MEASUREMENT AND STATISTICAL ISSUES IN MULTILEVEL RESEARCH ON SCHOOLING

Kenneth A. Sirotnik
and
Leigh Burstein

CSE Report No. 261

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

1986

The project presented, or reported herein, was performed pursuant to a Grant from the Office of Educational Research and Improvement/Department of Education (OERI/ED). However, the opinions expressed herein do not necessarily reflect the position or policy of the OERI/ED and no official endorsement by the OERI/ED should be inferred.

ABSTRACT

Measurement and statistical problems in the analysis of multilevel data should be approached through an identification of the appropriate set of substantive research questions at and within various levels and the specification of appropriate models for analyzing multilevel data. Simply selecting a unit of analysis and a level of analysis does not adequately address such problems as contextual or group effects, the selection of an aggregate statistic to represent group effects, and adoption of a statistical procedure that will consider effects at all pertinent levels. A further concern is that different multilevel approaches in analysis can change the order of importance of constructs and alter the total variation accounted for.

Measurement and Statistical Issues in Multilevel Research on Schooling 1

Kenneth A. Sirotnik
Leigh Burstein

Picture this abbreviated and hypothetical research scenario: There are 2 school districts with 8 schools (4 elementary and 4 secondary) in each district that are available in order to study the following research question:

What is the relationship between teacher job satisfaction, on the one hand, and the work climate in schools, the years of teaching experience of staff members, and school size, on the other?

The constructs are operationalized in a short teacher survey questionnaire containing agree-disagree-type (Likert scaled) items pertaining to their job satisfaction and their perceptions of various aspects of the organizational climate in their schools. Teachers are also asked to supply several pieces of demographic information including the number of years they have been teaching. The number of FTE faculty for each school is also recorded as a measure of school size.

After data are collected, psychometric analyses of the responses to the teacher satisfaction and organizational climate items are conducted. Since the fifteen work satisfaction items were conceptualized <u>a priori</u> as a single dimension and had prior empirical support, only an internal consistency check is done. Coefficient alpha and item-total scale correlations are computed for the 650 teachers in the study. The alpha reliability estimate is .82 and the correlations are all non-zero and positive, indicating a reasonably reliable (internally consistent) scale of teacher job satisfaction.

The organizational work climate items were intended to tap several dimensions (cooperation, autonomy, conflict resolution, motivation, communication, decision-making, trust, and so forth) and were newly written (or adapted) for this purpose. Thus, the responses of all teachers to the total pool of 100 climate items are factor analyzed, the resultant factors interpreted, and climate scales constructed in light of both the prior concepts and the empirical associations among the items. Three dimensions emerge in this analysis: "principal leadership" (e.g., the principal as an effective problem-solver, motivator, decision-maker, communicator, and so forth); "work facilitation" (e.g., the elimination of red tape, excessive rules, and availability of needed resources); and "staff

cohesiveness" (e.g., staff support, warmth, consideration, trust, and the like). Teacher scores on these dimensions are derived by summing their responses to the items on the 1-5 point agree-disagree scale. The resulting alpha re liabilities computed for the 650 teachers for each scale are all in the eighties.

To explore the research question, the following statistical analyses are carried out. Each teacher has six data points: a "job satisfaction" score, three "organizational work climate" scores, the number of years they have been teaching, and a school size score. First, simple descriptive univariate (means, standard deviations, frequency distributions) and bivariate (zero-order Pearson correlations) analyses are conducted to check distributional requirements and to get a feel for the directions and the magnitudes of relationships among the variables. To more adequately assess the research question in light of its multivariate nature, a regression analysis is conducted with job satisfaction as the "criterion" variable (the "Y") and the remaining five variables as "predictors" (the "Xs"). Overall, the relationship between these variables and job satisfaction is moderate (multiple R = .65), accounting for just over 40% of the variation in job satisfaction. School size is found to be the most powerful pred
ictor of job satisfaction, followed by work facilitation, staff cohesiveness, and principal leadership; years of teaching experience appears to have little predictive value. These results are basically supported by the bivariate correlations of the criterion with the predictor variables.

The Dilemmas

Now, what is wrong with this little hypothetical scenario? There are many conceptual, design, and analytical aspects of it that can be seen as problematic. Examples are the following: the conceptual basis for selecting only the few constructs involved and their definitions, the use of a paper and pencil survey instrument rather than more in-depth methods (e.g., observation and interview), and the use of rather simple analysis procedures instead of, say, structural equation modeling and the more sophisticated interpretations that can follow.

But these are not the points of our admittedly oversimplified example. What else is wrong? <u>The 650 teachers have been treated as though they are independent of their naturally occurring contexts.</u> To say it another way, an assumption implicit in the analyses is that the conceptually relevant organizational levels at which data have actually been collected - district, grade level (elementary or secondary), and school - are somehow inconsequential both for the <u>measurement</u> issues in scale development and for the <u>statistical</u> analyses that follow.

