A Critique of the Expert Report of Patricia Gurin in *Gratz v. Bollinger*

By Robert Lerner, Ph.D. and
Althea K. Nagai, Ph.D.

Prepared for the
Center for Equal Opportunity

Linda Chavez, President
815 15th Street, NW, Suite 928
Washington, DC 20005
Phone: 202-639-0803
Fax: 202-639-0827
http://www.ceousa.org

Table of Contents

Tables and Figures

Executive Summary

There are many design, measurement, sampling, and statistical flaws in this study. The statistical findings are inconsistent and trivially weak. No scientifically valid statistical evidence has been presented to show that racial and ethnic diversity in a school benefits students.

The bulk of Professor Gurin's analysis is based on the Cooperative Institutional Research Program (CIRP) dataset that compares schools and students across the country. This dataset is a random sample of neither schools nor student respondents. It is a non-probability "chunk" of volunteers. Findings from such a dataset must not be generalized statistically to the larger population, but Gurin does so.

There are two additional surveys of students from the University of Michigan. There is no school that serves as a comparison (i.e., control) group. These findings should be ignored.

The response rates are so low as to make the findings unreliable. There is a 14 percent response rate among schools asked to participate. There is a 28 percent response rate of students surveyed in the follow-up four-year survey.

Gurin's sample of respondents is incomplete. Asians are missing. Analysis is performed only on white, black and Hispanic respondents, although Gurin measures a school's diversity as the percentage of students of color—students of color being black, Hispanic, American Indian, *and* Asian.

There is no significant relationship between racially and ethnically diverse schools and students performing better in terms of their academic performance, their academic engagement and motivation, and their civic engagement.

Only some of Gurin's measures of civic engagement and racial/ethnic engagement are related to some of her combinations of her measures of on-campus diversity (taking an ethnic studies class, discussing racial issues, attending a diversity workshop, socializing with those from other racial/ethnic groups, and having close friends not of one's race/ethnicity). The relationships are statistically significant but very weak.

There is evidence that items comprising on-campus diversity and civic engagement outcomes are merely indicators of political liberalism, rather than support for Gurin's hypothesis that greater diversity causes greater civic engagement, but this evidence is not considered by Gurin. The measures of on-campus diversity correlate with students' desires to clean up the environment, influence the political structure, influence social values, join a community action program, etc.

The number of students is so large and so many variables are entered into the equations that any trivially small effect would be picked up as statistically significant, even if it is just a chance phenomenon.

Gurin acknowledges that diversity is beneficial *if* there is equal status contact among groups, but she does not test for the latter. Racial and ethnic preferences in admissions and gaps of 100 points in average verbal SAT scores and 130 points in average math SAT scores between black and white admittees (and 70 points between Hispanics and whites for both the verbal and math SATs) creates unequal status among groups at the University of Michigan.

Other studies show that emphasis on group identities and racial preferences can lead to more group stereotyping and more hostility between groups.

Acknowledgments

We would like to thank Linda Chavez and the Center for Equal Opportunity for giving us the chance to work on this critique. We would like to thank Thomas Wood and Malcolm Sherman of the National Association of Scholars for sharing with us their report, *Does Campus Diversity Have Educational Benefits?* And thanks once again to Roger Clegg for editing our work.

I. Introduction

For social science statistical analysis to support the idea that racial and ethnic preferences should be used in admissions to public universities and colleges, the social science study must be of the highest quality, because the principles at stake are so important. Patricia Y. Gurin, Professor of Psychology at the University of Michigan at Ann Arbor, was commissioned by the University to write a report defending the use of preferences in college admissions. She was to show that that racial and ethnic diversity on college campuses has substantial educational benefits.

Professor Gurin submitted her report, "The Expert Witness Report of Patricia Y. Gurin," purporting to make this point by means of a social science statistical analysis.[1] The document you are reading is our critique of her report.

Our overall conclusion is that the Gurin report is flawed in every major aspect—research design and method, measurement, sampling, statistics, and statistical interpretation. Failure to satisfy minimally the conditions of any one of these dimensions invalidates the conclusions. Gurin fails on all accounts.

Social science research is a complex process, but it follows a series of well-defined steps. Each of these steps must be carried out properly to obtain valid conclusions. Just as a chain is only as strong as its weakest link, the conclusions derived from any research study are only as reliable as its weakest part.[2]

The typical sequence of social-scientific research involves the following steps:

- Formulating concepts and research hypotheses
- Creating the research design
- Establishing measurements for important concepts
- Defining the sample and its selection procedures
- Collecting the data
- Performing statistical tests on the data analysis
- Based on the above, reaching valid conclusions

Books on research methods generally cover the same points, albeit sometimes in different order.[3]

It is not enough for a study in this area to be interesting or original, to raise important questions about a subject, or to be provocative. These features may be enough to get a study published or a dissertation passed, but it is doubtful that they justify dramatic alterations in the long-established constitutional principle of equal treatment under the law.

Our view is that, in the context of legal and public policy, social science studies and their findings should be robust enough that policymakers have faith in the study's reliability and validity. Above all, judges and policymakers must have confidence that more research is unlikely to overturn their findings.

This is not an unreasonable requirement. The National Institute of Health and the Food and Drug Administration set explicit research criteria, so that medical, food, and drug studies are required to pass very high standards, over a long period of time, over many clinical trials.

These standards have been developed because the consequences of relying on inadequate studies or insufficient empirical evidence can be devastating. Yet, in the areas of social and educational policy, the standards have been looser and, as a result, policy has been guided by research that follows academic fashions and fads rather than rigorous scientific study. When subsequent researches that are better disconfirms these earlier findings, the policies based on them must be rethought and even totally abandoned.

For example, in 1973, social scientists Elizabeth Herzog and Cecelia Sudia purported to find that the effects of growing up in fatherless homes were at most minimal and likely to be due to other factors. The authors did not stop here. They stated it might be a good idea to increase community support for single parents,[4] rather than developing policies that forestall the absence of fathers, or oppose easy divorce. This study was part of a larger current of expert opinion proclaiming that growing up in a one-parent family had no negative consequences for the children living in these arrangements. Sociologist Jessie Bernard, in her 1972 book, *The Future of Marriage*, went so far as to say that to be happy in a traditional marriage a woman must be mentally ill.[5]

With more rigorous research in the past decade, these results have been challenged, but the personal costs have been high. Research has demonstrated that divorce is not the costless exercise for children that many had proclaimed it to be in the sixties and seventies. The newer research shows that many children growing up in fatherless families do not do as well financially, academically, and emotionally, both as children and as adults, as those raised in families with their biological parents married to each other.[6]

Diversity studies in education in the 1990s look very much like the "divorce has no consequence" academic movement of the 1970s. Like others studying diversity and education, Gurin claims her statistics demonstrate conclusively that racial and ethnic diversity benefits students. Thus, universities and colleges should discriminate in favor of underrepresented minorities so students can all benefit.

We will show that if Gurin had followed the rules, she would have had to conclude that the statistics do not show that structural diversity has any benefits. Moreover, there is little evidence that other forms of teaching and promoting diversity have a positive impact. There is even some evidence that it may produce some harm.

We will take each of the standard research criteria for statistical studies in the social sciences to evaluate the expert report of Patricia Gurin. We will conclude with a review of alternative social science theories and studies on group formation, group identity, and group conflict. We will raise the alternative possibility, that racial preferences in admission may increase group hostility and group stereotyping on campus.

II. Formulating Concepts and a Research Hypothesis

All proper statistical studies begin with careful definitions of key concepts and careful delineation of the relationship between key concepts. Formulating the proper hypothesis is the crux of any scientific design, and its development requires special care for the hypothesis determines the main focus of the study and frames all subsequent research endeavors. In its simplest form, the proper hypothesis is expressed as an explicit conceptual relationship between two variables, whereby something (the independent variable) “causes” something else (the dependent variable).[7]

A. Gurin’s Independent Variables

Gurin contends that three independent variables—structural diversity, classroom diversity, and informal interactional diversity—contribute to positive academic and “democracy” outcomes.[8] While structural diversity and classroom diversity are single items, informal interactional diversity consists of several different measures that are discussed below.

Gurin’s theoretical argument is that late adolescence and early adulthood is the critical developmental stage during which individuals’ ways of thinking, the content of their thoughts, and their basic civic values can be changed under the proper institutional conditions. She cites the classic study of political socialization by psychologist Theodore Newcomb of the effects of attending Bennington College over a four-year period on students’ political and social attitudes, and several follow-up reports describing how these attitudes persisted into adulthood, as well as materials drawn from Piaget’s theories of cognitive growth as tied to Ruble, whereby cognitive complexity increases as a function of being in new and challenging environments.

Gurin claims, however, that the overwhelming majority of whites lives in predominantly white neighborhoods, as do students of color: “Increasing the numerical representation of various racial/ethnic and gender groups is the first essential step in the process of creating a diverse learning environment . . . . Structural diversity alone will present discontinuity for the vast proportion of college students who come from racially segregated[9] pre-college environments — students of color as well as white students.”[10]

Demographically, however, the vast majority of those growing up and then living in racially homogeneous neighborhoods are white. She notes, “Vast numbers of white students (about 92 percent) and about half (52 percent) of the African American students come . . . from segregated backgrounds. As groups, only our Asian American and Latino/a students arrive here already having encountered considerable diversity in their pre-college experience.”[11] In order to have a college setting where incoming white freshmen are exposed to new and different thoughts and lifestyles regarding race, colleges must have racially diverse campuses. Gurin views the college setting as critical places and times when racially diverse campuses result in changes in how whites think, what they think, and what they value, since the overwhelming majority grew up in what she calls racially segregated neighborhoods.

B. Gurin’s Dependent Variables

Gurin delineates two types of outcomes—academic outcomes and what she calls "democracy" outcomes—that are thought to result from both structural and campus diversity. In this section, we will discuss these as dependent variables at the level of theoretical concepts, focusing on academic outcomes and "democracy" outcomes as theoretical concepts. In later sections on operationalization and measurement, we will discuss how she takes these theoretical concepts of academic outcomes and democracy outcomes and translates them into "real world" concepts and measurable, quantitative variables.

1. Academic Outcomes as a Dependent Variable

Gurin contends that the appropriate academic outcomes from the college experience upon which one should judge the effectiveness of a collegiate education are essentially more "complex modes of thinking." Based on the recent work of Patricia King and colleagues (as cited in Gurin's Theoretical Foundations), Gurin argues that students "must think deeply and effortfully to take into account multiple points of view, evaluate evidentiary claims, and draw conclusions based on conceptual soundness, coherence, degree of fit with the data, and meaningfulness."

Gurin further notes that King promotes diversity and multiculturalism since presenting multiple perspectives from the points of view of race, class and gender foster fully reflective thinking." Gurin, however, admits that advocates such as King "have not measured the explicit effect of racial diversity" (Theoretical Foundations).

Gurin also finds that, in the literature on small group psychology, "members of heterogeneous working groups offer more creative solutions to problems than those in homogenous groups" and "show[] greater potential for critical thinking" (see Theoretical Foundations, where she cites the recent work of Cox, McLeod; Lobel, Cox; and Janus).

In this theoretical context, Gurin's study may be viewed as an early empirical test of King et al. as well as the heterogeneous small group studies of Cox and others. Because Gurin herself admits that little systematic quantitative research exists to support these claims made by King, Cox, and others as applied to racial diversity in the area of higher education and cognition, her study should be viewed as a preliminary study, and no more than that. It is a report based on some fairly recent and rather skimpy theoretical concepts.

It is worth noting what is not stressed as far as academic outcomes are concerned. These include the growth of student knowledge, college grade point average, student completion of the course of study, and, not the least, student performance on post-baccalaureate examinations such as the LSAT, the MCAT, the GRE exams, and other measure of academic achievement, ability, and knowledge. All of these variables are available in the Astin database, so presumably they would have been available for her analysis. In fact, two such outcomes, college GPA and completion of the course of study, are analyzed by Gurin and discussed in a later section of this paper.

2. "Democracy" Outcomes as the Other Dependent Variable

As in our discussion of the theoretical underpinning of cognition and academic outcomes, we will briefly examine the theoretical concept, "Democracy Outcomes," and the literature upon which Gurin develops her unusual concept. Gurin's presentation is one-sided and extraordinarily unconventional.

Gurin claims in her account of the theoretical underpinnings of American democracy that what "has prevailed in the United States is more akin to Plato's than Aristotle's concept [of democracy.]" She starts with Saxonhouse's presentation of democracy in ancient Greece, of Plato versus Aristotle, and incidentally cites Pitkin and Shumer. Gurin states further that this tradition, more akin to Plato than Aristotle, "is the Republican [sic] tradition, represented by Rousseau on through Jefferson . . . ." While there is no space for an extended analysis, we note that the traditional view of American democracy as based on Lockean liberalism as a doctrine of individual rights and self-government is not mentioned at all by Gurin.[12]

Gurin does recognize that there are conflicting views on the future of American society. "Little wonder that we are now facing cultural, disciplinary and political debates over the extent to which our American democracy can survive with so much heterogeneity and so many group-based claims in the polity." She sees the university's mission as pushing this new, group-based vision of democracy upon the rest of American society, since students "need to learn how to accept diversity, negotiate conflicts, and form coalitions with individuals and groups if they are to become prepared to be leaders in an increasingly heterogeneous and complex society."[13] It should be noted that this view is asserted as true; competing views are not discussed.

From this highly controversial and one-sided perspective, Gurin proceeds to lay out her Democracy Outcomes as basically consisting of ways in which students support civic activism and group-based action. We will discuss her application of the theoretical construct democracy outcomes in subsequent sections. For now, we present the basic theoretical hypothesis that Gurin claims she will test.

3. Gurin's Basic Hypothesis Statement Is Confusing

The hypothesis statement describes the relationship between the independent and dependent variables. For Gurin's study, the hypothesis is that increases in structural diversity, classroom diversity, and informal interaction diversity results in improvement in academic outcomes and increases in democracy outcomes.

While presenting the hypothesis seems straightforward, when the hypothesis is translated from its abstract theoretical concepts into the particular research setting, many problems of design and sampling emerge. Gurin corrects for none of these, thereby making her findings of little value. We will now turn to the subject of operational definitions and the primary units upon which the statistics will focus, or what researchers call the operationalization of concepts and specifying the units or levels of analysis.

4. Operationalization of the Concepts

The next step after formulating a theoretical hypothesis is to transform the variables into usable, real-world processes and actions. Who or what precisely is the unit being studied? At this step, Gurin and others who have executed similar research commit a major mistake. They misspecify the unit, or subject (in survey research, the respondent), under investigation.

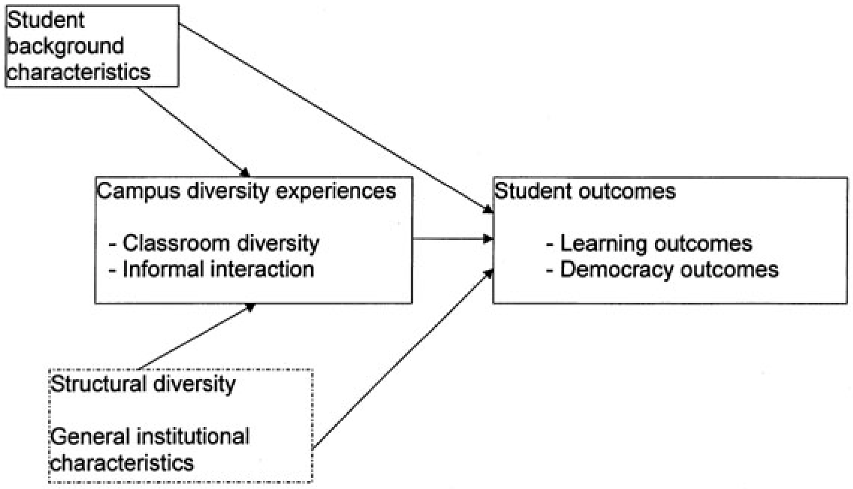
Gurin's hypothesis, which is really a series of hypotheses, contains several different levels or what are sometimes called units of analysis—the college or university, the class, and the individual. One of her independent variables, structural diversity, is a school-level variable. That is, it is a property of a particular school, and every individual in that school has the same value on it. Another variable, classroom diversity, is a classroom-level variable. The remaining variables are measured separately for each individual person regardless of what school they attended. Hypotheses that attempt to test effects on multiple levels of aggregation require a substantially more complicated analysis that must be taken into account in deciding what models to estimate and in deciding what kinds of statistical techniques should be used to test the hypotheses.

As discussed later, Gurin and others ignore the complexities introduced by formulating her hypotheses in this manner, and she subsequently selects the wrong models to estimate and the wrong statistical techniques to carry out the analysis.

In Gurin's Figure 1, she portrays her concepts as follows (see Figure 1 below, reproduced from Gurin's "D. The Studies: Method and Measures"):

Figure 1
Gurin's Conceptual Model[14]

Figure 1: General analytical approach used for the three studies



Note: The structural diversity and general institutional characteristics measures shown in the lower left of this figure are only relevant for the CIRP data analyses, as CIRP is a multi-institutional database. In contrast, the MSS and IGRCC databases are from a single institution (that is, the University of Michigan). As a result, institutional characteristics do not vary within these latter studies.

