Effects of High-School Size on Student Outcomes:
Response to Howley and Howley

Valerie E. Lee
University of Michigan

Abstract
I take issue with several points in the Howleys' reanalysis of "High School Size: Which Works Best and for Whom?" (Lee & Smith, 1997). That the original sample of NELS schools might have underrepresented small rural public schools would not bias results, as they claim. Their assertion that our conclusions about an ideal high-school size privileged excellence over equity ignores the fact that our multilevel analyses explored the two outcomes simultaneously. Neither do I agree that our claim about "ideal size" (600-900) was too narrow, as our paper was clear that our focus was on achievement and its equitable distribution. Perhaps the most important area of disagreement concerns non-linear relationships between school size and achievement gains. Ignoring the skewed distribution of school size, without either transforming or categorizing the variable produces findings that spuriously favor the smallest schools. Our recent involvement as expert witnesses on opposite sides in a court case may have motivated the Howleys' attempt to discredit our work. Finally, I argue that research attempting to establish a direct link between school size and student outcomes may be misguided. Rather, school size influences student outcomes only indirectly, through the academic and social organization of

schools. Considerable evidence links these organizational factors to student outcomes (especially learning and its equitable distribution).

In their article "School Size and the Influence of Socioeconomic Status on Student Achievement: Confronting the Threat of Size Bias in National Data Sets," Craig and Aimee Howley (Vol. 12 No. 52 in this journal) took exception to several issues in a paper I co-authored with Julia B. Smith about high school size. They also provided some evidence to support their claims of "size bias," as well as another study that claims benefits for smaller schools. My comments here are organized around four issues relevant to research on the effects of school size on student outcomes. First, I respond to specific criticisms the authors raised about the Lee and Smith (1997) study. The second issue concerns the evidence offered by the authors in their study using similar data to those used by Smith and me. Third, I describe the context within which the Howleys' and I have interacted recently, as it may have motivated their criticism of my work. Fourth, I briefly discuss a broader framework within which I suggest that research linking school size to student outcomes should be seen.

Issue 1: The Howleys' Critique of the Lee and Smith (1997) Study

The Howleys summarized our three major conclusions by citing our exact words. Though they agreed with the first conclusion ("high schools should be smaller than many are"), they took issue with the conclusion that "high schools can be too small." They also found our offering an "ideal" high-school size as problematic. I organize my response around five areas mentioned in their micro-analysis of our work: (1) that NELS is unrepresentative of small schools; (2) that we emphasized excellence at the expense of equity in our conclusions; (3) that we inappropriately drew conclusions about an "ideal" size for high schools; (4) that our use of weights did not adequately adjust for the non-random sampling of schools in NELS; and (5) that rural schools were undersampled in NELS. I address another area, implied but not stated directly: (6) that our results are incorrect because our analyses were structured so differently from other studies about school size. Though their discussion of our work faults both the data we used (over which we had no control) and our analyses of the data and the conclusions we drew from our results (both of which we did control), their critique seems aimed at undermining our work and the respect researchers and policy makers should afford it.

Area 1: NELS School Sample

The NELS:88 school sampling frame started in 1988 with U.S. schools including 8th grades; no high schools were sampled. According to the National Center for Education Statistics (NCES), schools with 8th grades in them were sampled from a national frame of about 39,000 schools (public and private) drawn from a school data list complied by Quality Education Data, Inc. (QED), which "contained information about whether a school was urban, suburban, or rural" (NCES, 1994, p.23). The longitudinal NELS:88 design, with students surveyed and tested every two years (i.e., in 8th grade, 10th grade, and 12th grade), needed to capture the phenomenon that virtually all students changed schools sometime between 8th and 12th grade.

One difficulty of the NELS design, one that had to be confronted by many analysts who wanted to follow the sample of NELS:88 students through secondary school, was that

secondary schools were not directly sampled by NELS:88. Rather, high schools in the NELS study were those the NELS-sampled students chose. Although rich survey data about the NELS high schools (from principals and teachers) were collected, the NELS data files never provided school weights for high schools in the study (although weights for the base-year schools weights were included).

Our 1997 study focused on NELS high schools, although the Howleys' discussion focuses exclusively on base-year (i.e., middle-grade) schools. Their comparison of public schools in the Common Core of Data (CCD) and the NELS base-year school sample in Table 1 of their does suggest some pattern of undersampling of the smallest public schools. Their title suggests that this underrepresentation of the smallest middle-grade public schools in NELS introduces bias into any studies (like ours) that used NELS to investigate the effects of school size on learning. Although the NELS study sampled schools at the outset (i.e., when students were in 8th grade), and didn't sample high schools, the underrepresentation of small schools at the high school level may have persisted. Although virtually all students went to a different high school than the middle-grade school they attended, and high schools are typically larger than middle-grade schools, it may be that small school size is related to the area where the schools are located -- so that "smallness" or "largeness" may be somewhat consistent as students move to secondary school. What is not clear (and possibly misleading) is that such undersampling of the smallest U.S. public middle-grade schools would bias the results of such analyses.

