REVIEW: THE SCIENTIFIC STUDY OF CAMPUS DIVERSITY AND STUDENTS' EDUCATIONAL OUTCOMES

JAMES H. KUKLINSKI

Abstract   This essay evaluates the scientific quality of two studies of campus diversity. Both gained prominence prior to the Supreme Court's 2003 *Grutter v. Bollinger* and *Gratz v. Bollinger* decisions. Using different data and different research designs, the studies reached contrary conclusions about the efficacy of campus diversity. The evaluation centers on conception formation, the reliability of measures, nested data and level of analysis, the demonstration of cause and effect, and the appropriateness of certain types of survey questions. One study generally fares better than the other in terms of these scientific considerations, although many interesting and formidable challenges to the study of campus diversity remain.

On June 23, 2003, the U.S. Supreme Court, in two related cases (*Grutter v. Bollinger et al.* [No. 02-241, 539 U.S. _ (June 23, 2003)] and *Gratz et al. v. Bollinger et al.* [No. 02-516, 593 U.S. _ (June 23, 2003)]), upheld the right of universities to consider race in their admission procedures while at the same time placed limits on that right. During the year or so preceding those final decisions, social scientists used survey research findings to influence elite and public opinion. Two dueling surveys gained especially high profiles. Psychologist Patricia Gurin and colleagues published a prominent article in the *Harvard Educational Review* (2003; also see Hurtado et al. 1999) and also submitted their survey evidence supporting racial diversity directly, as part of a Supreme Court legal brief.[1] About the same time, political scientist Stanley Rothman and colleagues published an article in the *International Journal of*

1. Gurin et al. submitted an amicus curiae brief on behalf of the University of Michigan.

Public Opinion Research (Rothman, Lipset, and Nevitte 2002) and another in *The Public Interest* (Rothman, Lipset, and Nevitte 2003). Also based on survey data, these studies showed that campus diversity has few positive and many negative effects on attitudes and educational outcomes. The authors promoted the second article with a media blitz reminiscent of the 1970 public debates surrounding busing (Coleman 1966).

This article evaluates the scientific study of campus diversity and college admissions, not its politics. The prominence of the Gurin et al. and Rothman, Lipset, and Nevitte studies before and during the Supreme Court deliberations makes them obvious choices to pursue this task. Representing the social science community as a whole, they served as its showcases. Did the studies, in this light, meet the highest academic standards? Or did they fall short and fail to represent the best the community has to offer?

Both studies are observational and survey based. That is where the similarities end. Most fundamentally, the authors define racial diversity in markedly different ways. The words—*racial diversity*—might be the same, but the shared terminology belies a divergence in meaning. Moreover, the authors adopt wholly different research designs, with one group using cross-sectional data and the other, longitudinal data. One research team determined the effects of diversity indirectly, arguing that students cannot identify those effects; the other used a variety of survey questions, some of which directly queried students on their perceptions of diversity's consequences. One study reaches a far more optimistic conclusion about the effects of diversity than the other.

Assessing the quality of the science on display in the two studies entails answering the following stock-in-trade questions: Do the scholars precisely and fully define the central concepts? Do they properly measure them? Do they analyze data at the proper level of analysis? Do they convincingly show a causal relationship between diversity and attitudinal and educational outcomes? Of course, these stock-in-trade questions must be fitted to the particular studies under consideration.

No researcher fully meets all the challenges that attend an observational study. Thus a critic can always find something to criticize, and this critic has. That is the easy part. The bigger challenge is not to exaggerate the problems by comparing them to an unrealistic standard. In the pages that follow, the working standard is this: Have the researchers recognized and acknowledged potential research problems and then explicitly shown how they tried to overcome them?

Different Studies, Different Conclusions

By all accounts, it was Justice Powell, in the well-known 1978 Supreme Court case *Regents of the University of California v. Bakke* (438 U.S. _ 312, 98 S. Ct. 2760 [1978]), who provided the initial impetus for the notion of on-campus

racial and ethnic diversity. Twenty years later, William Bowen and Derek Bok, two former Ivy League university presidents, elevated the words *campus diversity* to a new plane. In their book, *The Shape of the River* (1998), Bowen and Bok show that African Americans who attended selective institutions with active affirmative action programs outachieved their peers who attended less selective institutions. This held within every SAT interval. Bowen and Bok focus primarily on postcollege achievements such as earnings, educational attainments, and community service activities. However, they include responses to a few survey questions asking former students whether diversity had helped them get along better with members of other races and how much they supported affirmative action programs. Black and white alumni of elite institutions like Yale and Harvard both gave largely positive responses to these questions. Buoyed by the authors' prestige, as well as by other studies that reached a similar conclusion (see 1998: n. 4), the book immediately increased the legitimacy of diversity as a criterion in university admissions and thus intensified and broadened public discussion of it.

ROTHMAN, LIPSET, AND NEVITTE

Rothman, Lipset, and Nevitte (2002) take *The Shape of the River* as their point of departure and focus solely on the book's limited survey data. They contend that these data, along with those presented in other studies, are flawed. In some cases, they assert, the question wording is so biased as to push responses in the hypothesized (to be interpreted as desired) direction. In addition, the phrase "other races" means different things to different people, which undermines the assumption of subjective equivalency. Most crucially, social desirability effects likely motivate survey respondents to give "right" but insincere answers. Over time, social desirability and reported attitudes reinforce each other: increases in reported support for affirmative action and campus diversity motivate socially desirable responses, which in turn inflate reported support; and so the cycle goes.