Proper analytic strategy based on her model and multiple units of analysis require first and foremost proper operationalization of the schools variable and proceeding from that point.

Gurin operationally defines structural diversity as the percentage of undergraduates who were "students of color" (which she defines as African American, Asian, Hispanic, or Native American—that is, all groups but whites) at a particular school (see Gurin, Appendix C, The Studies, Methods, and Measures).

"Classroom diversity" is operationally defined as whether or not a student enrolled in an ethnic studies course during college. Here, the definition is in terms of an individual student's behavior, thereby transforming what should be a classroom level variable into a student variable. Classroom diversity should be measured as the percentage of students of color enrolled in certain classes. This would keep it definitionally consistent with the prior variable, "structural diversity." Following from this definition and the basic hypothesis, students enrolled in classes with greater percentages of students of color would have more positive learning outcomes and more positive democratic outcomes than students enrolled in more homogeneous classes. This would test whether classes with proportionately more persons of color (such as a civil rights class in political science or a social problems or race and ethnicity class in sociology), independent of how percentage of students of color in the school varied, has a positive effect on the dependent variables.

Or, in keeping with the theme of ethnic studies, "classroom diversity" could be operationalized as the percentage of courses on a particular campus that are multicultural in theme, the percentage of all classes that are part of an ethnic studies department, or the percentage of all classes that are part of a multicultural diversity requirement. Measuring classroom diversity as either a characteristic of the classroom population or as the percentage of the curriculum would be more true to the spirit of Gurin's approach. As it is, however, classroom diversity simply measures whether or not an individual student has taken an ethnic studies class or not.

The other independent variable is informal interactional diversity. Gurin operationally defines this for individual students in four different ways: whether the student has discussed racial issues, whether he or she attended a racial/cultural awareness workshop, whether the student socialized with someone of a different group, and the proportion of an individual's friends who were not of the same race as the respondent.

While structural diversity is relatively straightforward, Gurin's four measures of campus diversity, each measured for individual students, are more problematic. Gurin's classification of these variables makes little sense. To begin with, attending an ethnic studies class and participating in a racial/cultural awareness workshop really are outcomes of a similar kind and belong together in any analysis. Discussion of racial issues, socializing with someone of a different group, and the proportion of one's friends who are of a different racial group are somewhat similar as well. Unlike attending the class and the workshop, however, they all call on the individual students to make vague attributions about their surroundings. This is discussed more below.

While Gurin's discussion of these variables is sketchy and incomplete, a much bigger problem follows from Gurin's inattention to problems of operational definitions and the proper units of analysis. It is unclear what statistical model she purports to test. Testing the wrong model also means the findings, by definition, are false.

C. Other Internal Inconsistencies

Verbal theories, concepts, and formulations need to be translated into statistical language for empirical testing purposes. Sometimes they are first pictorially represented (Gurin, Figure 1). The verbal model, the pictorial model, and the statistical model should match in order for proper statistical testing to occur. A major problem with Gurin's formulation is its vagueness. There are two different statistical models that are suggested by her prose, charts, and actual regression (computer) runs.

The first statistical model implied by her words is as an interaction model and the second is a causal chain model. If Gurin really meant to use an interaction model, her statistical procedures, and thus her results, are wrong. If she really meant to use a chain model, her statistical procedures are still wrong but in different ways.

1. Gurin's Analysis As an Interaction Model

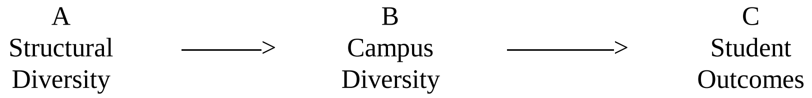
Theoretical reasoning that employs the language of necessary and sufficient conditions translates into statistical models that are called "interaction effects models."[15] Gurin's language explaining her view of the importance of diversity implies precisely this kind of statistical model. "Structural diversity alone will present discontinuity for the vast proportion of students who come from racially segregated pre-college environments" (Gurin, p. 21). And: "the impact of structural diversity depends greatly on classroom and informal interactional diversity. Structural diversity is essential but, by itself usually not sufficient to produce substantial benefits" (Gurin, p. 22). "Without this [on-campus diversity], students will retreat . . . to find settings that are familiar and that replicate their home environments" (Gurin, p. 23).

Thus, the first statistical model, the interaction model, can be gleaned from her verbal theorizing. In an interaction model, structural diversity is a precondition for positive learning and democracy outcomes, as well as the combination of positive structural diversity and positive campus diversity, by which is meant (a) classroom diversity (taking an ethnic studies class) and (b) interactional diversity (participating in an racial/ cultural awareness workshop, discussing racial issues, socializing with individuals of a different race, and having friends of a different race). For purposes of discussion, the five diversity measures of individual students are referred to as "campus diversity" measures. Positive structural diversity and positive campus diversity are both required and sufficient to produce these positive learning and democracy outcomes.

The problem is a fatal disconnect between Gurin's verbal theory and the models actually tested against her data. Gurin's statistical models do not correctly operationalize her concepts and words (as opposed to her pictures). Models of interaction effects should contain a multiplicative term; they are not additive. Gurin's statistics, according to her computer runs, are performed with additive and linear, not multiplicative, variables. If diversity is an interaction between structure and campus, then the variables should be: (1) Structural Diversity multiplied by Classroom Diversity; and (2) Structural Diversity multiplied by Interactional Diversity (each of the measures of interactional diversity—attending a diversity workshop, discussing racial issues, socializing with a person of a different group, and the inverse proportion of friends who were of the same race as the respondent).

2. Gurin's Analysis As a Chain Model

A second formulation of Gurin's statistical modeling strategy, one that is consistent with her diagrams and the models she estimated but not with her verbal formulations, is a chain model—that is, a model where greater structural diversity leads to greater on-campus diversity, which leads to improved learning and improved democracy outcomes. (See diagram below. As a diagram, it is a simplified version of Gurin's Figure 1.)



A Structural Diversity ———> B Campus Diversity ————> C Student Outcomes

In the diagram, “A” is structural diversity, “B” is on-campus diversity, and “C” is one of Gurin’s many outcome variables. The arrows show the assumed causal relationships between the variables. In the diagram above, “A” and “B” are direct causes of “C,” while, according to Gurin, “A” is an indirect cause of “C.” For simplicity’s sake, we combined the five measures of campus diversity into a single measure.

There are a number of problems with this formulation, even aside from the important fact that it is inconsistent with her necessary and sufficient condition reasoning. The first problem is that the effects of the five on-campus diversity measures are not clearly placed in the diagram. What is the relationship between them? Gurin does not say. The second problem is that her Figure 1 is unclear as to whether there is a direct path between structural diversity and student outcomes or whether there is only an indirect path from the box, which combines structural diversity and general institutional characteristics. Her text does not discuss the path explicitly.

The third problem is Gurin’s discussion of the link between structural diversity and on-campus diversity variables. When carrying out her statistical analysis in each of these regressions, Gurin controls for a series of individual characteristics but fails to control for *any* school characteristics, except for institutional selectivity. Since structural diversity is a collective property of an institution, one that is strongly shaped by the admissions office, additional institutional variables are needed as control variables in order to see if these correlations are spurious. For example, these relationships may be affected by type of institutional control, institutional size, geography, financial aid support, and other structural variables. In the absence of these controls, Gurin’s claim to have shown an indirect effect must be rejected as unproved.

The fourth problem has to do with the differing levels of analysis proposed to explain her outcome. Correct procedure regarding either model was not followed. Gurin fails to conduct a preliminary aggregate analysis, using the 184 schools in her sample as data points. Based on either model, the implications are, first, that the greater the percentage of students of color, the greater would be the percentage of ethnic studies classes taken, the greater the percentage of students enrolled in an ethnic studies class, etc. A simple correlation coefficient along with a graph containing one scatter plot could display the direction and strength of the relationship among the schools in her sample.

Second, Gurin should then have analyzed the relationship between the percentage of students of color among schools and aggregate measures of her outcome variables. To do this properly, Gurin should have conducted, but did not, a preliminary statistical analysis—the larger the percentage of students of color, the larger the mean or median student scores on learning and democracy outcomes.

This is a very simple two-variable analysis among colleges and universities. If it does not work, then any fancier analysis to see if there are structural diversity effects is likely to fail. This simple two-variable analysis at the school level was not done.

A final problem is with how Gurin operationalized structural diversity. There are better measures of diversity, called indices of dissimilarity, dispersion, or heterogeneity, that allow one to treat each racial or ethnic group as a separate component of the more complex measure. It permits establishing some kind of index that taps into the proportion of each group relative to each of the other three. This is superior to lumping Asians, African Americans, Hispanics, and Native Americans under the general rubric of persons of color (white versus non-white). For example, does it matter if there are more Asians, more Hispanics, or more blacks? Gurin's measure is unable to inform us.

To repeat, Gurin misspecifies her independent variables, and confuses important differences in the level of analyses. As we will discuss in later sections, awareness of the levels of analysis problem would have led to a better understanding of the structure of her samples and, in turn, would have avoided the misuse of subsequent statistical regression techniques in her subsequent analyses.

3. Gurin's List of Dependent Variables

In Appendix C, Gurin lists the various measures that constitute her learning outcomes and democracy outcomes. They are made up of many, many measures, including those that apply to four-year outcomes and those that constitute nine-year outcomes.[16] We outline them below.

Table 1
Dependent Variables in Gurin's CIRP Dataset[17]

Surveyed after Four Years
Learning Outcomes
Engagement and Motivation
Highest degree planning to complete
Self-rating of abilities compared to average person your age in terms of drive to achieve
Self-rating of abilities compared to average person your age in terms of self-confidence (intellectual)
Importance to you personally to write original works (poems, novels, short stories, etc.)
Importance to you personally to create artistic works (painting, sculptures, decorating, etc.)
Strength of preparation for graduate or professional school compared to when respondent entered college as a freshman
Academic and Intellectual Skills
Average undergraduate grade point average (self-reported)
Compared to when respondent entered college as a freshman, strength of respondent's general knowledge
Compared to when respondent entered college as a freshman, strength of respondent's analytical and problem-solving skills
Compared to when respondent entered college as a freshman, strength of respondent's ability to think Critically
Compared to when respondent entered college as a freshman, strength of respondent's writing skills
Compared to when respondent entered college as a freshman, strength of respondent's foreign language Skills
Self-rating of abilities compared to average person your age in terms of academic ability
Self-rating of abilities compared to average person your age in terms of writing ability
Self-rating of abilities compared to average person your age in terms of listening ability
Four-Year Democracy Outcomes
Citizenship Engagement

Importance to respondent personally of influencing the political structure
Importance to respondent personally of influencing social values
Importance to respondent personally of helping others in difficulty
Importance to respondent personally of being involved in programs to clean up the environment
Importance to respondent personally of participating in a community action program
Racial/Cultural Engagement
Importance to respondent personally of promoting racial understanding
Compared with when respondent entered college, strength of respondent’s cultural awareness and Appreciation
Compared with when respondent entered college, strength of respondent’s acceptance of persons from different races/cultures
Surveyed after Nine Years
Learning Outcomes, Nine Years Later
Engagement and Motivation
Self-rating of abilities compared to average person of same age in terms of drive to achieve
Self-rating of abilities compared to average person of same age in terms of self-confidence (intellectual)
Importance to you personally of writing original works (poems, novels, short stories, etc.)
Importance to you personally of creating artistic works (painting, sculptures, decorating, etc.)
Academic and Intellectual Skills
Average undergraduate grade point average (self-reported)
Self-rating of abilities compared to average person your age in terms of academic ability
Self-rating of abilities compared to average person your age in terms of writing ability
Self-rating of abilities compared to average person your age in terms of listening ability
Importance of general knowledge in respondent’s life today
Importance of analytical and problem-solving skills in respondent’s life today
Importance of ability to think critically in respondent’s life today
Importance of writing skills in respondent’s life today
Importance of foreign language skills in respondent’s life today
Democracy Outcomes, Nine Years Later
Citizenship Engagement
Hours per week spent on volunteer work/community service
Number of community service activities participated in
Type of service/volunteer activities participated in
As a reason for participating in community service/volunteer activities, importance of giving me a chance to work with people different from me
As a reason for participating in community service/volunteer activities, importance of influencing society as a whole
As a reason for participating in community service/volunteer activities, importance of improving my Community
As a reason for participating in community service/volunteer activities, importance of fulfilling my social responsibility
Importance to you personally of influencing the political structure
Importance to you personally of influencing social values
Importance to you personally of helping others in difficulty
Importance to you personally of being involved in programs to clean up the environment
Importance to you personally of participating in a community action program

Racial/Cultural Engagement
Importance to you personally of promoting racial understanding
Compared with when respondent entered college, strength of respondent’s cultural awareness and Appreciation
Compared with when respondent entered college, strength of respondent’s acceptance of persons from different races/cultures
Skills and Experiences Related to Living in a Diverse Society
How well undergraduate education prepares respondent for graduate schools
How well undergraduate education prepares respondent for respondent’s current or most recent job
Frequency of discussion of racial/ethnic issues
Frequency of socializing with someone of another group
How many current close friends are of respondent’s race/ethnicity
How many current neighbors are of respondent’s race/ethnicity
How many current work associates are of respondent’s race/ethnicity

In Gurin’s verbal elaboration, structural diversity and campus diversity should lead to better learning and democracy outcomes. She does not differentiate clearly, however, between subconcepts that are a part of learning outcomes and democracy outcomes, and actual measures and questionnaire items corresponding to them. As best as can be determined from Gurin’s Appendix C, this would be an increase in the following: student’s engagement and motivation, greater importance of writing and art, and intellectual and academic skills after four years of college (see four-year learning outcomes). Better learning outcomes would also show up after nine years—better engagement and motivation and better intellectual and academic skills, and a greater value placed on learning.

Structural and campus diversity should also lead to better democracy outcomes, which after four years are reflected in greater citizenship engagement, and greater racial and cultural engagement after four and nine years. Additionally, after nine years, there should be greater importance placed on community activism, racial and cultural engagement, and better skills in living in a diverse society.

These subconstructs are operationalized as follows. Regarding respondent’s engagement and motivation after four years, Gurin would expect for those in schools with greater structural diversity, having taken an ethnic studies course or experiencing greater interactional diversity, to have greater post-baccalaureate degree aspirations, higher self-rating for achievement drive and self-confidence, and greater importance placed on writing original works and creating original works of art after four years.

After four years, for those in schools with greater structural diversity, having taken an ethnic studies course, or experiencing greater interactional diversity is hypothesized to lead to better academic and intellectual skills, as measured by their self-reported undergraduate GPA, greater self-assessment of their general knowledge, analytical and problem-solving skills, better writing skills, and better foreign language skills, and, lastly, better academic ability, writing ability, and listening ability compared to those of the same age.

Students who are exposed to on-campus diversity and a diverse student body, in Gurin’s view, should also have a greater degree of citizenship engagement and racial and cultural engagement. She defines the relationship as follows: The more exposure to campus and school diversity, the more a respondent after four years should value influencing the political structure, influencing social values, helping others in difficulty, being involved in programs to clean up the environment, and participating in a community action program. Also, the more exposure to campus and school diversity, the more a respondent should think that promoting racial understanding is important, be more culturally aware and appreciate other cultures more, and be more accepting of persons from different races and cultures.

According to Gurin, taking this one ethnic studies course and exposure to interactional diversity at a school with structural diversity, even after nine years, should lead to greater learning engagement and motivation, greater academic and intellectual skills, greater citizenship and racial/cultural engagement, and better skills and experiences as related to living in a diverse society. The items presented by Gurin are similar to those she uses to measure democracy and learning outcomes after four years in school.

One difficulty with Gurin's elaboration of these items is that there is no conceptual discussion of how these many items link together to match the concepts, and not reflect other concepts. There is no explanation why originality in creative writing and the fine arts should be a function of school and campus diversity. Why would this not be reflected in originality in other spheres? More likely, the desire to create literary and fine arts projects reflect the student's major, a variable not tested. Astin's analysis of the original database suggests that a high degree of involvement in campus diversity activities is negatively correlated with majoring in business or engineering (p. 370 and p. 372, respectively).

Under academic and intellectual skills, students are asked where they stand after four years regarding writing and foreign language skills compared to those their age. They are also asked how much they've improved in terms of knowledge, analytical skill, writing, and problem solving skills. This could also reflect higher self-esteem, not real academic progress. Gurin does not use the test score data in the Astin database to test this conjecture.

Items for the four-year and then for the nine-year surveys under citizenship engagement appear to be proxies for political liberalism. There are no questions that discuss political activism that manifests a conservative tendency (e.g., attending an anti-abortion rally, the importance to you of reducing federal taxes, importance to you of reducing government regulation, importance of participating in a free speech movement, attending a Second Amendment rally, or bringing a property rights case, etc.).

Additionally, items under racial and cultural engagement are written so poorly that the socially desirable response is the only one that is likely to be given. Can anyone really imagine a respondent saying he wants to promote racial discord rather than racial understanding, or saying that her cultural awareness has declined and she has come increasingly to despise rather than appreciate persons from different cultures? It should be noted that there are no questions on affirmative action or racial preferences and apparently Gurin has never thought to ask students about these issues.