Area 2: Privileging Excellence Over Equity

The Howleys claimed that Smith and I inappropriately used "authorial privilege" in our conclusions about equity and excellence, in that we did not provide sufficient justification for our conclusions. We explored four outcomes in our multilevel study: gains in achievement in reading and mathematics over the four years the students were in high school and the relationship between SES and achievement gains in these two subjects. We characterized these measures as "excellence" (achievement gains) and "equity" (the SES/achievement gain slope). Although these outcomes were estimated simultaneously in the same multilevel models, our presentation of results in the body of the paper in graphic form may have suggested that we analyzed these outcomes separately (the text of the paper did explain this). Numerical results, both weighted and un-weighted, were included in Appendices B-2 and B-3 of the study. Readers who looked only at the graphs might assume that equity and excellence were separate outcomes, as separate graphs presented the achievement gains (Figure 2) and the SES/achievement slopes (Figure 3) as functions of school size. These results, estimated as Hierarchical Linear Models (HLM), included statistical adjustments for both student characteristics (gender, minority status, SES, initial ability) and school characteristics (school SES, minority concentration, sector).

For any study, authors must consider carefully the audience to whom the results might be relevant. In our case, two distinct audiences seemed reasonable: policy makers and school professionals on the one hand and researchers on the other. The technical expertise of these two audiences is rather different. Our purpose in presenting results in graphic form was to make analyses that were quite complex more accessible for non-technical readers. The many, many inquiries we have received from school people and policy makers, starting from our first presentation of the study's results (at the 1996 AERA meeting in New York) and continuing up to the present time, suggest that the graphs told our "story" to this audience well. However, the graphic presentation was perhaps misleading. We included all results in Appendices to allow for full scrutiny by reviewers and readers with more technical understanding. It is unclear whether the Howleys scrutinized the numerical results at the end of the paper.

Among the several criticisms about our study raised by the Howleys, their claim that we seemingly disregarded equity disturbs me the most. Identifying and encouraging educational structures and organizations that are simultaneously linked to excellence and equity has characterized almost all of my research, from my dissertation (Lee, 1985), through my work with Anthony Bryk focusing on Catholic schools (Bryk, Lee, & Holland; 1993; Lee & Bryk, 1988; 1989), including several studies about school restructuring (summarized in Lee, 2002), and guiding my recent research on young children (Lee & Burkam, 2002). School factors that are associated with a socially equitable distribution of achievement without also being linked to higher achievement would imply that in such schools students of different SES levels or minority groups would achieve equally -- at low levels. That is, *equity without excellence* is not something we should encourage in schools. Social equity in the distribution of outcomes is only useful if everyone -- high-SES or low SES, minority or non-minority -- does well.

Although the conclusions in our paper were drawn from our findings, we meant them to rise beyond the results. They represented the meaning we drew from our work. The evidence for our conclusions lay in our results. Drawing conclusions is, quite rightly, "authorial privilege." These conclusions were located in the Discussion section of the paper, where authors typically interpret their findings more broadly. Had the reviewers of this paper felt we had "gone beyond the data," they surely would have required us to scale back our conclusions. That the Howleys don't agree with some of our conclusions does not render them groundless.

Area 3: "Ideal" Size Too Narrowly Defined

The Howleys also took issue with our identifying an "ideal" size range (600-900 students) for three reasons: (1) that our outcome set was too narrow, (2) that the smallest high schools were not included in the ideal range; and (3) that private schools were included in our study. Regarding the first reason, they suggested that our use of the term "ideal" was inappropriate because our study was narrow, focusing only on size effects on achievement. Our focus in the 1997 study *was* on gains in achievement; we included only NELS students with test scores at both 8th and 12th grade who had remained in the same high school. Our analysis was admittedly narrow in that sense; we explored size effects on achievement gains only for students whose exposure to their schools was maximized.

Many other important educational outcomes surely could be influenced by school size, and I have pursued these in several studies. My colleague and I explored dropping out as a function of school social organization and structure (size and sector) in a subset of NELS high schools in urban and suburban areas (Lee & Burkam, 2003). Another colleague and I used multilevel methods to explore size effects on teachers' attitudes in Chicago elementary (K-8) schools (Lee & Loeb, 2000). A qualitative study compared large and small public high schools in terms of social relations and curriculum (Lee, Smerdon, Alfeld-Liro, & Brown, 2000). It is surely possible that different studies may come to different "ideal size" conclusions, based on the dependent variable of interest. We clearly defined the outcomes in the 1997 study: achievement gain (and its equitable distribution by SES) over the four years students spent in high school, and we selected our sample of students accordingly. Readers would recognize its focus on achievement. We suspect that school professionals and policy makers would "privilege" achievement over other outcomes (if, perhaps, "ideal sizes" differed for other outcomes), especially in the contemporary climate of achievement-related mandates from *No Child Left Behind.*

The second reason for the Howleys objection to our "ideal" size designation centers on our finding that secondary schools smaller than 600 were not "ideal" in terms of size. I believe that nationally representative longitudinal data provide an excellent (perhaps the best) venue for policy-

relevant research in education. The numbers of small high schools in our study using NELS data are reasonable. The numbers of schools in the various size categories (from Table 1 of Lee & Smith, 1997) do differ, but the smallest category (enrolling 300 or fewer students) contains 75 schools (and 912 students in those schools). The next-smallest category (301-600) contains 67 schools and 830 students. The Howleys statement that there is "much more error embedded in findings, and therefore, in conclusions, about smaller schools than is acknowledged" (p.10) seems groundless. Whatever error accrues is reflected in statistical testing (reported in Appendix B-2) and not in parameter estimates. If the Howleys are referring to sampling error, this doesn't seem problematic; the numbers of small schools and students are actually substantial.

