"New Political History"

Some Statistical Questions Answered

J. MORGAN KOUSSER
California Institute of Technology
ALLAN J. LICHTMAN
The American University

William G. Shade's (1981) "New Political History: Some Statistical Questions Raised" has two sometimes conflicting purposes: first, to remind historians to "think statistically" and to "give more self-conscious attention to the details and logic of research design," and second, to defend such ethnocultural historians as Ronald P. Formisano and Paul Kleppner against published criticisms. Too often confusing the former with the latter aim, Shade attains neither. His article is further compromised by distortions of other scholars' work and neglect of relevant literature published since 1974.[1] With Shade's two major prescriptions—plan research carefully and use genuinely multivariate methods—we have no quarrel. Often breached in practice, these familiar commandments can never be repeated too many times. To his lack of conceptual rigor, to his employment of a series of either meaningless or misleading "tests," and to several of his methodological dicta, we do take exception—thus the necessity for this note.

Underlying Shade's essay is a statistical pluralism that beckons historians to choose freely from among such techniques as

Authors' Note: *We wish to thank David Grether for giving this a close reading, and Jim Graham for many helpful editorial suggestions. We accept responsibility for all remaining errors.*

SOCIAL SCIENCE HISTORY, Vol. 7 No. 3, Summer 1983 321-344

multiple regression, bivariate correlation using either parametric or nonparametric measures, and homogeneous area analysis.[2] Yet he offers no guidelines for the appropriate use and interpretation of these procedures. Consider his discussion of the first of his five "questions" for analysis: "cross-level inference."

To put the discussion of cross-level inference into the proper context, let us first briefly review the "ecological fallacy" and some attempts to escape it. The ecological fallacy consists in naively inferring individual behavior from aggregate data. As Robinson (1950) showed, correlations at the two levels may differ. Benson (1961) and others responded by concentrating on those aggregate units that were more or less homogeneous in what they presumed were the vote-relevant individual traits. There are six problems with basing generalizations about whole states or regions on data from homogeneous areas.

(1) The technique excludes information from the much more numerous nonhomogeneous areas, thereby restricting the variation in both independent and dependent variables and decreasing the reliability of estimates of voter behavior.
(2) It assumes that individuals residing in homogeneous units have the same mean scores on other voter-relevant variables as do individuals living in heterogeneous units.
(3) The technique ignores "contextual effects," binding investigators to the supposition that people who live in different areas vote the same way, whether their peers are similar or dissimilar to them in demographic traits or voting proclivities.
(4) Investigators using homogeneous area analysis must implicitly treat all relationships between socioeconomic characteristics as linear.
(5) Despite heroic efforts, they can usually obtain data on only a few of the possibly relevant facts about voters. Of course, this is a more general problem, for we almost never have adequate measures on all the variables we want. Yet it is usually the case that county-level data are very considerably more plentiful than those at the township or precinct levels.
(6) Finally, the scrutiny of homogeneous areas is not compatible with the multivariate methods that Shade recommends, because to

isolate units that are simultaneously homogeneous on several variables of interest is virtually impossible. (For more extended discussions, see Kousser, 1976; Langbein and Lichtman, 1978; and Lichtman and Langbein, 1978).

Shade's only prescription for those still determined to analyze homogeneous units is to "choose homogeneous areas in a more systematic fashion in line with acceptable sampling procedures whenever the data permit." Unfortunately, no matter how circumspect the historian's sampling scheme, it cannot surmount deficiencies inherent in the method itself.

A more sophisticated method for bypassing the fallacy is "ecological regression," first used in the historical literature by Alexander and colleagues (1966) and first fully explained to historians in papers by Jones (1972), Kousser (1973), and Lichtman (1974).[3] In these articles the authors counseled historians to use multivariate, as well as (in appropriate cases) bivariate regression, and they outlined tests for the linearity of the relationships and procedures for dealing with apparent violations of the assumption crucial to both the ecological regression and homogeneous areas techniques—namely, that deviations from the predicated form of the relationship were random rather than systematic.[4]

As subsequent work has pointed out more clearly than did the early articles, a properly specified equation—that is, one that includes all the mutually correlated factors that influenced the dependent variable, and that captures the proper form of the relationships (e.g., linear, log-linear, quadratic, or interactive)—will, in many cases, correctly describe individual behavior[5] (see Hanushek et al., 1974; Lichtman and Langbein, 1978). As even the earliest articles argued, moreover, individual behavior is inferred more reliably from aggregate data through use of unstandardized regression coefficients rather than through the use of normalized measures, such as correlation coefficients or standardized regression measures (beta weights). Normalized measures are uniquely subject to distortions arising from changes in the standard deviations of variables produced by the aggregation process.

Shade's own discussion of ecological inference is especially confusing. Not only did he ignore distinctions between regression

and correlation measures, but he misled readers by repeating a rule of thumb offered by Dollar and Jensen (1971: 101). To simplify matters, Dollar and Jensen advised prospective users that the Pearsonian ecological correlation between two variables had to be "at least ± .7" to justify putting much confidence in ecological regression estimates. In fact, a high correlation is neither a necessary nor a sufficient condition for an unbiased ordinary least squares regression estimate. Such an estimate will be unbiased if, and only if, there are no variables excluded from the equation, which would have had nonzero coefficients had they been included, and if the form of the relationships is correctly specified. A high correlation may be a sign of a well-specified equation, but it is not necessary or sufficient. Suppose voting were really random across all possible groups, i.e., that an equal percentage of persons in every group voted for a certain party. Then both correlations and the slopes of the regression lines for the percentages in each group, class, or whatever and the vote would be zero (assuming no aggregation bias), but the estimates of how each group voted would be perfectly correct.