Rothman et al. offer an alternative although not entirely new strategy—indirect measurement:

> Instead of asking members of the university community directly how they felt about the effects of diversity on campus, we simply asked them to evaluate various aspects of their educational experience and campus environment. This was treated as the dependent variable. Then we correlated their attitudes with an independent empirical measure of enrollment diversity, which was treated as the independent variable. (2002, p. 13)

The authors define campus diversity simply, as the proportion of black students in a student body, and use National Center for Education Statistics data to measure it.[2] Their dependent variables come from data collected in a 1999–2000

---

2. There were too few Latino and Native American respondents to include in the analysis.

cross-national survey of students, administrators, and faculty members at U.S. and Canadian colleges and universities. The authors limit their analysis to data from U.S. schools, which they randomly selected from among three categories: doctoral, comprehensive, and liberal arts. Although the response rates among administrators and faculty members were high—70 and 72 percent, respectively—it was a much lower 53 percent among students, which, the authors argue, does not adversely affect their capacity to make accurate inferences about student attitudes and perceptions.[3]

Rothman et al. use seven survey items as dependent variables. Four measure perceptions of the general educational environment, and three measure perceptions of discrimination and treatment of minorities. The four items on education ask respondents how satisfied they are with their university experience (asked of students only), how good a job the school does educating students (asked of all groups), how hard students work at their studies (asked of all groups), and the proportion of students who have the academic preparation to succeed in their classes (asked of faculty and administrators only). The three items on minorities and discrimination (asked of all groups) ask whether minority students are treated better, worse, or about the same as white students; whether the respondent personally has been treated unfairly because of race or ethnicity; and the extent to which racial discrimination is a problem at the respondent's institution (Rothman, Lipset, and Nevitte 2002, p. 14).

The authors hypothesize that "increasing black student enrollment (enrollment diversity) in predominantly white student bodies will produce a better educational environment in general, greater attention to and satisfaction with the quality of education, and better relations between white students and students of color" (2002, p. 15). They use both simple correlations and multivariate regression analysis to test their hypotheses. In the case of the multivariate analysis, the authors include a large number of control variables that measure respondents' demographics (e.g., gender and economic status) and academic characteristics (e.g., major for students, field of study for faculty members) and their schools' institutional characteristics (e.g., public versus private, admission selectivity). They analyze the data separately for students, administrators, and faculty members.

Although the findings vary slightly across the three groups, the analyses reveal that enrollment diversity either does nothing to improve perceptions of campus life or affects them adversely. Few of the simple correlations reach statistical significance, and nearly all that do are negative. Take students: "As the proportion of black students enrolled at the institution rose, student satisfaction with their university experience dropped, as did assessments of the quality of their education, and the work efforts of their peers. In addition, the

---

3. This low rate of response among students raises the possibility of systematic censoring (Achen 1986). For example, students who did not participate might be less receptive to campus programs designed to enhance diversity. The authors do not report the evidence they used to ensure against a systematic bias.

higher the enrollment diversity, the more likely students were to say that they personally experienced discrimination" (Rothman, Lipset, and Nevitte 2002, p. 15). In the few cases where diversity has a positive effect, it is on perceptions of discrimination and black–white relations among administrators and faculty members (but not students). However, even these relationships become statistically insignificant among administrators when the authors later conduct their multiple regression analysis. On the other hand, the regression results show a consistently negative relationship between diversity and perceptions about the quality of education, especially among administrators and faculty members. The larger the proportion of African Americans on a campus, the more likely are administrators and faculty members to express dissatisfaction with students' overall work efforts, their academic readiness, and the overall quality of education. Only on administrators' and faculty members' assessments of racial discrimination does diversity have a positive effect; and even then, it is small.

The Rothman et al. findings contrast sharply with those of *The Shape of the River* and many other prior works.[4] Most of the time, Rothman et al. find diversity and perceptions of educational outcomes to be unrelated; when there is a significant relationship, it is almost always negative. Those who advocate affirmative action on the grounds that campus diversity affords educational benefits will find little to support their cause here.

GURIN, DEY, HURTADO, AND GURIN

Gurin and her collaborators (2003) assess the consequences of diversity differently. For one thing, they take an explicitly more theoretical approach to their work. For another, they adopt a more complex and encompassing conception of diversity itself. And, finally, Gurin et al. use the kinds of self-evaluation survey questions, along with many others, that Rothman et al. find problematic.

To develop the theoretical foundation of their work, Gurin et al. draw heavily on the still-influential research of developmental psychologists Erik Erikson (1946, 1956) and Jean Piaget (1965, [1975] 1985) as well as that of contemporary scholars Bargh (1997) and Langer (1978). They integrate this research to develop a single idea: discontinuities in life that break one from normal routines and disrupt mindless, habitual thinking spur mental growth. For students who have not formerly interacted with other racial and ethnic populations, a diverse campus represents an ideal situation for students to break out of past patterns of thought and behavior.

---

4. Perhaps most telling, Rothman et al. and Bowen and Bok draw on different sets of works. Rothman, Lipset, and Nevitte (2002) citations include Banks (1994), Bernstein (1994), Rowan (1996), Sacks and Thiel (1995), Sniderman and Piazza (1993), Thernstrom and Thernstrom (1997), and Trow (1999). Bowen and Bok (1998) citations include Marayama et al. (2000), Orfield and Kurlaender (1999), and Smith (1997).

Using their theoretical framework to guide them, Gurin et al. identify three categories of racial and ethnic diversity. The first, "structural diversity," is the numerical representation of different social groups on a campus. This category is identical to Rothman et al.'s conception of enrollment diversity. But whereas Rothman et al. equate numerical representation with diversity, Gurin et al. see it as only one component.

Arguing that a racially and ethnically mixed college or university does not guarantee meaningful interactions among members of the various groups, Gurin et al. delineate two other categories of diversity: "informal interaction diversity," which, as the term suggests, requires interactions among group representatives outside the formal classroom; and "classroom diversity," formal instruction on race and ethnicity in the form of courses designed to help students understand the politics, history, and sociology of racial and ethnic groups.[5] The authors view structural diversity as a necessary but not sufficient condition for maximal educational benefits. The other two types of diversity—informal interaction and classroom—produce the educational dividends.

The authors choose to explain two types of educational outcomes: learning and democracy. In their words, "*Learning outcomes* include active thinking skills, intellectual engagement and motivation, and a variety of academic skills. *Democracy outcomes* include perspective-taking, citizenship engagement, racial and cultural understanding, and judgment of compatibility among different groups in a democracy" (2003, p. 334). In simple terms, Gurin et al. ask whether the various types of diversity lead students to become more intellectually active and sophisticated and produce citizens who will be better able to participate in an increasingly heterogeneous society.

The measures come from two longitudinal databases: the Michigan Student Survey (MSS) and a survey that the Cooperative Institutional Research Program (CIRP) initiated. The MSS database consists of responses to questions that a sample of University of Michigan freshman answered in 1990. A follow-up survey was conducted four years later, about three years before individuals filed the affirmative action suits against the university. The initial sample includes 1,129 white, 187 African American, and 266 Asian American students. The much larger CIRP survey of 184 institutions includes 10,465 white, 216 African American, 496 Asian American, and 206 Latino students. To achieve time comparability across the two surveys, Gurin et al. use only those CIRP respondents who were in their fourth year of college in 1989 and had participated in the 1985 entry survey. The CIRP survey was conducted during a period of considerable racial upheaval on campuses across the nation.

