DOCUMENT RESUME

ED 129 881 95 TM 005 684

AUTHOR Burstein, Leigh
TITLE Assessing Differences Between Grouped and Individual-Level Regression Coefficients.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE [Apr 76]
CONTRACT NIE-C-74-0123
NOTE 43p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976); For related documents, see ED 100 958 and 108 984

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Analysis of Covariance; Analysis of Variance; Comparative Analysis; *Correlation; *Data Analysis; *Groups; Individual Characteristics; Individual Differences; *Mathematical Models; Multiple Regression Analysis; Prediction; Program Effectiveness; Schools; *Statistical Analysis; Student Characteristics

ABSTRACT

Two questions are investigated here: What should the unit of analysis be in investigations of educational effects and on what basis should the units be chosen Under what conditions can relationships among measurements on individuals be estimated from the relationships among measurements on aggregates of individuals? Models using standard analysis of variance, standard analysis of covariance, and standard regression analysis are compared as they would be applied to two different data aggregation levels. Implications for multilevel analysis are discussed. (Author/BW)

DRAFT

ASSESSING DIFFERENCES BETWEEN GROUPED AND INDIVIDUAL-LEVEL REGRESSION COEFFICIENTS*

Leigh Burstein
University of California,
Los Angeles

Paper presented at the annual meetings of
the American Educational Research Association,
San Francisco, April 1976

*The research reported here was partially supported by NIE research contract C-74-0123. The paper was stimulated primarily by work with Lee Cronbach and Michael Hannan.

The title of my presentation is somewhat misleading though it was more accurate last August when the proposal for this symposium was submitted. (In Appendix A, we have directly considered certain technical aspects of the problem suggested by the title.) The primary reasons for the shift in emphasis is that our thinking about the units of analysis in educational research has undergone rapid evolution and our interest in coming to grips with methodological problems in the identification of education effects is more pervasive than originally imagined.

The evolution in thought can be traced to the expanded consideration of two key questions which arise simply because schools are aggregates of their teachers, classrooms, and pupils and classrooms are aggregates of the persons and processes within them. These general questions can be stated as:

(1) What should the unit(s) of analysis be in investigations of educational effects and on what basis should the units be chosen?

(2) Given data from the multiple levels in the analysis hierarchy, when and how can one estimate relations generated by models involving one set of levels of aggregation from relations generated by models based on a different set of levels?

The second question may seem convoluted to the uninitiated, but in the simplest case, the question can be translated into

(2') Under what conditions can relationships among measurements on individuals (e.g., Pupil Achievement and Pupil SES) be estimated from the relationships among measurements on aggregates of individuals (e.g., school mean achievement and school mean SES)?

We now know a great deal about the answers to question 2 for the simplest cases involving comparisons of models purely at two distinct

levels of aggregation (Burstein, 1974, 1975a, 1975b; Burstein and Knapp, 1975; Hannan and Burstein, 1974; Hannan, et al., 1975) and are beginning to better understand what happens when the models mix variables from multiple levels (Burstein, 1975a, 1976; Burstein and Knapp, 1975; Burstein and Smith, 1975; Hannan, Freeman and Meyer, 1976). Furthermore, through creative applications of the general linear model (Keesling and Wiley, 1974; Rock, Baird, and Linn, 1972) and experimental designs (Glendening and Porter, 1974; Poynor, 1974), we have become more sensitive to the impact of correlated units inherent in hierarchically-nested school data.

Developing Interest in Issues Concerning Data Aggregation

Later on, I shall discuss the present state of the art in response to question 2 and provide an example of how work on the methodology of data aggregation has advanced both theoretically and substantively and has begun to take on a new degree of subtlety in its application. But first I want to provide some indication of how our thinking about and audience for units of analysis and data aggregation questions have changed in just two years.

As part of a Division D paper session at the 1974 AERA conventtion, I presented a paper entitled "Issues concerning the inferences from grouped observations" (Burstein, 1974) which, along with a joint paper with Mike Hannan appearing in the American Sociological Review that same year (Hannan and Burstein, 1974), reviewed, classified, interpreted, and hopefully expanded the work on data aggregation. In fact, the essentials of our answers to the simple cases of question 2 above were spelled out in these two efforts.