Taking seriously the possibility of profound effects due to aspects of these organizational (group) levels opens up a virtual Pandora's box of measurement and statistical dilemmas. These dilemmas eventually manifest themselves in choices regarding appropriate units and levels of analysis. For example, consider the data generated by teachers' responses to the job satisfaction items. One possible way is to analyze the individual teacher responses without taking into account grouping or level (see next section) variables such as school or district. This analysis can be referred to as a total or overall analysis, 2 and it is the one most often used and least often, if ever, relevant.

Another way of analyzing the data is to treat the individual teacher responses within each group (e.g., school) separately--a within group analysis. Moreover, one could average these analyses together for a single, hopefully representative, pooled within group analysis.

Another choice would be to use the teacher data in each group to calculate group level or aggregate statistics--for example, school means on teacher job satisfaction scores. These group means could then be used as units in what could be called a between groups analysis. Much useful information also occurs naturally as a group-level variable without aggregation--for example, the variable "school size" measured as the number of FTE faculty.

With these several distinctions in mind, it is now possible to be more specific about the measurement and statistical issues in this multilevel research context. For example:

* Are there contextual (or group) effects due to the different districts, schools, and grade levels that should be accounted for in the analyses?

* If yes, is this matter conceptually based or simply an empirical choice due to the likelihood of sizeable variance components associated with those grouping effects?

* In either case, what unit(s) should be used in the psychometric analyses (internal consistency and factor analyses) of the various instruments (work climate and job satisfaction): individual teachers across groups (e.g., schools or districts); individual teachers within groups; or the groups themselves (a "between" analysis) using aggregate measures (e.g., means) based on the individual data within the groups?

* Since different choices can lead to different results, how does one resolve, for example, a highly internally consistent job satisfaction scale using a between schools analysis with a rather small coefficient when analyzing the data from individuals within schools? Moreover, what does one make of findings within schools that suggest that, for some schools, teachers' responses to job satisfaction items are very consistent (highly correlated) whereas for other schools, the items are very heterogeneous (and uncorrelated)? In this case, would a pooled within-school analysis make any sense, and if not, how do we proceed with operationalizing the construct of teacher job satisfaction?

* Is the mean always the appropriate aggregate statistic to represent group level effects? Alternatively, are not other parameters of univariate distributions (e.g., variance, skewness) and bevariate distributions (e.g., regression slopes, standard errors of estimate) useful to investigate in multilevel analyses?

* Assuming measurement and aggregation issues are reasonably resolved and the multilevel nature of the problem truly acknowledged, how does one proceed statistically: separate between-group and within-group analysis or some kind of combined regression or structural modelling approach? Moveover, how are statistical dependencies between individual teacher responses within naturally occurring groups (e.g., schools) handled?

There is no intent to overwhelm the reader with these questions, but the fact of the matter is that issues about choosing units and levels of analysis are present whether one likes it or not. Moreover, these issues are of interest in research on educational administration for reasons amplified in recent articles by Miskel and Sandlin and by Knapp that appeared in the Educational Administration Quarterly 3. Analytical "resolutions" of units issues are not usually straightforward nor are the typically pro posed resolutions as satisfactory as they are convenient. Nevertheless, there are some measurement and statistical problems that one ought to be aware of, and it is these that will be discussed. First, however, it is necessary to briefly define the terminology that will be used in these discussions.

TERMINOLOGY

The uses of the terms total (overall), within, pooled within, and between groups analyses have already been noted,

but distinctions between units, groups, levels, and contexts have not yet been clarified. Since a number of researchers and methodologists are now aware of and writing about "unit of analysis" and "multilevel" issues, the terms are not always used in the same ways. Here it is only possible to clarify the way these terms will be used for the purposes of this discussion and to note that these usages are largely compatible with previous distinctions. 4

It is useful to distinguish between units of observation (UOs) and units of analysis (UAs). 5 The UOs are the smallest entitites on which data are directly collected. In the example at the outset, the UOs are teachers for variables like job satisfaction and perceptions of the work climate and schools for the variable of size in terms of faculty FTE. But the units of observation are not always the units of analysis. The UAs are those entities or objects whose behavior one is trying to understand or describe; they are the explanatory focus of an investigation. For example, if one were interested in analyzing the variation in teacher perceptions of their work climate within each school, then the UOs and UAs would actually be one and the same (teachers). But if the differences between school climates are of interest, and one conceptualizes schools as having these climate properties operationalized, say, by the mean of their teacher perceptions, then schools become the UAs, not individual teachers.

It is also useful to distinguish between global 6 and aggregated properties of group level units. For example, school size in terms of faculty FTE is a global property of the school since it cannot be meaningfully disaggregated (broken down) into properties of smaller UOs (e.g., teachers). School climate as defined above, however, is an aggregate property of the school since it is derived from individual teacher data. Although theoretically clear, these distinctions between global and aggregate properties are not always so in practice. The instructional resources of a school might be operationalized in part as a global variable equal to that part of the district's budget allocated to the school. However, if this chunk of the school budget can be mostly connected to the instructional programs of individual teachers, then it is in large part an aggregate property of the school. And the very same information may represent distinctly different instructional resource constructs when treated as a school level datum or as teacher level data.