In a related error, Shade misinterpreted Goodman (1959) by stating that the "no excluded correlated variables" assumption just discussed amounts to an assumption that "The patterns of settlement . . . produced a random distribution of ethnic groups" (Shade, 1981: 176). What is at issue in this assumption is not a random distribution across counties or townships, but random deviations of the points representing those units from the true regression line or surface in the population of interest.

Comparisons of the results of regression estimates with actual votes from poll book, survey, and other individual level data have uniformly found the estimates from carefully chosen equations based on aggregate data to be quite close to the actual behavior patterns of individuals in a fairly wide variety of cases. (See, for example, Irwin, 1967; Stokes, 1969; Kousser, 1973; Bourke and DeBats, 1980.) These tests increase one's confidence in the usefulness of the technique.

Shade proposed to "test" for the "pitfalls of ecological regression" by comparing a bivariate regression estimate of the proportion of Germans who voted Democratic in 1851 in Pennsylvania, calculated from data from all the state's counties, with two other estimates. The first was derived from an analysis of votes in homogeneous minor civil divisions in one county, and the second, from an ecological regression of the percentage of votes for each party on the percentage of Germans by townships in that county. No individual level data on German voting behavior for this election seem to have survived. Finding that the countywide regression and the homogeneous area estimates are equal to one another and differ from the statewide estimates by 20%, he concluded that the results of the ethnocultural historians for a number of states over a great many elections would probably not have differed had they used ecological regression, and that in the future, historians should feel entirely free to use either homogeneous areas or regression analysis "depending on the nature of the available data and the questions being asked" (Shade, 1981: 177). He offered no further guidance on which questions are appropriate for each technique.

Shade's test of ecological regression against homogeneous area estimation is not a proper test for seven reasons.

(1) The equation at the statewide level is misspecified unless all other determinants of voting behavior (for example, religion or class) were uncorrelated with ethnicity. His is one regression estimate of the behavior of individual German voters, but almost certainly not the best one that could have been made.[6]

(2) While a comparison of a statewide estimate based on data from all the counties against an estimate using all the civil divisions or townships for the whole state (not just for one county) might be an appropriate way to test for possible bias induced by the combination of smaller areas into counties, even this comparison would not uncover possible biases introduced by combining voters into townships.[7] The ecological fallacy affects the grouping of individuals into any units, i.e., townships, counties, states, and not just the grouping of townships into counties, as Shade seems

to believe. Furthermore, for ecological regression equations that measure temporal change in party voting, county level data may well be superior to information collected for wards or townships, since counties are less likely than these smaller units to experience the kind of massive population turnover that would substantially alter their demographic composition from one election to the next. (On township level turnover, see Winkle, 1983).

(3) If the estimates from one county are based on a small number of cases (Shade never said how many), they probably have larger standard errors than do the statewide estimates (he did not include standard errors or "t" tests for either regression in any version of his article). While they may not be biased, therefore, the Schuylkill county estimates are almost certainly less precise, in a statistical sense, than are the statewide estimates.

(4) Even if unbiased and precise, the estimates might have come from a county that for some reason deviated from the statewide relationship. Shade never made clear how much Schuylkill county deviated from the linear regression line. If Schuylkill were a randomly deviant case, the statewide estimates might still be correct, but the Schuylkill and statewide results would differ.

(5) Some of Shade's comments indicate that Schuylkill was a case that systematically deviated from the statewide linear relationship. The "premier mining area in eastern Pennsylvania" and later focal point for the Molly Maguires, it had a German plurality, a Welsh and English minority, and a "rapidly growing group of Catholic Irish" (Shade, 1981: 175). This ethnic stew might well have raised group conflict in voting to levels above that in the average county, thereby increasing the correlation between ethnicity and the vote. It is perfectly possible that 70% of the county's Germans, but only 50% of the state's voted Democratic. And if people in other counties that were similar to Schuylkill behaved in this manner, that fact would indicate that the statewide estimate should be based on a nonlinear, not a linear functional form. Further, if miners voted differently than non-miners, a properly specified equation for the state should include a measure of mining activity. Shade took neither of these factors into account in his statewide estimate.

(6) Even if ecological regression "failed" a test based on a representative county, that fact would not validate the homogeneous areas

method, because the data needed for comparison purposes are not those grouped by township, but actual individual level voting and ethnic records. To test how well two methods do at estimating individual behavior, one needs individual level records as a criterion. Shade had none; therefore, he did not test what he claimed to test.

(7) Even if he had made a proper test, it would not invalidate regression or support the use of estimates based on homogeneous areas for all the elections previously covered or to be treated by "new political historians." These conclusions would follow only if he were willing to argue that the same results held everywhere throughout all time, or at least in the nineteenth century in areas that the ethnoculturalists have studied.

In sum, Shade has used an almost certainly badly misspecified equation to make a statewide estimate, compared it with aggregate—not individual—estimates of voting behavior from one probably deviant county, and then used this one-county, one-election "test" to argue for the validity of estimates from homogeneous areas made by the ethnoculturalists in elections over most of the north for much of the nineteenth century! This parody of statistical procedure offers neither sanctions for the work of previous historians nor reliable guidance for that of future scholars.