Because the two data sets do not include identical items, the analyses differ slightly. When analyzing the CIRP data, the authors use classroom diversity and informal interaction diversity as independent variables and intellectual engagement, academic skills, citizenship engagement, and racial and cultural

---

5. Gurin et al. (2003) use the phrase "informal interactional diversity."

engagement as dependent variables. Most of the dependent and many of the independent variables take the form of indexes. When analyzing the MSS data, the authors use three independent variables—classroom diversity, informal interaction diversity, and participation in relevant campus events—and five dependent variables—active thinking, intellectual engagement, compatibility of differences, perspective taking, and racial and cultural engagement. Again, Gurin et al. use indexes when possible. The authors include a large number of control variables, which tend to be the same across data sets. Within each data set, Gurin et al. run separate regressions for the relevant racial and ethnic groups.

Gurin et al. report complicated results. Some independent variables are significantly related to some dependent variables among some racial and ethnic groups and not among others. Moreover, causal relationships vary across the two data sets. The strongest and most consistent finding in the national study is that "informal interaction diversity was especially influential in accounting for higher levels of intellectual engagement and self-assessed academic skills for all four groups of students" (Gurin et al. 2003, p. 351). On the other hand, classroom diversity, the other key independent variable, had a significant impact on the same dependent variables only among white and Latino students, and, even then, its impact was always less than informal interaction diversity's. The authors speculate that the greater impact of informal interaction diversity in the national data might be related to measurement; they use three indicators to measure informal interaction diversity and one to measure classroom diversity.

African American students display the most intriguing and perplexing patterns. Generally, the two forms of diversity have the fewest positive effects among this group, even when the authors use a lenient $p < .10$ significance level. Classroom diversity is *negatively* related to self-assessed academic skills among black students. Why African American students should evaluate themselves most harshly when enrolled in the most diverse classes is not self-evident.[6]

The Michigan data produce less consistent findings. In the authors' words, "All three kinds of diversity experiences were influential for at least one of the groups, and for at least one measure of learning outcomes" (2003, p. 352). To put it another way, diversity experiences frequently fail to be significantly related to the dependent variables. Overall, the results are strongest among white students, but even here the relationships do not consistently reach statistical significance. Among Asian American and African American students, estimated relationships reach statistical significance even less frequently. Nevertheless, Gurin et al. find enough positive relationships across the two data sets to reach this firm conclusion: "The actual experiences students have

---

6. Possibly African American students contrast their performance with fellow white students' and conclude, correctly or not, that their performance is subpar.

with diversity consistently and meaningfully affect important learning and democracy outcomes of a college education" (2003, p. 358).

Thus we have two studies, two very different methodologies, and two seemingly diametrically opposed conclusions about the effects of campus diversity. Is one conclusion more valid than the other? That is, does one study represent more compelling science than the other?

The Conceptualization of Campus Diversity

Although empirically oriented social scientists fuss primarily with variables and indicators, they communicate by way of concepts. Precisely defining and delimiting concepts pose formidable challenges, as any seasoned researcher knows. Moreover, communication across studies suffers when researchers share a label but not its meaning. Reaching conclusions about the quality of the two studies, therefore, even though the main story centers on survey data, requires attention to the researchers' uses and definitions of their key concept: diversity.

A SIMPLE (BUT NOT-SIMPLE) CONCEPTION OF CAMPUS DIVERSITY

On first thought, campus diversity seems a straightforward concept. Most ordinary citizens undoubtedly construe diversity as the proportions of racial and ethnic groups in a given student body: what Rothman et al. call enrollment diversity and Gurin et al. call structural diversity. However, even simple concepts raise questions, diversity not excepted.

Suppose that a campus student body consists of a white majority (M) and two other ethnic groups (m_1 and m_2). How should the researcher define diversity? One possibility is $m_1/(M + m_1 + m_2)$ for the first minority group and $m_2/(M + m_1 + m_2)$ for the second. An alternative is $m_1/(M + m_1)$ for the first group and $m_2/(M + m_2)$ for the second. The choice of denominator distinguishes the two cases; and that choice has obvious consequences. Unfortunately, neither Rothman et al. nor Gurin et al. explicitly state what conception they have in mind, nor can one infer it from their analyses.

There is a more fundamental problem: neither of the preceding computations provides a single numerical criterion that characterizes the campus's student body; the segmentation of diversity into individual ratios for each group necessarily masks the overall level of diversity. Imagine, for example, that one campus's student body is 60 percent M, 35 percent m_1, and 5 percent m_2, while the second's is 60 percent M, 20 percent m_1, and 20 percent m_2. Using either of the two calculations discussed above, the second campus would look more diverse with respect to m_1 but less diverse with respect to m_2. Yet, arguably, the second campus has the more diverse student body.

Measures of overall diversity exist. Lieberson summarizes some of them in his 1969 *American Sociological Review* article (and also laments their limited

use back then). He proposes an approach that applies permutations and combinations to a multinomial. It answers the question, If all students on the campus are paired together two at a time, what proportion of the pairs will consist of students with different ethnic backgrounds?

Take a campus whose student body consists of four ethnic groups, one of them being whites. If x_1, x_2, x_3, and x_4 represent the proportion of the total student body affiliated with each group, then $\Sigma x_i = 1.00$. The proportion of pairs with each possible ethnic combination is the following (Lieberson 1969, p. 851):

$$(x_1 + x_2 + x_3 + x_4)^2 = (x_1)^2 + (x_2)^2 + (x_3)^2 + (x_4)^2 + 2[(x_1 x_2) + (x_1 x_3) + (x_1 x_4) + (x_2 x_3) + (x_2 x_4) + (x_3 x_4)] = 1$$

The sum of the first four terms indicates the proportion of pairs with a common ethnicity, and the sum of the last six terms indicates the proportion of pairs without a common ethnicity. Subtracting the sum of the last six terms from 1 also indicates the proportion of pairs that do not share a common ethnicity.[7]

Consider, for example, three hypothetical ethnic mixes. In the first instance, there is a very dominant white group (70 percent of the total student body) and three small minority groups, each constituting 10 percent of the total student body. In the second, white students make up 50 percent of the student body, one ethnic group constitutes 30 percent, and the other two ethnic groups each constitute 10 percent. In the final case, each group makes up a third of the student body. The probabilities of two students randomly selected from the hypothetical campuses sharing a different ethnicity are .48, .64, and .70, respectively. The index shows, correctly, more diversity on the second and third campuses than on the first.[8]

Lacking such composite measures, researchers cannot identify the overall level of diversity on a single campus. Neither, of course, can they validly compare diversity across campuses. These problems beset both studies.