This leads to the last set of highly interrelated terms: group, level, and context. Group refers to a collection of UOs. Groups can occur by experimental manipulation (as in treatment and control conditions) or by naturally occurring classifications such as those of school,

sex, grade, district, and so forth. The level of analysis is a function of the UA selected; generally it is possible to have either group level or UO-level analyses. In the above illustration, for example, there can be individual teacher level and/or school or district level analyses. When analyses are done at more than one level--that is, more than one UA is involved--these are termed multilevel analyses. 7 In view of the complex ways in which schooling is organized and held accountable, it is unlikely that a single unilevel analysis will shed much light on phenomena studied in educational organizational and administrative research.

Finally, constructs that exist at the group level, and that can vary across groups, can therefore affect differentially the members in each group. In this sense, groups can define different contexts and can, thus, produce contextual effects. Either global or aggregated properties of groups can represent contextual effects. In the example, school size is a contextual variable, but individual teacher perceptions of job satisfaction within a given school are not. However, if some parameter of the distribution of these perceptions (e.g., median or range) were treated as a measure of the school, this measure would represent a contextual variable. Thus, in multilevel designs, both individual and contextual effects can be investigated in the same analysis.

Having clarified this terminology, it is now possible to summarize the basic measurement and statistical issues inherent in the above list of dilemmas.

MEASUREMENT ISSUES

The most salient measurement issues in multilevel research on schooling arise through efforts to operationalize the constructs of interest from the actual data collected on the UOs. Conceptually, one must determine and justify one's choice of the intrinsic property which a given measure acutely represents (fundamentally, a question of construct-indicator match--see below). In addition, where data are to be aggregated to form an indicator at a higher level, the question of the appropriate metric or index (e.g., a mean versus a proportion above some cut point) to represent the group-level property must be considered. Finally, when different observable measures are to be combined into a scale, the problem of selecting the appropriate levels for psychometric analyses employed in scale construction arises. In the remainder of this section, each of these measurement issues will be discussed.

Construct-Indicator Match

Suppose the following agree-disagree-type items were designed to measure the school climate construct of "trust," and the teachers in each of the 16 schools in the above example responded to each of the items:

(1) I am generally a trusting type of person.
(2) I trust the staff members at this school.
(3) We trust one another at this school.
(4) You trust one another at this school.
(5) Staff members trust one another at this school.

What characteristic is being measured by each of these items? Can there be more than one? Certainly teachers are responding to each item, but are these responses measuring something about the teacher, something about the school, or both? Suppose item means for each school are computed. Do these means represent measures of an intrinsic property of the school with averaging over teachers being incidental and merely a convenient operational device for getting at this property? Or, do these means represent the central tendency of distributions of measures of a property of teachers? Can these means represent both kinds of measures?

At first glance, the differences between these queries may appear subtle and perhaps even unimportant, but their resolution <u>is</u> the unit-of-analysis "problem" in the psychometric phase of a study. Understanding, both conceptually and empirically, what is being measured at the item level is the only way of unraveling these dilemmas.

It is easy to think of items as measuring something about individual teachers since the whole idea and use of measurement in the behavioral sciences has come about from this perspective. It is not so easy to think of items as measures of something systemic, i.e., endemic to the school itself (a global property). Understanding the multitude of variables (e.g., sex, beliefs, length of employment, internal-external personality type, and so on) and their interactions that account for the variation in teacher "trust" scores is clearly an interesting and worthy task. So also is understanding why some <u>schools</u> seem to have more potential for "teacher trust" than others. This implies a conceptual quality or attribute about the school (or rather the ambiance or climate of the school) called "trust." This attribute is obviously intangible and only exists because humans are operating in some kind of interactive process, which is precisely why we have to ask the humans about it to get at it.

What, then, is the best estimate of "teacher trust" for the school? The mean, being an overall level effect and cancelling out the various and sundry main effects and

interaction effects of other variables, would seem to be one choice for this purpose. It is extremely important, however, that the nature of this measure be understood as an estimate of the magnitude of an attribute of the school not as the magnitude of an attribute of the teachers in the school. The measure as it is now conceived as a mean, does not exist in different quantities for different people; rather, it exists in one quantity indicating the magnitude of a single property of the group. When one chooses to assess this property indirectly through the perceptions of the people in the group, one must be willing to assume that a constant indicant of this property somehow exists in each person as operationalized by the overall level effect, i.e., the mean. The teachers are merely a vehicle for getting the measure and to ascribe the school average as a measure of how any teacher perceives the school would be clearly fallacious given this line of reasoning.

But can one operationalize both individual and group properties in the same item? Perhaps only by fiat, hopefully based upon supportable conceptual and empirical rationales. There is simply no logical reason to suppose that the "something" being measured at the group level is the same thing that is being measured at the individual level. 8

What, then, are some if the item characteristics that might serve to distinguish between group-focused and person-focused measurement? Consider again the five items on trust. The first "I-form" item is the typical personality-type item and clearly has no group referent. The remaining four items have group referents, but they vary in degree of potential group effect with the "I-form" item most person-oriented, the "they-form" item most group-oriented, and the "we-form" item somewhere in between. In the organizational climate literature, this grammatical sequence (1st person singular to 3rd person plural) might be thought of as operationalizing the confounded continuums of psychological-to-organizational climate and effective-to-descriptive measures.

