Recent advances in fuzzy-set methods and their application to policy questions

Charles C. Ragin
Department of Sociology
University of Arizona
Tucson, Arizona 85721 USA

August 2003

Overview

Fuzzy sets have many potential applications in the social sciences. The ideas and suggestions presented in *Fuzzy-Set Social Science* scratch only the bare surface of their potential uses, for there are many ways to integrate fuzzy sets (and set-theoretic thinking more generally) into social research. In this paper, I sketch several recent advances in fuzzy-set methods, illustrating them with examples drawn from policy research. While these new fuzzy-set methods build on arguments presented in *Fuzzy-Set Social Science*, they also forge a strong link to some of the crisp-set principles presented in *The Comparative Method*, especially those concerning the issue of limited diversity. Thus, the techniques presented in this paper are generally relevant to both crisp-set and fuzzy-set analysis.

The first advance I present is the elaboration and refinement of the concepts of "consistency" and "coverage" in set-theoretic analysis. Specifically, I show how to assess the consistency and coverage of combinations of causal conditions. Consistency concerns the degree to which a combination of causal conditions is consistent with an argument of sufficiency; coverage concerns the relative importance of combinations of sufficient conditions in the effort to explain or "cover" instances of the outcome. In this discussion I emphasize the fuzzy-set analysis of sufficient combinations of causal conditions, but the two principles apply just as well to the fuzzy-set analysis of necessary conditions and to the analysis of necessity and sufficiency as set-theoretic relations using crisp sets.

The second advance I discuss is a new algorithm for the incorporation of "simplifying assumptions" into the results of applications of QCA and fs/QCA. This new algorithm allows the direct incorporation of theoretical and substantive knowledge into the evaluation of simplifying assumptions in situations of "limited diversity" (which is the rule in the study of naturally occurring social phenomena). I illustrate this algorithm with crisp sets, and then extend it to fuzzy sets. Along the way, I also introduce a new algorithm for the fuzzy-set analysis of social data. The new algorithm is more amenable to the analysis of limited diversity than the one presented in *Fuzzy-Set Social Science.*

I. A. Consistency

In *Fuzzy-Set Social Science*, the definition of the consistency of a set-theoretic relation is straightforward. In the plot of membership in an outcome (Y) against membership in a causal condition or combination of causal conditions (X), consistency is defined as the proportion of cases on or above the main diagonal of the scatterplot. If membership in X is consistently less than or equal to membership in Y, then all the cases will plot on or above the main diagonal of the scatterplot, yielding a consistency score of 1.0 (or 100%). In fs/QCA consistency scores are computed for different combinations of causal conditions, and these scores provide the basis for evaluating sufficiency, defined as a set-theoretic relation ($X_i \leq Y_i$). For example, if significantly greater than 80% of the cases plot on or above the main diagonal in the scatterplot just described, then the investigator might claim that the cause or causal combination X is "almost always" sufficient for the outcome Y. The program fs/QCA evaluates all logically possible combinations of causal conditions (3^k - 1 combinations, where k is the number of causal conditions) when deriving a logically simplified statement of the combinations of conditions linked to an outcome.

The procedures presented in *Fuzzy-Set Social Science* for the evaluation of the sufficiency of causal combinations are based on the simple categorization of cases as consistent or inconsistent and the computation of the raw proportion of consistent cases. One issue in the use of this procedure concerns the contrast between cases with strong versus weak membership in the causal condition or combination of causal conditions (X). Specifically, cases with strong and weak membership in the causal combination are weighted equally in the calculation, yet they differ substantially in their relevance to the set-theoretic argument. For example, a case with a membership of only .25 in the set of cases with the causal combination (X) and a score of .00 in the outcome set (Y) is just as inconsistent as a case with a score of 1.00 in the causal combination and a score of .75 in the outcome. (A membership of .25 indicates that a case is more "out" than "in" a set; .50 is the cross-over point.) In fact, however, the second inconsistent case, with full membership in X, clearly has more bearing on the set-theoretic argument because it is a much better instance of the causal combination. It thus constitutes a more glaring inconsistency than the first case.

The same reasoning holds for consistent cases. A consistent case with two high membership scores (e.g., .90 in the causal combination and 1.00 in the outcome) is clearly more relevant to the set-theoretic argument than a consistent case with two low scores (e.g., .10 in the causal combination and .20 in the outcome) or a consistent case with a low score in the causal combination (e.g., .15) and a high score in the outcome (e.g., .80). Yet all are counted equally in the formula for consistency used in fs/QCA. Imagine trying to support an argument in an oral presentation to colleagues using in-depth evidence on a case with only weak membership in the relevant sets. The common sense thinking that indicates that this presentation would be a waste of time is accurately formalized in fuzzy membership scores. Cases with strong membership in the causal condition provide the most relevant consistent cases and the most relevant inconsistent cases.

This common sense idea is operationalized in the alternate measure of the consistency of fuzzy-set data with set-theoretic arguments presented in this paper. This alternate procedure, like the first, differentiates between consistent and inconsistent cases using the main diagonal of the scatterplot. A case on or above the diagonal is consistent because its membership in the causal condition is less than or equal to its membership in the outcome. A case below the diagonal is inconsistent because its membership in the causal condition is greater than its membership in the outcome. However, rather than simply calculating the raw proportion of consistent cases, the alternate procedure uses fuzzy membership scores. Specifically, the sum of all the membership scores in the causal condition can be seen as a maximum value--the denominator, and the sum of the *consistent* membership scores (i.e., those that are less than or equal to the corresponding outcome membership score) can be used as the numerator. When all scores are consistent, the value is 1.00 (100% consistent); when no score is consistent, the value is .00 (0% consistent). This alternate calculation is much more sensitive to larger membership scores because a single inconsistent case with strong membership in the causal condition can outweigh many small consistent cases. As a general rule, this alternate calculation yields *lower* consistency scores than the one used in *Fuzzy-Set Social Science* (and fs/QCA), which is based simply on the raw percentage of consistent cases. Again, the benefit of this alternate measure of consistency is that cases with stronger membership in the causal combination have a greater impact on the measure, in line with common sense reasoning.

I. B. 1. Coverage

Coverage is not discussed directly in *Fuzzy-Set Social Science*. Thus, the ensuing presentation of the principle of coverage constitutes a more significant digression from *Fuzzy-Set Social Science* than the discussion of consistency, just presented.

