship between fuzzy-set scores on the outcome and on the hypothesized cause. R simply gives the proportion of empirical information consistent with a specific causal argument, relative to the total empirical information in the graph. This measure, while sharing some general features with the R^2 for linear models, has more in common with concentration and entropy measures used in conjunction with general likelihood-based modeling. See Eliason (1993) for details on these likelihood-based measures.

[28] There are parallels between this and the formalization of theoretical scope conditions as presented by Walker and Cohen (1985). Although Walker and Cohen elaborate their ideas in a very different context, it is nevertheless true that higher order conjunctural conditions necessarily restrict the scope of some theory beyond that posited by lower order conjunctions or main conditions and factors. Testing the higher order conjunction against the lower order conjunction or main conditions/factors thus provides a test of the scope of some theory. We leave elaboration and exploitation of this connection to further work.

[29] The choice of x_1 or x_2 is arbitrary, as the results subsequently apply equally to testing either x_1 or x_2 against the conjunction of x_1 and x_2.

[30] Corresponding sets or conditions are given in parentheses.

[31] See Stryker and Eliason (2003) for details on these fuzzy sets and for specific substantive hypotheses, elaborations, and comparisons of results including and excluding the British data. For an extended discussion of the potentially conflicting incentives on female labor supply built into maternity leave programs of different levels of generosity combined with different lengths of duration, see Eliason et al. (2008).

[32] While these fit statistics may be obtained using standard spreadsheets such as Excel, we have written and made available for free a computer program to facilitate calculation of these statistics. Please send an email to seliason@email.arizona.edu for more information and to download the program.

[33] This is true because conjunctions are represented by the minimum fuzzy-set score of the set of scores on the factors making up the conjunction. Thus, as unnecessary components (whether unnecessary lower order conjunctions or single factors) are stripped away from a higher order conjunction that gives a sufficient relation with the outcome, we necessarily get ever closer to a necessary and sufficient relation, were one to exist in the data.

[34] The relative consistency measures, R, are not presented in Table 2, although these are readily calculated from the distances given in Table 2.

[35] As welfare state researchers point out, maternity leave, day care, and other family policies were enacted at different times for different reasons in different countries. Already by the 1960s, Scandinavian countries were concerned with supporting female employment, but other countries were motivated more by population, health, or education concerns than with promoting or supporting mothers' labor force participation (for more discussion, see, e.g., Gauthier 1996; Eliason et al. 2008). For discussion of the causal import of day care for children 0-2 years, and comparison of the impact on female labor force participation of day care for older and younger children in the data we use here and also in a much larger set of advanced capitalist democracies over a longer time frame, see Stryker and Eliason (2003) and Eliason et al. (2008).

36. The numerator in the test statistic itself reflects the distance; the number of cases inconsistent with some hypothesis and the total number of cases are reflected in the test statistic as well as in the numerator and denominator degrees of freedom, respectively.

References

Almond, Russell G. 1995. "Fuzzy Logic: Better Science? Or Better Engineering?" *Technometrics* 37:267-70.

Arfi, Baderine. 2005. "Fuzzy Decision Making in Politics: A Linguistic Fuzzy-Set Approach (LFSA)." *Political Analysis* 13:23-56.

Arfi, Baderine. 2006. "Linguistic Fuzzy-Logic Game Theory." *Journal of Conflict Resolution* 50:28-57.

Black, Max. 1937. "Vagueness: An Exercise in Logical Analysis." *Philosophy of Science* 4:427-55.

Braumoeller, Bear F. and Gary Goertz. 2000. "The Methodology of Necessary Conditions." *American Journal of Political Science* 44:844-58.

Braumoeller, Bear F. and Gary Goertz. 2003. "The Statistical Methodology of Necessary Conditions." Pp. 197-223 in *Necessary Conditions: Theory, Methodology, and Applications*, edited by Gary Goertz and Harvey Starr. Oxford, UK: Rowan & Littlefield.

Degroot, Morris. 1986. *Probability and Statistics,* 2nd ed. Reading, MA: Addison-Wesley.