Nor does Shade's discussion on levels of measurement, in the second section of his article, provide useful direction. In an effort to defend prior work by some other historians, Shade stuffed and demolished a straw man, treated published criticisms selectively (consequently purveying misleading advice), set up another meaningless test and then misread his own results.

The "straw man" is that critics have charged that treating interval (numeric) data as ordinal (ranked) biased the results of the new political historians. So far as we know, this charge was never made, and it is certainly not present in the article that Shade cited (Kousser, 1976: 9-10; Shade, 1981: 178). Of the three points Kousser did make, Shade more or less admitted the first, which is that information on the exact extent of differences in variables is

squandered when analysts use rank-order correlations. He ignored the other two: that a resort to rank-order measures makes testing for nonlinear effects (e.g., polynomial, multiplicative interaction, logarithmic) impossible and that it forces one to rely on less powerful significance tests to assess whether two or more variables are associated or not. A fourth point, stressed throughout, but not explicitly stated during Kousser's discussion of nonparametrics, is that the use of Spearman's Rho or Kendall's Tau precludes historians from developing multivariate models.

Shade's test of whether the use of ordinal, rather than interval level measures leads to different conclusions is to calculate Spearman's Rho and bivariate Pearson's r coefficients for county level data on the relationships between the vote in a "Maine Law" (temperance) referendum in Pennsylvania in 1854 and ten social or economic variables. Ignoring criticisms of the use of zero-order, ecological correlations, Shade is content to eyeball two sets of flawed measures and conclude summarily that both lead to similar conclusions.[8]

But even accepting the validity of this test for purpose of argument, Shade's conclusion does not follow. Careful examination of the two sets of bivariate correlation coefficients reveals potentially important divergences in historical interpretation, not the "modest differences" that Shade claimed. Following his own procedure of rating the importance of variables according to their capacity to predict variation in the dependent variable, we note that the pattern of his Spearman's Rho coefficients supports a religious interpretation of temperance voting, whereas that of his Pearson's r coefficients suggests an ethnic interpretation. His column of Rhos indicates that the proportion of Presbyterians in a county was the most important determinant of temperance voting, followed by the proportion of Methodists. But in his column of Pearson's r the proportion of Pennsylvania Dutch, and the proportion of English in a county were the most crucial factors. The proportion of Presbyterians fell to fourth place and the proportion of Methodists to eighth place among the ten coefficients.

We begin the discussion of Shade's third topic, significance tests, by first setting out our position without reference to his. Two central priorities of the new historians have been to give numerical precision to such verbal expressions as "more," "less," and "most," and to assure that the quantitative analyses are as reliable as possible. Statistical methodology offers historians both a ready stock of numerical measures and a means of reasoning formally about error. Significance tests and related procedures are designed to avoid the confounding of substantive results with artifacts produced by random error—a form of error that, with equal probability, generates positive or negative discrepancies between measured and actual results. For example, statements that more Whigs than Democrats voted for temperance, or that a higher percentage did, or that 10% more did are not very interesting unless we can first reject the hypothesis that such observed differences reflect random error, rather than actual behavior. Although a significance test cannot, of course, establish the substantive importance of a result, it can, as David Gold (whom Shade approvingly cited on the subject of significance tests) has noted, provide a useful check against drawing conclusions from "an observed association [that] could be generated in a given set of data by a random process" (Gold, 1969: 46).

In Shade's view, however, the ethnoculturalists should be exempt from the usual canons of research procedure since "they generally did not deal with systematically drawn samples, but with 'total populations.'"[9] Yet numerous sources of random error can afflict measures computed for total populations, as well as those computed for samples of data.

Consider four examples in which significance tests can be useful correctives to unwarranted inferences even when studying total populations. First, random measurement error may arise either from the original collection of information or the historian's own processing of data. Second, the historian may be able to measure only proxies for the variables that truly are of interest. For example, an investigator might use occupation or education as a proxy for social standing, or church seats as a proxy for

religious affiliations. Measures computed from such proxy variables may differ both randomly and systematically from the true values for the variables that are really of interest.

Third, whenever historians infer conclusions about individuals from data collected for aggregate units, they are, in effect, engaged in sampling. In this case the units studied by the investigator represent a sample of cases drawn from the total population of individuals, according to the process by which they were arranged into geographical subdivisions. As in any sampling process, such grouping can generate both random and systematic error.

Finally, the total population studied by the historian may be only a subset of the population whose behavior he or she actually seeks to explore. For instance, the historian may have data for every legislator in a given statehouse for a roll call on a temperance law. But he or she may truly be interested in whether all Whigs and Democrats, voters as well as legislators, held different attitudes on temperance. The historian might also consider the roll call votes on specific proposals as samples of the legislators' attitudes on the general subject of liquor control. In these cases we can use significance tests to determine whether this sample of the population's underlying attitudes indicates that partisan views on the subject of drinking (which is the population that is really of interest) diverged or not.

Although Shade recognized the force of this fourth argument, he sidestepped it by asserting that "the significance test is a useful tool" only "if one is dealing with samples that have known probabilities," whereas "each of these authors conceived of his total population as a nonprobability sample of a larger universe." With this argument Shade inadvertently conceded that ethnocultural scholarship suffers from a more serious problem than random error. For if these historians were generalizing from nonprobability samples, their results may have been marred by systematic biases that yield results which are skewed in a particular direction and which, therefore, cannot be detected by the usual significance tests. The fact that historians should, as Shade admonished, "con-

sider the associations between their samples and the universes to which they wish to generalize," but that the ethnoculturalists, according to Shade, did not do so, reduces these historians' credibility even further.