In short, there is the possibility that one set of these items, under learning outcomes, may disproportionately bias responses towards favoring a non-science, non-quantitative, humanities orientation. On the democracy outcomes, the items may be mere proxies for political liberalism. To determine this, the variables of respondent's major field of study and the respondent's political ideology as extraneous control variables should have been added to the model. Unlike questions on affirmative action, which are not in the Astin database, these questions are contained in it and thus easily available to Gurin.

D. Extraneous Variables

If the goal of a study is to test a hypothesis, and one has operationalized the concepts used in the hypothesis, then the stage is set for the next step. One must impose various controls on the research design in order to eliminate false answers. While doing this can be complex, the basic concept is simple: If you want to show that A causes B, you need to get other causes out of the way. Possible methods for doing this are to use a comparison group, to use pair or group matching, or to use multivariate statistical tests in order to control for extraneous variables. [18] These methods build upon one another and most studies use at least some combination of these methods.

In the non-experimental context, there are a number of means of controlling potential effects of extraneous variables. The most widely used method today is some form of multivariate statistical analysis—statistical analysis of more than two variables whereby the extraneous variables are controlled for by means of statistics. This is what Gurin uses.[19] Typically, the investigator draws a random sample of respondents, and then statistically controls for the effects of extraneous variables. Gurin used multiple regression techniques for controlling her extraneous variables.[20]

On the question of using proper controls, Gurin's work is deficient. To begin with, Gurin has not worked any methods of controls into her research design regarding the Michigan Student Study (MSS) and Intergroup Relations, Community, and Conflict Program (IRGCC) studies at the institutional level because she has no other school with which to compare the University of Michigan. Gurin claims that the "MSS and CIRP [Cooperative Institutional Research Program] analyses are designed to be as parallel as possible" despite having no comparison school.[21] Parallel, however, is not statistically good enough. We will discuss this in our next section on control groups.

1. Control Groups

As an absolute minimum, a study of whether diversity affects student outcomes needs *a study group and a comparison group*.[22] There must be at least two groups. In her MSS and IRGCC studies, Professor Gurin has no comparable institution, only the University of Michigan. Whatever findings come from these databases cannot be considered scientifically valid, since there is absolutely no way to test for the critical independent variable, structural diversity. Without a control group, which, at the minimum, would include results from at least one other institution, it is logically impossible to draw any conclusions about the possible effects of structural diversity on anything. Attributes that do not vary are not variables and explain nothing.

Ideally, the study and comparison groups should differ solely on the single variable of structural diversity. The groups should be otherwise identical regarding the selectivity of the school, whether it is private or public, whether it is a college or a university, the size of the student population, whether it is recognized as a regional or national institution, its location, etc. These other features are extraneous variables [23] whose influence the researcher strives to eliminate as far as possible. This should increase the likelihood that any results uncovered by the investigator are actually based on differences found along the critical independent variable of structural diversity.

Gurin admits to having no comparable institution with which to compare her MSS and IRGCC studies. They cannot be considered as scientific studies measuring the impact of structural diversity.

2. Control for Extraneous Variables

The study group and comparison group should be identical except for the independent variable. Unless a controlled experiment is possible, however, this is nearly impossible to achieve.[24] Assuming that a controlled experiment is not feasible, either practically or ethically, the investigator should then use some form of control for extraneous variables. This will increase the probability that any changes found in the dependent variable are more likely due to changes in the independent variable, rather than to other variables.[25]

One common example used in statistics classes is the relationship between the number of storks and the birthrate in Swedish counties. Counties with more storks (the independent variable) also had higher birthrates (the dependent variable). If the researcher proceeded mechanically, he or she would erroneously conclude that there is casual connection between storks and babies. Yet there *is* a correlation. Where does it come from? A third variable—rural-urban differences. Rural areas have both a greater number of storks and a higher birth rate.

The only way we know that there is no relationship between the number of storks and the birthrate is because we controlled for urbanization. Controlling for extraneous variables, therefore, is absolutely critical in establishing any kind of causal inference in a non-experimental setting. If extraneous variables are not controlled for, or are improperly controlled for, the investigator cannot conclude that his or her findings have anything (or nothing) to do with structural diversity. Finding a causal relation may be the same as claiming that more storks cause more babies.

3. Critical Extraneous Variables at the Institutional Level That Are Missed

What is interesting here is that Professor Gurin considers the conditions under which racial diversity in higher education would lead to positive effects (C. Conceptual Model of the Impact of Diversity, last paragraph):

- When contacts between racial/ethnic groups are between those of equal group status;
- Where goals are held in common;
- Where there is intergroup cooperation;
- Where there is support of authorities for group equality; and
- Where there are opportunities for group members to know each other as

individuals

These conditions should be conceptualized as extraneous variables. That is, if these are, as Gurin claims, the necessary institutional conditions that must accompany structural diversity for structural diversity to have its intended effect, then these conditions must also be statistically worked into the model. They are not.

Empirically, this raises the questions, using Professor Gurin's own criteria, of whether racial and ethnic groups on campuses with diversity have equal group status, of whether racial and ethnic groups on campus hold common goals, cooperate with each other, and get to know each other as individuals, and whether campuses with diversity have authorities that support group equality.

Gurin does have information about a few institutional variables besides structural diversity: the selectivity of the school, as measured by the mean SAT composite score for the freshman class, whether the institution was a university or a college, whether it was private or public, and whether the school had what she calls an "institutional diversity emphasis" (that is, to what degree do the students perceive the institution to emphasize diversity) and a "faculty diversity emphasis" (that is, the degree to which faculty incorporate diversity issues into the curriculum).

For student background characteristics, which is an additional set of extraneous variables, Gurin includes five—the SAT composite scores of the student, the student's high school GPA, the ethnic diversity of the student's high school class, the ethnic diversity of the student's neighborhood, and the student's sex.

In reality, the race of respondent is also a variable, but it is not listed in any of Professor Gurin's compilations of variables. It is a critical variable because race of the respondent was used to select the three separate subsamples upon which the statistical analysis was applied, for all three different datasets (the CIRP, MSS, and the IRGCC). Results were reported separately for whites, blacks, and Hispanics. This makes race (minus Asians) another variable, but Gurin doesn't acknowledge this.

Probably the single most important potential extraneous variable, as discussed earlier, for both Gurin's campus diversity variables and her citizenship outcomes is political liberalism. If both are manifestations of political liberalism, then campus diversity does not "cause" positive citizenship outcomes. A liberal viewpoint encourages participation in diversity activities and a liberal viewpoint causes positive "democracy" outcomes. Student liberalism leads to taking ethnic studies courses, going to the diversity workshops, talking about race and ethnicity, cleaning up the environment, wanting to influence the political structure, participating in a community-action program, etc.

If one goes back to the original database and the original study by Alexander Astin, one finds several of Gurin's "democracy" outcomes to be significantly correlated with political liberalism. According to Astin, they are associated with promoting racial understanding, cultural awareness, and the other indicators of campus diversity and of citizenship outcomes; he also finds that they are related to political liberalism, social activism, and, a variable not included by Gurin, participating in campus demonstrations. The correlations are large and are highly statistically significant despite the introduction of numerous control variables.[26]

In other words, on the question of adequate control variables, Gurin herself provides a list of possible confounding variables, along with the analysis in the original dataset. These variables are never entered into any statistical analysis. The relationship between structural diversity, campus diversity, and various outcomes cannot be convincingly established because the necessary controls have not been introduced.

III. Measurement

In parts I and II we have seen that a well-formulated hypothesis is critical, and that one must use certain methods to control for unrelated effects that may skew a study's results. Proper measurement of concepts is part of proper operationalization of concepts. We treat measurement as a separate problem because specific topics arise that are not directly addressed by Professor Gurin. The kind of measurements a study uses is crucial. Regarding a study's measures, there are two questions that need to be answered: (1) Is the measure *reliable*? And (2) is it *valid*? We will provide a brief description of each topic, and look at how Gurin's studies fare. Once again, Gurin's methods offer us no confidence in her results.

A. Why Measures Matter

If the correct variables are properly measured, we can say with greater confidence that the observed relationship between A and B are likely due to real effects. If the variables are wrongly measured, we have a false impression based on errors of measurement.

Since no measure is perfect, however, there will always be some error in the results. Sometimes the errors are based on the state of the respondent. The respondent may be sick, or tired, or inattentive in some other way. The respondent may have previously answered a similar kind of survey and not be paying much attention. Of course, the respondent may not like the interviewer and give less-than-cooperative answers.

Another feature is very important regarding studies on race and measurement of racial attitudes. The respondent may give the socially desirable answer. Respondents have been known to conceal their true feelings, actions, and attitudes from the interviewer if it is thought that they are undesirable. Respondents have been known to exaggerate or even invent actions and attitudes to the interviewer that respondents believe put them in a favorable light. Another problem for evaluating the quality of measures is that one cannot really tell, given a respondent's answer, what *proportion* of an answer is true as opposed to false.[27]

Faced with the social desirability problem, checking for measurement errors, particularly in controversial areas, should be done, but should be done *indirectly*. Thus, an evaluator looks for indicators: Has the measure been used before? Does the measure work over and over again? Does it really measure the thing it claims to measure? Does the measure work in tandem with other measures?

In particular, one should be on the lookout for measurement errors that slant the responses consistently in one direction. After all, the researcher is supposed to eliminate as best as possible measurement errors that produce such *systematic bias*. Otherwise the measuring instruments will themselves increase the invalidity of the results.

1. Are the Measures Reliable?

Reliability is the extent to which repeated applications of the measure result in the same outcomes. No measuring instrument is perfectly reliable, but some measures are better than others. Established measures of physical variables such as height, weight, and body temperature are less prone to reliability errors than those in the social sciences.[28]

For example, if you use a ruler to measure a person's height at four different times during a week, the ruler should give you the same number of inches. In contrast, administering the SAT to a subject four different times will produce more varied results. The SAT is therefore a less reliable test, *compared to* using a ruler. On the other hand, if one compares SAT results to an individual's answers to a survey question, such as, "Should there be less regulation of the economy?," asked of the same person four different times, the SAT is more reliable compared to an individual's responses in a survey. Reliability is a matter of degree.

This means, for better or worse, that there is no standard level of acceptability when testing for reliability. But there are three basic methods of assessing the reliability of a measuring instrument: test-retest, parallel forms, and split halves. Of the three methods, experts agree that the test-retest index is the best measure of reliability.[29] In other words, pick a measure already established in the area and attempt to replicate it many times. Rossi and Freeman's rule of thumb is that, unless a measuring instrument yields the same results 75 to 80 percent of the time, it is not useful. We will spare the reader the technical aspects of assessing reliability, since these are easily found in Nachmias and Nachmias.[30] Professor Gurin appears to use none of them, since the test-retest results are not reported.

Context is important. It is one thing to conduct an exploratory study. It is another thing to set out to influence policymakers, including courts. When the goal is to affect the larger society with the findings obtained, researchers should be able to show that the measure has been widely used, over many studies, for a long period of time.

This is not the case with Gurin's research. There is no evidence of test-retest reliability regarding the 50-plus items in the questionnaires. The technique used is, instead, to reference the survey instrument (i.e., the CIRP questionnaire) and other studies that use the same instrument. This is not to say that the measures are unreliable; we just cannot say that the measures are reliable. At the minimum, readers should know that, when respondents take the same survey after a set period of time, their answers would be the same more than 70 percent of the time, as one possible way of measuring test-retest reliability. If we cannot say they are reliable, we ought not use them to recommend public policies.

2. Are the Measures Valid?

Validity is the other major concern regarding measurement. Being able to replicate a measurement helps, but the measurement also needs to actually measure what it purports to measure. Do readings on your oven thermometer truly measure the temperature of your oven? Do readings of pH levels from your soil testing kit really measure the degree of acidity or alkalinity in your lawn? Does an individual's astrological birth sign really measure a person's personality traits?

While there are a number of ways of thinking about validity, there are two ways that are most relevant here: construct validity and empirical validity. Construct validity evaluates whether the measure (the reading on the oven thermometer) is a valid indicator of the underlying construct (the temperature). Empirical validity (sometimes called predictive validity) evaluates to what degree a measure correlates empirically with other independent measures of the same construct.

Here are two examples of how validity might be tested. The first concerns tests of mathematical ability. Standardized scores on Test X should be the same as those on other measures of math ability. If scores on Test X correlate better with a measure that is seemingly unrelated to math ability—such as church attendance—Text X is likely an invalid measure of mathematical ability.

Second example: SATs are supposed to measure the theoretical construct, "academic ability." To a lesser extent, high school grade-point-averages do, too. SATs are highly correlated with high school GPAs. The SATs in turn are highly correlated with first-year college grades. That these two measures are highly correlated with each other increases the validity of the SAT as a measure of academic ability. In contrast, suppose we used another measure—the number of extracurricular activities in high school. This measure might have no validity regarding academic ability. If it has no validity, we would expect it not to predict academic performance in college. In other words, the SAT would have high predictive validity, but participation in high school extracurricular activities might have little or no such validity. The way to know if the number of extracurricular activities has validity regarding first-year college grades is to determine if it would correlate with (1) first-year grades but also (2) SAT scores and (3) high school GPAs.

Besides lacking proper reliability checks, Gurin's study also fails on issues of construct and empirical validity regarding her independent variable, "structural diversity," and her two campus diversity variables, and also regarding her two theoretical dependent variables, learning outcomes and democracy outcomes. These failures are discussed in the following sections.

B. Measuring Structural Diversity

Gurin measures the degree of structural diversity in a school by calculating the percentage of persons of color. This is the wrong measure. She should have used an index of dispersion or heterogeneity. An index of dispersion is sometimes used to talk of multiple racial and ethnic groups, multiple occupational groups, or multiple religions.[31] In the case of universities and colleges, the higher the number on the index, the greater the dispersion (i.e., ethnic/racial diversity) for that particular school. Hypothetically, the greater a school's diversity index number, the greater the mean or median student scores on learning and democracy outcomes. This was never tested.

Second, Gurin fails to answer how racial and ethnic classifications (African Americans, whites, Hispanics, Asians, etc.) were assigned. Classifying persons according to race and ethnicity is not as obvious as it first appears. The practice can be rather standardless and arbitrary. The questionnaire of the 1990 CIRP survey is reproduced in *The American Freshman*, a report that makes claims about how freshmen have changed since 1996.[32] The individual racial/ethnic categories used in the survey instrument are not identified in Gurin's report.

Although it is extremely difficult to obtain a copy of the questionnaire, it is available on microfiche. The questionnaire gives the respondent the following options when asked about race and ethnicity: "white," "black," "Asian," "American Indian," "Mexican-American/Chicano," "Puerto Rican American," and "other." Cuban, Central American, and South American Hispanics are not listed as options, and thus must not have been systematically included in the "Hispanic" category that CIRP researchers report. This is a serious flaw in the questionnaire that Gurin fails to mention.

Where to place Native Hawaiians is also problematic—"Asian Americans" or as a subgroup of "Native Americans."[33] How are they classified? Also, how are respondents labeling themselves "none", "no response," "just American," etc. classified? Are they dropped or (more likely but wrongly) included as "whites?" What about bi-or multi-racial respondents, and does that vary depending on the combinations? At minimum, the procedures used to classify respondents must be clearly stated, including how they handle missing data.[34] Gurin does not do this.

Additionally, we believe that "Structural Diversity" is in fact the wrong concept and the wrong measure because structural diversity exists as the consequence of school decisions. What is needed is a concept and its operationalization that directly tests what UM and other such admissions committees do—engage in racial and ethnic preferences in admission. Gurin's measure as a percentage of persons of color glosses over two things: first, that Asian applicants are not considered as underrepresented and thus favored groups at these schools that have preference policies in admission and, second, that structural diversity could come about through intentional discrimination based on race or through other means, through the ordinary process of admissions. As we have shown in our separate statistical analyses of public colleges and universities, odds ratios for various groups display how much preference an institution grants blacks and Hispanics over whites and Asians. The University of Michigan, in our analysis, has one of the largest odds ratios favoring blacks over whites, controlling for test scores and grades, among all schools where such data were made available.[35]

Using the school as the unit of analysis, one additional structural variable should be the degree of preference favoring blacks, operationalized as the odds ratio of black to white applicants, while another variable would be degree of preference favoring Hispanics, operationalized as the odds ratio of Hispanic to white applicants. A third could be the Asian-white odds ratio as an institutional check to see if the particular school engages in preferences favoring (or against) Asians and by how much. A study of how this measure is correlated with institutional diversity and various other measures of diversity emphasis among faculty and students is itself an important preliminary task of analysis completely neglected by Gurin.

C. Measuring Classroom Diversity

Using the CIRP database, Gurin measures classroom diversity as taking an ethnic studies course. There is no explanation why this should be the only measure of classroom diversity used. In any case, it is plainly inadequate. Why not measure the classroom diversity of other courses in other departments? Perhaps one reason is that the idea that the racial/ethnic composition of a calculus class will influence a student's knowledge of calculus is plainly ridiculous.