A COMPLEX CONCEPTION OF DIVERSITY

To their credit, Gurin et al. take pains to delineate their three categories of diversity—structural, classroom, and informal interaction—and the possible

7. This index assumes sampling with replacement.
8. Students of comparative politics, who face the challenge of characterizing party systems on the basis of a common numerical criterion for purposes of comparing those systems, use a similar index. Especially noteworthy is Rae's (1967) index of fractionalization, which incorporates both the number of parties and their relative sizes. Rae's index is computed as $F = 1 - \Sigma(s_i)^2$, where s_i is the proportion of parliamentary seats of party i. The closer the fractionalization score comes to 1, the more fractionalized is the party system. Conversely, a score of close to 0 indicates that one party gets nearly all of the seats.

relationships among them. Especially interesting is the authors' (probably inadvertent) introduction of necessary and sufficient conditions. They designate structural diversity, the subject of the preceding section, as "a necessary but insufficient condition for maximal educational benefits" (2003, p. 333). In other words, structural diversity in the absence of classroom and informal interaction diversity will not bring about significant educational outcomes. On the other hand, the presence of all three diversity types presumably meets the sufficiency condition.

The introduction of this language underlines two key questions that arguably constitute the guts of the real-world diversity debate: First, what racial and ethnic group proportions meet the necessary condition for diversity? Does a campus with 2 percent African Americans meet the necessary condition for structural diversity? Or does it take 20 percent? Second, what levels or combinations of all three diversity types meet the sufficiency condition? Does the University of Michigan student profile meet it? Do other campus profiles also meet it?

Necessary and sufficient logic, then, challenges the implicit assumption that more is always better than less. At the least, it turns the assumption into an empirical question. But answering the question requires other than ordinary, main effects regression models designed to identify covariation. Only recently have scholars begun to develop statistical methods commensurate with theoretical questions framed in necessary and sufficient terms (Braumoeller and Goertz 2000). Scholars seeking to address real-world debates about affirmative action and campus diversity will want to consult this emerging work.

Reliability of Measures

Measurement encompasses many difficult and complicated problems, of which the most basic are validity and reliability. Whether the two studies use valid measures is difficult to ascertain, especially in light of the previous discussion about the proper conceptualization of campus diversity. But the authors clearly differ in their attention to the reliabilities of their key dependent and independent variables. As unreliability attenuates relationships, understanding the differences is crucial to any assessment of the two studies' relative merits and contributions.

The simple verdict is this: Gurin et al. (2003) make a considerable effort to construct multiple measures of their concepts; Rothman et al. (2002) do not. Recall that the latter report many statistically insignificant relationships. The almost-certain unreliability of their single-item measures could and probably does explain many of them. In any event, their inattention to reliability reduces confidence in their reported findings.

In contrast, Gurin et al. create indexes, when possible, by summing responses to individual items. They then include the indexes in their regression

models. Although this is far superior to the Rothman et al. approach, Gurin et al. ideally would have provided more information on the indexes' reliabilities. Better yet, they would have used structural equations models to model measurement error explicitly, thus producing unbiased estimates of the relationships among their concepts. In short, the University of Michigan authors do not maximize the value of the many measures they so impressively collected.

Levels of Analysis and Research Questions

Both Rothman et al. and Gurin et al. use nested data to explain the attitudes and educational outcomes of individual students (and, in the former case, administrators and faculty members as well). One set of covariates characterizes the individual students (e.g., age, gender, SAT scores), and the other provides contextual information, primarily about the colleges or universities they attend (e.g., private versus public, college versus university, percentage students of color) or the high schools they attended earlier (e.g., quality, racial composition). Both research teams regress each of their dependent variables, measured at the level of the individual student, on all of the covariates. In other words, following in the footsteps of past work, they disaggregate all of the higher-order contextual variables to the individual level.

Nested data, while potentially powerful explanatorily, create their own set of challenges. Researchers have made great strides in documenting the complexities that inhere in analyzing such data. Gurin et al. (2003) acknowledge some of the problems in footnotes, but neither study fully addresses them.

COMPLICATIONS ARISING FROM NESTED DATA

Students sampled from a particular campus presumably interact with each other. One would expect this to be especially true for small liberal arts colleges and large universities, such as the University of Michigan, that encourage if not require diversity-oriented interaction and instruction. From the Gurin et al. perspective, in fact, meaningful diversity—classroom and informal interaction diversity—*requires* interaction, which raises interesting and challenging implications for statistical analysis.

Student interaction on a campus decreases the likelihood that the sample of observations from that campus will be independent. A lack of independence causes an underestimation of standard errors in an ordinary regression model, which, in turn, often results in coefficients achieving statistical significance when they should not (a Type I error). There are ways to test for the lack of independence, and Gurin et al. (2003) report in a footnote that they employed at least one of them, the intraclass correlation coefficient. However, they offer

no details about the nature of their test.[9] Rothman et al. fail to address the potential problem.

Both teams of researchers could have aggregated to campuses rather than disaggregated to individuals. This would have entailed computing mean scores (or something similar) on the dependent variables for each campus and regressing them on the campus diversity scores. The question then takes a different form: Are *campus* educational outcomes related to *campus* diversity?[10] Not only does the aggregation approach change the outcome variable, it creates a problem: the elimination of all within-campus variation, which is likely to be considerable. As a result, an aggregate study will likely reveal a stronger association between diversity and performance than an individual-level study will (see, for example, Robinson's classic 1950 study). In fact, individual- and aggregate-level analyses often produce different if not contradictory findings, thus creating a quandary for the researcher.