In some ways, however, researchers in the area of organizational climate have unwittingly camouflaged the measurement issue by creating the dichotomy between "affective" and "descriptive" items using what are essentially perceptual data. 9 On the "affective-descriptive" continuum, an "I-form" item like number 2 is somewhere near the "effective" end point, and a "they-form" item like number 5 is somewhere in the middle. Items which are nearer the "descriptive" end point might be something like these: "Most of the staff in this school are over the age of 21," or "The curriculum structure of this school is departmentalized." Obviously, such data as these can be gathered by observation, document review, and simple

demographic survey. But as soon as the constructs are such that one has to question the humans in the organization, sizable between and within school variance components are opened up. It is difficult if not impossible to write items reflecting constructs such as "trust," "influence." "role definition," "morale," and so forth that could be placed near the descriptive end of the continuum. This is the dilemma faced by organizational climate researchers who demand high degrees of consensus on items which are every bit as "effective" as they are "descriptive." 10

Alternative Indices of Group Level Constructs

A related measurement concern is the choice of the aggregate measure (metric and/or index) to operationalize a group-level construct. Selecting the group mean should not be based solely on its ease of computation or its relatively nice statistical behavior. Certainly, the central tendency in a distribution of, for example, teacher perceptions of a school attribute can be useful as noted above. But for different constructs, such as consensus or cohesiveness of staff attitudes, a measure of variability would be more important. Or, if asymmetry in perceptions is important, then a measure of skewness of the percent of teachers above a certain cutoff score might be more relevant. 11 Moreover, in looking at multivariate relationships in multilevel designs, particularly when one variable is conceived as an outcome dependent upon the others, within group regression coefficients can be used as the measure of the group-level property of interest. 12 To be sure, using statistics (other than the mean) can present some rather sticky analytical issues. Nevertheless, if averages do not fit the constructs being measured, then there is no point pretending that they do.

Psychometric Decisions in Scale Construction

Now, how should one go about constructing a scale based on the item responses from teachers (the individual level in our example)? Ordinarily, the correlation (or covariance) matrix for the items is computed and one or more internal consistency-type analyses performed (e.g., coefficient alphas, factor, or cluster analysis), or perhaps, the data might be fit into one or another latent trait-type model. In any case, what are the UAs? Obviously, the authors would advocate (a) a clear, conceptual understanding of the multilevel nature of the problem and (b) resisting the temptation of a total analysis of the item intercorrelations computed across individual teachers irrespective of school, elementary and secondary distinctions, and perhaps, even district affiliation. What are left, then, are the substantive interpretations of within and between approaches.

An example of the within approach is as follows: The teachers' item responses within a particular school are correlated and factor analyzed, and scales are developed in the usual fashion. This process is then repeated for each of the 15 remaining schools resulting in 16 psychometric "case studies." The "good news" would be that these case studies are sufficiently similar so as to warrant statistically averaging (pooling) these within school results into a single analysis for the purpose of scale construction. The "bad news" would be finding sufficient dissimilarities in the multivariate structure of the item interrelationships--that is, some items "hang together" forming a useful climate dimension (e.g., "trust") at one school but are nearly uncorrelated at another. In this case, the researcher has at least two options: (1) treat the "problem" as a substantive finding or result or (2) rethink the conceptual basis for what it was he/she was trying to measure.

The between approach proceeds quite differently. Suppose that a case can be made for using the mean of teachers' responses to any item as a measure of an aggregate property of a particular school. For the school work climate items, then, the school can be given 100 "scores," the 100 item means based on teacher responses. This process is then repeated for each school. Then, using schools as replications (i.e., as UAs), the items can be intercorrelated and scaled as above. There is no statistical reason to expect the results to be at all similar to those from the within analysis. As Robinson demonstrated years ago, between, within, and total correlations can be quite different in both magnitude and direction (sign). [13] Indeed, researchers who have pursued analyses of these types have found sizable and interpretable differences. [14]

In these examples, only bilevel analyses, individuals within schools, and schools themselves have been dealt with. But it is quite likely that climate constructs would change depending upon whether elementary or secondary schools were being considered. In effect, the above analyses would all have to be recomputed separately for elementary and secondary schools and then pooled only if empirically warranted. (The authors will not even bother to suggest possible district level effects, department-level effects within secondary schools, and so forth. Clearly the problem can grow in complexity and it cannot be wished away.)

This discussion has also been restricted to measurement issues involving the scaling of items. The same basic principle, however-that the same observable variable can measure different constructs at different levels of aggregation-occurs for measures such as age, SES, years of teaching experience, and so forth. As a change of pace, consider the example of socioeconomic background measures on students typically used in the school effects research. At the student level within schools, these measures may convey the "parental investments" in the child's learning. Once aggregated to the school level, however, they may reflect the community context (e.g., wealth, urbanism, and the like) which can condition resource allocations to schools. 15

STATISTICAL ISSUES

The discussion thus far has been developing the idea that in research on schooling, a multilevel perspective regarding the collection, analyses, interpretation, and reporting of data is essential. 16 Consistent with this perspective is the necessity of employing statistical methods that examine data from the different levels of the school system and their interconnections in an effort to disentangle effects from a variety of sources and levels.