Gurin acknowledges this limitation of the CIRP data in her discussion of measuring classroom diversity in her MSS study. She measures classroom diversity in her MSS study using an index of classroom diversity that is based on two questions: (1) the extent of exposure in a student's class to information/activities devoted to understanding other racial/ethnic groups and inter-racial/ethnic relations and (2) the respondent's self-reported view of whether he or she had taken a course that had an important impact on his or her views of racial/ethnic diversity and multiculturalism.

Exactly how these items are combined into an index is not described. There is also no check on whether these are valid or reliable measures of classroom diversity. In this vein, if it is argued that taking an ethnic studies course, classroom exposure to materials concerning racial/ethnic groups and relations, and taking a course that had an impact on views of race, ethnicity, and multiculturalism are measures of classroom diversity, they should statistically hang together—each should correlate with the others, and none should be more correlated with some item assumed to be unrelated to classroom diversity such as the degree of mathematical skill.

One item available in the CIRP database that should always be included as a test-factor in ascertaining any relationship between diversity measures and student outcome measures is political ideology. That is, how do we know that all these measures of classroom diversity—taking an ethnic studies course, taking a class in race/ethnicity or race relations, etc.—are not behavioral sub-components of political liberalism? If the latter is actually the master theoretical construct, then all these items would correlate with political ideology. This brings us back to Newcomb's study of Bennington College that Gurin cites extensively in her "Theoretical Foundations for the Effects of Diversity." The Bennington College study was explicitly designed to measure whether students became more liberal the longer they were at Bennington, which they did. If these classroom diversity indicators are not subsets of political liberalism, then they should not correlate with ideology. It is incumbent on the researchers to show that these items in fact do not and, unfortunately for Gurin, as noted above Astin reports precisely such correlations.

Additionally, Gurin asks no questions about the racial/ethnic composition of the class, the classroom activities required (e.g., group work), the student's grade in the class, whether the class was easy/hard, etc. These indicators of classroom diversity should correlate with measures (1) and (2) of her MSS study, as well as whether a student has taken an ethnic studies course or not.

In short, a preliminary validity check on "classroom diversity" as a measure would take these individual items and statistically correlate them with each other. They should hang together statistically as parts of a single concept, classroom diversity.

The technique often used to see whether items "hang together" is factor analysis. Such checks are conducted because what the researcher hopes hang together as indicators of a larger concept sometimes do not. For example, in studies on the ideology of American elites, responses to questions on social and political issues do not align neatly into an ideological index based on a conservative-liberal scale, but statistically factor into distinctive sub-types of political ideology.[36] Elites' responses to economic questions—such as attitudes toward government regulation, reducing the size of the income gap between the rich and the poor, and the need for environmental regulation—statistically hang together, while social questions on abortion and homosexuality cluster separately. Attitudes on social policy questions did not predict well attitudes towards the economy. Responses to questions on affirmative action, which for some reason are not part of the CIRP database, did not correlate with either cluster.

D. Measuring Outcomes

Using the CIRP database, Gurin examines 56 outcome measures (we assume this means 56 questions) divided into four subgroups—4-year learning and 4-year democracy outcomes, and 9-year learning and 9-year democracy outcomes (see Table 1 in the previous discussion on operationalization of Gurin's dependent variables).

We have the same criticisms of these 56 outcome measures as those for "classroom diversity." There is no test-retest reliability check or any other kind of test of reliability.

For the items that are listed as measuring learning outcomes, there is no analysis of whether or not these items are correlated. Why do the items under engagement and motivation in learning outcomes favor the fine arts but not the sciences? Why include writing original poems, novels, or short stories or creating painting, sculptures, or decorating matter as indicators of engagement and motivation but not building robots or writing computer software?

On intellectual and academic skills, respondent's skills are self-reported. No independent checks are built into the dataset. Undergraduate GPAs are reported by the respondents, rather than relying on administrative records. Also, the respondents self-assess their increase in general knowledge, problem-solving skills, critical thinking ability, writing skills, and foreign language skills, among others. There are no comparisons with objective aptitude and achievement test scores or course grades, even though these latter measures are available in the CIRP database.

On the items that measure the "democracy" outcomes, there is no analysis of whether these items are correlated with each other either. Nor, as mentioned earlier, is there any attempt to correlate these measures with student political ideology. We suspect that some are merely indicators of political liberalism. For example, under Gurin's indicators of four-year democracy outcomes, she has an index of citizenship engagement, where components include, among others, being involved in programs to clean up the environment and participating in a community action program. The index does not include conservative forms of participation, such as an anti-abortion or Second Amendment rally.

Gurin’s measures of democracy outcomes also suffer from the problem of social desirability bias. That is, the questions are likely to prompt a positive response because the answers are socially approved (i.e., politically correct). For example, under racial/cultural engagement, each one of these responses would involve a tremendous degree of political correctness. It is very conceivable that a student answering this questionnaire would not answer truthfully. Indeed, how likely would it be for a respondent to say that it is extremely unimportant to her personally to promote racial understanding, or to say that he is more culturally unaware and unappreciative, or to say that she rejects persons from different races/cultures? It is unlikely that fourth-year students, on the verge of graduating, would say anything on a university survey that would indicate that they would promote racial hatred and conflict.

One should be on the lookout for measurement errors that slant the responses consistently in one direction. This need not be intentional, but the research is supposed to correct for, as much as possible, measurement errors that produce a systematic bias. The systematic bias inherent in the measuring instrument (e.g., the questionnaire) increases the invalidity of results.

E. Measuring Control Variables

Attention should also be paid to properly measuring extraneous or what are sometimes called controlled variables. Of course, if the researchers fail to think of the proper controls at the conceptual stage, these missing variables can’t be properly controlled for. There are two types of extraneous variables not controlled for by Gurin. One set is individual student backgrounds variables; the other is school-level variables.

1. Student-Level Extraneous Variables

Gurin includes the following student-level controls: self-reported verbal plus math SAT scores, the high-school GPA, the degree of ethnic diversity of the respondent’s high school classmates, the ethnic diversity of the respondent’s neighborhood while growing up, and the student’s sex and race.

The following were not included at the student level as socio-demographic background variables in the CIRP although they are standard background variables in quantitative survey research: religion, father’s occupation, mother’s occupation, student’s estimate of family income, student’s birth date, and region where student lived before college. Controls relevant to students that are not socio-demographic variables but are related to the college experience include whether a student is on financial aid or not, student’s expected major or general field (humanities, natural science, fine arts, etc.) in 1989, and student’s major or field after four years. Lastly, at the student level, the respondent’s occupation after 9 years and respondent’s personal income, job title, and marital status were not included for the 9-year follow-up.

These extraneous variables should be measured following the format used by the large survey research firms (e.g., the National Opinion Research Center, Gallup). It is common practice to take the wording of the questions and possible responses from annual surveys done by Roper, Gallup, or NORC, where they have gone through test-retest and validity checks.

Without these standard controls, the likelihood still exists that the relationships uncovered by the investigator are in fact spurious and the conclusions inferred from the data are false.

2. School Level Variables: Equal Group Status, Common Goals, and Intergroup Cooperation

At the school level, there are other extraneous institutional-level variables that should have been included. These come from Gurin's list of several conditions that must be met for institutional diversity to have positive outcomes.

One condition is that the racial/ethnic groups be of equal status. Gurin does not define what equal group status means, however. One way to measure it is by the gap in SAT scores between racial and ethnic groups. That is, increasing the gap between groups increases the inequality between groups. In our report for the Center of Equal Opportunity (CEO) of the 47 public universities and colleges studied, the black-white gap in median SAT scores varied among schools, but at Michigan the gaps were some of the largest we have found between whites and blacks (but not for other groups). The largest gaps were at Berkeley before Proposition 209 and the University of Michigan at Ann Arbor, with gaps in median SAT verbal scores of 150 points at Berkeley and 100 points at Michigan. The gap in median SAT math scores was 180 points at Berkeley and 130 points at Michigan. The white-black gap in high school GPAs was 0.58 of a grade-point at Berkeley and 0.40 at Michigan.[37]

At Michigan, the gaps between Hispanics and whites were smaller. There was a 60-point gap in median SAT verbal scores, a 70-point gap in math scores, and a 0.30 gap in high school GPA. Between whites and Asians, the median Asian verbal score was 10 points higher than the white score and 40 points higher for median math scores. There was no gap in median high school GPAs between Asians and whites. The test score differentials among admittees should have been a part of Gurin's data analysis.

Gurin states that sharing common goals is also a necessary condition for institutional diversity to lead to positive student outcomes. Common goals are not defined. Providing percentage differences among the racial/ethnic groups as to their responses to simple series of survey questions would be a start, but this kind of analysis was not undertaken by Gurin.

Intergroup cooperation is another variable that is supposed to be a necessary condition. Gurin does not include it as a variable. The control variable could be measured in several different ways: Hate-crimes reported? Students' ranking of campus intergroup cooperation? Percentage of student participation in diversity workshops? There are several possibilities.

Two variables—support of authorities, and opportunities for group members to know each other as individuals—are covered by Gurin's independent variables. Support of authorities for "group equality" is implied if one would use racial and ethnic preferences as an independent variable instead of structural diversity. Operationally, the larger odds ratio would mean the greater support of authorities for all kinds of diversity policies. One could also research the admissions statements, whereby the researchers would rank the school as being very supportive, somewhat supportive, neutral, somewhat unsupportive, or very unsupportive, but this was not done. As for "Opportunities for group members to know each other as individuals," this factor is measured by Gurin's interactional diversity measures.

As will be shown in the section on statistical testing, Gurin's conceptualization of the relationship between this variable and structural diversity and the statistical meaning of the relationship are two different things. Her statistical procedure suggests a certain model, if one is only to look at the mathematical procedures. Her statistical model, however, does not match her verbal model. She actually tests some other model. This is a problem.

So far, we have shown that there are many problems associated with proper measurement of variables. Reliability statistics are not provided—raising the possibility that reliability checks were never done. Concept and empirical validity of the variables were called into question, and lastly, many variables suffer from the problem of social desirability bias in respondents' answers.

As we stated earlier, each time major errors are introduced into design, operationalization, measurement, or statistics, we increase the likelihood that the final answers give a false relationship, no matter what the level of statistical significance.

The next section looks at systematic bias and issues of sampling and sampling error. Further error is introduced as a function of how the sample was constructed and the critical problem of low response rates.

IV. Sampling

"Sampling" is a simple concept: choosing cases to include in your study. The key issues are whether the researcher used a method from which he or she can reasonably generalize and that is not subject to bias.

A. What Sampling Is and Why It Matters

Sampling is the systematic means by which cases are selected for inclusion in a study. There are two basic types of samples: probability and non-probability samples. *The distinction is critical because one cannot generalize from a non-probability sample.*

Probability versus non-probability sampling is a fundamental distinction in research. The most important fact about the CIRP database in this context is that it is a non-probability sample. Any generalizations made from it cannot be assessed for their statistical accuracy. One must not generalize to all colleges and universities in the United States using results from the CIRP database. The data may give us interesting leads, or suggest possible insights, but nothing reliable can be inferred from them outside of the individuals actually included in the database itself.[38]

1. Probability Sampling: The Key to Valid Research

In a probability sample, each unit of the population studied has a known probability of being included in the sample. These studies use randomization methods to select the respondents for a study for which population estimates may validly be made.[39] There are three types of probability samples: the simple random sample, the stratified random sample, and the cluster sample.

In the simple random sample, which is the sampling design assumed by many of the standard statistical tests, each unit in the population has an equal chance of being included in the study sample.

In the stratified random sample, the population is divided into strata, and each stratum must be represented in known proportions within the sample. Independent samples are selected by a random procedure within each stratum, and each unit must appear in one and only one stratum. This technique is used to make sure that important groups are included in sufficient numbers for statistical analysis. The technique works best to the extent the strata are homogeneous.

The third type of probability sample is the cluster sample. Unlike random samples and simple stratified samples, this technique is used when there is no easily available or reliable list of elements to be sampled. The population is divided into groups. A sample of these groups is drawn by random procedure, and elements within each of these samples are in turn selected by random procedure. For example, cluster-sampling households in a major city might first involve starting with census tracts. The investigator would then randomly select a sample of census tracts (first-stage cluster sampling), then randomly select city blocks within the sample of census tracts (second-stage cluster sampling), and finally select randomly the households within the sample blocks (third-stage cluster sampling) to obtain his final sample of households (there are random procedures used to select individuals within the household). Cluster sampling is used when no list of the population exists but where lists of units at higher levels of geography are available.

2. Non-Probability Sampling

Non-probability samples are used because the costs of obtaining a probability sample are too high for the researcher, because the researcher does not know any better, or because the researcher does not expect to make his data available to other users. The following are kinds of non-probability samples that are sometimes used.

“Convenience sampling” is just what it sounds like. One selects whoever is available, such as students in an introductory psychology class.

“Purposive sampling” involves selecting cases that the investigator believes are representative of the larger population, such as selecting “representative precincts” for election forecasting. The investigator hopes that these pre-selected precincts will mirror the state election returns.

The investigator may also resort to depending on a social network for study volunteers, where members of the network identify others in the same network until a sufficient number of cases is reached (“snowball sampling”). With “quota sampling,” the investigator tries to select a sample as similar as possible to the sampling population. An investigator may seek an equal number of men and women, for instance, if the investigator thinks the population from which the sample is drawn will have an equal number of men and women. Quota sampling requires the investigator to use his or her judgment to identify all the important features that might affect the sampling.

Magazine volunteer polls are a common form of non-probability sampling. In a typical magazine poll, the magazine reports the results of those who voluntarily respond to a questionnaire in a magazine. The respondents almost certainly differ in systematic ways from the non-volunteers, first of all by showing a strong interest in the subject of the questionnaire. There are also biases inherent in those reading the magazine itself, as compared to the general population. No matter how large the number of respondents, findings from such magazine, television, or Internet surveys cannot be generalized to the larger population.

The study of human sexual behavior is especially plagued by over-reliance on non-probability sampling. The famous Kinsey reports, for example, are non-probability samples relying substantially upon volunteers and heavily sampling highly unrepresentative locales like prisons.[40]

3. The CIRP As a Two-Stage "Chunk" of Volunteers

The original CIRP data set is a "chunk," or what some call a non-probability sample of volunteers. According to CIRP's developers, "The CIRP Norms sample is derived from students attending a group a institutions that voluntarily chose to participate in the CIRP."[41] There is no detailed description of the CIRP sampling design in either Astin's *College and Beyond* or Gurin's expert report. Nowhere is it mentioned that the CIRP is a non-probability sample of schools and freshmen. The details of the sampling design were found in a technical appendix to a monograph, *The American Freshman: Twenty-Five Year Trends, 1966-1990*.[42]

Gurin's CIRP dataset is a subsample of the original CIRP dataset. She took a subsample of the original chunk of schools. The exact procedures by which she took a subsample are not revealed. It should have, at minimum, been described in a technical appendix.

From this subsample, Gurin apparently obtained a sample of freshmen, the four-year follow-up sample, and the nine-year follow-up sample—or perhaps she just relied upon the ones that were already in the CIRP dataset. This still makes it a non-probability sample from which one cannot scientifically generalize to the larger population. Yet Gurin and others do so.

The non-probability nature of the CIRP database is not typical of large-scale academic data collection efforts. According to one author of survey research methods, "The federal government generally will not fund survey research efforts designed to make estimates of population characteristics that are not based on probability sampling techniques. Most academic survey organizations and many nonprofit research organizations have a similar approach to sampling."[43]

The first problem with non-probability sampling has already been mentioned. One must not generalize from a non-probability sample, although the diversity studies repeatedly do so. *Gurin generalizes from a non-probability chunk of volunteers. This is wrong.*

The second problem with this non-probability sample is that its size may confuse the unwary into thinking that generalizations can be reliably drawn from it. The size of the sample is irrelevant for making estimates. Population estimates based on non-probability samples are not scientific, despite appearing to be so.[44] A large non-probability sample does not provide better population estimates than a small non-probability sample. The size of the sample is only relevant for probability samples, where larger samples allow greater precision of population estimates (in extremely large samples, the tiniest effect is statistically significant).