Finally, nested data cry out for cross-level inference. How, for example, do an individual's race and the campus's proportion of minority students interact to affect attitudes and outcomes? Does the diversity of the campus affect African Americans' satisfaction more than whites'? In traditional contextual analysis, which both studies approximate, cross-level inference takes the form of conditioning an individual-level variable on a contextual one; the estimated relationship between the individual characteristic and the dependent variable is allowed to vary as a function of the value of the contextual variable. This traditional regression-based approach, scholars now recognize, places all of the random error in a single individual-level error term. It improperly assumes deterministic contextual effects.

For all the reasons just enumerated, traditional regression analysis does not excel at analyzing nested data. It does not compensate for the lack of independent observations; it causes the researcher to choose between individual- and aggregate-level analyses, a decision that likely will affect conclusions; and it does not provide for the complex error structure required for cross-level inference. Do all of these problems apply to the Rothman et al. and Gurin et al. studies? The answer is yes. Could any one of these problems undermine the reported findings? Yes, but a reader cannot determine whether this is the case. Gurin et al. buffer themselves from these concerns to a far greater extent than Rothman et al. do, but uncertainties still remain.[11]

---

9. Apparently Gurin et al. (2003) measured the intraclass correlation as part of a multilevel analysis that they cite in a footnote but whose results they do not report. In a communication, Gurin, Dey, and Gurin say they are completing a hierarchical linear analysis of the nine-year CIRP data, which they do not include in the present study.

10. An aggregate analysis would require Gurin et al. to compute an overall measure of classroom and informal interaction diversity for each campus, not an easy task or one that would likely reveal much of interest.

11. In a communication, Gurin, Dey, and Gurin report that a more appropriate analysis is under way.

Future studies into the impact of student diversity should address the level of analysis problem more fully than either of the current pieces does. In particular, researchers must look to multilevel modeling, which is designed for the analysis of hierarchical phenomena such as campus diversity. Others have discussed the technical and statistical advantages of the method, which need no repeating here.[12] Suffice it to say that such models first estimate an individual-level equation and then regress the estimated slopes and intercepts on the relevant second-level variables. Most crucially, the models allow for the necessary complex error structures that regression models cannot incorporate.

DIVERSITY AS A HIGHER-LEVEL CONDITIONING VARIABLE

Less obvious than the statistical advantages of multilevel modeling is the subtle, important, and arguably improved reformulation of the research question that accompanies it. Rather than function as, essentially, an individual-level variable, diversity now takes the form of a contextual factor that shapes individual-level processes. In Gurin et al.'s case, classroom and informal interaction diversity are second-level, group phenomena while structural diversity is a third-level, campus-wide phenomenon. For Rothman et al., enrollment (structural) diversity is a second-level phenomenon, as they do not consider the intermediate levels of diversity. Gurin et al.'s more comprehensive formulation will focus the following discussion.

The researcher would first ask, in the tradition of classic contextual analysis, how classroom and informal interaction diversity shape individual-level processes and relationships. One question, for example, might be this: Does relatively high exposure to classroom and informal diversity reduce the effects of race on support for democratic values? One might also consider whether these diversity experiences mediate the relationship between high school and college performance. In both instances, the researcher would first estimate the appropriate individual-level parameters and then include them as dependent variables in a second equation that uses classroom and informal diversity as independent variables. The next step would entail determining whether structural diversity, a third-level variable, conditions the conditioning effects of classroom and informal interaction diversity on the individual-level relationships.[13]

---

12. The advantages of multilevel analysis include the following: it provides statistically efficient estimates of the regression coefficients; it provides the right standard errors; it allows the researcher to determine the relative importance of individual- and higher-level factors; it facilitates estimation of cross-level effects; and it uses data from all the units to obtain estimates for any one unit. The leading works are Raudenbush and Bryk (2002) and Kreft and De Leeuw (1998). Jones and Steenbergen (2002) present an excellent overview.

13. This sounds complicated, and it is. To simplify the interpretation of the coefficients in models with cross-level interactions, researchers often express their variables as deviations from their respective means (Hox 2002).

This approach, while potentially powerful, requires additional data on student interactions in formal courses and informal settings. In fact, without them, a researcher cannot confidently say whether classroom and informal interaction diversity enhance educational outcomes. Gurin et al.'s introduction of the two second-level forms of diversity (from a multilevel perspective) represents a gigantic step in the right direction; it cannot be the final one.

Interestingly, Gurin et al. appear to have some conception of multilevel modeling in mind when they run separate regression equations for whites, Latinos, Asian Americans, and African Americans.[14] Notice, though, that diversity, the central campus characteristic, functions as an independent variable in their analysis, while race/ethnicity, a key individual-level characteristic, serves as the conditioning variable. This precludes answering the important question: Does increased diversity change the individual-level relationship between race/ethnicity and educational outcomes?

It is too easy and not very productive to invoke the latest statistical technique and criticize prior research for not using it. It is equally easy to imply that authors do not care about level of analysis problems. Neither is intended. Gurin et al., in particular, by including some (unreported) interaction effects, by using step-wise regression, and by apparently applying multilevel analysis to their data, show an awareness of the across-level complexities. However, researchers who continue to study the effects of diversity on educational outcomes should consider multilevel modeling as an alternative to ordinary regression. Not only does it eliminate otherwise troublesome statistical problems, but it leads the researcher to construe diversity for what it is: a contextual factor that does or does not change the mental dynamics that students bring to campus.

Cause and Effect

Arguably the most critical question in assessing the two studies, especially because they offer contradictory conclusions, is this: How effectively does each demonstrate cause and effect? The answer depends on everything discussed thus far. Proper conceptualization, valid measurement, and the selection of proper units of analysis precede all else. The following discussion puts these matters aside and evaluates the two studies in terms of eliminating selection effects and identifying change across time. Each is essential to showing cause and effect.

SELECTION BIAS

Users of observational data face no potentially more severe problem than selectivity. Rothman et al. do not raise the possibility of a selectivity problem.

---

14. Gurin et al. (2003) report standardized coefficients, which cannot be compared across groups. Moreover, they do not report the intercepts, which would reveal whether the various racial and ethnic groups begin with different values on the dependent variables.