The two papers included, among other things, reviews of work from sociologists' problems with ecological inference (the prime example

being Robinson (1950)) and change in units of analysis (Blalock, 1964; Hannan, 1971), the statisticians' concerns for measurement error (e.g., Mandansky (1959)), the political science treatment of missing data (e.g., Kline, Kent and Davis (1971)) and economists' treatments of economy of analysis (Cramer, 1964; Prais and Aitchinson, 1954) and confidentiality of data (Feige and Watts, 1972). To my knowledge prior to that time (later proven incorrect when Haney's (1974) enlightening paper on the unit of analysis in Project Follow Through was uncovered), there had been no discussion of any of the problems mentioned above by educational research methodologists with perhaps possible exception of early papers by Walker (1928) and Burks (1928) and an insightful note by Thorndike (1939).[1] The response to my AERA presentation was not in the least bit overwhelming (no interested audience members and about 10 requests for the paper). In contrast, the Hannan-Burstein ASR paper, though riddled with errors in printing and, in retrospect, with confusing notation, continues to receive attention (sometimes critical) from sociologists. In summary, in 1974, questions about data aggregation as addressed by other social scientists were just about non-existent from educated researchers.

In the summer of 1974, NIE funded two projects on methodology for aggregating data in educational research (Mike Hannan and I are co-principal investigators for one of the contracts). In my opinion this investment by NIE helped expand the thinking about the effects of grouping to more complex research situations as evidenced in presentations on applications and recent developments in data aggregation in educational research at the 1975 AERA meeting (Burstein 1975a; Hannan, Young and Nielson, 1975). The Hannan, et al. paper dealt with the effects of grouping in multivariate and longgitudinal models and my own paper focussed firmly on units of analysis issues arising in the large-scale, regression-based studies

of the effects of schooling. Despite being scheduled for a noon meeting on the last day of the convention, the symposium drew an audience of 20 or more and we have received over 100 requests for copies of these papers.

Later that same year, the students and faculty from the educational measurement and statistics programs of the New York universities devoted their annual meeting to the consideration of questions about the units of observation, the units of analysis, and the independence of observations. More recently, Lee Cronbach presented preliminary results of his project on multilevel analysis of educational data to the May 12th group, consisting mainly of educational evaluation theorists.

So we are already reaching some audience with the results of our efforts. Whether the increased awareness leads to improved research with multilevel data remains to be seen.

Issues in the Analysis of Hieracrchical Data

For the present, I would like to restrict the discussion to the questions that arise when there are potentially two levels of observation and analysis in a school setting. Assume that we have measures on pupils (e.g., family background, ability, opportunity to learn, achievement; attitude towards school), on teachers (e.g., subject matter, training, education, experience; teacher behavior), and on the classroom as a whole (e.g., climate, class size; facilities). We might also introduce an educational treatment of some sort (e.g., drill vs. meaningful learning strategies) either to intact classrooms or to pupils within classrooms.[2]

Focusing on a single outcome, a single input, and a single treatment, we get the following full model for accounting for the pupil-level outcome of a study in the educational setting previously described:

(1) $$O = b_1T + b_2G + b_3I + e$$

where

O = Outcome variable (e.g., achievement)

T = Treatment/Control

G = Identification of Class Membership

I = Input Variable (e.g., entering ability)

b_1, b_2, b_3 = Estimates of the effects of the respective explanatory variables.

(NOTE: T, G, and I may all be dummy variables sets distinguishing among the categories of a nominal variable.)

An example description of a model depicted in equation (1) is "the pupil performance in mathematics tests (O) is a function of whether he received drill or meaningful learning instruction (T), the classroom in which he received the instruction (G), and his entering ability as measured by a mathematics pretest (I)."

The classroom-level analogue of model I is:

(2) $\bar{O} = \bar{b}_1 T + \bar{b}_2 G + \bar{b}_3 \bar{I} + \bar{e}$

An example of model (2) explanation is: "class mean performance ($\bar{O}$) is a function of the type of instruction received, the class receiving the instruction and the mean performance of students on the pretest ($\bar{I}$)". Note that T and G are measured the same at both the pupil and the class level.