Eliason, Scott R. 1993. *Maximum Likelihood Estimation: Logic and Practice* (Sage Quantitative Applications in the Social Sciences, No. 96). Newbury Park, CA: Sage.

Eliason, Scott R., Robin Stryker, and Eric Tranby. 2008. "The Welfare State, Family Policies and Women's Labor Force Participation: Combining Fuzzy-Set and Statistical Methods to Assess Causal Relations and Estimate Causal Effects." Pp. 135-95 in *Method and Substance in Macrocomparative Analysis*, edited by Lane Kenworthy and Alex Hicks. New York: Palgrave Macmillan.

Gauthier, Anne. 1996. *The State and the Family: A Comparative Analysis of Family Policies in Industrialized Countries*. Oxford, UK: Clarendon.

Goertz, Gary. 2003. "The Substantive Importance of Necessary Condition Hypotheses." Pp. 65-94 in *Necessary Conditions: Theory, Methodology, and Applications,* edited by Gary Goertz and Harvey Starr. Oxford, UK: Rowan & Littlefield.

Goertz, Gary. 2006. "Assessing the Trivialness, Relevance, and Relative Importance of Necessary or Sufficient Conditions in Social Science." *Studies in Comparative International Development* 41:88-109.

Goertz, Gary and James Mahoney. 2005. "Two-Level Theories and Fuzzy Sets." *Sociological Methods & Research* 33:497-538.

Goertz, Gary and Harvey Starr, eds. 2003. *Necessary Conditions: Theory, Methodology, and Applications*. Oxford, UK: Rowan & Littlefield.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith 1999. "The Economics and Econometrics of Active Labor Market Programs." pp. 1865-2097 in Handbook of Labor Economics, Volume 111, edited by Orley Ashenfelter and David Card. New York: Elsevier.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945-60.

Imbens, Guido W. and Donald B. Rubin. 1997. "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance." *The Annals of Statistics* 25:305-327.

Kandel, Abraham, Alejandro Martins, and Roberto Pacheco. 1995. "On the Very Real Distinction Between Fuzzy and Statistical Methods." *Technometrics* 37:276-81.

Klement, E. P., M. L. Puri, and D. A. Ralescu. 1986. "Limit Theorems for Fuzzy Random Variables." *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, Predictability in Science and Society* 407:171-82.

Laviolette, Michael, John W. Seaman Jr., J. Douglas Barrett, and William H. Woodall. 1995. "A Probabilistic and Statistical View of Fuzzy Methods." *Technometrics* 37:249-61.

Lieberson, Stanley. 1991. "Small N's and Big Conclusions: An Examination of the Reasoning in Comparative Studies Based on a Small Number of Cases." *Social Forces* 70:307-20.

Lieberson, Stanley. 1994. "More on the Uneasy Case for Using Mill-Type Methods in Small-N Comparative Studies." *Social Forces* 72:1225-37.

Loginov, V. J. 1966. "Probability Treatment of Zadeh Membership Functions and Their Use in Pattern Recognition." *Engineering Cybernetics*: 68-69.

Mahoney, James. 2003. "Long-Run Development and the Legacy of Colonialism in Spanish America." *American Journal of Sociology* 109:50-106.

Manton, Kenneth G., Eric Stallard, Max A. Woodbury, H. Dennis Tolley, and Anatoli I. Yashin. 1987. "Grade-of-Membership Techniques for Studying Complex Event History Processes With Unobserved Covariates." *Sociological Methodology* 17:309-46.

Manton, Kenneth G., Max A. Woodbury, Eric Stallard, and Larry S. Corder. 1992. "The Use of Grade-of-Membership Techniques to Estimate Regression Relationships." *Sociological Methodology* 22:321-81.

Marini, Margaret Mooney and Burton Singer. 1988. "Causality in the Social Sciences." *Sociological Methodology* 18:347-409.

McNeill, Daniel and Paul Freiberger. 1993. *Fuzzy Logic*. New York: Simon & Schuster.

Mill, John Stuart. [1843] 1967. *A System of Logic: Ratiocinative and Inductive*. Reprint, Toronto, Canada: University of Toronto.