Yet that observation would not free them from using significance tests. It merely implies that they and other historians should try to reason about the biases involved in their data and, if possible, redefine the significance tests accordingly. If their samples were skewed toward finding, for example, more cohesive German Lutheran and Pietist political behavior than was typical for the state or region, they might increase the required significance level on a test for differences in the two groups' voting records, in order to take the sampling bias into account. Because homogeneous areas probably exhibited more uniform voting patterns than did heterogeneous areas, one could reasonably infer that two groups' behaviors differed in the total population only if a very great divergence—more than one would expect using a conventional level of significance—existed in voting returns drawn exclusively from such areas.

Finally, Shade raised questions about the choice of particular levels of statistical significance, in the process confusing scientific convention with "subjectivity." It is surely true, as is invariably noted in the first few weeks of any introductory statistics course, that there is nothing sacred about .05, .01, or, for that matter, .75. The reason for using low significance levels is that it would otherwise be too easy to reject null hypotheses, and science would become violently unstable, as every new study overturned a previous one. While any particular level is arbitrary, it at least provides a precise and, in that sense, "objective" decision rule for accepting or rejecting a finding, and one which may well be widely agreed upon by scholars of diverse disciplines and interests.[10] Thus, for someone to publish coefficients significant at the .25 level would raise numerous eyebrows.

Consider Table 1, which is based on an 1840 roll call in the lower house of the Michigan legislature on banning railroad trains from running on Sunday (Formisano, 1971: 123-24). To

Table 1 1840 Michigan House Sabbatarian Vote

Attitude on Banning Sunday Travel	Party: Whig	Party: Democrat
For	19 (66%)	7 (54%)
Against	10 (34%)	6 (46%)
TOTAL	29 (100%)	13 (100%)

decide whether this difference in party behavior, taken by Formisano as indicative of the "central tendencies of the parties" on such issues, is sufficiently large for reliable inference, we computed a chi-square statistic. The chi-square value is .518, which is significant only at the .47 level. While statisticians often urge analysts to publish the actual significance levels of their parameters, and not just to denote which of them pass the .05 or .01 barriers, few social scientists would feel comfortable printing ".47" as an attained significance level at the bottom of a table, and few readers would put much credence in conclusions based on the contention that, despite a tiny chi-square, the relationship in question was actually strong.[11] While significance levels are only conventions, they are useful ones.

Although Shade's warnings about the unthinking use of significance tests and the confusion of statistical with substantive importance are worth heeding, he confused more than he guided on this topic. If he is interpreted as encouraging historians to abjure tests of significance altogether, which is not an unreasonable reading of some of his remarks, the result will be a move away from instead of toward proper methodological practice.

Nor is Shade's discussion of "synoptic measures" a positive aid to understanding. Suppose a historian believes that the Democratic vote (D) in some election depended on the number of Irish (I), Germans (G), and Episcopalians (E) in each county in some

state. Then he or she might formulate and test the hypothesis in the usual multiple regression manner as

$$D = B_0 + B_1I + B_2G + B_3E + u \qquad [1]$$

where the Bs are coefficients to be estimated and u is an error term. But since many analysts are interested less in the actual vote than in the percentage the party received, they would find it more natural to express their hypotheses as the greater the *proportion* of Irish and so on, the greater the *proportion* of Democrats. This second hypothesis would take the form

$$D/P = B_0 + B_1I/P + B_2G/P + B_3E/P + u \qquad [2]$$

where P is the number in the eligible voting population, or perhaps the number who actually voted.

The reader will note that P appears as a denominator for the variables on both sides of the equation. The dependent variable is, therefore, being regressed on independent variables that are, by definition, partly functions of itself. Will this fact artificially inflate the estimates of the regression parameters (the Bs)? The short answer is that it depends on which hypothesis the historian believes—that given in equation 1 or that encompassed in equation 2. If the predicated relationship is between proportions, then the coefficients will not be higher than they "should" be; on the other hand, if the theory properly relates numbers of people, and the error term meets the usual assumptions, then there is no particular reason to normalize by population or by any other quantity.[12]

Recognizing this point, Shade argued that population size may still be important as a proxy for urban/rural differences in voting; he urged historians to "control for size statistically" by introducing population as an independent variable in equations such as 1; he performed another test to see whether the ethnoculturalists' failure to introduce such a control distorted

their findings; he concluded that it did not; and he closed the section by repeating his homily about the necessity for "carefully formulated hypotheses" (Shade, 1981: 184-186). His discussion is flawed on several counts.

First, if a historian thinks that an urban/rural split was an important determinant of voting, the percentage living in cities or towns would appear to be the natural proxy to choose.[13] If an investigator concerned with the United States in the 1850s, say, used population instead, it might well be that a geographically large and densely populated rural county would be judged, by population size, more urban than a geographically smaller county where nearly everyone lived in a town.

Second, Shade erred in suggesting that the proper adjustment for differences in voter turnout is to "control for [population] size statistically." Such controls index only the influence of differences in the number of potential voters, not in voter turnout. Investigators can take turnout into account by measuring both independent and dependent variables, using the potential voting population rather than the vote cast as the denominator for percentages. Analysis of such variables reveals the relative support given candidates and parties by groups within the total potential electorate. Historians can also gain insight into turnout effects by using the proportion of voters in the potential electorate as a dependent variable and by employing regression techniques for measuring transition probabilities between voting and nonvoting.