Regarding the third reason, the Howleys suggest that many of the schools in our "ideal size" range (600-900) are private schools, and that this might bias our findings that favor schools in that range. There *are* more private schools in that size category; but the large majority (75.5 percent) of the 148 schools in that category are public. Moreover, schools in the smaller size categories are almost all public (95 percent of schools enrolling 300 or fewer students are public; 92.5 percent of schools with 301-600 students are public -- see Lee & Smith, 1997, Table 1). The Howleys argue that "the issue of size is arguably confounded with sector" (p.13); I disagree. All of our HLM analyses included statistical adjustments for school sector (Catholic and elite independent schools each compared to public schools). Moreover, our HLMs also included statistical adjustment for school average SES and minority composition, on which public and private (as well as small and large) schools differ.[1] The reason to include such controls is precisely to avoid such a bias.

Area 4: Weighting

The concept of weighting in multivariate analysis is theoretically simple: weights are the inverse of the probability of being sampled. Weights adjust for non-random sampling; over-sampled units get weighted down and under-sampled units get weighted up. The concept is simple, but the process of creating weights is not. Researchers typically rely on those who collect the data to supply weights. Virtually all NCES longitudinal datasets require the use of weights for multivariate analysis, to compensate for non-random sample selection. Although NELS students as 8th graders were selected close to randomly *within schools,* the original sample of schools was not random. Not only was the original 8th-grade school sample stratified by location, certain types of schools were purposely oversampled (i.e., private schools). All documentation that accompanies NELS data (e.g., NCES, 1994) suggests that analyses must be weighted. Multilevel analyses (in our case, students nested in schools) allow weighting at different levels. Because of the original near-random sampling of students within schools, we assumed that samples of students within high schools was also close to random (without evidence to the contrary). Thus, the within-school portion of our HLMs were unweighted. However, we needed weights for the between-school HLM analyses.

As quantitative researchers like Smith and me recognize, the great value of nationally representative longitudinal data in strengthening generalizable causal inferences and the also necessity of using multilevel methods to conduct school-effects studies, we faced a serious dilemma. In our several published studies using NELS secondary schools, we described several decisions in choosing our samples of students and schools. For the 1997 study, we selected only high schools with at least 5 original NELS students in them.[2] We also included only students who were 12th graders in 1992 (i.e., those who had neither dropped out, transferred, nor repeated a grade in high school), and we constructed our own school weights (which we used in all of our high-school studies with NELS data). Not being sampling statisticians ourselves, we sought advice from colleagues at the University of Michigan's Institute for Social Research (ISR), which is internationally recognized for expertise in sampling theory. After the publication of our first NELS high-school

study (Lee & Smith, 1995), other NELS researchers asked us to "lend" them our weights; we declined. Rather, we explained how we had constructed NELS school weights and suggested they make their own.

The Howleys stated that "the National Center for Education Statistics has in fact recommended against using school-level weights for any but school-level analyses" (pp.10-11); exactly what we did. They also stated (p.10) that "despite weighting and adjustments of mean standard errors for design effects, much more error is embedded in findings, and therefore conclusions, about smaller schools than is acknowledged." Why? We included no adjustment for design effects; 2-level HLMs render the need for design effects unnecessary with NELS (because of the parallel between students- within-school sampling and analysis). If there were larger errors accruing to estimates for smaller schools, as the Howleys suggest (but which I question), this would influence statistical testing rather than parameter estimates. The major results in our study, presented in graphic form, did not report statistical testing. However, the p-values associated with statistical testing of size comparisons are available in the Appendices (to which the Howleys do not refer). The Howleys imply that somehow we have tried to mislead readers; this I disagree with most strenuously.

Neither Smith nor I are sampling statisticians, nor to my knowledge are the Howleys. Thus, we all should follow the recommendations from NCES about analyses of their datasets. We were certain that school weights were necessary, and we did our best to create weights based on the information available about the high schools in first and second follow-ups of NELS. We checked our procedures with colleagues who knew more about sampling and weights than we did. We weighted our analysis at the school level, within a multilevel analysis framework.[3] Although researchers could surely question the method we used to create our school weights, we have not heard such criticism. Moreover, as we worried that our results might be influenced by the school weights we created, we reported the size effects from unweighted HLMs in Appendix B-3 of our paper. The pattern of results did not change, although the magnitude of some coefficients did.

Area 5: Why So Few Small Rural Schools?

The Howleys' discussion of base-year NELS schools is actually not directly relevant to our study, in that we did not examine base-year school effects in this study. Julia Smith and I did publish a study of that NELS students as 8th graders (Lee & Smith, 1993). In that case, we felt it was inappropriate to explore school size directly, as the variation in the grade-level composition of the base-year NELS schools clouded the issue (e.g., K-8, K-12, 6-8, 7-9). In that study, we captured "size" with the number of 8th graders in the school. In their analyses of NELS base-year data, the Howleys also used 8th-grade cohort size.