Most analyses using QCA or fs/QCA culminate with a specification of the different combinations of causal conditions linked to some outcome (typically described as a pattern of "multiple conjunctural causation"). These combinations of conditions are understood as alternate routes or pathways to the outcome, based on the principle of equifinality. Generally, these alternate paths are treated as logically equivalent because they are substitutable. However, it is usually instructive in crisp-set analyses to assess how many instances displaying the outcome also display (or "follow") each combination of causal conditions (or "path"). (Of course, it is important to allow for the possibility that a given case may conform to more than one path, a common finding.) The relative number of cases on a path is a direct indicator of the empirical importance of a causal combination. Thus, a simple measure of the "coverage" of a causal combination in crisp-set analysis is the proportion of instances of the outcome that display it. Furthermore, by calculating the proportion of instances of the outcome that can be credited to each causal combination, the investigator can assess the relative importance of causal combinations. A causal combination than covers only a small proportion of the instances of an outcome is not as empirically important as one that covers a large proportion.

Coverage is distinct from consistency, and often the two work against each other because high consistency may yield low coverage. Complex set-theoretic arguments involving the intersection of many sets can achieve remarkable consistency but low coverage. For example, 100% of the adults in the U.S. who combine excellent school records, high achievement test scores, college educated parents, high parental income, graduation from Ivy League universities, and so on, may avoid poverty as adults. (Perfect set-theoretic consistency is unusual with individual-level data, but certainly not impossible.) There are, however, relatively few individuals with this specific combination of highly favorable circumstances among those who successfully avoid poverty. From the viewpoint of policy discourse, therefore, this high level of set-theoretic consistency is not compelling because the causal combination is so narrowly formulated that its coverage is quite low. In other words, if a path is traveled by relatively few cases (i.e., if it has very low coverage), then it is probably neither substantively important nor policy relevant. (Of course, its theoretical relevance might still be considerable.)

For further illustration of the general idea of coverage consider Table 1, which shows a hypothetical cross-tabulation of poverty status (in-poverty versus not-in-poverty) against school achievement test scores (high versus low/average), using conventional sets. This crude analysis using binary data supports the argument that individuals with high test scores are able to avoid poverty. This set-theoretic argument is substantiated by the high percentage of cases in the second column who are not in poverty (cell b divided by the sum of cells b and d yields a consistency score of 96.4%). But how important is this path when it comes to avoiding of poverty? The simplest way to answer this question is to calculate the percentage of the cases not in poverty that have high test scores--that is, cell b divided by the sum of cells a and b, which is 32.6%. This calculation shows that the path in question covers almost a third of the cases not in poverty, which is substantial.

For comparison purposes consider Table 2 which has the same total number of cases as Table 1, but some of the cases have been shifted from cell b to cell a and from cell d to cell c. The percentage of cases consistent with the set-theoretic argument in Table 2 is 96.7%., about the same as in Table 1 (96.4%). Thus, from a set-theoretic point of view, the evidence is again highly consistent. But how important is this path, using the hypothetical frequencies presented in Table 2? This can be ascertained by computing the percentage of cases covered by the set-theoretic argument, which is only 3.25% (147/4520). Thus, in Table 2 the set-theoretic pattern is highly consistent, about the same as in Table 1, but coverage is very low, indicating (hypothetically) that having high test scores is not an important path to the outcome (avoiding poverty).

The procedures for calculating coverage using fuzzy sets parallel the computations for crisp sets just presented. Another way to understand the calculation of coverage using conventional binary sets (cell b divided by the sum of cells a and b) is to visualize Table 1 as a Venn diagram showing a subset relationship, as in Figure 1. The basic idea behind the calculation of coverage is to assess the degree to which the smaller set (the set of cases with high test scores in this example) physically covers the larger set (the set of cases not in poverty). Thus, coverage, our gauge of importance, can be seen as the size of the overlap of the two sets relative to the size of the larger set (the outcome). Using fuzzy sets, the size of the larger set is given by the sum of the membership scores in that set (e.g., the sum of the membership scores in the set of cases with the outcome). This calculation parallels the simple counting of the number of cases in a set (e.g., in the set of cases not in poverty) using crisp sets. The calculation of the overlap using fuzzy sets is given by the sum of the *consistent* memberships scores (cases where $X_i \leq Y_i$), applied to the set representing the causal condition (or combination of conditions). To calculate the size of the overlap, it is necessary simply to sum the membership scores in the causal condition that are less than or equal to their corresponding membership scores in the outcome. The coverage of a sufficient causal condition using fuzzy sets, therefore, is the sum of the consistent causal membership scores divided by the sum of the outcome membership scores.

For purposes of illustration, consider some evidence from a fuzzy-set analysis of individual-level policy data. The data set is the National Longitudinal Survey of Youth (better known as the *Bell Curve* data). The sample is white males, interviewed as young adults; the outcome is the fuzzy set of cases not in poverty (~P); the three causal conditions are the fuzzy set of cases with high achievement test scores (T), the fuzzy set of cases with high parental income (I), and the fuzzy set of cases with college education (C). Applying fs/QCA to these data yields two combinations of conditions that are sufficient for the avoidance of poverty, namely, the combination of high test scores and high parental income (T·I) and the combination of high parental income and college education (I·C). The calculation of their coverage of the outcome, the set of cases not in poverty (~P), is shown in Table 3. The first row reports the coverage calculation for the combination of high test scores and high parental income (T·I). The sum of the consistent T·I scores is 181.287. (These are the T·I membership scores that are less than or equal to the outcome score--membership in the set of cases not in poverty.) The sum of the memberships in the outcome is 949.847. Thus, this combination covers about 19.1% of the total membership in the outcome (181.830/949.847 = .191). Using these same procedures, combination I·C covers about 23.9% of the total membership of the outcome (see row two of Table 3). Thus, both combinations cover a substantial proportion of the outcome. However, the combination of high-income parents and college education (I·C) covers more.

For comparison purposes, Table 3 also shows the coverage of the two combinations conceived as alternate paths (the maximum of the two paths, reported in the third row). Recall that when there are multiple paths to an outcome, it is possible to calculate how close a case is to the outcome by finding its highest membership score among the possible paths. In fuzzy algebra, this is accomplished using logical *or* (+), which dictates the use of the maximum score (the greater of T·I and I·C). The coverage of this maximum score, in turn, can be calculated using the same procedures applied separately to the two component combinations. Using the maximum of these two, the coverage is 26.7%, greater than the coverage of either of the component combinations (compare row three of Table 3 with the first two rows). However, the fact that the coverage of the two-path model (26.7%) is only modestly superior to the coverage of the best single path (path I·C, 23.9%) indicates that the two paths overlap considerably.