Thus, once a multilevel perspective is adopted, a salient concern is how to conduct an analysis that considers effects at all pertinent levels. Multilevel analyses, i.e., separate analyses at two or more levels or a combined analysis containing explanatory variables at two or more levels, are typically necessary. As in the case of psychometric analyses, statistical analyses conducted during the study phase at between and within levels can show very different results and interpretation. 17 Aside from substantive considerations, then, the central technical issue focuses on the development of strategies that can combine the features of statistical analyses at more than one level.

Multilevel Procedures

While earlier forays into estimating educational effects at multiple levels concentrated on the estimation of variance components or proportions of variation, 18 current emphasis is on the decomposition of relationships (covariances, correlations, and regression coefficients). Certain methods are basically direct extensions of widely used regression methods for handling multilevel data. Cronbach, 19 on the one hand, decomposes the individual-level regression relationships into between group and within group components and recommends that between group and pooled within-group regressions be separately estimated. Burstein, Keesling, and Wiley, 20 on the other hand, are concerned with the potential confounding of compositional effects in the analysis of means when individuals are non-randomly assigned to groups. They offer alternative analytical procedures that purportedly adjust estimates of group effects for within-group composition. Whether these adjustments are the proper ones is, however, the subject of continuing debate. 21

A set of more elaborate multilevel procedures has recently been proposed. 22 These methods are intended to model multilevel processes and outcomes more conscientiously and they involve more powerful estimation routines. Further conceptual and empirical work on these methods is clearly warranted as they have yet to be subjected to the kinds of empirical tests that could identify their properties, much less their range of utility.

Dependencies Among Observations

Finally, the issue of statistical dependencies among UOs needs to be addressed. Again, this issue has not been resolved definitively and its resolution depends, in part, on the nature of the research design, i.e., experimental or non-experimental. The requirement for independence of UOs derives from the classical concept of experimental units as the smallest units (lowest level) which can receive different treatments or different replications of the same treatment. Experimental canons caution that "experimental units should respond independently of one another...(there should be) no way in which the treatment applied to one unit can affect the observation obtained from another unit." 23 Dependencies among units of the type described confound treatment effects and complicate the estimation of the within treatment error.

Obviously, both types of design concerns are likely to arise in the typical non-experimental investigations that characterize most of the research done on schooling, particularly in administration and organizational development. Although the consequences of dependencies among obervations and contaminated treatments are certainly real,

their role in educational research and, thus, the means of handling them are best understood by a specification of their statistical and substantive manifestations. For example, the statistical consequences of dependency have to do primarily with complications in specifying the correlation structure among disturbances (errors) which in turn yield spuriously liberal tests of group effects in many instances. 24 The most direct way to resolve this problem is to devote more attention to specifying an appropriate error structure (as in the multilevel estimation methods noted in footnote 22) and adjusting one's analysis accordingly.

The substantive manifestation is that dependencies among observations within groups (classes, schools, and so forth) are a function of the treatment or processes under investigation as well as the manner in which groups are formed and their composition. Thus within group dependency itself is information about substantive educational processes and should be examined accordingly. 25

CONCLUDING REMARKS

In summary, treating measurement and statistical problems in the analysis of multilevel data as simply a matter of selecting an appropriate unit and level of analysis is much too narrow a conception of the issues that are involved. Rather, the focus should be on the identification of the appropriate set of substantive research questions at and within various levels and the specification of appropriate models for analyzing multilevel data. In other words, to the extent possible, the resolution of dilemmas such as the ones in the above illustration cannot rest upon arbitrary decisions and must be based on both conceptual arguments and empirical consequences.

Readers may be frustrated at this point, since serious dilemmas were raised at the outset for which no definite "answers" have been provided. The authors share in this frustration, since it is one likely outcome of taking multilevel issues seriously in the psychometric and statistical phases of research on schooling. It is not a settling thought, returning to the initial example, to realize that the principal leadership, staff cohesiveness, and work facilitation climate dimensions might not have existed if the psychometric analyses were conducted (as they should have been) using between and within approaches. Moreover, once climate constructs were identified appropriately, it is further distressing that their order of importance and the total variation accounted for in predicting job satisfaction could have been quite different depending on the multilevel approaches used in the statistical analyses.

The hope is that when significant numbers of educational researchers recognize the multilevel nature of the phenomena they typically encounter, they will be more apt to brave the analytic terrain and confront the problems that rightfully dominate the examination of interrelations among units at and within various levels of the educational system. As a consequence, progress towards the resolution of still unanswered measurement and statistical concerns will accelerate, as will, most importantly, progress in the understanding of educational phenomena.

NOTES

1. The work reported herein was in part performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

2. In most technical discussions, an analysis conducted on units at the lowest level (teachers in this case) is called an individual-levelanalysis. However, the terms total and overall are equally acceptable, and perhaps preferable here, because of the tendency in educational research to confuse individual with student.