Third, Shade used yet another misleading test to determine whether previous historians' failures to control for population size distorted their results. In his test Shade estimated a series of equations such as

$$\begin{aligned} T &= B_0 + B_1 I + B_2 P + u \\ T &= B_0 + B_1 G + B_2 P + u \\ T &= B_0 + B_1 M + B_2 P + u \end{aligned} \qquad [3]$$

where T stands for the vote in the 1854 Maine Law referendum in Pennsylvania, M for Methodists, all the rest of the variables are as

defined earlier, and the data are aggregated at the county level.[14] He then compared the partial correlation coefficients, not the partial regression coefficients, for the variables I, G, M, and so on (but not for P) with zero-order correlation coefficients computed from equations of the form

$$\begin{aligned} T/P &= B_0 + B_1 I/P + u \\ T/P &= B_0 + B_1 G/P + u \\ T/P &= B_0 + B_1 M/P + u \end{aligned} \qquad [4]$$

Since those two sets of correlation coefficients are roughly similar, he again exonerated the ethnoculturalists, who, when they computed statewide statistics, used equations like 4, and not, as he favored, equations like 3.

As with his other tests, this one is seriously deficient. Every bivariate or trivariate equation he used is surely misspecified, and the parameters are therefore biased. A comparison of two sets of biased coefficients is hardly conclusive evidence that one of them is not biased. Even ignoring bias, correlation coefficients are, for reasons detailed in our 1973 and 1974 articles and reiterated above, inferior statistics for aggregate data analysis. Furthermore, by comparing equations like 4 to equations like 3, Shade was, in effect, contrasting two rather different though related theories—one based on votes and the number in each group, and one based on proportions.[15] It would therefore be a bit difficult to know what to expect from such a comparison or what to make of the results, even if a meaningful test had been performed.

Shade's results are also difficult to interpret since his zero-order correlations are reported only for "Pro-Temperance" (1981: Table 1) voting and his partial correlations (controlling for population) for both "Pro-Temperance" and "Anti-Temperance" voting (1981: Table 2). The differences are potentially important. The partial correlation for "Farm Value" is only –.0785 for Pro-Temperance voting, but .5007 for Anti-Temperance voting. Unfortunately, we are not supplied the information necessary, either for explaining this difference or for determining whether it

is also present in the zero-order coefficient. Nonetheless, the ethnoculturalists once again emerge from Pennsylvania in fine shape, as Shade emphasized "that the basic relationships remain the same," whether or not population is controlled.

It is not only his test, however, that is inadequate here: his whole section on synoptic measures misleads. The chief problem with leaving out population is that doing so often causes "heteroscedasticity," or unequal variances of the error terms for each unit of observation.[16] While heteroscedasticity does not produce biased parameter estimates in ordinary least squares regression, it does increase the variance of the estimates and it invalidates the usual significance tests. The standard solution, as outlined more extensively in Kousser (1980), is to weight each variable in equations such as equation 2 by the square root of P. In this, as in other sections of his article, then, Shade's analysis raises important points without clarifying them, his test is deceptive, and his suggestions for future work are counter-productive.

Shade's fifth and final "question" is whether the ethno-culturalists' use of bivariate correlation, rather than what Shade implicitly admitted in this section of his article is the superior technique of multiple regression, might have led them to adopt an ethnic or ethnoreligious, rather than an economic interpretation of American politics. He tests this possibility by regressing the county level election returns in the 1854 Pennsylvania temperance referendum in a stepwise fashion, on one ethnic (percentage English), one religious (percentage Pennsylvania Dutch), and one economic (farm value) variable and determining how much additional variance in the wet and dry percentages each type of variable explains.

With its exclusive focus on a temperance referendum, Shade's procedure cannot determine whether ethnocultural divisions pervaded Pennsylvania's partisan contests in the 1850s, but can only shred a straw man of his own creation. That liquor laws divided nineteenth-century ethnic groups was no discovery of

Benson, Hays, et al., nor did they claim such originality. Historians have long known that nineteenth-century Germans liked their beer, Irish their whiskey, Yankees their cold water. It is precisely in the response to temperance laws that one would expect ethnic and religious variables to count most heavily as determinants of voter choice.

Shade's operational measures and specific procedures also raise serious questions about the logic and execution of his test. "Farm value" has the wrong sign in one part of his Table 3, no doubt because of a misprint. Shade's index of "Pennsylvania Dutch," based on church seats, is (oddly) almost perfectly positively correlated with his "German Orthodox" variable and nearly perfectly negatively correlated with his percentage "English" (Shade, 1981: Tables 1 and 3). After all his strictures about "holding population constant statistically" earlier in his article, he did not include population in his multiple regression. Several variables with relatively high zero-order correlations with temperance in Table 1 were not entered into the multiple regression in Table 3. The statistician's model does not demand, as he implied on page 191, orthogonal independent variables in multiple regression. Indeed, if all independent variables were mutually orthogonal, bivariate methods would suffice.