The third and most serious problem with a non-probability sample is the likelihood of unknown biases in the results.[45] While properly selected probability samples systematically eliminate this problem of bias, non-probability samples do not. Have the investigators overlooked any obvious biases in their process of sample selection? For the CIRP database, investigators did not use geographical location or urban-rural location, to name two factors, as considerations when drawing up their initial categories of potential schools that could participate. The same problems and questions arise with the chunk of freshmen volunteers within the schools that chose to participate in CIRP. In 1990, there were 2727 schools in the eligible universe of schools of which 574 provided some information (21 percent) and of which 382 provided enough information for the "national norms" (14 percent of the total). Apparently the questionnaires are distributed to the schools by CIRP and administered by the schools using methods that are not described in the CIRP publication we have cited. The size of the volunteer sample for CIRP's national norms database for 1990—194,000—does not matter, since it is not randomly selected.[46]

B. The Attrition Problem As a Coverage Problem

Another problem associated with the CIRP database is what is called undercoverage. That is, the investigator fails to include all the target population in the universe of those that could be sampled. For example, despite its exemplary record the U.S. Census fails to enumerate all adult persons in the U.S. Estimates for the 1980 Census show a net undercount of black males by roughly 8 percent. Surveys that used this data as their sampling frame in the 1980s thus failed to include roughly 8 percent of black males in the United States population. The problem of less than perfect coverage started even before the survey began.[47]

The CIRP situation is far worse than the U.S. Census. The sample is made up of roughly 380 schools that voluntarily participated in 1990.[48] There is roughly a 90 percent yearly repeat participation rate. This means that, over time, roughly 90 percent of the same schools participate in the CIRP on a year-by-year basis. Schools were stratified by type (2-year college, 4-year college, university), selectivity, public versus private, and whether or not they are nonsectarian, Protestant, or Catholic, or historically black colleges and universities—for a total of 37 stratification "cells" or types of school.[49]

The volunteer schools participating in the CIRP distribute the survey to all incoming freshmen. Over 194,000 freshmen volunteers returned the surveys to CIRP. The number of respondents was weighted by sex.

The procedures used in the CIRP data in the Gurin report are as follows. Using freshman year respondents as the population to be sampled from, the CIRP and Gurin's other studies use a four-year follow-up survey of freshmen. The coverage error here is clear. Students who did not complete the initial survey could not be sampled for the follow-up survey. This introduces the same kind of coverage error described above. But the overall situation is worse for two reasons. First, the original surveys are based on only a small non-random percentage of colleges and universities in the United States. Out of a total of 2727 predominantly white institutions, 380 schools participated, for a response rate of roughly 14 percent. Students not part of this sample chunk cannot be included in the follow-up survey.

Second, those students in the selected schools who did not respond to the freshman survey cannot be included in follow-up survey. Although there is only limited information available, it appears that the four-year colleges could have response rates of 85 percent and universities could have response rates of 75 percent and still not have the freshmen that responded in the survey.[50]

Additionally, in any follow-up study, there is systematic bias introduced when some of the freshmen later drop out of school and, thus, the study. Dropouts differ in significant ways from those who complete their higher education. So attrition, as a type of coverage error, introduces another source of systematic bias in the CIRP database and, thus, into Gurin's analysis as well.

C. Only a 28 Percent Response Rate

While it is the case that the original CIRP chunk of schools had a very low response rate, the within-school student responses were quite variable from institution to institution, and there is no information available on the conditions under which the questionnaires were administered, these are not the worst problems faced by users of CIRP data. The most serious problem with the CIRP database and with Gurin's subsample of it is that the four-year follow-up survey samples had a combined response rate of only 29.8 percent, so 70.2 percent of fourth-year students who had previously filled out a questionnaire were asked to participate but did not.[51] We use the word "samples" in the plural because CIRP included students from two distinct samplings: a follow-up survey that was based on the normative sample and a second follow-up sample that was selected separately. This blending of data sources makes for considerable confusion in understanding exactly what was done, and is itself poor procedure.

The confusing in-sample selection procedures should not be allowed to deflect attention from the extraordinarily low 29.8 percent response rate of the combined samples. This is a very low response rate for such a major study.

Babbie, in *Survey Research Methods,* sets 70 percent as an acceptable response rate, unless it is a difficult population to survey, for which he sets the rate at 50 percent. The General Social Survey has response rates of roughly 80 percent. The U.S. Census has an overall response rate of over 95 percent—still deemed low and unacceptable to many. Dillman, in his studies of mailed questionnaires, estimates his methods of mail questionnaire surveys result in response rates of roughly 70 percent. Lauman et al., in their random sample of Americans and their sexual behaviors and partners, obtained response rates of roughly 80 percent. In surveys of American elites (made up of difficult target populations to sample), response rates were above 50 percent, except for the philanthropic elite, which was slightly lower. [52] The CIRP falls well short of these rates.

Moreover, Gurin and others do not ask, What might be a consequence of such low response rates? Why don't students want to participate? What makes non-respondents different from respondents? What about their political attitudes or their fields of interest? What about political ideology—are more liberal students more likely to participate?

The low response rate compared to other national surveys is not commented upon, by either Gurin or CIRP. No questions are raised as to whether this low response rate systematically biases the questions, perhaps in a liberal direction. Perhaps respondents are disproportionately civic activists, while non-respondents are not. Or respondents could be in the humanities or social sciences or disproportionately majoring in relevant fields (e.g., ethnic studies, sociology, etc).

To be fair, the designers of the CIRP do attempt to correct for non-response bias by employing a sophisticated weighting scheme. But this is insufficient. It is impossible to correct for all the missing data bias in this manner because it is impossible to know all the factors that might bias the selection of respondents. This is the major reason that most statisticians strongly advocate using only random sampling techniques in major national studies.

In short, the sampling problems of the Gurin report are immense. The CIRP is a non-probability sample, which means any findings must not be generalized to the larger population. It is a volunteer sample, and no one knows how these volunteers differ in significant ways from the general population. This seriously biases the sample.

Also, there is an attrition rate from first to fourth years that is not discussed. No one knows how those that finish college after four years differ in ways from those that leave college after one, two, or three years, or those that are still on campus but will not graduate. Again, this seriously biases the sample.

Lastly, there is only about a 14 percent response rate among schools,[53] and a 29.2 percent response rate among freshmen in the follow-up survey. These response rates are extremely low, also raising serious problems regarding the validity of the CIRP findings.

These problems associated with sampling and the sample design make generalizations drawn from Gurin's study scientifically invalid. At a minimum, these sampling factors make it not possible for Gurin to prove her hypotheses.

V. Statistical Analysis

We come now to the culmination of the social-scientific process: statistical testing. For a non-social scientist this may sound like a valley rather than a mountaintop. But if hypotheses are properly conceptualized, if extraneous variables are properly controlled, if concepts are properly measured, and if populations are properly defined and samples properly drawn, then we are ready for the process of statistical hypothesis testing. It should be quite straightforward. But in the Gurin study it is not.

A. Missing Asians

Professor Gurin essentially leaves Asian Americans out of her study. There are no regressions computed for Asian students the way they are for black students, Hispanic students, and white students. In fact, there are no data of any kind reported for Asian students using any of the databases in Gurin's report, again unlike for all the others. Yet Asian students are part of her definition of structural diversity.

Gurin defines structural diversity as the percentage of students of color in an institution. In her section entitled "Measures," in Appendix C: The Studies, Methods, and Measures, Gurin defines students of color in the CIRP database to include "African-American, Asian, Hispanic, or Native American" students (Appendix C, p. 14). In her Michigan Student Study, students of color include Asian Americans (p. 18). Among the five multi-campus events attended, a variable in the latter database is included for "Asian American Awareness Week events" (p. 19), and under the heading of democracy outcomes "familiarity with Asians" is included (p. 20).

Gurin's neglect cannot be for the reason that Asians do not exist on campus in sufficient numbers to be studied. They do. In our analyses of various public universities and colleges, Asians were often a larger percentage of the first-year enrollees than were blacks and Hispanics. At the University of Michigan, 12 percent of enrollees were Asian, 10 percent were black, and 5 percent were Hispanic (72 percent were white).[54] For this reason alone, they should have been analyzed. But they were not.

In addition, our own statistical analysis of data provided by the universities and colleges themselves shows that Asians receive little or no preferences in admission over whites when controlling for SAT scores and high school grades. In contrast, black applicants frequently receive such preferences, as do Hispanics to a lesser extent. At the University of Michigan, the 1995 data show that, with the same test scores and grades, the odds ratio of a black applicant being admitted over a white applicant was 174 to 1 in favor of the black applicant. The odds ratio of a Hispanic over a white applicant with the same scores and grades was 131 to 1, while the odds ratio of Asian to white applicants was roughly 1 to 1.[55]

Overlooking Asian students is a critical flaw. Diversity studies leaving Asians out of their analyses should be discounted. As there is evidence at Michigan and elsewhere that they are discriminated against relative to blacks and Hispanics, this is a serious ethical as well as a statistical issue. Are analyses on Asians not run because they undercut some hypotheses? Might they, for instance, have refuted the supposition that structural diversity is necessary to enhance student outcomes, both academic and "democratic"? We cannot know because Professor Gurin failed to look, nor did she mention that she did not look.[56]

B. Other Research with the Same Database Shows No Effect

A major problem for Gurin's hypothesis on the effects of structural diversity is that other research shows that Gurin's major contention, namely that the balance between whites and "persons of color" produces positive learning and democracy outcomes, is largely false. This is part of the findings from the larger dataset from which Gurin's comparative report is taken. Instead of Gurin's single variable (percentage of students of color), the principal investigator of the CIRP data, Alexander Astin, treated each ethnic group separately. He considered the percent of the student body that is black, the percent Hispanic, and the percent Asian, as three separate variables.[57] Astin reported that "with few exceptions, outcomes are generally not affected by these measures of diversity, and in all but one case, the effects are very weak and indirect." He concludes from his own statistical analyses, "[N]one of the three measures produces any direct effects, and practically all the indirect effects are very weak."[58]

Moreover, Astin's statistical analysis of the data finds that the proportion of Hispanics in a school is negatively related to the typical student's likelihood of graduation. That is, the greater the proportion of Hispanics enrolled, the greater is the probability that the typical individual student will drop out.[59]

Astin's findings must be taken as seriously as Gurin's. Those that believe in diversity may find reasons why Astin's analysis is flawed. Nevertheless, it cast strong doubt on Gurin's hypothesis, even before Gurin's project commenced. As we shall see in following sections, and not surprisingly, Gurin's statistical output reveals the same basic pattern—no relationship.

We have discussed previously how Gurin's explanations do not fit her models. In this review of her statistics, we will present a picture of what Gurin really tested with her statistical procedures and what she really found.

We pointed out earlier that proper operationalization of the necessary and sufficient formulation required that the structural diversity measure be multiplied by the classroom/interaction diversity measure in order to provide a correct translation of her verbal theory. The equations failed to test her formulations properly.

It is, however, still possible to draw some conclusions from her analyses. Looking at her results, we find that structural diversity has no significant relationship with any outcome measures. A model that is formulated in terms of necessary and sufficient conditions means that the dependent variable takes on a positive value only when both independent variables take on positive values. If one of them takes on zero value, then the result is a zero value regardless of what value the second independent variable, campus diversity, takes on. Multiplying by zero yields zero and statistically non-significant results in this context mean that the true value cannot be differentiated from zero.

Since structural diversity is not statistically significant in almost all of Gurin's equations, it is highly unlikely that the product of structural diversity and on-campus activity diversity would produce any statistically significant student benefits. Her statistical analysis fails to support her structural and campus interaction formulation. In short, diversity does not matter.

There are problems with the "indirect" effects of structural diversity on student outcomes as well. If Gurin intends a chain model, three preliminary conditions must be met: (1) there is a simple correlation between A and B; (2) there is a simple correlation between B and C; and (3) there is a simple correlation between A and C.[60]

First of all, there is no simple correlation presented between structural diversity and any student outcome, learning or democracy.[61] Second, there are statistically significant correlations between structural and the on-campus diversity variables for white students, but critical school-level variables such as institutional size, type, geography, and endowment were not controlled for in those equations. This means that the relationship between structural and campus diversity may be a function of other institutional variables, most likely geography, endowment, and school reputation.

Even more important is that there is no correlation between structural diversity and on-campus diversity for Hispanics and for blacks. The correlations between discussed racial/ethnic issues and structural diversity were statistically significant but negative among Hispanics. Having no correlation between structural and campus diversity for blacks or Hispanics or both means the model does not work. Structural diversity cannot be an indirect effect if there is no link between the two variables.

Gurin's initial models are not supported by the statistics. What the computer output tells us is that campus diversity, or at least some form of campus diversity, can have an effect—but that it may not require structural diversity to do so.

C. The Real Meaning of Gurin's Statistical Output

To recapitulate, Gurin finds no correlation between structural diversity and student benefits in practically all of her regression equations. Her statistical output, however, shows that campus diversity, or some aspects of it, is correlated with learning and democracy outcomes at least some of the time. But the campus diversity variables are only weakly correlated with structural diversity among whites, and not at all among blacks or Hispanics.

Gurin misreads her statistical output. With all her background controls, the campus diversity variables, in some cases, produce statistically significant effects, but structural diversity generally produces no statistically significant effects. Since her statistical output is calculated with independent variables as additive, not multiplicative (i.e., structural diversity is not necessarily a requirement for the other types of diversity), at the minimum, her statistical output shows that one can have on-campus diversity effects without structural diversity, all other things being equal.

On the (dubious) assumption that her analysis is otherwise valid, and unless it can be shown that structural diversity correlates directly with student outcomes, it appears that colleges and universities do not have to change the ratio of whites and students of color in order to produce the benefits from on-campus diversity. Instead they might increase the number of ethnic studies courses and diversity workshops (the educational wisdom in doing so is another matter).

1. Negative Relationship between Diversity and Some Outcomes for Blacks, but Not for Hispanics

Gurin's computer runs show that 17 of the regression coefficients for structural diversity among black students were negative and statistically significant. The greater the structural diversity, the worse the outcome measures for black students.

This is a very significant problem. The major independent variable should not show nonsignificant associations with the dependent outcomes. It is also a problem when an independent variable shows a negative relationship with the dependent variable, and it is an even bigger problem when the relationship is both negative and statistically significant. Unfortunately, the results of statistically analyzing structural diversity and its impact on black respondents can be determined only if one examines the original computer output, not Professor Gurin's tables.[62]

Gurin fails to mention this in her expert report. Among blacks, there is also little correlation between structural diversity and enrollment in ethnic studies classes, discussion of racial/ethnic issues, socialization with someone from another race, and having friends of a different race or ethnicity. Blacks were statistically significantly less likely to attend workshops, but for the black sample, being on a campus with a faculty diversity emphasis was positively correlated with the percentage of students of color (beta=0.29).

Gurin's findings regarding black students cast doubt on the utility of structural diversity in improving education, since the remaining coefficients for structural diversity among black students were not statistically significant. She does note that interactions among individuals of the *same* race benefit African Americans but that "classroom content on issues of race and ethnicity provides a less novel perspective [than for white students],"[63] so that the lack of correlation between structural diversity and campus diversity, to her, can be explained away for blacks. This is a weak explanation.

We will turn briefly to the negative results for black respondents when statistically analyzing the effects of the campus diversity variables (ethnic studies, diversity workshops, discussing racial issues, socializing with others not of one's race, and having close friends not of one's race). The results are summarized in Gurin's Table D-2, in Appendix D of her report.

As for academics, in almost all instances there was no significant relationship between the campus diversity variables and academic measures for black respondents. There was one equation where, all other things being equal, taking ethnic studies was significantly but negatively correlated with college GPA (beta= –0.16). In other words, if black, taking ethnic studies may lead to a lower GPA in college (or perhaps a lower GPA propels a student to take ethnic studies classes).

Taking ethnic studies along with discussing racial issues was negatively related to the (self-rated) ability to think critically for blacks; again the results were statistically significant. If we use Gurin's statistical rules, then, ethnic studies classes plus talking about racial issues results in less ability to think critically.

Taking ethnic studies and a diversity workshop, discussing racial issues, socializing with non-blacks, and having non-black friends was negatively and significantly correlated with analytical and problem-solving skills for the nine-year group—that is, if black, the more one participated in campus diversity activities, the lower one's analytical and problem solving skills after nine years.

Regarding democracy outcomes, in general there are no significant relationships between ethnic studies and influencing the political structure, influencing social values, etc. (Taking ethnic studies, however, is related to promoting racial understanding and appreciating cultural and racial awareness.) If black, taking ethnic studies has no relationship to acceptance of persons from different races and cultures after four years. Nor does taking a workshop, discussing racial problems, or having friends of different races. The only item related here is socializing with non-blacks. As noted above, if Gurin were interested in understanding educational outcomes among black students, it is puzzling why she dropped the historically black colleges from her sample. In fact, they would provide the basis of a most useful comparison among outcomes for blacks students.

In the larger CIRP analysis done by Astin, he finds that the college experience tends to make blacks more activist than when they enter, and divides the races politically, which is exacerbated by their tendency to segregate themselves.[64] This is consistent with the negative relationship discovered by Gurin as applied to blacks but not whites and Hispanics, and consistent with the predicted effects of racial preferences on campuses as discussed by many critics.

Among Hispanics,[65] there is no correlation between structural diversity and having enrolled in ethnic studies classes, attended racial workshops, socialized with someone from another race, or having made close friends of a different ethnicity. The only correlation that was significant was between structural diversity and discussing racial/ethnic issues (and it was negative), although a faculty diversity emphasis was powerfully correlated with the percentage of students of color (beta =0.48).

Moreover, there are practically no correlations between campus diversity measures and academic outcomes. It does not make students significantly more motivated and engaged in learning, nor improve their general knowledge, writing, learning, and analytical problem-solving abilities. There are a few correlations between ethnic studies and higher GPA as well as a higher assessment of academic ability and foreign language skills.[66]

There are relatively few correlations between campus diversity measures and citizenship measures, after four years and after nine. Ethnic studies, diversity workshops, talking about race, socializing with non-Hispanics, and having close non-Hispanic friends, however, are generally related to promoting racial understanding, cultural awareness and appreciation, and accepting persons of different cultures—not a surprise. Taking an ethnic studies course, however, does not seem to be related to greater acceptance of those from different races/cultures for Hispanic respondents.