However, even at the high-school level in our study, there were sufficient numbers of schools in even the smallest size categories to sustain analysis. It is unclear why an under-representation of smaller high schools (if it exists) would bias the results of our study. Their use of the word "bias" in the title of their paper suggests that results of such a study would not be correct. Were that the case, I wonder why the Howleys themselves used the NELS data for analyses. Perhaps there is an under-representation of small middle-grade schools in the NELS base-year school sample. Why, however, would this lead to biased results?

From the totality of their paper (particularly the Discussion), I infer that that the Howleys believe that small rural schools are actually quite different from (and probably much better than) other small schools. They imply that the effects of school size might be different for rural than suburban or urban schools. This hypothesis, which the descriptive results presented in Table 6 of their study suggest, could be tested directly using NELS data to explore size-by-urbanicity

interactions. With the same data and structure of our 1997 study, one could create a series of interaction terms for the size categories and test them, just as we tested size-by-school SES and size-by-minority composition interactions.

In their own analyses of base-year NELS data, they did not include school-level urbanicity-by-grade cohort size interactions, nor did their analysis include even a first-order dummy-variable indicator for rural and small-town schools. It is not appropriate to proclaim as fact an interesting and testable hypothesis. Small rural schools may, indeed, differentially influence students' achievement gains, the social distribution of achievement, or many other outcomes. The technology to test interactions is well developed (e.g., Cohen, Cohen, West, & Aiken (2003), Chapters 7 and 9). The Howleys obviously understand interactions, as they included them in their own study. If small school size is hypothesized to be differentially effective for schools in rural areas, the data should support this statistically.

Area 6: Structure of Our Analyses

Multilevel questions, multilevel methods. A large volume of the research on the size of educational units has explored data aggregated to the school level. That is, such studies have chosen to structure their analyses with "school" (or perhaps "district") as the single unit of analysis. In such analyses, student outcomes (e.g., achievement, achievement gains, dropout rates) have also been aggregated to the school level, as have other student characteristics (e.g., student-SES, gender, ability, minority status). In several instances, SES has been captured as many schools and districts do, by the proportion of students in the school receiving lunch subsidies. Though this approach may seem to make intuitive sense -- after all, school size is inherently a school characteristic — a school-level analysis is actually inappropriate for several reasons. First, student outcomes (and background characteristics) accrue to individuals. When these variables are aggregated to the school level, they mean something different (creating a mistake that is called either "ecological fallacy" or "aggregation bias"). More importantly, aggregation essentially discards the large majority of the variance in the outcome of interest (in U.S. data on achievement, typically only 20-25 percent of the total variance lies systematically between schools). Using only school-level aggregates essentially discards 75-80 percent of the variation. Moreover, by doing that, researchers are unable to explore within-school relationships between achievement and student background -- essentially relegating all exploration of inequality to between-school analyses. More than three decades ago, Jencks and his colleagues informed us that the large majority of the inequitable distribution of educational resources lies within, not between, schools (Jencks et al., 1972). Arguments about the proper structure of what has come to be called "school effects research" have been made frequently in other venues, as well as in the Lee and Smith (1997) study. Readers who are interested in this issue should surely consult the major source (Raudenbush & Bryk, 2002).

To me, the question of appropriate methodology is simple: if you are asking a multilevel question, you need multilevel methods. Many questions in educational research are inherently multilevel; children experience their education in groups: reading groups, classrooms, schools, districts. The question of how school size influences student outcomes is inherently multilevel. Thus, statements about the consistency of findings in school-size studies rings a bit hollow. Almost all of those studies were conducted using data aggregated to the school level. Exceptions are the Howleys study described in their paper and the Bickel and Howley (2000) study.

Distribution of school size. Perhaps a more intuitive (but equally important) technical issue surrounds the form of the independent variable of focus. Most school size studies (especially those that focus on schools in a particular state), use size as a continuous variable. However, school size is rarely normally distributed. Rather, it is positively skewed, with a long right-hand tail (similar to the

distribution of family income). There are generally more small than large schools (even though most students attend larger schools). Such a non-normal distribution typically results in a non-linear relationship between size and achievement (even if achievement is normally distributed, which it usually is). A glance at Figure 1 in the Lee and Smith (1997) study shows a distinct non-linear relationship. Multivariate analysis techniques such as OLS regression and HLM assume normally distributed continuous variables (or dummy-coded independent variables) and linear bivariate relationships.

        Quantitative researchers exploring the size/achievement relationship have three options. They can either (1) transform the school size variable to make it normally distributed (typically a logarithmic transformation will do the trick); or (2) create a series of categories and use them as dummy-coded indicators in the analysis; or (3) leave the continuous variable non-transformed and include a quadratic term in the analysis to test for non-linear effects. In our 1997 paper we pursued the second option, precisely because we wanted to know "which size high school works best?" In other studies (Lee & Smith, 1993; 1995; Lee, Smith, and Croninger, 1997) we chose the first option, using size in its logarithmic transformation. Many other studies of school size have used school size (or grade cohort size or even school district size) without correcting for the non-normal distribution. To non-technical readers, this may seem like an esoteric point, but to me it is not. Many of these studies have also used data aggregated to the school or district level (e.g., Howley, 1995).