In any case, the assessment of the influence of different factors through stepwise regression and the "additional variance explained" (incremental increase in R^2) criterion is a misleading procedure that attibutes all the variance mutually explained by correlated variables to whichever variable is first entered in the equation. For example, any part of the correlations with "dry" sentiment explained by both farm values and the percentage English is chalked up entirely to the English. The economic variable (or any variable entered later than the percentage English) gets a "chance" to explain only the residual variation in the dependent variable left after the English variable has explained everything it could. Whatever its actual importance, the contribution of R^2 of the n^{th} variable entered in a regression equation

cannot be greater than $1 - R^2_{n-1}$ where R^2_{n-1} is the value of R^2 attained prior to the inclusion of the n^{th} variable. Since Shade's technique treats variables, in this sense, asymmetrically, it is a deficient method for comparing the importance of the contributions of different variable to explanations of voting behavior.

Aside from problems of asymmetrical measurement, an exclusive focus on explained variance is misguided for cases of cross-level inference, such as Shade's example. Since the most spirited and sophisticated defender of such a focus is John Hammond, and since Shade provided no such defense, we will consider Hammond's arguments. In two thoughtful articles and a book, Hammond (1973, 1979a, 1979b) advocated the use of standardized regression coefficients for rating the importance of variables according to their contribution to explained variance. In particular, he recommended using beta weights for aggregate data in which the variables are measured with error that has a particular (multiplicative) structure or in which the variables are assumed to be arbitrarily scaled indicators of underlying attitudes (Hammond, 1979a: 478-483). He also points out that beta weights (which are identical to Pearsonian correlation coefficients in the bivariate case) allow a comparison of the effects of variables that do not have a common scale, such as most measures of ethnicity and economic welfare.

Although Hammond's points are well argued, on balance, we reject his advice for six reasons.

(1) As Hammond himself admitted (1979a: 485), standardized coefficients for aggregate data may be biased estimates of individual level relationships in cases where unstandardized coefficients are unbiased. The values of beta weights, like other normalized measures, are functions both of individual level relations between independent and dependent variables and of differences in the relative variance of competing independent variables produced entirely by the process of aggregating individuals into groups.[17] Indeed, such differences in relative variances are precisely what would be expected in virtually every case of interest to historians, as groups are almost invariably distributed differently across geographical units.

(2) As Hammond showed in his 1973 article, if one group is more geographically concentrated than another, the aggregate estimate of the individual level standardized coefficient will be more inflated for the more segregated group, although the unstandardized estimates may well be unbiased for each. Thus, a comparison of the relative magnitudes of the two standardized coefficients will exaggerate the importance in explaining the dependent variable of the more segregated, relative to the less concentrated, group in many cases in which the same comparison for unstandardized coefficients will not.

(3) Since aggregation processes are likely to produce different changes in relative variations, the values of beta weights will depend on the particular set of units chosen for analysis. This means that even for properly specified models, ecological inference will be unstable as the analysis shifts from one level of aggregation to another.

(4) For multivariate equations, the standardized regression coefficients have no natural interpretation. In particular, they are not truly measures of the percentages of variance explained by independent variables. As Hubert Blalock noted, "The partial correlation is a measure of the *amount of variation explained* by one independent variable. . . . The beta weights, on the other hand, indicate *how much change* in the dependent variable is produced by a standardized change in one of the independent variables" (Blalock, 1972: 453).

(5) We see no reason to believe that, in general, measurement error for aggregate level variables is multiplicative. Indeed, most of the examples of multiplicative error cited by Hammond (such as using the percentage of residents born in Scandinavia to infer the behavior of all generations of Scandinavian-Americans) can be conceptualized as specification error and treated accordingly.

(6) As a sociologist, Hammond may wish to ignore the specifics of the historical situation, such as how referendum questions were posed or differences in the demographic details of various groups—their age and sex composition, recency of migration, voting turnout, and so on (Hammond, 1979a: 483-484). As historians, we believe all these specifics are potentially important, and we want to avoid using a method that would make it easy to ignore such factors.[18]

Although Shade is to be commended for raising important questions, some of which had seldom been bruited in the historical literature, his discussion generally confuses more than it clarifies, his specific prescriptions for future work are usually misleading or wrong, and his defenses of the conclusions of previous scholars are seriously flawed. While we agree with Shade's call for better and more self-conscious research designs and for the adoption of genuinely multivariate methods, we think that the misconceptions of his article underline to an even greater extent the necessity for much more thorough statistical training for quantitative social scientific historians. Now that some of the misunderstandings have been cleared up, a more productive debate on these and related issues may proceed.

NOTES

1. Shade's list of 66 references includes only two post-1974 citations, neither of which he consulted for methodological guidance. In fact, citations to the Hanushek et al. (1974) and two Lichtman and Langbein (1978) articles were even deleted between early drafts and the published version of Shade's article.

2. While multiple regression, on which Shade concentrates, is undoubtedly useful, he might also have mentioned more advanced techniques, which have recently been introduced into the historical literature, such as logit and probit analysis. On these, see Knoke and Burke (1980), Kousser (1980), and Goldin (1981).

3. Shade distorted the work of Alexander et al. (1966) and McCrary et al. (1970) when he said that the contrasts in their findings, both based on ecological regression, "can only be resolved by a methodological 'leap of faith' " (Shade, 1981: 173). While the McCrary results were based on multiple regression analyses for all Alabama counties, Alexander's rested on regressions with data drawn from nonrandom surviving beat (the Southern equivalent of township) returns in only 15 of the state's counties. It is hardly surprising that analyses based on different universes of data led to different results.