3. C. Miskel and T. Sandlin, "Survey Research in Educational Administration," Educational Administration Quarterly17, 4(Fall 1981): 1-20; and T.R. Knapp, "The Unit and the Context of the Analysis for Research in Educational Administration," Educational Administration Quarterly 18, 1(Winter 1982): 1-13.

4. L. Burstein, "Analysis of Multilevel Data in Educational Research and Evaluation," in Review of Research in Education, Vol.8, D. Berliner, ed. (Washington, D.C.: American Educational Research Association, 1980): 158-233; W. Haney, "Unit and Levels of Analysis in Large-Scale Evaluation," in Issues in Aggregation, No. 6, New Directions for Methodology of Social and Behavioral Sciences, K.H. Roberts and L. Burstein, eds. (San Francisco: Jossey-Bass, 1980): 1-16; W. Glick, "Problems in Cross-Level Inferences," in Issues in Aggregation, No. 6, New Directions for Methodology of Social and Behavioual Sciences,T.R. Knapp, "The Unit and the Context" ; and L. Burstein, "Units of Analysis," International Encyclopedia of Education Research and Studies (Oxford: Pergamon Press, in press).

5. D.E. Wiley, "Design and Analysis of Evaluation Studies," in The Evaluation of Instruction, M.C. Wittrock and D.E. Wiley, eds. (New York: Holt, Rinehart & Winston, 1970); W. Haney, "Units of Analysis Issues in the Evaluation of Project Follow Through" (Unpublished report, Cambridge, Mass.: The Huron Institute, 1974); L.J.Cronbach, "Research on Classrooms and Schools: Formulation of Questions, Designs and Analysis" (Occasional paper, Stanford Evaluation Consortium, Stanford, Calif., 1976); W. Haney, "The Follow Through Planned Variation Experiment," Vol. V., A Technical History of the National Follow Through Evaluation (Cambridge, Mass.: The Huron Institute, 1977): T.R. Knapp, "The Unit-of-Analysis Problem in Applications of Simple Correlation Analysis to Educational Research," Journal of Educational Statistics (Fall, 1977): 171-186; L. Burstein, "Analysis of Multilevel Data" L. Burstein, "Analyzing Multilevel Educational Data: The Choice of an Analytical Model Rather than a Unit of Analysis," in Design, Analysis, and Policy in Testing and Evaluation, E. Baker and E. Quellmalz, eds. (Beverly Hills, Calif.: Sage Publications, 1980): 81-94; L. Burstein, "The Role of Levels of Analysis in the Specification of Educational Effects," in Analysis of Educational Productivity, Vol. 1: Issues in Microanalysis, R. Dreeben and J.A. Thomas, eds. (Cambridge, Mass.: Ballinger Press, 1980): 119-190): and T.R. Knapp, "The Unit and the Context."

6. The term "global" is due to P.F. Lazarsfeld and M. Rosenberg, The Language of Social Research (Glencoe, Ill.: Free Press, 1955: 287-288). The authors are not particularly satisfied with this term, but others that come to mind (e.g.,collective, macro, structural, systemic, or group) are equally ambiguous.

7. Technical discussions of levels of analysis and multilevel analysis employ more elaborate distinctions than are pertinent to our present treatment of these topics. In typical educational research employing regression methods, investigations in which separate regression analyses are performed at two or more levels or in which the set of explanatory variables in the regression involve UOs from two or more levels are both considered to be multilevel analyses. See L. Burstein, "Units of Analysis".

8. For more in-depth discussions of the construct-indicator match issues in multi-level research on schooling, see L. Burstein, "Analysis of Multilevel Data"; K.A. Sirotnik, "Psychometric Implications of the Unit-of Analysis Problem (with Examples from the Measurement of Organizational Climate)," Journal of Educational Measurement 17 (Winter 1980); 245-281; K.A. Sirotnik, M.A. Nides, and G.A. Engstrom, "Some Methodological Issues in Developing Measures of Classroom Learning Environment," Studies in Educational Evaluation, 6(1980): 279-289; L. Burstein, Using

Multilevel Methods for Local School Improvement: A Beginning Conceptual Synthesis (Center for the Study of Evaluation, University of California, Los Angeles, April, 1983); and L. Burstein, "Units of Analysis."

9. See, for example, reviews such as those of J.G. Howe and J.F. Gavin, Organizational Climate: A Review and Delineation, technical report no. 7402 (Fort Collins, Col., Colorado State University, Industrial Psychological Association of Colorado, 1974); A.D. Jones and L.R. James, "Psychological Climate: Dimensions and Relationships of Individual and Aggregated Work Environment Perceptions," Organizational Behavior and Human Performance 23(1979): 201-250; and synthesizing attempts such as that of R.L. Payne, S. Finemen, and T.D. Wall, "Organizational Climate and Job Satisfaction: A Conceptual Synthesis," Organizational Behavior and Human Performance 16(1976): 45-62.