Table 3 also reports the coverage of the intersection of these two paths, which involves the combination of all three favorable conditions. This model asserts that in order to avoid poverty, individuals must combine high test scores, high parental income, and college education. The coverage of the one-path model, shown in the last row, is the lowest reported in the table, but still non-trivial at 16.3%. Note, however, that the two-path model shown in the third row of Table 3 is considerably more powerful (26.7%) than the one-path model. (In set-theoretic terms, the one-path model is the overlap or intersection of the two paths, while the two-path model is their union.) Overall, the comparison of the coverage calculations across the four rows indicates that there is substantial gain derived from differentiating the two paths (T·I and I·C), even though the two paths do overlap.

I. B. 2. Partitioning Coverage

A further benefit of the coverage measure introduced here is that it is possible to partition the measure into its components in a manner that is somewhat analogous to the partitioning of explained variation in multiple regression. To assess an independent variable's separate or unique contribution to explained variation in a multiple regression analysis involving many predictor variables, researchers calculate the decrease in explained variation that occurs once the variable is removed from the multivariate equation in question. The calculation of the unique coverage of a path in fuzzy-set analysis is parallel. For example, the coverage of the outcome (avoiding poverty) that is uniquely due to path T·I is the difference between the coverage of the two-path model (26.7%) and the coverage level that is obtained once this path (T·I) is removed from the two-path model, equivalent to the coverage of the other path (I·C) examined by itself. Thus, the unique coverage of path T·I is 2.8% (26.7% , the combined coverage of the two paths, less 23.9%, the single coverage of path I·C). Likewise, the coverage of the outcome that is uniquely due to path I·C is the difference between the coverage of the two-path model (26.7%) and the coverage of path T·I (19.1%), or 7.6%. These calculations reveal that the unique coverage of path I·C is much greater than the unique coverage of path T·I.

The balance of the coverage of the two-path model is confounded between the two paths. This portion can be calculated as the difference between the coverage of the two-path model (26.7%) and the sum of the two unique portions (2.8% + 7.6% = 10.4%), which is 16.3%. This difference (16.3%) is the same as the coverage of the one-path model involving the combination of all three causal conditions, T·I·C (shown in the fourth row of Table 3). Thus, the partitioning of the total coverage of the two-path model has three components, the two components that are unique to each path, and the confounded component, which is the same as the coverage of the intersection or overlap of the two paths (i.e., T·I·C). About three-fourths of the difference between the coverage of the one-path model (16.3%) and the two-path model (26.7%) is due to the unique contribution of path I·C, the combination of high parental income and college education.

This example demonstrates not only that is it possible to analyze data in terms of alternate combinations of causally relevant conditions, but also that it is possible to assess the relative importance of paths. Further, I have presented procedures for partitioning coverage in a manner that is somewhat analogous to the partitioning of explained variation using conventional quantitative methods such as multiple regression. The key difference is that the fuzzy-set procedures assess the relative importance of paths representing combinations of causally relevant conditions, not the relative importance of independent variables considered in isolation from one another, as in regression analysis.

II. A. Limited Diversity and Simplifying Assumptions

If the empirical world presented social scientists with cases exhibiting all logically possible combinations of relevant causal conditions, then social research would be much more straightforward. In QCA this would correspond to a situation where the truth table for an outcome would have cases on each of the 2^k rows of the truth table (where k is the number of presence/absence causal conditions). A truth table this complete is rare, especially when k is greater than 5, because naturally occurring social phenomena are almost always limited in their diversity.

As explained in *The Comparative Method*, when the diversity of social phenomena is limited, researchers can use causal combinations lacking cases (called "remainders" in fs/QCA) to simplify the results of the analysis of the causal conditions linked to an outcome. This process is usually described as one of making "simplifying assumptions" because the researcher must assume that if the nonexistent combinations could be observed, they would have the outcome, as indicated in the equation derived from the truth table. A simple example: Suppose a researcher can show that the causal combination a·b·C·d results in the outcome. (Uppercase letters indicate that a condition is present; lower case letters indicate that it is absent; the "·" symbol indicates combined conditions.) He has no cases of a·b·C·D, but believes that if he did, they would also exhibit the outcome. By adding the simplifying assumption that if a·b·C·D existed, it would display the outcome, he can simplify his solution to a·b·C. (Please remember that it is always important to document any simplifying assumptions that have been incorporated into a solution.)

While there are important exceptions to this generalization, most researchers who use QCA and fs/QCA either incorporate as many simplifying assumptions as possible or they avoid them altogether (both strategies are easy to implement in QCA and fs/QCA). Rarely do they do what they "should" do, which is to evaluate each simplifying assumption to see if it is plausible and then incorporate only those that are. The reason this preferred practice is relatively rare is that it is difficult and tedious. The greater the number of causal conditions, the greater the number of logically possible combinations. The greater the number of logically possible combinations, the greater the number of combinations lacking cases and thus the greater the number of remainders to evaluate for possible inclusion as simplifying assumptions. Plus, most researchers are likely to examine many different specifications of the relevant causal conditions in an investigation. It is simply not possible to examine such a large number of potential simplifying assumptions across multiple specifications of the relevant causal conditions.

Fortunately, there is a relatively simple solution to this problem, which I will now sketch. I first summarize the logic of the procedure and then provide an example.

Usually, but not always, researchers have one-sided or directional expectations regarding the relation between a causal condition and the outcome they are studying. A directional expectation has the following form: "the presence of X, and not its absence, should be linked to the presence of the outcome." For example, we might argue that the presence of widespread urban poverty, and not the absence of urban poverty, should be linked to mass protest against government austerity programs. This statement is directional because the expectation is that there should be *no* instances where the *absence* of widespread urban poverty is a contributing factor to protest against government austerity programs (i.e., a part of a sufficient combination of causal conditions). In other words, if it is causally relevant at all, widespread urban poverty, not its absence, should appear in combinations of conditions generating austerity protests.