Gurin et al. do: "Certain kinds of students might be predisposed to take courses that deal with race and ethnicity and to interact with students from varied backgrounds. For example, it is plausible that students who entered college with greater exposure to diverse peers because they lived in racially heterogeneous neighborhoods and attended heterogeneous high schools might seek diversity experiences in college" (2003, p. 361). One could not wish for a more precise and effective definition of the problem as it applies to diversity and educational outcomes, except, perhaps, to add that these same students might self-select into institutions known for their emphasis on diversity.

Not only do Gurin et al. express concern for the problem, they try diligently to overcome it. For example, the authors control for factors such as the racial composition of students' neighborhoods and high schools and students' initial positions on some of the outcome measures. At the same time, the authors acknowledge that "controls for these predisposing influences do not remove all sources of selection bias" (2003, p. 362).

Most crucially, the University of Michigan authors recognize that they do not "control for correlated error in the predisposing and outcome measures" (p. 362), especially when using the national data. The authors might have generated separate selection and outcome equations (Achen 1986). In the selection equations, neighborhood and high school composition, along with other variables, would predict, separately, classroom and informal interaction diversity. In the educational outcome equations (the equations the authors report), the predicted scores from the selection equations would substitute for the original classroom and informal interaction diversity values, supposedly producing unbiased estimates of the effects of classroom and informal interaction diversity on educational performance.

Gurin et al. address selection effects in the Michigan study in an especially creative way. The College of Literature, Sciences, and the Arts required all of its majors to take a race or ethnicity course at the time the researchers collected their data. Because these students constitute 70 percent of the total sample, the college requirement eliminates much of the potential force of selection. Ideally—presumably the data were unavailable—Gurin et al. would have distinguished those liberal arts and sciences majors who enrolled solely because of the requirement from those who intended to enroll anyway. Nevertheless, the authors show impressive ingenuity.

And what about Rothman et al., who fail to raise the problem of selection bias? One can imagine how selection might explain their negative findings between diversity and attitudes. It goes like this. Many young people who live in large cities and whom the best colleges and universities do not recruit enroll in urban institutions that use relatively low admission standards. The result is a diverse student body. These campuses also lack first-rate services and facilities. The result is student dissatisfaction. The combined result is a diverse and dissatisfied student body. Controlling for factors such as public versus private and standards of admission, which Rothman et al. do, might help to eliminate the selection bias, although one would be wise not to bet on it.

Rothman et al. state their hypothesis as follows: "Increasing black student enrollment (enrollment diversity) in predominantly white student bodies will produce a better educational environment in general, greater attention to and satisfaction with the quality of education, and better relations between white students and students of color" (2002, p. 15). Note that this research question centers on change: What happens when universities increase the proportions of minority students? It is the right question. However, Rothman et al. conduct their analysis using data collected from 140 campuses in 1999–2000. In other words, theirs is a cross-section analysis. Although in the tradition of 40 years of public opinion research, their research design cannot answer the question the authors pose; indeed, it can produce a wrong answer.

Assume there are four universities (A, B, C, and D), and all use admission policies to increase the level of diversity over time (see figure 1). At the first observation, taken just prior to the implementation of affirmative action, the universities hold identical scores on diversity and student performance. During the ensuing years, all four universities increase student diversity, and all see performance levels rise, although at different rates. By the second observation, and thereafter, the ranking of the universities on diversity is, from high to low, A, B, C, and D. The rank ordering on student performance throughout the remainder of the study is the reverse: D, C, B, and A. Minus evidence that
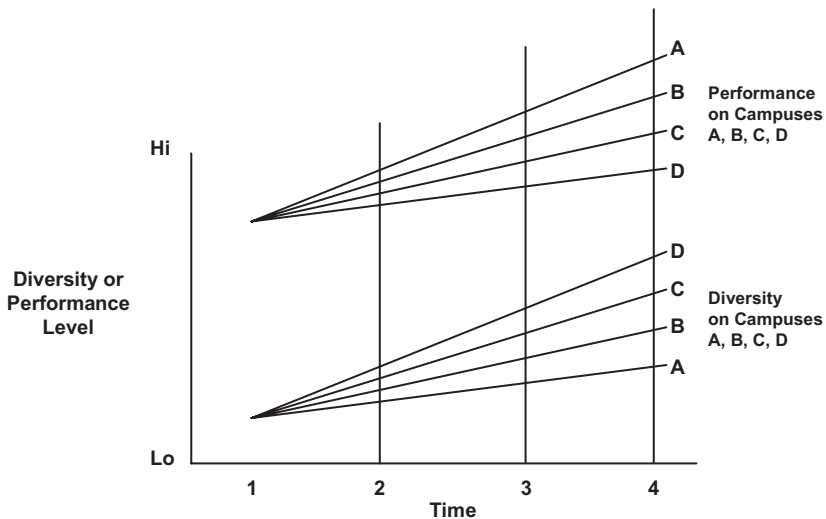
Figure 1. Positive Longitudinal and Negative Cross-Sectional Relationship between Diversity and Performance.

some other factor is driving the upward movement in both diversity and performance, the researchers would correctly conclude that an increase in campus diversity led performance to improve.

But now suppose that data exist for only one point in time, after the campuses implement affirmative action—which time point is immaterial—and the researchers use those data to calculate the cross-sectional association between diversity and performance. It will be *negative*, leading the researchers to conclude that diversity reduces performance. This is an extreme and simplified example, to be sure, and it does not prove that Rothman et al. have reached a wrong conclusion; but without evidence to the contrary, the possibility remains strong.

Gurin et al. use panel data. The Michigan survey interviewed the same respondents upon admission (1990) and four years later, whereas the CIRP survey includes interviews taken in 1985 and 1989. The authors do not use data from two follow-up interviews. Gurin et al. offer no information on attrition rate and the possible consequences of that attrition. In fact, they do not report the number of subjects included in their analysis of the CIRP data.

When possible, the authors include scores on outcome measures collected at the time students entered college as controls. This assists greatly in accounting for initial individual-level differences. That they often could not include such controls makes for a mixed bag of results. Nevertheless, Gurin et al. make good use of the data at their disposal. Even more to the point, they show an acute awareness of the obstacles to showing cause and effect.