10. R.M. Guion, "A Note on Organizational Climate," Organizational Behavior and Human Performance 9(1973): 120-125).

11. References to alternative indices of group-level constructs include D.E. Wiley, "Design and Analysis"; P. Lohnes, "Statistical Descriptors of School Classes," American Educational Research Journal 9(1972): 574-586; B.W. Brown and D.H. Saks, "The Production and Distribution of Cognitive Skills Within Schools," Journal of Political Economy 83(1975): 571-593; R.W. Klitgaard, "Going Beyond the Mean in Educational Evaluation," Public Policy 23(1975) 59-79; W.W. Cooley and P.R. Lohnes, Evaluation Research in Education: Theory Principles and Practices (New York: Irvington Publishers, Inc., 1976); R.L. Linn and L. Burstein, Descriptions of Aggregates, CSE Report Series (University of California, Los Angeles: Center for the Study of Evaluation, 1977); L. Burstein, "Analyzing Multilevel Educational Data"; L. Burstein and R. L. Linn, Analysis of Educational Effects from A Multilevel Perspective: Disentangling Between- and Within-Class Relationships on Mathematics Performance, CSE Report Series 172 (University of California, Los Angeles: Center for the Study of Evaluation, 1981); L. Burstein, Using Multilevel Methods; B.D. Spencer, "On Interpreting Test Scores as Social Indicators: Statistical Considerations,: Journal of Educational Measurement 20(Winter, 1983): 317-344; and L. Burstein, "Information Use in Local School Improvement: A Multilevel Perspective" (Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana, April 1984.

12. See, for example, D.E. Wiley, "Design and Analysis"; L. Burstein, R.L. Linn, and F.J. Capell, "Analyzing Multilevel Data in the Presence of Heterogeneous Within-Class Regressions," Journal of Educational Statistics

3(4)(1978): 347-383; and L. Burstein and M.D. Miller, "Regression-Based Analyses of Multilevel Data," in Reanalyzing Program Evaluations, R.F.Boruch, P.M. Wortman, D.S. Cordray, and Associates, eds. (San Francisco: Jossey-Bass Publishers, 1981): 194-211.

13. W.S. Robinson, "Ecological Correlations and the Behavior of Individuals," American Sociological Review 15(1950):351-357. See also M.T. Hannan, Aggregation and Disaggregation in Sociology (Lexington Mass.:D.C.Heath, 1971): M.T. Hannan and L. Burstein, "Estimation from Group Observations," American Sociological Review 39(June 1974): 374-392; L. Burstein, "The Use of Data from Groups for Inferences About Individuals in Educational Research" (Unpublished doctoral dissertation, Stanford University, 1975); L.F. Cronbach, "Research on Classrooms and Schools"; T. R. Knapp, "The Unit-of-Analysis"; L. Burstein, "Analysis of Multilevel Data" and "The Role of Levels of Analysis"; W. Glick, "Problems in Cross-Level Inferences"; and T.R. Knapp, "The Unit and the Context."

14. L.J. Cronbach, "Research on Classrooms and Schools"; K. Harnquist, "Primary Mental Abilities at Collective and Individual Levels," Journal of Educational Psychology 70(1978): 706-716; J.E. Gustafsson, "Attitudes Towards the School, the Teacher, and Classmates at the Class and Individual Level," British Journal of Educational Psychology 49(1979): 124-131; M.D. Miller and L. Burstein, Multilevel Properties of Test Items: An Exploratory Study, CSE Report #181 (University of California, Los Angeles: Center for the Study of Evaluation, 1979); K.A. Sirotnik, "Psychometric Implications"; K.A. Sirotnik, M.A. Nides, and B.A. Engstrom, "Some Methodological Issues"; and M.D. Miller, Measuring Between-Group Differences in Instruction (Unpublished Doctoral Dissertation, University of California, Los Angeles, 1981).

15. The references that consider multilevel scaling issues specific to socioeconomic background measurement include L. Burstein, "Analysis of Multilevel Data" and "The Role of Levels of Analysis"; L. Burstein, K. Fischer, and M.D. Miller, "The Multilevel Effects."

16. References that more thoroughly discuss the concept of a multilevel perspective in research on schooling include R. Barr and R. Dreeben, "Instruction in Classrooms," in Review of Research in Education, Volume 5, L.S. Shulman, ed. (Itasca, Ill.: F.E. Peacock, 1977): 89-162; D. Rogosa "Politics, Processes, and Pyramids," Journal of Educational Statistics 3(Spring 1978):79-86; L. Burstein, "Analysis of Multilevel Data," "Analyzing Multilevel Educational Data," "The Role of Levels of Analysis," Using Multi-level Methods, "Information Use on Local School Improvement", and "Units of Analysis"; R. Barr and R. Dreeben, How Schools Work

(Chicago: University of Chicago Press, 1983);R.J. Shavelson, N.M. Webb, and L. Burstein, "Measurement of Teaching," in Third Handbook of Research on Teaching, M. Wittrock, ed. (Macmillan, in press).