Directional expectations can be used in conjunction with "thought experiments" applied to existing cases. For example, in the hypothetical study of austerity protests, assume that there are protest cases that have many of the relevant causal conditions, as specified in the truth table, but clearly lack widespread urban poverty. It would be reasonable to assume that if these cases had widespread urban poverty, in addition to presence of these other causes, they still would have austerity protests. The thought experiment is the imagination of these cases *with* instead of *without* widespread urban poverty. Using standard Boolean notation, we have the empirical case:

A·B·C·d **---->** Y

where A through D represent the four causal conditions (the lower-case D indicates the absence of widespread urban poverty), and Y represents the outcome. Based on our directional expectation with respect to urban poverty (D), we can construct the following hypothetical case, by way of the thought experiment just described:

A·B·C·D **---->** Y

In this example, we have assumed that there are no empirical instances of A·B·C·D and that we have been able to draw this inference (that A·B·C·D **---->** Y) by bringing together the empirical case (A·B·C·d) and the directional expectation (that only D can contribute to Y). Using crisp-set logic, we can then make the following logical simplification:

A·B·C·d + A·B·C·D **---->** Y
A·B·C·(d+D) **---->** Y
A·B·C **---->** Y

In other words, we have used a remainder (A·B·C·D) as a simplifying assumption, but we have done so selectively, based on the marriage of an empirical case with theoretical knowledge (i.e., the directional expectation). Any scholar who has conducted a small-N comparative investigation would not find anything I have presented here especially shocking or novel, for this procedure mimics and formalizes routine practices in comparative research.

This marriage of directional expectations and thought experiments can be applied to all the rows of a truth table with empirical instances. In this way, it is possible to "spawn" all the "acceptable" simplifying assumptions, based on the existing cases and the investigator's directional expectations. Consider the following simple example, where the investigator has empirical instances of only four of the eight (2^3) logically possible combinations of three causal conditions:

Positive Cases:
A·B·c ----> Y
A·b·C ----> Y

Negative Cases:
A·b·c ----> y
a·b·c ----> y

Remainder terms:
A·B·C ----> ?
a·B·C ----> ?
a·B·c ----> ?
a·b·C ----> ?

First, it is worth noting that without simplifying assumptions (i.e., without using any remainders), no simplification is possible. The two cases exhibiting the outcome differ on two causal conditions (B and C); thus, no pair-wise simplification is possible. Without simplifying assumptions, the solution is simply:

A·B·c + A·b·C ----> Y

Next, consider the solution using all possible simplifying assumptions. This solution can be derived by hand or using either program (QCA of fs/QCA). The investigator simply sets all the remainder terms to the "don't care" outcome. The solution that follows from this procedure (i.e., allowing all possible simplifying assumptions) is:

B + C ----> Y

It is tidy and parsimonious, and much simpler than the solution without simplifying assumptions.

These two solutions can be seen as establishing the range of possible solutions, for the two most extreme solutions have been derived--one with no simplifying assumptions and one with as many as possible. Selective uses of simplifying assumptions will yield solutions that are supersets of the first solution and subsets of the second.

Before moving on to the marriage of directional expectations and thought experiments, take a close look at the two solutions. In the first, condition A is important; it appears in both terms and could be seen as a necessary condition--if the investigator chose to take this additional interpretive step. However, condition A is completely absent from the solution with simplifying assumptions. The use of simplifying assumptions has masked a causal condition that is found in each instance of the outcome! Thus, it is clear from this simple example that to allow all possible simplifying assumptions is to use an analytic strategy that is fraught with potential hazard. A careful case-oriented investigator would not eliminate from his or her analysis a causal condition that appears in every instance of the outcome.

To implement the new approach to simplifying assumptions, it is important to state directional expectations, based on theoretical and substantive knowledge. For simplicity's sake, assume that the directional expectations in the present analysis are:

1. The presence of A (and not its absence) should be linked to the outcome (Y).
2. The presence of B (and not its absence) should be linked to the outcome (Y).
3. The presence of C (and not its absence) should be linked to the outcome (Y).

Based on these directional expectations we can return to the existing cases and use them to spawn *acceptable* simplifying assumptions. The first case (A·B·c) inspires simplifying assumption A·B·C because only the presence of C should be linked to the outcome. In other words, it is reasonable to assume that if the first case exhibited the presence of C instead of its absence, then Y would still follow. The second case (A·b·C) inspires the same simplifying assumption (A·B·C) using parallel reasoning because only the presence of B should be linked to the outcome. Thus, the marriage of directional expectations and existing cases sanctions only one of the four possible simplifying assumptions, which in turn yields the solution:

A·B + A·C **---->** Y

Notice that the causal condition which is shared by instances of the outcome (condition A) is not eliminated from the solution when directional expectations and existing cases are used to generate simplifying assumptions. Notice also that, as indicated previously, this solution is a superset of the solution without simplifying assumptions and a subset of the solution with all possible simplifying assumptions.

II. B. Crisp Set Implementation

It is a fairly simple matter to implement this new approach to simplifying assumptions in QCA. To illustrate the procedure, I use a data set created by Olav Schram Stokke which is currently available as a working paper through www.compasss.org. Stokke studied the conditions under which "shaming" succeeds in international regimes. He focuses explicitly on noncompliance with international fishing agreements. He examines 10 cases of attempted shaming--five were successful; five were not. His causal conditions were:

1. Advice (A): Whether the shamers can substantiate their criticism with reference to explicit recommendations of the regime's scientific advisory body.
2. Commitment (C): Whether the target behavior explicitly violates a conservation measure adopted by the regime's decision-making body.
3. Shadow of the future (S): Perceived need of the target of shaming to strike new deals under the regime--such beneficial deals are likely to be jeopardized if the criticism is ignored.
4. Inconvenience (I): The inconvenience of the behavioral change that the shamers are trying to prompt.
5. Reverberation (R): The domestic political costs to the target of shaming for not complying (i.e., for being scandalized as a culprit).

His original truth table (eliminating redundant cases) is reported in Table 4. Analysis of this truth table without simplifying assumptions yields very little logical simplification. The results are:

A·S·I·R + A·C·S·i·r + A·c·s·i·r **---->** Y

This relatively complex result follows from the fact that there are 32 (2^5) logically possible combinations of conditions, but only eight exist, and only four of these eight combinations display the outcome.

By contrast, the use of all possible simplifying assumptions results in what is perhaps an overly parsimonious solution:

i + S·R **---->** Y

This solution states that shaming works when it is not inconvenient (i) for the target of shaming to change behavior or when the "shadow of the future" and "domestic reverberations" combine (S·R) to produce a conforming response to shaming. While some might take great delight in such parsimony, notice that this solution eliminates a causal condition that is present in all instances of successful shaming--the supportive advice of the regime's scientific advisory board (A). Parsimony is desirable, of course, but there can be too much of a good thing. Some scholars might justifiably conclude, based on simple inspection of the positive cases in the truth table, that the support of a regime's scientific advisory board is a necessary condition for successful shaming.