        Although I have discussed some of these issues at length, my purpose here is not to engage in a lengthy debate about the best (or acceptable) way to investigate the effects of school size on student outcomes. Rather, I have responded to what I consider to be several inappropriate criticisms directed to a study I stand behind strongly. I contend also that these two methodological issues undermine the validity of many school-size studies. Later in this paper, I offer a possible explanation for what I consider to be unwarranted criticisms raised about our study, when I describe the context of my contact with Craig Howley.

Issue 2: Research Provided in Their Article

School and Grade--Cohort Size in Middle-Grade Schools

        Similar to our study with the base-year NELS data (Lee & Smith, 1993), in the study described in their article, the Howleys used the indicator of the number of 8th graders in the NELS' base-year schools, rather than the total enrollment of the school (i.e., school size). *However*, they refer often to *small schools*, when they mean *schools with small 8th grades*. I can think of contexts where a seemingly small 8th grade cohort might exist in a relatively large school: if the school offered a wide grade range (e.g., K-12 or K-8). The distribution of grade grouping by school enrollment size in NELS base-year schools (including private schools) is described elsewhere (see Figure 2.2, p.23 in Lee, 2002). Clearly, the base-year NELS schools offered many different grade configurations.

        The authors focused only on the public NELS middle-grade schools, grouping them into those they labeled "small schools" and "large schools," using the cut-point of 84 (i.e., they used the CCD to determine that the average middle-grade public school in the U.S. enrolled 84 8th graders in 1987-88). However, they then referred to "smaller or larger school size" (p.19). More accurately, they should refer to "schools with smaller or larger 8th-grade cohorts." My point here is simple: grade cohort size and school size are different structural features of schools. Either is interesting, but they are not the same thing. They are especially different in schools that include 8th grades, as the grade groupings are so varied.

Rather than offering policy conclusions about the size of schools that enroll young adolescents, the results of the Howleys' study might be more useful to policy makers interested in decisions about how the schools that young adolescents attend should be configured (i.e., the grades they should include). Their results say something positive about schools with fewer 8th-graders; quite likely these *are* schools that include more grade levels. Such schools are more likely to be located (and results more positive) in rural areas and small towns. Much has been written recently about troubled large middle schools or junior high schools, many of which are located in large cities. There is new research supporting the K-8 organizational form.

Process vs. Structure

In their cross-sectional analysis of base-year data from NELS:88 collected on 8th grade students in middle-grade schools, the Howleys tell us that they are interested in the "structural ramifications of size" rather than "hypothetical influence of size on process" (p.14). To me, that means that rather than attempting to investigate how students who attend schools of different sizes are influenced by their schools' sizes, they are simply exploring issues of selectivity. i.e., which types of students attend schools of different sizes (or with 8th grades of different sizes). Because they explore data from the base year of NELS, they may not investigate achievement gains.

However, they have quite appropriately included a statistical control as a proxy measure of students' ability -- their self-reported grades since 6th grade -- the same statistical control that Smith and I used in our 1993 study using NELS base-year data. They refer to this as "prior achievement" (p.16), which it is not. The majority of research on school size has used such a design -- cross-sectional data with schools or districts as the unit of analysis. The distinction between process and structure, given their multilevel analyses and inclusion of a proxy control for ability, is unclear. They seem to be backing away from inferring causality in the introductory sections of their paper, but their analyses and conclusions seem to me to be constructed to infer causality. Which is it?

Centering Decisions in Multilevel Models

For their multilevel analysis, the Howleys used the SPSS mixed-models analysis methodology, whereas we made use of HLM (Raudenbush & Bryk, 2002). They also included adjustments for design effects, something that is not needed with HLM; the stratification in sampling (students within schools) is the same as the stratification in analysis. As I am unfamiliar with this particular SPSS procedure, I do not make direct comparisons between their analyses results and ours. However, in their text and in footnote *a* of Table 10, they suggest that they followed the same centering procedures as we did in our 1997 study. As recommended by Raudenbush and Bryk (2002), we centered the intercept and the SES/achievement gain slopes around the grand mean, and other control variables (gender, ability, minority status) around the school means.

In their analyses they have investigated as outcomes at Level-2 not only the intercept (8-grade achievement) but two social distributional outcomes: the SES/achievement slope and the self-reported grades/achievement slope. If these slopes are to be investigated at Level-2 as functions of school size (as their models suggest), then these slopes must be centered around the grand mean and they must be allowed to vary between schools. These are standard centering decision rules in HLM. As they claimed to have followed the same procedures we have, one would assume that their models would be similar (which they seem not to be).