Montgomery, James D. 1998. "Toward a Role-Theoretic Conception of Embeddedness." *American Journal of Sociology* 104:92-125.

Montgomery, James D. 2000. "The Self as a Fuzzy Set of Roles, Role Theory as a Fuzzy System." *Sociological Methodology* 30:261-314.

Morgan, Stephen L. and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge, UK: Cambridge University Press.

Puri, Madan L. and Dan A. Ralescu. 1985. "The Concept of Normality for Fuzzy Random Variables." *Annals of Probability* 13:1373-79.

Ragin, Charles C. 1987. *The Comparative Method*. Berkeley: University of California.

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

Ragin, Charles C. 2003. "Fuzzy-Set Analysis of Necessary Conditions." Pp. 179-96 in *Necessary Conditions: Theory, Methodology, and Applications*, edited by Gary Goertz and Harvey Starr. Oxford, UK: Rowan & Littlefield.

Ragin, Charles C. 2006. "Set Relations in Social Research: Evaluating Their Consistency and Coverage." *Political Analysis* 14:291-310.

Ragin, Charles C. and Paul Pennings. 2005. "Fuzzy Sets" [special issue]. *Sociological Methods & Research* 33.

Rindfuss, Ronald R. and Futing Liao. 1988. "Medical and Contraceptive Reasons for Sterilization in the United States." *Studies in Family Planning* 19:370-80.

Sobel, Michael E. 1995. "Causal Inference in the Social and Behavioral Sciences." Pp. 1- 38 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, edited by G. Arminger, C. C. Clogg, and M. E. Sobel. New York: Plenum.

Sobel, Michael E. 1996. "An Introduction to Causal Inference." *Sociological Methods & Research* 24:353-79.

Stojakovic, Mila and Zoran Stojakovic. 1996. "Support Functions for Fuzzy Sets." *Proceedings: Mathematical, Physical and Engineering Sciences* 452:421-38.

Stryker, Robin and Scott R. Eliason. 2003. "The Welfare State, Gendered Labor Markets and Aggregate Political Orientations in France, Belgium, Germany, Italy, Denmark and Britain, 1977-1994." Robert Schuman Centre Advanced Studies Working Paper RSC2003/20, European University Institute, Florence, Italy.

Stuart, Alan and Keith Ord. 1987. *Kendall's Advanced Theory of Statistics*. Vol. 1, *Distribution Theory*. 5th ed. New York: Oxford University Press.

Stuart, Alan, Keith Ord, and Steven Arnold. 1999. *Kendall's Advanced Theory of* Statistics. Vol. 2A, *Classical Inference & the Linear Model*. 6th ed. New York: Oxford University Press.

Walker, Henry A. and Bernard P. Cohen. 1985. "Scope Statements: Imperatives for Evaluating Theory." *American Sociological Review* 50:288-301.

Weisberg, Sanford. 1985. *Applied Linear Regression*. 2d ed. New York: John Wiley.

Western, Bruce. 1998. "Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach." *American Journal of Political Science* 42:1233-59.

Western, Bruce. 2001. "Bayesian Thinking About Macrosociology." *American Journal of Sociology* 107:353-78.

Western, Bruce and Simon Jackman. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88:412-23.

Winship, Christopher and Stephen L. Morgan. 1999. "The Estimation of Causal Effects from Observational Data" *Annual Review of Sociology* 25:659-706.

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

Zadeh, Lotfi. 1995. "Probability Theory and Fuzzy Logic Are Complimentary Rather Than Competitive." *Technometrics* 37:271-76.

Scott R. Eliason is associate professor of Sociology, faculty affiliate of The BIO5 Institute, and regular member of the Statistics Graduate Interdisciplinary Degree Program at the University of Arizona. He teaches and does research in the areas of inequality and stratification, causal inference, and categorical data modeling.

Robin Stryker is professor of sociology and affiliated professor of law at the University of Arizona. She has written broadly on theory and methods, the politics of law, legitimacy, organizational and institutional change, and the welfare state and social policy. She currently has a John Simon Guggenheim Foundation fellowship for her project "Social Science in Government Regulation of Equal Employment Opportunity."