The new approach to simplifying assumptions involves first of all listing the directional expectations. I use four in this example:

1. The presence of scientific advice (A), not its absence, leads to successful shaming (Y).
2. The presence of commitment (C), not its absence, leads to successful shaming (Y).
3. The presence of the shadow of the future (S), not its absence, leads to successful shaming (Y).
4. The presence of domestic reverberations (R), not its absence, leads to successful shaming (Y).

As explained previously, these directional expectations are used to generate acceptable simplifying assumptions. For example, the seventh row of the truth table is:

A·C·S·i·r **---->** Y

By using the fourth directional expectation, it is possible to generate the following simplifying assumption:

A·C·S·i·R **---->** Y

These two expressions can be simplified to:

A·C·S·i **---->** Y

using simple Boolean algebra.

The quickest way to implement these directional expectations as thought experiments is simply to edit the truth table, transforming codes that are the opposite of expectations into the "don't care" value. For example, the seventh row is recoded from:

A	C	S	I	R	Y
1	1	1	0	0	1

to:

A	C	S	I	R	Y
1	1	1	0	-	1

where the dash code indicates that a term has been eliminated through logical reduction. This coding shortcut takes the place of logical reductions that would inevitably be made in QCA. Using this simple procedure, the entire truth table can be quickly recoded, as shown in Table 5. (Notice that, for now at least, thought experiments have been applied only to the cases displaying the outcome--the positive cases. I will address negative cases subsequently.) This new table can then be analyzed using QCA (not fs/QCA). When performing the analysis, it is important to specify in the set-up that remainder terms should NOT be assigned the "don't care" outcome. Simplifying assumptions have already been encoded into the positive cases in the table. Additional assumptions are not warranted. The results of this analysis show:

A·i + A·S·R **---->** Y

Once again, using the new approach to simplifying assumptions has the benefit of retaining what could be interpreted as a necessary condition. As I demonstrate subsequently, this is not the sole benefit of using the new approach, but it is an important one.

These procedures also can be applied to the truth table rows with negative cases. The important difference is that instead of coding the *absence* of a condition that is expected to contribute to the occurrence of the outcome (according to directional expectations) to don't care (-), with the negative cases the investigator codes the *presence* of a condition expected to contribute to the outcome to don't care (-). This procedure follows the same "thought experiment" logic as before: The case lacks the outcome, but has one or more conditions expected to contribute to the outcome. Imagine the same case *without* these contributing conditions, and still the expectation would be that the outcome should not occur. Following these guidelines, the rows with negative cases would be recoded as shown in Table 6, using the same four directional expectations applied to the positive cases. (I leave the analysis of the negative cases as an exercise for the reader.)

II. C. Fuzzy-Set Implementation

The crisp-set implementation of this new approach to simplifying assumptions is straightforward; the investigator simply recodes the truth table using the two rules just presented (one for rows with positive outcomes and one for rows with negative outcomes) and then simplifies the recoded truth table without using any additional simplifying assumptions (i.e., without assigning remainder terms the "don't care" output). Using the fuzzy-set algorithm presented in *Fuzzy-Set Social Science* and implemented in fs/QCA, however, there is no truth table, per se, to recode. Thus, there is no direct way to implement these new ideas about simplifying assumptions using the current version of fs/QCA. The current fuzzy-set algorithm uses the containment rule instead of a truth table to simplify causal combinations, and simplifying assumptions are examined after the fact (i.e., after a solution has been generated). This feature of fs/QCA poses a substantial obstacle to the use of directional expectations to encode simplifying assumptions.

The solution to this problem is to modify the fuzzy-set algorithm so that a conventional Boolean truth table is produced as an intermediary step. I first demonstrate how to assess limited diversity using fuzzy sets and then show how to use fuzzy sets to produce a conventional truth table.

Limited diversity and fuzzy sets. Recall that with fuzzy sets, the researcher examines the multidimensional vector space defined by the causal conditions. Cases are arrayed in this k-dimensional space according to their fuzzy memberships in the k causal conditions. This multidimensional space has 2^k

corners, the same as the number of rows in a truth table with k causal conditions. (The vector space is "cornered" because fuzzy membership scores have a maximum of 1 and a minimum of 0.) As explained in *Fuzzy Set Social Science* almost every case has a *maximum* membership score in *only one* of the 2^k corners of this vector space, even though a case may have nonzero membership in all 2^k corners. Each case "belongs" to the corner it is closest to because its highest membership score (and only membership score greater than .5--the cross-over point) is in that corner. The exceptions to this rule are cases with membership scores of .5 in one or more of the fuzzy sets that define the vector space. These cases have maximum membership scores of .5 in two or more corners of the vector space. (These principles are explained in detail in *Fuzzy-Set Social Science.*)

This property of fuzzy sets is important because it provides the key to specifying patterns of limited diversity in fuzzy-set analysis. If there are no cases with greater than .5 membership in a given corner, then, in essence, there are no cases that are "more in than out" of the sector of the vector space adjacent to that corner. In other words, if there are no cases with greater than .5 membership in a corner, then there are no "good empirical instances" of the combination of characteristics defined by the corner. Thus, it is a relatively simple matter to determine which corners of the multi-dimensional vector space formed by the causal conditions lack instances. This assessment of degree of membership in the corners of a vector space provides the basis for determining which combinations of conditions "exist" and which are "remainders" in fuzzy-set analysis.

For illustration of these ideas, consider individual-level data (limited in this example to Black males, interviewed as young adults) on causal conditions relevant to avoiding poverty. These data are from the U.S. National Longitudinal Survey of Youth. There are five causal conditions, three represented as fuzzy sets and two as crisp sets. The outcome is degree of membership in the set of cases not in poverty. The causal conditions are:

1. Degree of membership in the set of cases with high achievement test scores.
2. Degree of membership in the set of cases with high-income parents.
3. Degree of membership in the set of cases with college education.
4. Married (yes/no).
5. Children (yes/no).