To summarize, there are many non-significant findings for blacks and Hispanics, and there are a number of significant negative results. In more than one instance, for black and Hispanic respondents, taking ethnic studies courses is negatively related to academic and democracy outcomes.

2. No Relationship between Diversity and Academic Performance

This section focuses on the regression output describing the relationship between the diversity measures and academic outcomes: college GPAs and highest degree earned. It is based on copies of the actual output in Professor Gurin's step-wise regression procedure. She performs the regression analyses on white, black, and Hispanic respondents separately. Not surprisingly, her statistical output shows that the two most important variables predicting college GPA are SAT scores and high school grades. This is consistent with Astin's statistical results.[67] This is also consistent with the SAT validity studies routinely released by the Educational Testing Service (ETS). Since Gurin divided her sample by race, it is useful to discuss her results in the same manner.[68]

College GPAs. For blacks, high school GPA (HSGPA) and the SAT composite score (SATCOMP) predict college GPA quite well. Not only are they statistically significant in every case, but they are also by far the strongest effects in all the regressions as measured by the betas or standardized regression coefficients (0.38, 0.28).

For black respondents, structural diversity is never significantly correlated with college GPA. Moreover, the direction, when it comes close to reaching standard levels of statistical significance, is uniformly negative (but small with betas of 0.13, 0.16). Similarly, campus diversity is not significantly associated with college GPA. The only exception is the equation that included taking ethnic studies classes and discussing racial issues along with the outcome measures. Taking ethnic studies classes is statistically significantly correlated with college GPA but the relationship is in the wrong direction while the latter is not statistically significant.

For Hispanics, an individual's SATCOMP and HSGPA (with betas of 0.35 and 0.28, respectively) are the most important predictors of his or her college GPA. Structural diversity is consistently negative but not statistically significant, with the exception that for one of the equations it is negative and statistically significant. Two of the four equations containing the variable instances of taking an ethnic studies course find the variable's coefficients to be statistically significant predictors (although quite weak) of college GPA and none of the other on-campus diversity variables are statistically significant. It is likely that, if all four variables were included in the equation, none of the variables would be statistically significant.

For whites, the variables SATCOMP and HSGPA are very strongly related to college grades (the betas are about 0.37 and 0.29). Once again, structural diversity is never a statistically significant predictor of college grade point average. Similarly, of the five campus diversity measures—taking ethnic studies, attending a workshop, discussing racial/ethnic issues with those of other groups, socializing with other groups, and having friends not of one's race/ethnicity—discussing issues is the only statistically significant campus diversity variable related to college GPA.[69]

Highest Degree Earned. A second academic outcome is highest degree earned. Gurin found structural effects are not positively associated with highest degree earned, for members of all three groups. Additionally, there are no equations where college GPA and highest degree earned are simultaneously entered, since these are likely to be strongly related. There were no statistically significant findings regarding campus diversity measures. Astin found the expected strong effects of high school grades and college SAT scores.[70] In addition, student undergraduate GPA is the strongest predictor as to the highest degree earned by respondents. Once again, no structural or on-campus diversity measures are mentioned as having any effect whatsoever.

The most striking finding is that the greater the racial/ethnic diversity of a college or university, the less likely that black students were to graduate.[71] These findings were statistically significant and fairly strong (beta= –0.25, –0.27, –0.26, –0.25). This may be one of the consequences of preferential admission policies that promote diversity but often leave its intended beneficiaries floundering. None of the on-campus diversity variables are statistically significant except socializing with those of other groups, which may be a chance artifact.

For Hispanics, none of the usual academic variables is statistically significant and the effects of structural diversity are negative but not statistically significant. In three of the four times it appeared, taking an ethnic studies class for Hispanic respondents is statistically significant, but since no equation is presented where all the on-campus diversity variables are statistically significant, and none of these equations include college GPA in their analyses, these results are highly limited in scope.

For whites, high school GPA is strongly related to finishing the degree, but structural diversity is a negative and statistically significant predictor of finishing the degree. The greater the structural diversity, the less likely whites are to finish college. This is hardly the positive effect that Gurin claims to be finding. Among the campus diversity variables, “workshop,” “discussion,” “socialize,” and “friends” are statistically significant, but taking an ethnic studies course is not statistically significant. Once again, these results do not utilize all indicators of structural diversity in a single equation and do not take into account the results of college GPA in predicting degree completion.

3. Campus Diversity Measures As Statistical Proxies for Political Liberalism

For whites, Gurin finds that these campus diversity measures are generally related to several democracy outcomes—citizenship engagement and racial/cultural engagement. This is the case for both the four-year and nine-year respondents. The significance of these results, however, is questionable in light of the relationship between these measures and student political ideology.

We believe that these campus diversity measures and democracy outcome questions are proxies for political liberalism, a concept that Gurin ignores but is discussed extensively by Alexander Astin. He shows that these variables are all in fact closely related to each other.[72] Attending a racial or cultural awareness workshop is strongly related to promoting racial understanding, participating in campus demonstrations, self-reported cultural awareness, social activism, not believing that racial discrimination is no longer a problem in America, and cleaning up the environment, and all are highly correlated with each other and with political liberalism. Of course, the fact that cleaning up the environment is correlated with campus diversity and other such measures suggests that these correlations reflect the underlying factor of political liberalism and nothing else.

There are other interesting findings turned up by Astin that Gurin also ignores and that cast some additional doubts as to the benefits of campus diversity. Astin finds a correlation between participating in campus demonstrations and various measures of political liberalism and social activism; for example, the standardized regression coefficient between participating in a racial/cultural awareness workshop and participating in a campus protest is 0.21.[73] Other variables similarly correlated include taking ethnic studies courses and women's studies courses. These latter variables, moreover, are negatively related to majoring in business and majoring in engineering.[74] Additionally, the variables "discussed ethnic issues" and "socialized with someone from another ethnic group" in Astin's statistical analysis are correlated with a propensity to protest.[75]

Campus protests are strongly related to the usual diversity variables.[76] Social activism, in turn, is positively related to what he considers to be diversity measures (taking an ethnic studies course, participating in a diversity workshop, socializing with persons of another group, not having friends who are of the same race/ethnicity, and participating in a campus demonstration),[77] but is a measure Gurin fails to include in her subset of CIRP data.

In statistical analyses of white respondents, Gurin generally found the same patterns. Taking an ethnic studies class or a diversity workshop, socializing with persons of color, and having close friends of color are positively and significantly related to a desire to influence the political structure, influence social values, help others in difficulty, clean up the environment, and participate in a community action program.[78] After nine years, they correlate with doing volunteer work; doing community service to work with people different from the respondent, to improve society, to improve the community, to fulfill social responsibility, to influence the political structure, to influence social values, and to help others in difficulty (a redundancy with community service items above—ergo, their correlation is to be expected); cleaning up the environment; and participating in community action programs. The relationship between campus diversity measures and cleaning up the environment is another bit of evidence that these campus diversity items are related to political liberalism. The positive relationship between campus diversity measures and cleaning up the environment is a strange relationship conceptually, except that these items are all surrogates for political liberalism.

Ethnic studies classes, diversity workshops, discussions on race, socializing with persons of color, and having friends who are not white are also all significantly correlated (positively) to promoting racial understanding, having greater cultural awareness of other races and ethnicities, and accepting persons of different racial and ethnic groups, for the four-year survey respondents and nine-year survey respondents.[79] Since these last measures seem to be alternative ways to measure socializing with non-whites and having non-white friends, it is not surprising that they correlate.

In sum, Gurin's campus diversity measure and democracy outcome items probably measure political liberalism for whites. One would expect them to correlate and as a result they are of questionable educational significance.

D. Fewer Significant Results Than Professor Gurin Concludes

Since Professor Gurin defines her mission as finding statistically significant results, her study stands or falls with how many statistically significant findings she can generate. Her method, however, ignores what statisticians call Type I error. If Gurin properly corrected in her many equations, most statistically significant findings that currently appear would likely be nonsignificant. Why is this the case?

1. Reducing the Likelihood of Chance Findings: An Overview

The goal of statistical testing is to rule out findings that are likely due to chance. The essential logic of these tests can be found in any applied statistics textbook.[80] In brief, the tests involve three steps: (1) formulating the “alternative hypothesis,” (2) formulating the “null hypothesis,” and (3) testing the hypotheses.

The Research Hypothesis. The proper way to use statistics in the analysis of data is to formulate a research hypothesis *before* collecting and analyzing data. This was discussed earlier. The statistical testing literature refers to this as the alternative hypothesis. The alternative hypothesis can take many different forms depending upon the statistical model tested.[81] In all cases, however, the alternative hypothesis must take the form of an *affirmative* research hypothesis. That is to say, the hypothesis must posit that some relationship among variables *does* exist.

The Null Hypothesis. The researcher should then formulate the appropriate null hypothesis. The precise form of the null hypothesis also depends on the statistical test used.[82] Statistically, the null hypothesis is the actual hypothesis tested on the data—what is the likelihood that the results are based on chance?

Testing the Hypotheses. Once the data are collected and the test hypotheses (affirmative versus null) are formulated, the actual process of testing is as follows. The investigators select a level of statistical significance. This is usually 0.05 or 0.01 in the typical research situation where the investigators, to be statistically accurate, seek to *reject* the null hypothesis. Let us assume the significance level is the conventional 0.05. Under the null hypothesis, the investigator (or, more typically, statistical software on a computer) calculates a test statistic and computes an accompanying *p*-value for each variable in the statistical equation. If the *p*-value is less than or equal to the preset level of significance (e.g., $p \leq 0.05$), the null hypothesis is rejected, and the results are said to be statistically significant. If it is not, then the investigator fails to reject the null hypothesis.

It must be stressed here that the significance level of 0.05 is a probability level, or what statisticians call an error rate. The 0.05 level of statistical significance means that 5 per cent of the results may be due to chance even when the “true” effect is in fact zero. (Alternatively put, the error rate is 5 per cent.) If we generate 20 estimates with 20 corresponding *p*-values, 1 out of these 20 (5 per cent of the results) will be statistically significant due to chance alone. We just do not know which of the 20 results is a function of chance and which are true results. We can be more stringent, by setting our significance level at 0.01, whereby 1 per cent of the results are due to chance, or at 0.001, whereby 0.1 per cent (1 out of 1000) are due to chance. We can never be 100 per cent sure that our results are not due to chance. We can be more certain of not making this kind of error if the significance level is set at .01 compared to .05, and 0.05 compared to 0.20, whereby 20 per cent of the results generated (and we don't know which ones) are due to chance.

Gurin, however, in her analysis of black and Hispanic respondents, sets the significance level at 0.10. Doing so means more significant results by definition. Unfortunately, we have also increased the chances to one in ten where we reject the null hypothesis based on our calculations even if the null hypothesis is true. That is, we increase the chances that we really have no relationship.

Critical to setting levels of statistical significance is the problem of false positives and false negatives, or what statisticians call Type I and Type II errors. The false positive is what statisticians refer to as Type I error. It applies to results found to be statistically significant. That is, the test yields a positive result that is in reality false. The investigator rejects the null hypothesis *despite* the null hypothesis being true. The higher the investigator sets the level of significance, the greater will be the probability of committing a Type I error, so normally it is set to be quite low.[83]

The problem with Gurin's statistics frequently occurs when running equations with so many variables. Wood and Sherman's count of Gurin's statistical equations concluded the following: Only 50 of the 276 statistical equations that Gurin estimated show any statistical significance at all for the structural diversity variable and of these 18 are negative, not positive. Of the 32 of 276 regression equations that find the regression coefficient for structural diversity to be positive, none is positive for the regressions estimated for black students.[84]

In total, only 11 percent of all the equations Gurin estimated find any positive statistically significant effects of structural diversity. Five percent are bound to be statistically significant for white respondents (since she sets the significance level at 0.05) strictly as a function of chance. Ten percent are wrongly categorized due to chance for blacks and Hispanics. One must correct for this problem by using special statistical techniques sometimes referred to as adjusting for multiple comparisons.[85] When correcting for this error, the number of statistical significant relationships will certainly be reduced as it will be for any result to achieve the standard level of statistical significance.

Findings fewer statistically significant relationships, however, is not as bad as finding some that are statistically significant but in the wrong direction. This casts significant doubt on the investigator's thesis that structural diversity benefits all students. A policy that only benefits white students, if it even does this, is inadequate and wrong.

2. Significant, but Trivially Weak, Positive Effects

Statistically significant effects can be so trivial, especially as samples get larger, that they have no policy relevance whatsoever. Moreover, variables with such small effects may merely reflect the presence of unmeasured but competing extraneous variables. Gurin's analysis suffers from both problems.

We will start with an analysis of small, medium, and large effect sizes and apply the standards to Gurin's analysis. Mere statistical significance does not indicate substantive significance. Trivial effects can be statistically significant if the sample used for the test is large enough, since the statistical significance of a variable is a function of the pre-set significance level (e.g., 0.05), pre-set statistical power (conventionally 0.80), sample size, and the size of the effect for which one tests.

Effect sizes are usually defined as follows. A 0.10 correlation is considered the minimum cut-off for a small effect size. Less than 0.10 is trivial. A 0.30 correlation is considered to be the cut-off for a medium effect size, and 0.50 is used as the cut-off for large effects.[86]

Gurin, however, seeks to emphasize statistical significance wherever she can, regardless of its justification.

The problem is that, with her large sample, statistical significance is itself easy to achieve and is of little use in delimiting scientific importance. With her total sample of 9316 cases, a correlation as low as 0.03 will be statistically significant, but smaller than Cohen's small effect size. With a sample size of 7542, our estimate sample size for the number of whites in her sample, a correlation of 0.04 is statistically significant. With a sample of 466, an estimate of the number of blacks in the sample, a correlation of 0.13 is statistically significant at the 0.05 level and correlation of 0.07 is statistically significant at the $p \leq 0.10$ level, the significance level she prefers to use for blacks. (A critic should not have to estimate these figures, since the racial/ethnic composition of the sample should be reported as a matter of course.)

Besides reporting trivial effects, Gurin's approach presents additional statistical problems: an unconventional and inaccurate method of entering her campus diversity variables, not conducting initial (and simpler) correlations of all diversity variables with each other, and, as a result of not performing the prior two steps, and as already pointed out, not investigating whether her campus diversity variables are essentially the result of their correlations with the larger master variable, political liberalism.

3. Incorrect Method of Entering the Independent Variables

Additionally, Professor Gurin's strategy of not entering all four on-campus diversity variables into her regressions at the same time is odd and wrong. This is not standard statistical practice and Gurin fails to explain it or cite any sources that describe the validity of such procedures. Either all these variables should be entered at once, or an index of these variables should be created under the rubric of on-campus diversity. Upon creating this index based on the campus diversity variables, it should then be used in all of her equations. Given the way she computed her regression equations, one cannot know what the effect was of each of these variables controlling for all the other variables. This is a serious failing, which largely vitiates even her findings among whites.

4. Failure to Perform Any Preliminary Factor Analysis

As discussed previously, Gurin neglects to correlate statistically all campus diversity variables with each other. Such would be a normal procedure in any preliminary investigation before engaging in the more complex regression analyses with many independent, dependent, and extraneous variables. A preliminary correlational analysis would have allowed her to answer the question, Are these items related? Astin's prior statistical analysis of the CIRP database shows that the five measures of campus diversity—taking an ethnic studies course, attending a diversity workshop, socializing with persons of a different group, discussing racial issues, and having close friends who are not the same race as the respondent —are indeed closely related to each other.

5. How the Effects of Structural Diversity Should Have Been Studied

Professor Gurin fails to acknowledge the general problem posed for statistical analysis in having multiple levels of analysis in theory, models, and data. One can statistically utilize data that contain multiple levels of analysis, but OLS regression (the standard type of multiple regression model) is inadequate for this purpose. [87] The kinds of models that take the differing levels of analysis into account in a proper statistical manner are hierarchical linear models.[88] This technique is widely used in educational research.[89] These are models where structural or aggregate characteristics of institutions and the individual characteristics of individual students are thought to affect individual performance both within the institution and subsequently. This is the kind of statistical model, for example, that is used when attempting to ascertain whether Catholic schools are better than public schools in producing K-12 educational achievement.

Many of the problems discussed in earlier parts of our critique stem from the units of analysis problem. Structural diversity is an institutional characteristic, not an individual student characteristic, and the data analysis itself ought to be organized around this fact. That means that Gurin ought to present aggregate results of her analysis for her sample of 184 colleges and universities first before she analyzes individual effects. This was not done.

Considering the multiple levels of analysis situation model explicitly shows how problematic these analyses can be. This is because individuals within the unit are not independent of each other when it comes to structural characteristics. Consider two colleges, X and Y. Each student at X is characterized by the same degree of structural diversity as his or her fellow classmates. At the same time, each student at Y is in a similar situation. The degree of structural diversity varies only from school to school. To estimate a relationship between the structural diversity of colleges and individual student characteristics requires taking this restriction of variation into account. There is no within-group variance on the structural characteristic, only between-group variance.