Structure of Their Multilevel Models

         In Table 10, the authors present results of a multilevel analysis of 8th-grade mathematics
achievement. Although I am very familiar with multilevel analyses (and teach courses in this
methodology), I find it difficult to make sense of their results. For example, what is the within-
school model, and what is the between-school model? Perhaps these results could be presented
more clearly. Do PRIOR2 and WHITE2 represent school-level aggregates of within-school variables
that measure students' race and prior achievement? From footnote *c* of Table 10, I surmise that
"size" is divided into deciles (absent decile 1) and treated as a continuous variable. Is this still grade
cohort size? Why use the deciles rather than the continuous measure? What is the distribution of this
9-level measure? In our NELS study, our decision to use school-size categories was made because
(a) the distribution of high-school size was definitely not normal and (b) we wanted to identify an
"ideal" size. The Howleys have also categorized school size (9 categories), but they have used this as
a continuous variable. They report that this is the same measure they have used in their analyses in
Tables 8 and 9 (footnote *c*, Table 10). However, the analyses in Tables 8 and 9 did use school size
categories, whereas in the results from the multilevel analyses presented in Table 10 they appear to
have used this as a continuous variable. We have no idea whether this variable, used this way,
satisfies the distributional requirement of their methodology.[4]

Summary of Questions About Their Analyses

         Query 1: Might there be non-linear size effects? Actually, this question is at the heart
of the Lee and Smith (1997), and our findings on this issue are those that the Howleys objected to
most strongly. Readers would not know the answer to this question from the analyses offered here.
The Howleys used a 9-level continuous variable to represent grade-cohort size in their study. Why
were these categories used? They did not show us the distribution of this variable, nor did they
explore the possibility of a non-linear cohort size effect.
         Without knowing if the quasi-continuous variable they used as an indicator of 8th-grade
cohort size is normally distributed, we cannot judge whether estimating a linear effect of grade-
cohort size on achievement is appropriate, or whether this unusual variable has in fact masked a
possible non-linear effect. The distribution of school size in U.S. schools (elementary, middle-grade,
or secondary schools) is definitely non-linear; there are many more small schools than larger schools.
Given that our 1997 study indicated a definite a non-linear effect, and because the Howleys were
particularly critical about that finding from our study, I believe that this issue must be addressed
before we can be confident in their results and conclusions. They state (p.26), "contrary to the
assertion of Lee and Smith (1997), these results do not disclose any lower limits for school size."
First, we did not *assert* this; rather, we supported our conclusions on this issue with empirical results.
Second, the Howleys study surely did not disclose any lower limits for schools size (a) because they
did not structure their analysis so such disclosures would be manifested, and (b) because they didn't
actually study school size.
         Query 2: Is school size equivalent to grade-cohort size? Although the issue of the
link between grade-cohort size and student achievement is interesting, particularly in middle-grade
schools, it is a different issue from school size. Several studies of school size have used grade-cohort
size as a size proxy, precisely because schools could contain different grade configurations (or to
combine elementary, middle, and high schools in a particular state in the same analysis). The
Howleys made this decision in their study, reasonable one given the substantial variation in grade
levels in U.S. middle-grade schools sampled in NELS. However, their equating of grade cohort

effects with school size effects is inappropriate. They should change their language, and also discuss the policy implications based on *different* grade configurations for U.S. schools that enroll 8th graders.

Query 3: Are policy conclusions about school size appropriate? Even if we could have confidence in the Howleys results, are "efforts to build and sustained smaller schools... warranted on the basis of these findings," as they state (p.26)? Their study was not focused on school size or small schools, it was focused on schools with different sized grade cohorts. Moreover, the focus of their study was on middle-grade schools, but the conclusions offered would seem to apply to schools of all levels. It could very well be the case that size effects at one level of schooling were not generalizable to another. To their credit, the final paragraph of their paper does discuss grade-cohort size; however, it refers to high schools rather than the middle-grade schools they studied.

Issue 3: The Context

Normally, in the academic world we take critiques of our published work in stride -- believing that reasonable people can disagree. The Lee and Smith (1997) paper has been cited widely, and I have been asked about it often by school and district personnel who are in positions to make important decisions about how big or small their high schools should be. These queries have led me to recognize that high school size (and research about it) is more relevant to policy makers than much of my research on other topics. In fact, the relevance of this issue has extended most recently into another policy arena: the courts.

Within the last year, the Howleys and I were invited to serve as expert witnesses on opposite sides of a lawsuit focusing on high school size in Lincoln County, West Virginia. I agreed, quite reluctantly, to serve as an expert for the defense. The State of West Virginia had taken control of the schools in Lincoln Country in 2000 due to extreme poverty in the county and very weak school performance in the county's schools compared to the rest of the state. Last year the state recommended that four very small high schools be closed and one larger higher school (with a projected enrollment of about 800 students) be constructed -- a classic case of school consolidation.

An advocacy group, "Challenge West Virginia," sued the State to enjoin it from pursuing these actions, and the Howleys agreed to serve as expert witnesses for prosecution. Even though depositions have been collected and the trial postponed several times, the case may be over without going to trial. Earlier this year the judge assigned to the case ruled in favor of the State, and construction of the new high school is underway (scheduled to be opened for business in the 2005-06 school year). The Lee and Smith (1997) study was offered by the State in support of their actions. The Howleys work (including this new article) was offered as evidence. A few of my other studies on the topic of schools size were also offered in evidence.