4. Shade's charge (1981: 172) that we recommended only the use of bivariate regression is incorrect.

5. It is misleading to put as much stress on the assumption of constant behavior across geographical units as, for example, Vinovskis (1980) has done. That assumption is relatively easy to test and correct for in practice. The much graver difficulties in deciding whether individual level inferences are right or not arise from the possibility of specification error and aggregation bias (see Lichtman and Langbein, 1978).

6. In addition to the criticisms offered in the text, it must be noted that even if Shade were convincing, his analysis here would support neither an ethnocultural nor ethnoreligious hypothesis, but merely a simple ethnic hypothesis.

7. For an interesting test of the differences between county- and civil-division-level regression estimates for Iowa in 1924, see Waterhouse (1983).

8. These criticisms of ecological correlations, which pervade the literature, were reiterated in the paragraph preceding the discussion of nonparametrics in Kousser's 1976 article. Shade noted and admitted such criticisms earlier in his article, but conveniently ignored them in his "levels of measurement" section.

9. Few if any of the controversialists in Morrison and Henkel (1970), Shade's main source of criticisms of the use of significance tests, discussed the question of running such tests on "total populations." Indeed, it is a question that has attracted little systematic theoretical attention. Shade also ignored the effective criticisms of the 1957 article by Selvin (which sparked the debate in sociology and which is reprinted in Morrison and Henkel), in numerous other articles in the book, as well as the fact that the whole controversy has seemingly died away in sociology since 1970. Indeed, sociologists are increasingly turning to such techniques as log linear modeling, which involves wholesale computations of chi-square values to evaluate different hypotheses (see Goodman, 1978).

10. Of course it is possible to imagine situations in which standard significance tests and conventional levels of certainty ought to be jettisoned. Since if one is working with a very large number of observations, chance alone will often produce apparently significant relationships between variables at the .10 or .05 levels, one ought in such cases to impose more stringent criteria for significance. If one's sample were skewed in known ways or if one had a sharply peaked "prior" belief about some outcome, a redesigned test or perhaps an unusual significance level would provide a more appropriate decision rule. Alas, historians generally operate in a world of diffuse priors and samples of unknown bias. Armed only with a rough set of hypotheses, they are presented with a bunch of numbers whose representativeness they can determine only approximately, and they must say to themselves: "On the basis of this collection of data, how much credence should I give to this hypothesis?" As a practical matter, conventional significance tests are often the only stop this side of relativism.

11. Formisano ran no such significance test and this is the only legislative vote on "moral" issues for which he provided a party breakdown. If one adds the two Whig and four Democratic abstainers to the table, the chi-square rises to 3.412, which is significant at the .18 level.

12. Bollen and Ward's 1980 article provides a good introduction to the literature on this highly controversial subject. Because the problems of multicollinearity and misspecification, discussed below in the text, seem to us especially grave in the equation 1 form of the hypothesis, we take a somewhat different position on the use of ratio variables in this particular case than they do for the general case.

13. To the extent that Shade's article is a defense of the ethnoculturalists, this point seems a curious one for him to raise, for neither their (homogeneous area) methods nor their ethnoreligious theory seems compatible with urban-rural differences in voting behavior.

14. It is possible that the actual set of equations Shade used is of the form $T/P = B_0 + B_1 I/P + B_2 P + u$. His discussion is not clear on this point.

15. If, of course, population size per se is an independent "contextual" influence on aggregate level voting, then its exclusion from a regression equation will bias parameter estimates whether theory calls for specification in ratios or raw numbers. In this special case, population should be entered as an additional control variable. If, however, equations are specified in terms of raw numbers rather than ratios, population size will be highly collinear with the numbers of people in the larger population groups, creating severe problems of multicollinearity in regression estimates.

16. There is a good discussion of hetereoscedasticity and autocorrelation (Shade confused the two) in Hanushek and Jackson (1977: 142-146).

17. This follows directly from the formula defining the standardized regression coefficient, which is the unstandardized regression coefficient multiplied by the ratio of the standard deviation of the independent to the dependent variable—for example, the three-variable case: $B_{yx.z} = b_{yx.z} S_x/S_y$, where $B_{yx.z}$ is the beta weight, $b_{yx.z}$ the unstandardized regression coefficient, S_x the standard deviation of X, and S_y the standard deviation of Y. Even when $b_{yx.z}$ at the aggregate level is an unbiased estimator of its individual level counterpart, $B_{yx.z}$ will not be an unbiased estimator of the individual level beta weight, except in the unlikely event that S_x/S_y remains unchanged after aggregation. Hammond was certainly aware of how grouping alters relative variance, but oddly concluded that such changes bias unstandardized, but not standardized, coefficients (Hammond, 1979a: 478-482). For more detailed discussion and empirical examples see Langbein and Lichtman (1978: 36-38).

18. In the execution of multivariate analysis, we would stress not only the estimation of particular parameters, but also the proper form (e.g., multiplicative, linear, interactive) of the multivariate model itself (see Broder and Lichtman, 1983).

REFERENCES

ALEXANDER, T. B., P. ELMORE, F. LOWERY and M. SKINNER (1966) "The basis of Alabama's two-party system." Alabama Rev. 19: 243-276.

BENSON, L. (1961) The Concept of Jacksonian Democracy: New York as a Test Case. Princeton, NJ: Princeton Univ. Press.