Questions about Survey Questions

Scholars who study the effects of campus diversity sometimes use students' responses to survey questions, about which Rothman et al. raise three concerns. First, questions that ask students to assess how diversity shaped their own outlooks and educational performances demand an impossible task. Students never experience the counterfactual—being part of a student body that is more or less diverse than their own—and thus cannot know what difference diversity makes. Second, students will give socially desirable rather than sincere answers to questions about race and ethnicity. Rothman et al. note: "One wouldn't evaluate the success of a program to reduce teenage pregnancy on the basis of an attitude survey of teenagers, rather than from empirical data on changes in the pregnancy rate" (2002, p. 22); neither should one expect survey respondents to express their attitudes openly on topics laden with racial and ethnic overtones. Finally, students differently interpret questions that ask for subjective assessments. Consequently their answers to such items lack commensurability, rendering any analysis that compares them suspect.

THE COUNTERFACTUAL PROBLEM

Students, Rothman and colleagues claim, when asked if a diverse student body shaped their attitudes and educational performance, cannot know the answer if they have not also experienced life on a campus—in an ideal world, the same campus—whose level of diversity differs.[15] Only by comparing the two sets of experiences can they determine the effect of increased diversity. So when scholars ask students to assess the impact of diversity on their own lives—neither of the two studies discussed here does—they impose a task that students cannot meaningfully fulfill.

Rothman et al. make a good point. Such survey questions ask students to reach causal judgments even though they lack the requisite information. From the perspective of students, and in social science language, the independent variable lacks variation. Moreover, the logic of survey research expects the researcher, not the respondents, to identify the causal relationship. So, for example, a researcher might track students' attitudes and perceptions over a four-year (or longer) period and demonstrate an association with changes in diversity. In fact, both Rothman et al. and Gurin et al. adopt this very strategy, except that the former use cross-sectional rather than longitudinal data.[16] However, the authors' claim stands: the researcher who asks students to judge the causal effects of campus diversity demands more than they can legitimately deliver.[17]

Although important, the criticism that Rothman et al. levy applies in a limited way to the study of diversity. They (unwittingly?) exaggerate the extent to which researchers rely on the types of question they criticize. Even studies that rely heavily on survey data only rarely ask students to ponder a counterfactual.

SOCIAL DESIRABILITY

In contending that social desirability can invalidate survey studies of diversity and affirmative action, Rothman et al. identify a problem that transcends their own research interest. Social scientists recognize that respondents might give "socially right" answers on sensitive topics such as race, abortion, and homosexuality, but they have not fully mastered the problem. They have tried (see Kuklinski, Cobb, and Gilens 1997).

Suppose that social desirability motivates some students to give the responses they do. Then students from those campuses that most strongly

15. Rothman et al. (2002) also criticize the tendency by researchers to pose the questions in a positive direction, which precludes students saying that diversity had a negative effect. Their observation appears to be well grounded.
16. In fairness to Rothman et al., social scientists routinely adopt the cross-sectional strategy, which requires them to assume unit homogeneity (King, Koehane, and Verba 1994). Scholars increasingly are recognizing the problematic nature of this time-honored strategy.
17. This is not to say that students refuse to answer such questions. One can only speculate about the calculus they use to determine their answers.

emphasize diversity will feel the greatest pressure to express the "right" answers, which will inflate the relationship between diversity and the dependent variables. Neither of the current two studies appears to be overly vulnerable to this worrisome problem. Gurin et al. use two items from the CIRP data that, on their face, could elicit socially desirable responses: how much the student had changed in "cultural awareness and appreciation" and how much the student had changed in "acceptance of persons from different races/cultures."[18] Otherwise, both sets of authors steer clear of social desirability problems.

THE INCOMMENSURABILITY OF SURVEY QUESTIONS

Rothman et al. contend that students do not share a common meaning of terms such as "diverse student body" and "race or ethnicity different from your own." This, in their view, creates incommensurability problems: what the researcher assumes to be a single variable that measures the same phenomenon across respondents is not. The indictment, if right, casts a cloud over most research on diversity.

Rothman et al. presumably would extend their criticism to Gurin et al., who use students' self-assessments as dependent variables. The latter's CIRP data include the following: drive to achieve, aspirations to attend graduate school, intellectual self-confidence, importance of good writing, tendency to consider other people's points of view, amount of change in cultural awareness, and amount of change in acceptance of persons from other races and cultures since entering college (p. 347). Some of the Michigan survey items are variants of the CIRP questions.[19] From the Rothman et al. perspective, students will differ in their definitions of "good writer," "intellectually self-confident," and the like.

Although well taken, Rothman et al.'s criticism is too narrow. Incommensurability inflicts all survey research, not just studies of diversity. If there is any doubt, consider that Rothman et al. do not overcome the incommensurability problem either. They use responses to questions that ask respondents whether minorities are treated better, worse, or about the same as white students; whether they had personally been treated unfairly because of their race or ethnicity; and the extent to which racial discrimination is a problem at their institution. Members of different racial and ethnic groups probably do not share a meaning of "minorities," "treated unfairly," and "racial discrimination" any more than they share a meaning of "a race or ethnicity different from your own." Unless social scientists are ready to throw out the whole survey enterprise, therefore, they must accept, if grudgingly, the incommensurability that doubtless contaminates the study of campus diversity.

18. Note that both of these questions also ask students to judge how much they changed over the course of their college years.
19. Gurin et al. (2003) themselves apparently feel a need to defend their use of self-assessments, referring to two education studies that justify them.

It would be fair to impose a higher standard on future studies of diversity and affirmative action. A few in-depth and fine-grained analyses of how students interpret specific items would be a useful first step. Do some white students, for example, define racial discrimination as reverse discrimination? Do members of one minority group construe discrimination and unfair treatment more liberally than members of another? Longer term, researchers might find potential remedies to the incommensurability problem in the recent work by King and his colleagues (2004), who propose anchoring vignettes as a way to standardize survey responses.

Concluding Comments

Argument and persuasion form the foundation of both politics and social scientific research. Politicians seek to persuade citizens that they hold the right policy positions; scholars try to convince their colleagues that they have reached the right conclusions. Politicians frame policy debates, and social scientists frame research questions, both with an end in mind. In short, values motivate politicians and social scientists alike. Consider, in this light, the following: many of Rothman et al.'s (2002) citations include the term *multiculturalism*, while many of Gurin et al.'s (2003) include the term *diversity*. Presumably the latter term has more general and favorable appeal than the former. Gurin and colleagues reach a far more positive assessment of campus diversity than Rothman and colleagues do.