With five causal conditions, there are 2^5 (or 32) corners in the vector space. The number of cases with greater than .5 membership in each corner is reported in Table 7. The table lists the 32 corners of the 5 dimensional vector space, the number of cases with nonzero membership in each corner (a membership score of 0 indicates that a case is completely out of the set defined by the corner), the number of cases with greater than .5 membership in each corner, and the consistency of the nonzero cases with the statement, "Black males with this combination of characteristics are not in poverty." (The computation of consistency in Table 7 follows the procedures discussed previously in this paper.)

The most striking feature of Table 7 is the limited diversity of Black males with respect to the five causal conditions. Of greatest immediate interest is the column showing the number of cases with greater than .5 membership in each corner. The five most populated corners consume 644 of the 758 Black males (85%) included in this analysis. By contrast, the 19 least populated corners (those with fewer than 5 cases each) consume a mere 25 of 758 Black males (3.3%). Clearly, Black males in the U.S. are limited in their diversity when it comes to conditions that shield individuals from poverty. Indeed, 327 of 758 Black males (43.1%) share a single combination of conditions: low achievement test scores, low parental income, not college educated, not married, no children.

It is important to point out in passing that there is an underlying similarity between small-N and large-N data sets which scholars routinely overlook: they both may exhibit extreme limited diversity. Discussions of comparative methodology lament the fact that small-N investigations have too many variables relative to the number of cases, resulting in limited diversity and a limited capacity for comparative analysis. Table 7 shows that the problem of limited diversity is *not* limited to small Ns. Indeed, having 758 cases, a gargantuan N by the standards of comparative research, does not resolve the problem. Instead, there is simply an abundance of cases that are redundant from a configurational viewpoint. Thus, Table 7 underscores the fact that limited diversity is inherent in the study of naturally occurring social phenomenon. The problem of limited diversity is masked in most quantitative studies because researchers restrict their analyses to causal models that assume additivity and linearity, and they rarely, if ever, examine the distribution of their cases across configurations of causal conditions, as in Table 7.

In most small-N studies, if a researcher has a single instance of a configuration, then the configuration is included in the analysis and not treated as a remainder (i.e., as a combination of conditions lacking cases). In large-N analyses, such as the one framed in Table 7, by contrast, it is often prudent to establish a "relevance threshold." For example, a researcher might want to include a configuration in the analysis as an "existent" configuration only if it has at least five instances. After all, with large-N, individual-level data sets, measurement error is common, and cases are sometimes assigned to the wrong configuration. When there are only a few instances populating a particular sector of the vector space, it is reasonable to safeguard against "false positives" resulting from measurement error by establishing a minimum frequency. The specification of this frequency threshold can be based on a convention (such as a minimum of five cases) or on a probabilistic assessment (for example, the threshold might be set by examining which frequencies are significantly less than "expected" using an assumption of equiprobability; see Von Eye 1990). In the analysis that follows, configurations with fewer than five instances are treated as remainders. These low-frequency configurations together embrace only 25 of the 758 cases (3.3%).

From Fuzzy Sets to Truth Tables. The last column of Table 7 reports the consistency of each configuration with the statement: "Black males with this combination of characteristics are not in poverty." Before describing how this information can be used to construct a conventional truth table, it is important to review how consistency is calculated. Consider the first row. The cases in this row combine characteristics that are least favorable for the outcome, avoiding poverty. There are 72 cases with nonzero scores in this combination; thus, 686 (758-72) Black males have a score of 0 on at least one of the characteristics in this row. (For example, a Black male with a score of 1 on high-income parents would have a score of 0 on "out" of the set of cases with high-income parents.) Consistency is defined in terms of causal sufficiency, which in turn is defined as a set-theoretic relation. The key question is: Do the cases with this combination of conditions constitute a subset of the cases not in poverty? Are the membership scores in this combination consistently less than or equal to membership scores in the outcome? The calculation of consistency is straightforward: the sum of the membership scores in the causal combination that are consistent with the set-theoretic relation ($X_i \leq Y_i$) is expressed as a proportion of total membership scores in the combination of conditions (i.e., the sum of X_i).

For the first combination of conditions (reported in the first row of Table 7), the proportion consistent is only .039, which indicates that almost all the cases with nonzero membership in this combination have scores in the causal combination that exceed their outcome membership scores. In other words, the evidence indicates almost complete *in*consistency with the set-theoretic relation in question. However, the eighth row is highly consistent, with a proportion of .863. This row represents the evidence for the 29 Black males with nonzero membership in the combination of not-high test scores, not-high-income parents, college educated, married, with no children. The consistency score of .863 indicates that these cases constitute a rough subset of the cases not in poverty. Altogether, the consistency column of Table 7 summarizes findings from the assessment of the set-theoretic relations between each of the 13 existent combinations of conditions and the outcome, all represented as fuzzy sets.

The key question at this point is which of the 13 existent combinations are consistent with the fuzzy set-theoretic relation? The coding of these rows as consistent versus not consistent is dependent upon the cut-off value used to define "roughly consistent." This determination is parallel to the use of benchmarks to assess the quasi-sufficiency of combinations of causal conditions, as detailed in *Fuzzy-Set Social Science*. For example, a cut-off value of at least 65% consistent might be used to define "usually" sufficient. The important difference between the present analysis and the procedure described in *Fuzzy-Set Social Science* is that the new procedure uses a more precise measure of consistency, based directly on fuzzy membership scores. In the truth table analysis that follows, I use a benchmark of .70 and an alpha of .05, which qualifies 5 of the 13 existent configurations as "consistent."

The evidence on fuzzy set-theoretic relations presented in Table 7 can now be represented as a conventional Boolean truth table. The thirteen existent rows (reflecting evidence on the corners of the k-dimensional vector space) are coded as presence/absence dichotomies; consistency is recoded as consistent versus not consistent, using the .70 benchmark. The evidence is summarized in Table 8, which retains row numbers from Table 7. Keep in mind that this truth table summarizes evidence on 733 cases, 96.7% of the original sample of Black males. It is a relatively simple matter at this point to simplify this truth table and derive the minimal combinations of conditions linked to avoiding poverty. For purposes of illustration, I present three analysis. The first incorporates no simplifying assumptions; the second incorporates as many as possible; and the third uses the new techniques sketched in this paper--the application of directional expectations to existent cases.