Put more formally in the language of regression analysis, both the observations and the errors are not independent within institutions. This produces bias. In this situation, “standard regression analysis,” which is exactly what Gurin did in her expert report, “is inappropriate.”[90]

The procedure suggested above would allow the proper statistical testing of an aggregate analysis of the relationship between structural diversity and the on-campus diversity measures (for example, do schools with a high degree of diversity exhibit a larger proportion of students taking ethnic studies courses?) and an aggregate analysis of the relationship between the outcome measures and structural diversity. This might be followed by 184 separate regression analyses of individual characteristics to ensure that these results are not biased by school effects. Then, using the hierarchical linear modeling technique, it is possible to combine these results into a single unified model.

It should be noted that this approach dramatically reduces the effective sample size for statistical testing from about 9000 (the number of students in Gurin’s sample) to

184 (the number of colleges in Gurin's sample). This would, of course, result in fewer statistically significant results. This shows that Gurin's regression analyses produced an inflated level of statistical significance because, in these situations, the risk of Type I error increases.[91]

Professor Gurin's study displays, unfortunately, an unreflective application of statistical techniques. While she states that she has found what she was looking for because her findings are statistically significant, we have shown in this section that this assertion is simply wrong. Statistical significance is a question of whether the findings are due to chance. They do not tell you if what you say you intend to find and what you find are in fact the same thing. Nor do they tell you how important (that is, substantively significant) the findings are. Gurin dredges through her output until she finds something she believes she can report, and ignores Type I error problems. Data-dredging to confirm prior theory is not an accepted statistical practice.

Finally, these tests and conclusions are presented as if they were performed on random samples of schools and students but, as shown in the previous section, this is not the case and to draw inferences beyond the "chunk" of schools and respondents is wrong and misleading.

The substantive importance of a policy is measured by other factors, not statistical significance. What we have shown is how Gurin's effects, assuming they were in fact statistically significant, are miniscule. Gurin and others, however, are content to argue that these miniscule effects between some types of campus diversity and some types of student outcomes, based on a volunteer non-random sample of schools and students, of dubious statistical significance, should justify continuing policies of racial and ethnic preferences in undergraduate admission. Statistically, however, Professor Gurin cannot make this case.

VI. Theoretical Literature on Group Conflict

While Professor Gurin quotes the phrase "equal status contact," citing among others psychologist Gordon Allport, she transforms the meaning of this phrase to something other than what it should mean. In fact, if one applies the notion of equal status contact properly, achieving racial diversity by means of racial and ethnic preferences will not only fail to foster intergroup cooperation but will enhance mutual suspicion and hostility between racial and ethnic groups.

A. The Power of Group Identification, Membership, and Ties

The sociological and psychological literature on the importance of groups is very important, basic, and noncontroversial. These fundamental concepts in social science have been developed by a wide variety of social scientists, beginning with William Graham Sumner and including Muzafer Sherif, Gordon Allport, Robert K. Merton, Peter I. Rose, Robin Williams, William Beer, J.M. Yinger, Robert LeVine and Donald Campbell, Henri Tajfel, and, more recently, David M. Messick and Diane M. Mackie, Judith Rich Harris, Byron Roth, and Paul Sniderman and Thomas Piazza.[92] These authors make much of the easily invoked distinction between the in-group and the out-group as the basic unit of social life.[93]

Group identifications are ubiquitous, easily established, easily maintained, and difficult to change, and are a fundamental part of any society. These are linked to liking for one's own group and usually to disliking for other groups. In-groups and out-groups are thus characterized by a degree of group preference usually referred to as ethnocentrism.[94] As sociologists and social psychologists have long known, similarity leads to liking while difference and strangeness leads to disliking.

While the study of ethnocentrism, in-groups and out-groups, and prejudice and discrimination has often focused on "real" preexisting (racial and ethnic) groups, a well-developed experimental research tradition has evolved in social psychology of highlighting the minimal group phenomenon or "groupness" among completely artificial groups. Psychologist Henri Tajfel has shown that the mere act of categorization, whether or not there is a name for the group, and whether or not individuals know who other members of the group are, is by itself sufficient to create in-group favoritism and out-group lack of favoritism. Groups of boys were asked individually (and apart from one another) to count the number of dots in an experimental setting, After this was done, the boys were randomly assigned into two subgroups—the overestimators and the underestimators—and told which group they belonged to. Then they were asked to determine the amount of reward that each participant in the experiment was to receive as characterized by their identification number and which group they belonged to. Individuals favored members of "their" group over members of the other group, even though this had no effect on their own personal reward situation.[95]

Groupness requires self-categorization, which also requires recognition of basic similarities or common fate among putative members but little else. [96] This effect is remarkably robust. It has been replicated again and again in many different experimental settings and is a widely accepted phenomenon.[97]

These in-group and out-group identifications, these "we groups" and "they groups," these insiders and outsiders, contain a belief component. Accentuated or even created de novo simply by the act of categorization, this can exaggerate the differences between groups. The exaggeration of differences between groups is the stereotype. The in-group and out-group categorization produces the tendency to see two juxtaposed categories as more different than they really are, or what psychologists call "group contrast effects."[98]

Categorization leads to both group contrast effects and assimilation within groups. In other words, a perceptual phenomenon of group contrast effects is created when in-group and out-groups are created, but what is also created is in-group conformity. In-group conformity further increases the tendency to stereotype.[99]

As Harris points out (citing a review article in the *Annual Review of Psychology* by Hilton and von Hipple), stereotypes as generalizations are similar to other generalizations. In-group stereotypes are simply favorable generalizations; out-group stereotypes are unfavorable generalizations.[100]

These in-group and out-group categorizations, feelings of liking/disliking, and beliefs about out-groups and in-groups are associated with an action component. Hostile actions include racial discrimination, either unilaterally or in a reciprocal and mutual manner.

B. The Importance of Equal Status Contact in Reducing Prejudice

The conditions under which in-groups and out-groups are hostile have been well studied. Probably the most famous of these are the Robbers Cave experiments, classic social-psychological studies conducted by Muzafer Sherif and his colleagues.[101] These authors show how easy it is to take a group of boys and give them a separate identity and liking for their fellow group members. It turned out to be easy to bring forth another group of boys, with no initial differences between the two groups, and to set them competing with each other as teams. The resulting series of competitions produced intense mutual dislike by members of each group of boys for the other group. The dislike was acted out also. The competitive behavior went from name calling to fighting so that the researchers had difficulty defusing the conflicts that they had set into motion, a result that was only achieved with some considerable difficulty.

Defusing the conflict required more than mere contact. Sherif and his collaborators tried seven different interaction situations without making any other changes. These failed to bring about a positive result. Members of the two groups continued to be antagonistic towards each other. They then contrived a number of tasks that, in order to be accomplished, required the cooperation of both groups. These superordinate goals, as Sherif called them, were required in order to reduce out-group antagonism. It was during this phase of the experiment that the boys began to like members of the other group, and the relative preferences for members of each group to favor their own group declined nearly to the vanishing point by the time the experiment had ended.[102]

The literature on prejudice reduction is very clear. Mere interaction between members of different groups is not sufficient to reduce prejudice and hostility, but only interaction under certain conditions.

The most important of these conditions is equal status contact between members of different racial/ethnic groups. Equal status contact is a precondition for establishing friendship between two individuals who are members of different ethnic groups and a cause of reduction of prejudice. The equal status contact hypothesis is critical in explaining how to reduce frictions and hostilities and increase friendliness and amity between members of different groups.

C. Empirical Research on Affirmative Action

1. General Studies

There appears to have been relatively little research done on affirmative action itself. About 15 years ago, sociologist William Beer lamented the lack of research on affirmative action. But there are some studies that bear directly on the subject.[103]

Psychologist Stephen D. Johnson found that policies of reverse discrimination in favor of blacks against whites increased hostility towards blacks. In one study, white subjects were told they had lost a (staged) competition to a black competitor. These subjects were more aggressive towards him than they were towards their white counterparts, leading the researcher to conclude "reverse discrimination leads towards more prejudice towards black people" than when they lost arbitrarily to a poor white. But the study also found that, when the subjects were told that the black competitor won because of his superior performance, white feelings of hostility actually decreased. In his second study, which repeated the experimental manipulation of the first, Johnson explored some of the reasons for this difference. He found that reverse discrimination was seen as unjust because it violated the principle of equal treatment, the sense of equity, and the belief that his need was equal to that of his black competitor. Unlike in the previous experiment, however, the respondents were less likely to express hostility toward their black competitors. Johnson suggests the reason for the difference with the earlier experiment is that the individuals in question had had face-to-face dealings with the black competitor (actually the experimenter's confederate) and felt they were unable to express their hostility and preserve their appearance as a good and proper non-prejudiced person.[104]

Research in a different discipline provides support for Johnson's findings. Political scientists Sniderman and Piazza devised a very different kind of experiment, which they call the "mere mention" experiment. This experiment demonstrated that merely mentioning the subject of affirmative action increases hostility towards blacks. Unlike Johnson's studies, these experiments were part of a field survey. A field survey of whites was divided randomly into two groups. Members of the first group were asked their view of affirmative action and then their images of blacks. The other half-sample was asked exactly the same questions except in reverse order. Since these are random half samples, any difference between the two groups would be due to the question ordering. Sniderman and Piazza found that 43 percent of whites that were asked about affirmative action first described blacks as irresponsible, as compared with 26 percent who were asked their opinion about blacks before the subject of affirmative action had been raised. The authors conclude that while affirmative action did not create the problem of white racial prejudice, it can and does aggravate it.[105]

The salience of group boundaries themselves is itself of direct relevance to the question of the impact of racial preferences on attitudes. Brewer and Miller propose that intergroup contact will be successful in improving intergroup relations when group and category memberships are as inconspicuous as possible, there is differentiation between out-group members, and the interaction is intimate.[106] In their extensive review of the minimal group literature, psychologists Messick and Mackie make the point that, in order to change the out-group stereotype held by members of the in-group, group boundaries needed to be weakened (such as with cross-cutting ties that create multiple group memberships) and the instrumental importance of group membership needs to be reduced—both of which in turn will reduce the salience of in-group identification. Lastly, they find that suppressing intergroup categorization impedes generalization—if individuals are not classified into certain groups, this reduces the tendency of people to stereotype by these groups.[107] In reacting to their point, psychologist Byron Roth contends that, to the extent group boundaries are strengthened and are made more salient, this will increase hostility towards out-groups.[108]

2. Equal Status Contact versus Heightened Racial and Ethnic Group Salience

In the case of college students from different ethnic groups, admission by means of racial and ethnic preference does not constitute the preconditions for equal status contact. Students are students but, because they are not selected with the same criteria, they are unequal academically. Table 2 displays the SAT scores for blacks, Hispanics, Asians, and whites who were admitted in 1995.

Table 2
1995 Admittee SAT Scores at the University of Michigan, Ann Arbor

	Verbal SAT Scores			*Math SAT Scores*		
	25th Percentile	50th Percentile	75th Percentile	25th Percentile	50th Percentile	75th Percentile
Blacks	430	480	550	470	540	622
Hispanics	460	520	590	530	600	670
Asians	530	590	640	660	710	740
Whites	530	580	640	620	670	720

The University of Michigan admitted blacks and Hispanics with significantly lower SAT scores than whites and Asians. This is true for both verbal and math scores.

White and Asian verbal SAT scores are higher at the 25th, 50th, and 75th percentiles compared with blacks and Hispanics. The white-black gap in median SAT scores is 100 points. The white-Hispanic gap is 60 points. The Asian-black and Asian-Hispanic gap in median verbal scores is 110 points and 70 points, respectively.

White and Asian verbal SAT scores at the *25th percentile* are higher than the *median* verbal SAT scores for blacks and Hispanics. Thus, 75 percent of whites and Asians admitted to the University of Michigan at Ann Arbor had higher verbal SAT scores compared to the average black and Hispanic admittee.

Like their verbal SAT scores, white and Asian math scores are higher at the 25th, 50th, and 75th percentiles compared with blacks and Hispanics. The median white math SAT score is 130 points higher than the median black math score, and 70 points higher than the median Hispanic score. The Asian-black gap in median scores is 170 points, while the Asian-Hispanic gap is 110 points.

Black scores at the *75th percentile* are roughly the same as Asian and white scores at the *25th percentile*. This means that approximately 75 percent of blacks admitted to the

University of Michigan at Ann Arbor have lower math SAT scores than roughly 75 percent of Asian and white admittees.

This discrepancy is noticed by students when they associate classroom performance with racial/ethnic identification. And some white and Asian students will also remember their friends back home who were not admitted despite having better credentials than many of their black and Hispanic classmates.

3. Group Stereotyping and Preferences

The effect of increasing diversity by means of preferential admissions, therefore, is likely to be the exact opposite of what the diversity advocates hope for. Even worse, when administrators heighten group formation, by insisting on racial/cultural awareness workshops and similar measures, group boundaries and group identification are heightened.

When instrumental benefits (e.g., admission status, scholarships, etc.) are distributed based on group membership, group membership is given even greater importance. Preferential admission and a gap in academic credentials operationally mean that blacks and Hispanics obtain rewards denied to their white and Asian counterparts of equal credentials. As a result, dissimilarity, not similarity among groups, is strengthened, and cultural differences among the groups are reinforced. Racial and ethnic preferences in admission and the emphasis on ethnic studies courses, workshops, discussions, etc., may increase, not decrease, the salience of “we versus they.”

Finally, these predictable effects of racial and ethnic preference policies have led to the growth of speech codes, political correctness, and other attempts to paper over the heightened “groupness” and perceptions of unequal treatment by some and white-male intolerance by others. Institutionalizing a policy of racial and ethnic preferences in the guise of diversity is likely to lead to permanent antagonism and a permanent “P.C. police.” The academy moves farther away from the center of intellectual debate and free, open inquiry it was meant to be.

VII. Conclusion

There are no statistically significant relationships between a school’s racial/ethnic diversity and any outcome measures. The survey of universities and colleges is a non-random sample from which one cannot scientifically generalize. There are very low response rates, critical extraneous variables are not controlled for, and the statistical effects are in general very weak. For these and the other reasons we have documented, we conclude that Dr. Gurin has not shown statistically that racial/ethnic diversity at a school yields educational benefits.

Endnotes

[1] The report is on the University of Michigan's web site at www.umich.edu /~urel/admissions /legal/expert /gurintoc.html.

[2] For example, one can have a perfectly selected sample but concepts that are so badly defined and poorly measured that one is unable to conclude anything from the results of the study.

[3] Earl Babbie, *Survey Research Methods* (Belmont, CA: Wadsworth, 1990), divides survey research methods into three basic units: survey research design (which includes types of study design, sampling, and levels of measurement), data collection (including questionnaire construction and administration as well as data processing, pretesting, and pilot studies), and statistical analysis. See also Fred N. Kerlinger, *Foundations of Behavioral Research* (New York: Holt, Rinehart, Winston, 1973); Chava Frankfort-Nachmias and David Nachmias, *Research Methods in the Social Sciences* (New York: St. Martin's Press, 1996).

[4] Cited and discussed in David Popenoe, *Life Without Father: Compelling New Evidence That Fatherhood and Marriage Are Indispensable for the Good of Children and Society* (Cambridge, MA: Harvard University Press, 1998), pp. 59-61. See also Sarah MacLanahan and Gary D. Sandefur, *Growing Up With a Single Parent* (Cambridge, MA: Harvard University Press, 1994), pp. 13-14.

[5] Quoted in Barbara Dafoe Whitehead, *The Divorce Culture: Rethinking Our Commitments to Marriage and Family* (New York: Vintage Books, 1998), p. 51.

[6] For detailed discussion of the extensive research literature, see the following works: Linda Waite, "Does Marriage Matter?," *Demography* (32, 4, 1995), pp. 483-507; MacLanahan and Sandefur, 1994; Popenoe, 1998; Whitehead, 1996.

[7] For a discussion of research design, see Morris Rosenberg, *The Logic of Survey Analysis* (New York: Basic, 1968); Floyd J. Fowler, Jr., *Survey Research Methods* (Beverley Hills, CA: Sage, 1984); Peter Rossi and Howard E. Freeman, *Evaluation: A Systematic Approach,* 5th ed. (Newbury Park, CA: Sage, 1994); and Nachmias and Nachmias, especially pp. 52-73.

[8] In "B. Theoretical Foundations for the Effect of Diversity," Gurin presents the theoretical background from which she derives her abstract constructs and research hypothesis. See Gurin, "C. Conceptual Model of the Impact of Diversity," for her basic hypotheses.

[9] Gurin does not present data showing exactly how she “defines” segregation.

[10] “C. Conceptual Model of the Impact of Diversity,” second paragraph.

[11] “Theoretical Foundations for the Effects of Diversity.” See also Gurin’s Appendix E.

[12] See Robert Lerner, Althea K. Nagai, and Stanley Rothman, *American Elites* (New Haven: Yale University Press, 1996) for a discussion of this point in more detail in the context of contemporary American ideology.

[13] “Theoretical Foundations.”