17. F.J. Capell, "Interpreting Multilevel Data from Program Evaluation" (Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, Calif., April 1979); L. Burstein, "The Role of Levels of Analysis"; L. Burstein, "Analyzing Multilevel Educational Data"; L. Burstein, K. Fischer, and M.D. Miller, "The Multilevel Effects f Background on Science Achievement at Different Levels of Analysis: A Cross-National Comparison," Sociology of Education 53,4(1980): 215-255; K.A. Sirotnik, "Psychometric Implications"; K.A. Sirotnik, M.A. Nides, and B.A. Engstrom, "Some Methodological Issues"; L. Burstein, "Explanatory Models Using Between and Within Class Regression: Basic Concepts and an Example" (Paper presented at the Second International Mathematics Study Data Analysis Workshop, Toronto, Canada, December 1981); K.A. Sirotnik, "The Contextual Correlates of the Relative Expenditures of Classroom Time on Instruction and Behavior: An Exploratory Study of Secondary Schools and Classes," American Educational Research Journal 19(Summer 1982): 275-292.

18. D.E. Wiley and D. Bock, "Quasi-Experimentation in Educational Settings: Comment," School Review 75(1967): 353-366; G.F. Madaus, T. Kellaghan, E.A. Rakow, and D.J. King, "The Sensitivity of Measures of School Effectiveness," Harvard Educational Review 40(1979): 207-230.

19. L.J. Cronbach, "Research on Classrooms and Schools."

20. J.W. Keesling and D.E. Wiley, "Regression Models for Hierarchical Data" (Paper presented at the Annual Meeting of the Psychometric Society, Stanford University, 1974); L. Burstein, "Analysis of Multilevel Data," "Analyzing Multilevel Educational Data," and "Explanatory Models."

21. The most recent attempt at technical resolution of this choice is R.L. Tate and Y. Wongbundhit, "Random Versus Nonrandom Coefficient Models for Multilevel Analysis," Journal of Educational Statistics 8(Summer 1983): 103-120.

22. Much of this new work has yet to appear in journals. The list of recent efforts include L. Erbring and A.A. Young, "Individuals and Social Structure: Contextual Effects as Endogenous Feedback," Social Methodological Research 7(Winter 1979): 396-430; J.W. Keesling, "Some Explorations in Multilevel Analysis" (Paper presented at the Annual Meeting of the American Educational Research Association, March 1978); J. Wisenbaker and W.J. Schmidt "The Structural Analysis of Hierarchical Data". (Paper

presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1979); M. Aitkin, D. Anderson, and J. Hinde, "Statistical Modeling of Data on Teaching Styles," Journal of Royal Statistical Society Series 144(1981):419-461; H. Goldstein, "Multilevel Mixed Model Analysis Using Iterative Generalized Least Squares" (Unpublished manuscript, 1983); W.M. Mason, G.Y. Wong, and B. Entwisle, "Contextual Analysis Through the Multilevel Linear Model," Sociological Methodology 1983/1984 (San Francisco: Jossey-Bass, 1984); L. Burstein, "Units of Analysis"; C. Chou, "Examination of Contextual Effects Through an Endogenious Feedback Model," (Unpublished doctoral dissertation, University of California, Los Angeles, 1983); D. Rachman-Moore and R.G. Wolfe, "Robust Analysis of a Nonlinear Model for Multilevel Educational Survey Data," Journal of Educational Statistics 9(Winter 1984): 277-293; W. Schneider and B. Treiber, "Classroom Differences in the Determination of Achievement Changes," American Educational Research Journal 21(Spring 1984); 195-211; and B.W. Brown and D.H. Saks, "An Economic Approach to Measuring the Effects of Instructional Time on Student Learning" (Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada, 1983); and S. Raudenbush and A. Bryk, "A Hierarchical Model for Studying School Effects" (Unpublished manuscript, 1984).

23. D.R. Cox, Planning of Experiments (New York: John Wiley and Sons, 1958); D.E. Wiley, "Design and Analysis"; K. Hopkins, "The Unit of Analysis: Group Mean vs. Individual Observation," American Educational Research Journal 19(1982): 5-18; T.R. Knapp, "The Unit-of-Analysis" and "The Unit and the Context"; W. Haney, "The Follow Through Planned" and "Unit and Levels of Analysis."

24. L. Burstein, "Analysis of Multilevel Data"; K. Hopkins, "The Unit of Analysis" L. Glendening, "The Effects of Correlated Units of Analysis: Choosing the Unit" (Paper presented at the Annual Meeting of the American Educational Association, San Francisco, April 1976); M. Aitkin et al., "Statistical Modeling of Data."

25. R. Barr and R. Dreeben, How Schools Work; N.M. Webb, "Group Process: The Key to Learning in Groups" in Issues in Aggregation, No. 6, New Directions for Methodology of Social and Behavioral Sciences, K.G. Roberts and L. Burstein, eds. (San Francisco: Jossey-Bass, 1980): 77-88; L. Burstein, "Analysis of Multilevel Data" and "The Role of Levels of Analysis"; L Burstein and M.D. Miller, "Regression-Based Analyses"; and C.E. Bidwell and J.D. Kasarda, "Problems of Multilevel Measurement: The Case of School and Schooling," in Issues in Aggregation, No. 6, New Directions for Methodology of Social and Behavioral Sciences, K.H. Roberts and L. Burstein,eds. (San Francisco: Jossey-Bass, 1980): 53-64.