The results using no simplifying assumptions are:

~T·I·~C·M + ~T·~I·C·M·~K + ~T·I·C·~M·~K + T·~I·C·M·K **---->** ~P

where:

· indicates intersection (combinations of conditions)
+ indicates union (alternate combinations)

~ indicates "not" (negation)
T is the set of cases with high achievement test scores
I is the set of cases with high-income parents
C is the set of college educated
M is the set of married cases
K is the set of cases with one or more children
P is the set of cases in poverty

Very little simplification occurs. Three of the four combinations have all five conditions, indicating that they are the same as in the truth table. The other combination has four conditions and thus combines only two rows of the truth table. This general type of solution, where very little simplification of the truth table takes place, is common in situations of limited diversity without simplifying assumptions. Unfortunately, it is all too common in the study of naturally occurring social phenomena, even when Ns are large.

At the other extreme is the solution that incorporates all possible simplifying assumptions. This solution is derived by setting all remainder terms to the "don't care" output. The solution incorporating all possible simplifying assumptions is the "most minimal" (simplest) possible and is a superset of the solution using no simplifying assumptions. Usually, but not always, this type of solution is the most consistent with the results of conventional quantitative analysis of the same data, where there is little or no recognition of or allowance for limited diversity. With all possible simplifying assumptions, the solution is:

$$T + I \cdot M + I \cdot C + C \cdot M \cdot {\sim}K \text{ ----> } {\sim}P$$

which can be factored to show:

$$T + I \cdot (M + C) + C \cdot M \cdot {\sim}K \text{ ----> } {\sim}P$$

According to the equation, having high test scores, by itself, shields Black males from poverty; having high-income parents combined with either college education or marriage is also sufficient for avoiding poverty, as is the three-way combination of college education, marriage, and no children. The authors of *The Bell Curve* would be pleased with these results, for they indicate that having high test scores, by itself, is sufficient for avoiding poverty. Their main theme is that we live in a world of greater and greater technical sophistication and that success in this new world requires substantial cognitive ability. In the end, they argue that cognitive ability is more important than family background when it comes to a variety of policy-relevant social outcomes, such as avoiding poverty. The result using all possible simplifying assumptions lends some support to this argument and delineates several alternate paths to the outcome, as well.

Before jumping on the *Bell Curve* bandwagon, take another look at Tables 7 and 8. Out of 32 logically possible combinations of conditions, thirteen reach the relevance threshold of at least five cases with greater than .5 membership. Of these thirteen, twelve involve combinations that include not-high test scores; only one combination involves high test scores. The one with high test scores is "consistent" with the set-theoretic relation; however, the other fifteen logically possible combinations involving high test scores are all remainders. Thus, when an analysis allowing all possible simplifying assumptions is computed, this one combination (shown in the last row of Table 8) reduces to a single causal condition--the presence of high test scores. In fact, however, as Table 8 shows, the black males who have high test scores (and avoid poverty), are also college educated and married. These two conditions are strongly linked to avoiding poverty, to put it mildly, so it is a real stretch to state that this truth table supports the claim that having high test scores, by itself, is sufficient for avoiding poverty. Clearly, the use of all possible simplifying assumptions can be distorting!

This problem is solved by applying directional expectations to existent cases and then reanalyzing the truth table. The relevant directional expectations are:

1. Having high test scores, not low, should be linked to avoiding poverty.
2. Having high-income parents, not low, should be linked to avoiding poverty.
3. Having a college education, not its absence, should be linked to avoiding poverty.
4. Being married, not unmarried, should be linked to avoiding poverty.
5. Not having children, instead of having them, should be linked to avoiding poverty.

To implement these directional expectations, it is necessary simply to recode the truth table using the two rules described previously. The recoded truth table is presented in Table 9. The reduction of this truth table yields:

T·C·M + I·M + I·C·~K + C·M·~K **---->** ~P

This solution is a superset of the first (no simplifying assumptions) and a subset of the second (all possible simplifying assumptions). It is intermediate in complexity and shows, as it should, that high test scores *combined with* college education and marriage is sufficient for the avoidance of poverty. This result casts serious doubt on the *Bell Curve* claim of the supremacy of test scores (interpreted as cognitive ability) for policy relevant outcomes such as poverty. The equation also shows three other important paths to the avoidance of poverty: having high-income parents combined with marriage, having high-income parents combined with college education and no children, and college education combined with marriage and no children. It is worth noting that all four paths involve household composition in some way. This finding is much more policy relevant than the findings for test scores.

As these results show, the use of directional expectations provides a middle path between avoiding simplifying assumptions altogether and making as many as possible. Using QCA 3.0 (not fs/QCA 0.96), the procedure is easy to implement. Additionally, this new approach to simplifying assumptions requires the clear statement of directional expectations, which means, in turn, that the investigator must consciously use his or her substantive and theoretical knowledge and must state directional expectations clearly and openly so that others may evaluate them.

The next step in this analysis would be to calculate the coverage of these four causal combinations and then to partition coverage into its separate components. Because this is a paper on methods and I have already demonstrated these procedures, I save the examination of the relative importance of these four combinations for later.

III. Conclusion

While my title is "Recent Advances in Fuzzy-Set Methods," the techniques I present have broad application in all set-theoretic analyses of social data. The most important advance presented in this paper is the demonstration of new techniques for incorporating theoretical and substantive knowledge into the construction of representations of naturally occurring social phenomena. While I have neither time nor space to develop the argument, it is my contention that these procedures lead to more useful and more policy-relevant representations. Furthermore, the techniques I present have very broad implications for the analysis of social data in general, especially for conventional approaches to the (largely unrecognized) problem of limited diversity.

Table 1: Crosstabulation of Poverty Status and Test Scores: Original Frequencies

	Low/Average Test Scores	High Test Scores
Not In Poverty	a. 3046	b. 1474
In Poverty	c. 625	d. 55

Table 2: Crosstabulation of Poverty Status and Test Scores: Altered Frequencies

	Low/Average Test Scores	High Test Scores
Not In Poverty	a. 4373	b. 147
In Poverty	c. 675	d. 5

Table 3: Calculation of Coverage

Causal Conditions	Sum of Consistent Scores	Sum of Outcome Scores	Coverage
T·I	181.830	949.847	.191
I·C	226.792	949.847	.239
T·I + I·C	253.622	949.847	.267
T·I·C	155.000	949.847	.163

Table 4: Original Truth Table for Causes of Successful Shaming in International Regimes

Advice (A)	Commitment (C)	Shadow (S)	Inconvenience (I)	Reverberation (R)	Success (Y)
1	0	1	1	1	1
1	0	0	1	0	0
1	0	0	1	1	0
0	0	0	1	0	0
1	1	1	1	1	1
1	1	1	1	0	0
1	1	1	0	0	1
1	0	0	0	0	1