[14] From Gurin, Figure 1 (www.umich.edu./~urel/admissions /legal/expert /gurintoc.html.).

[15] See Hubert Blalock, *Theory Construction: From Verbal to Mathematical Formations* (Englewood Cliffs, NJ: Prentice-Hall, 1969), generally, but especially pp. 153-165.

[16] There is no information available on these latter measures.

[17] We have classified Gurin’s items as best we could based in part on her description of them in Appendix C and a copy of her survey instrument.

[18] We will not discuss pair and group matching, which emerged out of the experimental tradition. Methods of matching are used when populations are rare, costly, and hard to reach through usual study methods.

[19] Statistical analysis of more than two variables usually means multiple regression or its close relative, the analysis of variance as statistical tools.

[20] They are not, however, the most important controls. We will discuss these later when we go over what extraneous variables should or should not be included in this kind of study.

[21] Appendix C, p. 5.

[22] Sometimes these are called the “experimental” or “treatment” group and the “control” group.

[23] These are sometimes called confounding variables.

[24] In this case, the two groups are equated by randomly assigning individuals to the experimental group or the control group. This is the best available research design, because any statistically significant differences between the two groups can be plausibly attributed to the experimental manipulation.

[25] Stated slightly differently, controlling for extraneous variables reduces the probability that the relationship between the independent variable and the dependent variable is spurious. Spurious relationships are those that are not true causal relationships, but are due to the presence of a third variable. The existence of spurious relationships, caused by extraneous variables, accounts for the substantial degree of truth in the old adage "correlation is not causation."

[26] Alexander W. Astin, *What Matters in College: Four Critical Years Revisited* (San Francisco: Jossey Bass, 1993), pp. 379-380, 386.

[27] This problem is discussed by Sniderman and Piazza, who show that there are three different racial issues: the equal treatment issue, the social welfare issue, and the race-conscious issue (p. 9).

[28] Peter Rossi and Freeman, p. 230; Nachmias and Nachmias, pp. 170-171.

[29] Delbert C. Miller, *Handbook of Research Design and Social Measurement,* 5th ed. (Newbury Park: Sage, 1991), p. 580.

[30] Pp. 170-175.

[31] On elite occupations and ideological dissensus, see Robert Lerner, Althea K. Nagai, and Stanley Rothman, "Elite Dissensus and Its Origins," *Journal of Political and Military Sociology* (18, 1, Summer 1990), pp. 25-40, especially pp. 29-32. See Gene Lutz, *Understanding Social Statistics* (New York: Macmillan, 1983), for examples of how an index of dispersion is used**.**

[32] Eric L. Dey, Alexander W. Astin, and William Korn, *The American Freshman: Twenty-Five Year Trends, 1966-1990* (Los Angeles: Higher Education Research Institute, Graduate School of Education, UCLA, September 1991), ERIC, ED 340 325.

[33] The American Association of Medical Colleges until 1997 classified Native Hawaiian applicants as Asian/Pacific Islanders (along with such groups as Japanese, Chinese, Filipinos, Pakistanis, Guamanians, and Samoans, to name a few). After that, Native Hawaiians, but not other Pacific Island groups, were placed with Native Alaskans and American Indians as part of the Native American group. Moving into that latter group made them part of AAMC's recognized Underrepresented Minorities. See Association of American Medical Colleges, *AAMC Data Book: Statistical Information Related to Medical Schools and Teaching Hospitals, January 1999* (Washington, D.C.: AAMC), p. 15.

[34] The method used for the CEO studies relies on the institution's own racial and ethnic classifications, while dropping "missing," "unknown," or "other" from statistical analysis. The Census Bureau uses its own method, as does the National Opinion Research Center at the University of Chicago for the General Social Survey. All methods are documented; Gurin's should be, too.

[35] Robert Lerner and Althea Nagai, *Pervasive Preferences: Racial and Ethnic Discrimination in Undergraduate Admission Across the Nation* (Washington, D.C.: Center for Equal Opportunity, 2001), pp. 42-44.

[36] Lerner, Nagai, and Rothman, *American Elites*.

[37] Lerner and Nagai, *Pervasive Preferences*, pp. 13-31.

[38] This is reflected in the CIRP developers' recommendation that standard errors for their "national norms" not be reported and relied upon to analyze trends in the data. See Eric L. Dey, Alexander W. Astin, and William Korn, *The American Freshman: Twenty-Five Year Trends, 1966-1990* (Los Angeles: Higher Education Research Institute, Graduate School of Education, UCLA, September 1991), pp. 185-186.

[39] Nachmias and Nachmias, pp. 185-195, provide a textbook discussion of types of probability sampling, and examples of drawing a nationwide sample.

[40] Laumann et al. provide an explanation as to how the Kinsey 10-percent figure became accepted as the "right" proportion of homosexuals in the U.S. population in an insightful discussion, "The Myth of 10 Percent and the Kinsey Research." See E.O. Laumann, J.H. Gagnon, R.T. Michael, and S. Michaels, *The Social Organization of Sexuality: Sexual Practices in the United States* (Chicago: The University of Chicago Press, 1994), pp. 287-290.

[41] Dey, Astin, and Korn, *The American Freshman*, p. 185.

[42] Dey, Astin, and Korn, especially Appendix A, pp. 129-138, and Appendix E, p. 185.

[43] Fowler, *Survey Research Methods*, p. 56.

[44] In a probability sample, the sample size does affect the power of the statistical tests used to detect statistical significance (this is discussed in far more detail in subsequent sections on the logic of statistical testing).

[45] Fowler, pp. 55-58, describes the possible types of bias associated with non-probability samples. See also Sharon L. Lohr, *Sampling Design and Analysis* (Pacific Grove, CA: Duxbury Press, 1999), pp. 4-8.

[46] Dey et al., p. 135.

[47] Lohr, p. 5. Since then, the Census has done even better. The net undercount for blacks in 1990 was 4.57 percent, and in 2000 it was estimated to be only 2.17 percent. Cited in Howard Hogan, *Accuracy and Coverage Evaluation: Data and Analysis to Inform the ESCAP Report* (Washington, D.C.: U.S. Census Bureau, March 1, 2001), pp. 13, 15.

[48] Dey, Astin, and Korn, Appendix E, p. 137, Table A-3.

[49] Ibid., Figure A1, p. 133.

[50] Ibid., p. 132.

[51] Astin reports a final sample size of 24,847 students. See Astin, pp. 22-24.

[52] Babbie, *Survey Research Methods*, p. 182; Don A. Dillman, *Mail and Telephone Surveys* (New York: Wiley, 1978), p. 27; National Opinion Research Center, *General Social Survey, 1972-1996 Cumulative Codebook* (National Opinion Research Center, University of Chicago, 1996), pp. 975-976; Laumann, et al., pp. 555-557; Lerner, Nagai and Rothman, *American Elites*, pp. 141-143.

[53] This is based on the total number of institutions in the 1990 normative sample divided by the total number of institutions in the total universe (Dey et al., p. 135).

[54] Lerner and Nagai, *Racial Preferences in Michigan Higher Education* (Washington, D.C.: Center for Equal Opportunity, 1998), p. 28.

[55] Lerner and Nagai, *Pervasive Preferences: Racial and Ethnic Discrimination in Undergraduate Admission Across the Nation* (Washington, D.C.: Center for Equal Opportunity, 2001), pp. 42-44.

[56] Although there is not space to discuss it here, Gurin also omits historically black colleges from her study, which is a serious problem for anyone interested in studying the effects of differing racial/ethnic environments on black student achievement and attitudes. But, unlike the Asian situation, at least she notes this omission, although without really justifying it.

[57] Astin, p. 362.

[58] Ibid.

[59] Ibid.

[60] Rosenberg, *The Logic of Survey Analysis*, pp. 54-83, especially pp. 68-69.

[61] This is where one would use factor analysis and create indices so that the many items listed as outcome measures would be reduced to (ideally) the two outcome variables as laid out in Gurin's model.

[62] Table D-2 in Gurin's Appendix D summarizes the effects of campus diversity variables on her list of academic and democracy outcomes for the four-year and nine-year surveys

[63] Gurin, p. 37.

[64] Astin, p. 407.

[65] Recall from the earlier chapter on measurement that this omits Cuban Americans and South Americans from the category of Hispanic since they were omitted from the intake questionnaire.

[66] Table D-3 in Gurin's Appendix D.

[67] Astin, p. 188. As Astin does not mention diversity variables in his discussion of academic and cognitive development as having any impact whatsoever, there is no reason to assume that diversity, either structural diversity or on-campus diversity, has any effect on academic performance. See Astin, pp. 186-244.

[68] Gurin does not display any correlations between college GPA and students' self-assessment of their academic skills (under which is included all sorts of self-reported areas of knowledge, general knowledge, analytical and problem-solving skills, ability to think critically, writing skills, foreign language skills, as well as self-ratings of abilities compared to the average person of respondent's age, including their academic ability, writing, and listening ability). Unlike many most of the other cognitive and "democracy variables," college GPA and college graduation are relatively "hard" measures of academic performance.

[69] The procedures for entering the campus diversity variables, however, are odd. They are entered as follows: In every equation, taking an ethnic studies course is entered, along with one of the other three measures. At no time are the other three measures and ethnic studies entered in as sole measures of campus diversity, nor are all four entered in simultaneously. It is likely that if all four variables were included in the equation simultaneously, none of the variables would be statistically significant.

[70] Astin, pp. 191-197.

[71] For blacks, high school GPAs are statistically significant (beta= 0.27), but combined SAT scores are not.

[72] Astin, pp. 379-380, 386.

[73] Ibid., p. 380.

[74] Ibid., pp. 370, 372.

[75] Ibid., p. 379.

[76] Ibid., pp. 176-177.

[77] Ibid., pp. 137-138, 151, 160.

[78] Gurin, Table D-1, "Detailed Regression Summary Tables, CIRP database, White students" in Appendix D, Regression Summary Details, under the category, "Four-year democracy outcomes: Citizenship engagement."

[79] Ibid., Table D-1, "Detailed Regression Summary Tables, CIRP database, White students" in Appendix D, Regression Summary Details, under the category, "Four-year democracy outcomes: Racial/cultural engagement."

[80] See, e.g., Alan Agresti and Barbara Findlay, *Statistical Methods for the Social Sciences,* 3rd ed. (San Francisco: Dellen Publishing, 1996); Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed. (Hillsdale, NJ: Erlbaum, 1988); David Freedman, Robert Pisani, and Roger Purves, *Statistics* (New York: WW Norton, 1985).

[81] For example, a t-test statistically compares the difference between the mean of the study group versus the mean of the comparison group. A simple ANOVA involves a test of the hypothesis that the means of all the groups are not equal to zero. Multiple regression involves a test of the hypothesis that all the regression coefficients are not equal to zero or that one or more of them is either positive or negative.

[82] The null hypothesis for a t-test is that the difference between the means of the two groups is equal to zero. The null hypothesis for the ANOVA is that all means are equal to each other. The null hypothesis for correlations is that the correlation coefficient, which measures the linear association between two variables, is equal to zero; for multiple regression, the null hypothesis is that one or more of the regression coefficients is equal to zero. Of course, it is very unlikely that the value of these coefficients will be exactly zero when they are computed using an actual data set. The statistical testing question is whether they are statistically significantly different from zero.

[83] There is also the problem of finding the false negative, or the problem of Type II Error. Type II Error occurs when one fails to reject the null hypothesis when it *should* be rejected and where in fact the null hypothesis (e.g., "no difference") is false. It applies to results when the researcher intends to find no difference. These are major statistical problems in studies where investigators hypothesize from the outset that they expect to find no difference among groups or no effect due to a certain treatment. The logic of statistical inference, statistical power, and sample size is substantially different from issues of the false positive.

[84] Thomas E. Wood and Malcolm J. Sherman, *Is Campus Racial Diversity Correlated with Educational Benefits?* (Princeton, NJ: National Association of Scholars, April 4, 2001), p. 44. See www.calscholars.org /race&highered.html.

[85] There are several possible methods. See "One-Way ANOVA Post Hoc Tests," *SPSS Base 9.0 User's Guide* (Chicago: SPSS, 1999), p. 257, for a list of possible tests, and also "Table 3.3. Factor Scores with Scheffe Confidence Intervals for Four Ideological Dimensions," in Lerner, Nagai, and Rothman, *American Elites*, pp. 52-53.

[86] Cohen, p. 89.

[87] Of course, this is a major problem for Astin's analysis as well and, like Gurin, he does not appear to realize it.

[88] Anthony S. Bryk and Stephen W. Raudenbush, *Hierarchical Linear Models: Advanced Quantitative Techniques in the Social Sciences I* (Newbury Park, CA: Sage Publications, 1992).

[89] Ibid., pp. 60-83.

[90] Ibid., pp. 15, 84.

[91] Ibid., p. 86.

[92] For an overview, see Robert A. LeVine and Donald T. Campbell, *Ethnocentrism: Theories of Conflict, Ethnic Attitudes and Group Behavior* (New York: John Wiley & Sons, 1972); see also Muzafer Sherif and Carolyn W. Sherif, *Groups in Harmony and Tension: An Integration of Studies on Intergroup Relations* (New York: Harper, 1953); Muzafer Sherif, O. J. Henry, B. Jack White, William R. Hood, and Carolyn W. Sherif, *Intergroup Conflict and Cooperation: The Robbers Cave Experiment* (Norman, OK: The University

Book Exchange, 1961); Gordon W. Allport, *The Nature of Prejudice* (New York: Doubleday, 1958); Robert. K. Merton, “The Perspectives of Insiders and Outsiders,” *The Sociology of Science: Theoretical and Empirical Investigations* (Chicago: University of Chicago Press, 1973); Peter I. Rose, *They and We: Racial and Ethnic Relations in the United States* (New York: Random House, 1981); Robin M. Williams, Jr., *Mutual Accommodation: Ethnic Conflict and Cooperation* (Minneapolis: University of Minnesota Press, 1977); Henri Tajfel, “Social Psychology of Intergroup Relations,” *Annual Review of Psychology* (33, 1982), pp. 1-31; David M. Messick and Diane M. Mackie, “Intergroup Relations,” *Annual Review of Psychology* (40, 1989), pp. 45-81; Byron M. Roth, *Prescription for Failure: Race Relations in the Age of Social Science* (New Brunswick, NJ: Transaction Press, 1994); Paul M. Sniderman and Thomas Piazza, *The Scar of Race* (Cambridge, MA: Belknap Press of Harvard University Press, 1993); and Judith Rich Harris, *The Nurture Assumption* (New York: Free Press, 1998).

[93] The group can be a social category, collectivity, or organization. The first is simply a grouping that is either a folk category or an administrative classification. It need not have any subjective reality. The second is a grouping whose members exhibit consciousness of kind and identify with the group. There is, however, no reason to suspect that individuals identifying with a collectivity have personal acquaintance with one another. The third has all the properties of the second and, in addition, has members who identify with each other as such and are capable of joint or collective action. The key point is that the group, however defined, serves as a reference group for right conduct and proper performance.

[94] Some suggest that human nature is by nature social and that ethnocentrism may well have sociobiological roots. See Harris, 1998; Roth, 1994.

[95] Tajfel, 1982; Roth, pp. 199-201; and Harris, pp. 125-133, 136-145. This is what Tajfel calls “groupness” (Harris, p. 166).

[96] Harris, pp. 166-167; Campbell and LeVine, pp. 104-108.

[97] Messick and Mackie, 1989; see also Harris, p. 145.

[98] See Harris, p.132, for a review of the literature.

[99] Campbell and LeVine, pp. 156-175.

[100] Harris, p. 224.

[101] Sherif and Sherif, 1953; Sherif et al., 1961; Sherif and Sherif, 1966.

[102] Sherif et al., 1961. In an earlier study, also based on two groups of boys, Sherif found that creating a common enemy worked quite well in promoting solidarity between members of the two formerly competing groups. The two groups formed a single baseball team, which played and beat a team from a nearby town. See Sherif and Sherif, 1953, pp. 285-286.

[103] William R. Beer, "Resolute Ignorance: Social Science and Affirmative Action," *Society* (24, 4, May/June, 1987), pp. 63-69.

[104] Stephen D. Johnson, "Reverse Discrimination and Aggressive Behavior," *The Journal of Psychology* (104, 1980), pp. 11-19; and "Consequences of Reverse Discrimination," *Psychological Reports* (47, 1980), pp. 1035-1038.

[105] Sniderman and Piazza, pp. 103-104. These authors make the useful point that there are really three separate agendas when talking about race: the equal treatment agenda, the social welfare agenda, and the race-conscious agenda.

[106] Summarized in Messick and Mackie, p. 67.

[107] Messick and Mackie, pp. 66-71.

[108] Roth, p. 205.

About the Authors

Robert Lerner and Althea Nagai are statistical consultants based in Rockville, Maryland, a suburb of Washington, D.C. They are co-authors of three books, several monographs, and numerous papers. Lerner and Nagai both received their Ph.D.'s from the University of Chicago—he in sociology, she in political science. He has a B.A. in economics and sociology from Oberlin College, and she has a B.A. in psychology and political science from the University of Hawaii. Lerner is also an applied statistician for the Congressional Monitoring Board of the U.S. Census.