Obviously, a legal setting is by nature adversarial. In this context, it is difficult for me to overlook both the timing and the unusually critical nature of the Howleys 2004 article. I have seldom experienced such micro-level criticism of my work. I appreciate the effort by *Education Policy Analysis Archives* and its editor, Gene V Glass, to present readers with different viewpoints about what seems to have developed as a contentious debate about an issue of educational policy. In fact, I would like readers to see this issue in a somewhat different context.

Issue 4: A Causal Link?

It is quite appealing in educational research to focus on issues that translate into direct policy levers over which schools, districts, states, and nations actually have control. This is the

essence of policy-related research. The enrollment size of a school represents such a lever, in that schools are built (and money allocated) based on student head counts. Thus, it may seem reasonable for policy makers to ask, "Which size school works best?" Of course, this requires that those exploring the issues define what "works" means; not unusually, this has been defined in terms of student achievement, or even more appropriately, student learning. If one wants to explore a relationship between school size and student learning, moreover, it may be reasonable to define learning in terms of how much the same student's achievement changes over the period he or she has been enrolled in his or her school.

However appealing might be the policy issue that links school size and student learning, researchers might challenge the validity of such a question. Is it really appropriate to posit a causal link between these two factors? I agree with the Howleys suggestion that research and writings that focus on small schools often confound issues of pedagogical and curricular changes and size per se. However, this suggestion raises an even more important and appropriate question: "Why would anyone think that school size would exert a direct effect on student achievement or learning?" Julia Smith and I raised this same issue toward the end of our 1997 article. We stated: "...we suspect that size acts as a facilitating or debilitative factor for other organizational forms or practices that, in turn, promote student learning" (p.218).

I teach several courses that focus on quantitative methodology for conducting social science research. From almost the first day in any of the courses I teach (or those I took in graduate school), fledgling researchers are cautioned that "correlation does not imply causality." This caution is typically followed with a few examples that illustrate this point, usually with an obvious "third variable" that might explain a spurious link between the two variables in question. We researchers try to keep these cautions in mind, even as we frequently conduct solid correlational research. We are mindful of the need to discount alternative explanations for our findings -- by introducing appropriate statistical controls, using longitudinal designs, employing appropriate statistical methods, and many other ways to increase the validity of our studies.

In the case of efforts to link school size with student outcomes (particularly learning), we would be wise to revisit the cautions about correlation and causality. Were we really to identify a residual causal link between school size and student learning (i.e., gains in achievement over the time students have attended the schools), we might want to control for other school and classroom characteristics that might be confounded with size -- variables that describe, for example, the curriculum, instruction, student engagement, or social relations among school members.

Our 1997 study did not include statistical adjustment for such forms or practices. Our controls were limited to those describing demographic characteristics of students (SES, race/ethnicity, gender) and structural or compositional measures of schools (average SES, minority concentration, school sector). That is, we mainly included statistical controls for *selectivity bias*. In other research (Lee, Burkam, Chow-Hoy, Smerdon, & Geverdt, 1998; Lee & Smith, 1995; Lee, Smith, & Croninger, 1997), we did find residual school size effects even after taking into account many other factors that captured the social and academic organization of schools. However, we never claimed that our research models were exhaustive. Our major focus in those studies was on issues other than size.

Why would we expect school size to influence student learning (or other student outcomes)? It seems logical to think that basic organizational structures are different in smaller than larger schools. School members may relate to one another through more productive and sustained encounters in smaller schools. The ability to offer a full curriculum may be constrained in very small schools. Small schools in rural areas may have trouble attracting faculty with sufficient expertise to prepare students for a productive future. It may seem reasonable, even logical, to differentiate

students by ability in larger schools, thus facilitating social stratification through ability grouping and tracking. The list could go on and on.

The important issue in such studies is unlikely to be school size per se. Rather, size facilitates or constrains how people relate to one another, the offerings that schools can muster, the web of human relationships that surrounds adults' efforts to facilitate the academic development of the young people they serve. The very fractious court case in West Virginia may be missing the point. And we who study school size as though it influences student outcomes directly may also be missing something very important.

Notes

1. Because the private school effects are captured by two dummy-coded variables (one coded 1 for Catholic schools, 0 for public schools, another coded 1 for elite private schools, 0 for public schools), technically the size effects in our study are for schools who are coded 0 on all school-level control variables (average SES, school minority concentration, the two sector dummies). That is, the size effects reported in our study are for public schools with average SES and minority enrollments below 40 percent. Even if the private schools were smaller than the public schools (but mostly not the very smallest schools), the size effects in our study are estimated net of school sector, average SES, and minority concentration.

2. NCES made the identical decision when they created the *High School Effectiveness Study* (HSES), that included only high schools attended by NELS students that were (a) in the 30 largest MSAs (Metropolitan Sampling Areas) in the U.S. (i.e., rural schools were excluded), and (b) enrolled at least 5 original NELS students. In these high schools, NCES staff increased within-school sample sizes (which they tested and surveyed). See Scott et al. (1996) for more detail about HSES sampling.