Table 5: Revised Truth Table on Shaming in International Regimes, Applying Directional Expectations to Positive Cases Only

Advice (A)	Commitment (C)	Shadow (S)	Inconvenience (I)	Reverberations (R)	Success (Y)
1	-	1	1	1	1
1	0	0	1	0	0
1	0	0	1	1	0
0	0	0	1	0	0
1	1	1	1	1	1
1	1	1	1	0	0
1	1	1	0	-	1
1	-	-	0	-	1

Table 6: Revised Truth Table on Shaming in International Regimes, Applying Directional Expectations to Both Positive and Negative Cases

Advice (A)	Commitment (C)	Shadow (S)	Inconvenience (I)	Reverberations (R)	Success (Y)
1	-	1	1	1	1
-	0	0	1	0	0
-	0	0	1	-	0
0	0	0	1	0	0
1	1	1	1	1	1
-	-	-	1	0	0
1	1	1	0	-	1
1	-	-	0	-	1

Table 7: Distribution of Black Males Across the 32 Corners and the Consistency of Each Combination

Row	High Test Scores	High Parental Income	College Educated	Married	No Kids	Non-Zero	Greater than .5	Percent Consistent
1	out	out	out	out	out	72	65	0.039
2	out	out	out	out	in	375	327	0.219
3	out	out	out	in	out	172	154	0.165
4	out	out	out	in	in	50	41	0.595
5	out	out	in	out	out	13	7	0.227
6	out	out	in	out	in	126	57	0.586
7	out	out	in	in	out	74	24	0.594
8	out	out	in	in	in	29	13	0.863
9	out	in	out	out	out	36	1*	#
10	out	in	out	out	in	170	14	0.649
11	out	in	out	in	out	93	10	0.745
12	out	in	out	in	in	28	5	0.913
13	out	in	in	out	out	9	0*	#
14	out	in	in	out	in	81	10	0.873
15	out	in	in	in	out	49	3*	#
16	out	in	in	in	in	23	4*	#
17	in	out	out	out	out	1	1*	#
18	in	out	out	out	in	3	1*	#
19	in	out	out	in	out	6	2*	#
20	in	out	out	in	in	1	0*	#
21	in	out	in	out	out	0	0*	#
22	in	out	in	out	in	10	3*	#
23	in	out	in	in	out	20	6	0.852
24	in	out	in	in	in	8	1*	#
25	in	in	out	out	out	1	0*	#
26	in	in	out	out	in	1	0*	#
27	in	in	out	in	out	4	0*	#
28	in	in	out	in	in	1	0*	#
29	in	in	in	out	out	0	0*	#
30	in	in	in	out	in	9	4*	#
31	in	in	in	in	out	14	2*	#
32	in	in	in	in	in	7	3*	#

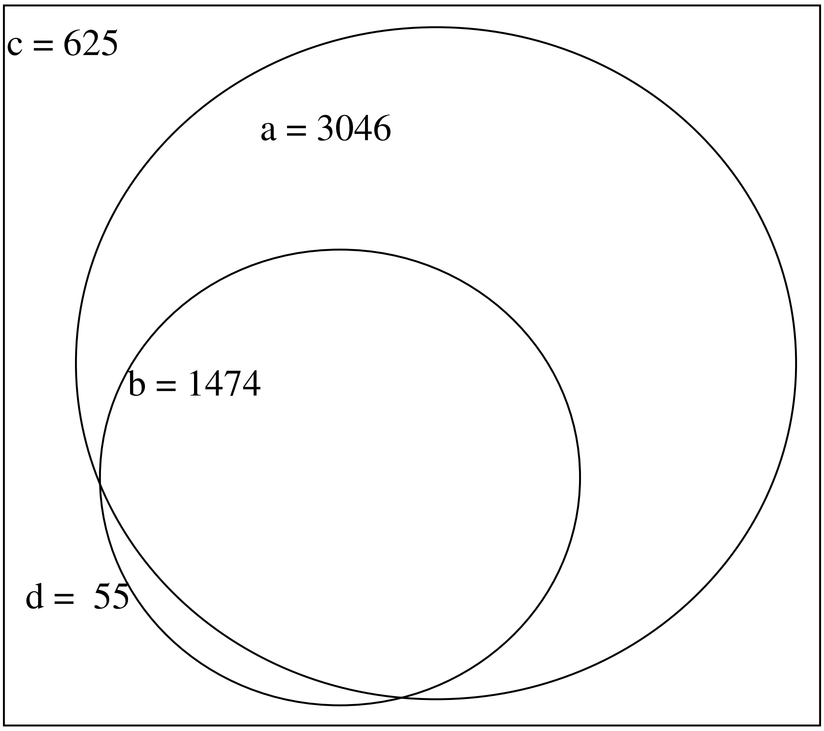
Table 8: Truth Table for Black Males Not in Poverty

Row	High Test Scores	High Parental Income	College Educated	Married	No Children	Not in Poverty
1	0	0	0	0	0	0
2	0	0	0	0	1	0
3	0	0	0	1	0	0
4	0	0	0	1	1	0
5	0	0	1	0	0	0
6	0	0	1	0	1	0
7	0	0	1	1	0	0
8	0	0	1	1	1	1
10	0	1	0	0	1	0
11	0	1	0	1	0	1
12	0	1	0	1	1	1
14	0	1	1	0	1	1
23	1	0	1	1	0	1

Table 9: Revised Truth Table for Black Males Not in Poverty, Showing Impact of Directional Expectations

Row	High Test Scores	High Parental Income	College Educated	Married	No Children	Not in Poverty
1	0	0	0	0	0	0
2	0	0	0	0	-	0
3	0	0	0	-	0	0
4	0	0	0	-	-	0
5	0	0	-	0	0	0
6	0	0	-	0	-	0
7	0	0	-	-	0	0
8	-	-	1	1	1	1
10	0	-	0	0	-	0
11	-	1	-	1	-	1
12	-	1	-	1	1	1
14	-	1	1	-	1	1
23	1	-	1	1	-	1

Figure 1: Venn Diagram Depicting Coverage



Area a =	Cases with low/average test scores, not in poverty
Area b =	Cases with high test scores, not in poverty
Area c =	Cases with low/average test scores, in poverty
Area d =	Cases with high test scores, in poverty