Given the highly controversial nature of diversity and affirmative action, the obvious influence of values on the two studies neither surprises nor reduces the contribution of either study. For when all is said and done, each of the studies is a *scientific* study. Scientists use generally agreed-on scientific criteria to judge each other's work. Scientific evaluation, in other words, can—and must—proceed independently of any explicit or implicit normative perspective the authors of a scientific study adopt.

This review has evaluated the two studies with respect to their science: the approach to causal analysis, the appropriateness of the research design, the attention to conceptualization and measurement. Rothman et al. and Gurin et al. might or might not agree with these criteria; they likely will not agree fully with their applications. Scientific evaluation does not imply consensus. Indeed, the lack of consensus drives scientific progress.

So how do the two studies fare? Avoiding the question would comfort, but it would not serve the purpose of this discussion. Gurin et al. consistently raise and address issues and problems that distinguish good science from not-so-good science. Their never-ending attention to these matters is paradigmatic. Whether it is selection effects, causal direction, or the weakness of cross-sectional data, the authors anticipate criticisms and act accordingly. To be sure, they often fall short of the highest standards, but so does just about every

other published study. As noted at the outset, it is the awareness of potential problems that counts most, and here they excel.

Rothman et al. show less sensitivity. Perhaps their concentrated focus on question wording and social desirability effects explains their relative inattention to other matters. In the final outcome, however, these other matters determine research quality.

Assuming that Gurin et al. have set the standard for research on campus diversity, should future investigators willingly accept it? No. Scholars must raise the standard, especially when the topic has so many real-world implications. Part of this task entails addressing the weaknesses identified throughout this article or otherwise showing why they are not weaknesses. Until this happens, declaring campus diversity a success or failure will be premature.

References

Achen, Christopher H. 1986. *The Statistical Analysis of Quasi-Experiments*. Berkeley: University of California Press.

Banks, James A. 1994. *Multiethnic Education: Theory and Practice*. 3d ed. Boston: Allyn and Bacon.

Bargh, John A. 1997. "The Automaticity of Everyday Life." *Advances in Social Cognition* 10:2–48.

Bernstein, Richard. 1994. *Dictatorship of Virtue: Multiculturalism and the Battle for America's Future*. New York: Knopf.

Bowen, William G., and Derek Bok. 1998. *The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions*. Princeton: Princeton University Press.

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

Coleman, James S. 1966. *Equality of Educational Opportunity*. Washington, DC: U.S. Governmental Printing Office.

Erikson, Erik. 1946. "Ego Development and Historical Change." *Psychoanalytic Study of the Child* 2:359–96.

———. 1956. "The Problem of Ego Identity." *Journal of American Psychoanalytic Association* 4:56–121.

Gurin, Patricia, Eric L. Dey, Sylvia Hurtardo, and Gerald Gurin. 2003. "Diversity and Higher Education: Theory and Impact on Educational Outcomes." *Harvard Educational Review* 72:330–66.

Hox, Joop 2002. *Multilevel Analysis: Techniques and Applications*. Mahwah, NJ: Erlbaum.

Hurtado, Sylvia, Jeffery Milem, Alma Clayton-Pederson, and Walter Allen. 1999. *Enacting Diverse Learning Environments: Improving the Climate for Racial/Ethnic Diversity in Higher Education*. San Francisco: Jossey-Bass.

Jones, Bradford S., and Marco R. Steenbergen. 2002. "Modeling Multilevel Data Structures." *American Journal of Political Science* 46:218–37.

King, Gary, Robert O. Koehane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

King, Gary, C. J. C. Murray, J. A. Salomon, and A. Tandon. 2004. "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98:191–207.

Kreft, Ita G. G., and Jan De Leeuw. 1998. *Introducing Multilevel Modeling*. Thousand Oaks, CA: Sage.

Kuklinski, James H., Michael D. Cobb, and Martin Gilens. 1997. "Racial Attitudes and the 'New' South." *Journal of Politics* 59:323–49.

Langer, Ellen J. 1978. "Rethinking the Role of Thought in Social Interaction." In *New Directions in Attribution Research*, vol. 3, ed. J. Harvey, W. Ickes, and R. Kiss, pp. 35–38. Hillsdale, NJ: Erlbaum.

Lieberson, Stanley. 1969. "Measuring Population Diversity." *American Sociological Review* 34:850–62.

Marayama, Goeffrey, Jose F. Moreno, Roxanne H. Gudeman, and Paul Martin. 2000. *Does Diversity Make a Difference? Three Research Studies on Diversity in College Classrooms*. Washington, DC: American Council on Education.

Orfield, Gary, and Michael Kurlaender. 1999. "In Defense of Diversity: New Research and Evidence for the University of Michigan." *Equity and Excellence in Education* 32:31–35.

Piaget, Jean. 1965. *The Moral Judgment of the Child*. New York: Free Press.

———. [1975] 1985. *The Equilibration of Cognitive Structures: The Central Problem of Intellectual Development*. Chicago: University of Chicago Press.

Rae, Douglas. 1967. *The Political Consequences of Electoral Laws*. New Haven: Yale University Press.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis*. 2d ed. Thousand Oaks, CA: Sage.

Robinson, William S. 1950. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 15:351–57.

Rothman, Stanley, Seymour Martin Lipset, and Nigel Nevitte. 2002. "Does Enrollment Diversity Improve University Education?" *International Journal of Public Opinion Research* 15:8–26.

———. 2003. "Racial Diversity Reconsidered." *The Public Interest* 151 (spring):25–38.

Rowan, Carl. 1996. *The Coming Race War in America: A Wake-Up Call*. Boston: Little, Brown.

Sacks, David, and Peter Thiel. 1995. *The Diversity Myth: "Multiculturalism" and the Politics of Intolerance at Stanford*. Oakland: Independent Institute.

Smith, Daryl G. 1997. *Diversity Works: The Emerging Picture of How Students Benefit*. Washington, DC: Association of American Colleges and Universities.

Sniderman, Paul M., and Thomas Piazza. 1993. *The Scar of Race*. Cambridge, MA: Belknap Press of Harvard University.

Thernstrom, Stephan, and Abigail Thernstrom. 1997. *America in Black and White: One Nation Indivisible*. New York: Simon and Schuster.

Trow, Martin. 1999. "California after Racial Preferences." *The Public Interest* 135:64–68.