3. NCES did provide school weights with the HSES data (see footnote 2)- - in fact they provided three of them. My research team and I were asked to conduct a study using HSES data and write a working paper for them (Lee et al, 1998). Because the HSES data specifically excluded rural schools, we believed that they were not ideal for studying the full range of school size effects. The Lee and Burkam (2003) paper used the HSES data as well, where size was also explored.

4. Although it is not relevant to the issue of school or grade-cohort size, I find the Howleys' interpretation of the SES2 X PRIOR1 interaction confusing. Because this interaction effect is positive, I would interpret this as indicating that schools with higher average SES are particularly stratifying, in that the relationship between 8th-graders' self-reported grades and their mathematics achievement is even stronger than in schools of lower average SES.

References

Bickel, R. & Howley, C. (2000). The influence of scale on school performance: A multi-level extension of the Matthew Principle. *Education Policy Analysis Archives, 8*(22), May 10, 2000.

Bryk, A.S., Lee, V.E., & Holland, P.B. (1993). *Catholic Schools and the Common Good.* Cambridge, MA: Harvard University Press.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Science* (3rd Ed.). Mahwah, N.J.: Erlbaum.

Howley, C. (1995). The Matthew Principle: A West Virginia replication? *Education Policy Analysis Archives, 3*(18), November 15, 1995.

Howley, C. B. & Howley, A. A. (2004, September 24). School size and the influence of socioeconomic status on student achievement: Confronting the threat of size bias in national data sets. *Education Policy Analysis Archives, 12*(52). Retrieved [date] from http://epaa.asu.edu/epaa/v12n52/.

Jencks, C., Smith, M., Acland, H., Bane, M.J., Cohen, D.K., Gintis, H., Heyns, B., & Michelson, S. (1972). *Inequality: A Reassessment of the Effect of Family and Schooling in America.* New York: Basic Books.

Lee, V.E. (2002). *Restructuring High Schools for Equity and Excellence: What Works.* New York: Teachers College Press.

Lee, V.E. (1985). *Investigating the Relationship Between Social Class and Academic Achievement in Catholic and Public Schools: The Role of the Academic Organization of the School.* Cambridge, MA: Harvard Graduate School of Education, unpublished doctoral dissertation.

Lee, V.E. & Bryk, A.S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education, 62*(3), 172-192.

Lee, V.E. & Bryk, A.S. (1988). Curriculum tracking as mediating the social distribution of high school achievement. *Sociology of Education, 62* (2), 78-94.

Lee, V.E. & Burkam, D.T. (2003). Dropping out of high school: The role of school organization and structure. *American Educational Research Journal, 40*(2), 353-393.

Lee, V.E. & Burkam, D.T. (2002). *Inequality at the Starting Gate: Social Background Differences in Achievement as Children Begin School.* Washington, D. C.: Economic Policy Institute.

Lee, V.E., Burkam, D.T., Chow-Hoy, T., Smerdon, B.A., & Geverdt, D. (1998). *High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates.* National Center for Education Statistics Working Paper Series (Working Paper No. 98-09). U.S. Department of Education, Office of Educational Research and Improvement.

Lee, V.E. & Loeb, S. (2000). School size in Chicago elementary schools: effects on teachers' attitudes and students' achievement. *American Educational Research Journal, 37*(1), 3-31.

Lee, V.E., Smerdon, B.A., Alfeld-Liro, C., & Brown, S.L. (2000). Inside large and small high schools: Curriculum and social relations. *American Educational Research Journal, 22*(2), 147-171.

Lee, V.E. & Smith, J.B. (1997). High school size: Which works best, and for whom? *Educational Evaluation and Policy Analysis, 19*(3), 205-227.

Lee, V.E. & Smith, J.B. (1995). Effects of school restructuring and size on gains in achievement and engagement for early secondary school students. *Sociology of Education, 68*(4), 241-270.

Lee, V.E. & Smith, J.B. (1993). Effects of school restructuring on the achievement and engagement of middle-grade students. *Sociology of Education, 66*(2), 164-187.

Lee, V.E., Smith, J.B., & Croninger, R.G. (1995). How high school organization influences the equitable distribution of learning in mathematics and science. *Sociology of Education, 70*(2), 128-150.

National Center for Educational Statistics (1994 September). National Education Longitudinal Study of 1988. Second Follow-Up: Student Component Data File User's Manual (NCES 94-374. Washington, D. C.: U.S. Department of Education, Office of Educational Research and Improvement.

Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd Edition). Newbury Park, CA: SAGE.

Scott, L.A., Ingels, S.J., Pulliam, P., Sehra, S., Taylor, J.R., & Jerovic, D. (1996). *National Education Longitudinal Study of 1988. High School Effectiveness Study, Data File User's Manual.* Washington, D. C: National Center for Education Statistics, U.S. Department of Education, March 1996.

About the Author
Valerie E. Lee
School of Education
University of Michigan
Email: velee@umich.edu

Valerie E. Lee is Professor of Education at the University of Michigan and a Faculty Associate at the University's Institute for Social Research. She teaches courses in quantitative research methods (including multilevel analysis) and the sociology of education. Her research focuses on issues of social equity in education, and her most recent work has considered these issues in the early elementary grades, using a multi-method approach to address the same research questions.

Education Policy Analysis Archives   http://**epaa.asu.edu**