BLALOCK, H. M. (1972) Social Statistics. New York: McGraw-Hill.

BOLLEN, K. A. and S. WARD (1980) "Ratio variables in aggregate data analysis: their uses, problems, and alternatives," pp. 60-79 in E. F. Borgatta and D. J. Jackson (eds.) Aggregate Data: Analysis and Interpretation. Beverly Hills, CA: Sage.

BOURKE, P. F. and D. A. DeBATS (1980) "Individuals and aggregates: a note on historical data and assumptions." Social Sci. History 4: 229-250.

BRODER, I. and A. J. LICHTMAN (1983) "Modeling the past: a note on the search for proper form." J. of Interdisciplinary History 13: 489-502.

DOLLAR, C. M. and R. J. JENSEN (1971) Historian's Guide to Statistics: Quantitative Analysis and Historical Research. New York: Holt, Rinehart & Winston.

FORMISANO, R. P. (1971) The Birth of Mass Political Parties: Michigan, 1827-1861. Princeton, NJ: Princeton Univ. Press.

GOLD, D. (1969) "Statistical tests and substantive significance," Amer. Sociologist 4: 42-46.

GOLDIN, C. (1981) "Family strategies and the family economy in the late 19th century: the role of secondary workers," pp. 277-310 in T. Hershberg (ed.) Philadelphia: Work, Space, Family, and Group Experience in the 19th Century. New York: Oxford Univ. Press.

GOODMAN, L. S. (1978) Analyzing Qualitative/Categorical Data: Log-Linear Models and Latent Structure Analysis. Cambridge, MA: Abt Books.

———(1959) "Some alternatives to ecological correlation." Amer. J. of Sociology 64: 610-625.

HAMMOND, J. L. (1979a) "New approaches to aggregate electoral data." J. of Interdisciplinary History 9: 473-492.

———(1979b) The Politics of Benevolence: Revival Religion and American Voting Behavior. Norwood, NJ: Ablex.

———(1973) "Two sources of error in ecological correlations." Ameri. Soc. Rev. 38: 764-777.

HANUSHEK, E. A. and J. E. JACKSON (1977) Statistical Methods For Social Scientists. New York: Academic.

——— and J. F. KAIN (1974) "Model specification, use of aggregate data, and the ecological correlation fallacy." Political Methodology 1: 89-107.

IRWIN, G. A. (1967) "Two methods for estimating voter transition probabilities." Ph. D. Dissertation, Florida State University.

JONES, E. T. (1972) "Ecological inference and electoral analysis." J. of Interdisciplinary History 2: 249-262.

KNOKE, D. and P. J. BURKE (1980) Log-Linear Models. Beverly Hills, CA: Sage.

KOUSSER, J. M. (1980) "Making separate equal: integration of black and white school funds in Kentucky."J. of Interdisciplinary History 10: 399-428.

———(1976) "The new political history: a methodological critique." Reviews in Amer. History 4: 1-14.

———(1973) "Ecological regression and the analysis of past politics." J. of Interdisciplinary History 4: 237-262.

LANGBEIN, L. I. and A. J. LICHTMAN (1978) Ecological Inference. Beverly Hills, CA: Sage.

LICHTMAN, A. J. (1974) "Correlation, regression, and the ecological fallacy: a critique." J. of Interdisciplinary History 4: 417-433.

———and L. I. LANGBEIN (1978) "Ecological regression versus homogeneous units: a specification analysis." Social Sci. History 2: 172-194.

McCRARY, P., C. MILLER, and D. BAUM (1978) "Class and party in the secession crisis: voting behavior in the Deep South, 1856-1861." J. of Interdisciplinary History 8: 429-457.

MORRISON, D. E. and R. E. HENKEL (1970) The Significance Test Controversy—A Reader. Chicago: Aldine.

ROBINSON, W. S. (1950) "Ecological correlations and the behavior of individuals." Amer. Soc. Rev. 15: 351-357.

SHADE, W. A. (1981) " 'New political history': some statistical questions raised." Social Sci. History 5 (Spring): 171-196.

STOKES, D. E. (1969) "Cross-level inference as a game against nature," pp. 62-83 in J. Berns (ed.) Mathematical Applications in Political Science. Charlottesville: Univ. of Virginia Press.

VINOVSKIS, M. A. (1980) "Problems and opportunities in the use of individual and aggregate level census data," pp. 53-70 in J. M. Clubb and E. K. Scheuch (eds.) Historical Social Research: The Use of Historical and Process-Produced Data. Stuttgart, Germany: Klett-Cotta.

WATERHOUSE, D. (1983) "The estimation of voting behavior from aggregate data: a test." J. of Social History 16: 35-53.

WINKLE, K. J. (1983) "A social analysis of voter turnout in Ohio, 1850-1860." J. of Interdisciplinary History 136: 411-436.

J. Morgan Kousser is Professor of History and Social Science at the California Institute of Technology. With James M. McPherson, he edited Region, Race, and Reconstruction: Essays in Honor of C. Vann Woodward *(Oxford University Press, 1982), and he coauthored (with Gary W. Cox and David W. Galenson) an article on log-linear analysis of contingency tables, which was published in the fall issue of* Historical Methods.

Allan J. Lichtman is Professor of History at The American University. He has recently published an article on realignment theory in Historical Methods *and an article entitled "Political Realignment and 'Ethnocultural' Voting in Late Nineteenth-Century America"* (Journal of Social History *spring 1983).*