I would argue that in general, models (1) and (2) answer different questions. Furthermore, different modifications (e.g., inclusions/deletions of T, G, or I) lead to different questions being addressed and have, to this point, lead to different decisions about the appropriate units of analysis. We explore some of these differences below.

Standard ANOVA. In the typical experimental study, the usual model has been:

(3) $O = b_1 T + e$ (pupil-level ANOVA);

where b_1 now represents the treatment effect[3].

Or, if the investigator were more sensitive to the problem of independence of observations, either

(4) $\bar{O} = \bar{b}_1 T + \bar{e}$, (Class-level ANOVA)

or

(5) $O = b_1 T + b_2 G + e$ (Nested ANOVA, or Pooled Within-class ANOVA)

The choice among (3)—(5) has been the subject of discussion dating back to at least Lindquist (1940), Cochran (1947), and McNemar (1940) and has also been considered by Peckham, et al. (1969), Glass and Stanley (1970), Wiley (1970) and more recently by Poynor (1974) and Glendening and Porter (1974).

The weight of the evidence is that most educational research using intact classrooms employs model (3) though model (5) and, perhaps, model (4) are superior (Glendening and Porter (1974); Poynor (1974)). Both models (4) and (5) take into account the fact that there are groups of individuals whose responses are correlated and are thus more attuned to the realities of the situation.

My primary reason for preferring model (5) is that "independence of observations" is a matter of degree rather than existence in research and if group differences are small, more powerful tests of effects may be possible with the nested model. (Glendining's and/or Poynor's presentation may have more to say with regard to this particular model.)

Standard ANCOVA. In the typical analysis of covariance problem, the usual model is

(6) $O = b_1 T + b_3 I + e$

where b_3 now represents the pooled within-treatment regression coefficient.

The adjustment, $b_3 I$, is the appropriate one when the assumptions of parallelism of regression slopes and independence of treatments (T) and covariate (I) can be met. For example, in aptitude-x-treatment interaction

research, the assumptions require that the relationship of outcome achievement to entering aptitude be the "same" (not significantly different) for each treatment group, and the treatment should be uncorrelated with entering aptitude.

As Cronbach (1976; Cronbach and Webb, 1975) has pointed out, model (6) is highly likely to be inappropriate when intact classrooms are sampled (whether or not students within classrooms are randomly assigned to treatments). He urges that between-class and within-class analyses be conducted instead by examining the following models:

$$(7) \qquad \bar{O} = \bar{b}_1 T + \bar{b}_3 \bar{I} + e \quad \text{(between-class)}$$

and

$$(8) \qquad O_w = b_1 T + b_{3w} I_w + e \quad \text{(within-class)}$$

(Note that the within-class analysis might also be done using our model (1).)

The greater sensitivity implicit in Cronbach's proposed analyses is an important step forward in the use of analysis of covariance with hierarchical data. The between-class and within-class analyses do not remove the need for concern about homogeneity of regression; in fact, they should increase the investigator's wariness regarding this problem and add the need to watch for lack of independence between classrooms and covariate. The startling reversals that Cronbach and Webb found are perhaps warning enough.

The methodology on the application of ANCOVA in non-equivalent groups designs is in an expansionist phase. Cronbach's comments (1976) suggest that something might be gained from recognizing the parallels between the analysis of covariance model and the models we (Burstein, 1974, 1975a,b; Burstein and Knapp, 1975; Hannan, 1971; Hannan and Burstein, 1974) have proposed for identifying the effects of grouping (Fennessey (1968) and

Werts and Linn (1969, 1971) deserve the credit for first noting this parallel.). The model incorporating the grouping variable discussed in the next section from regression-based studies takes the form:

(6') $O = b_2G + b_3I + e$ (see equation (11) below).

Equation (6') could be viewed as (6) where group membership represents the treatment, or (6) can be considered to be (6') when the group membership is defined by treatment. In any case, though the relationship of interest is different -- effects of treatment (T) in ANCOVA; relationship to input (I) in regression -- for the two models, the same phenomenon should sanction or invalidate their use. So persons working on either problem should be able to learn from the work on the other.

Standard Regression. In the typical regression-based or correlational study, we generally find the model:

(9) $O = b_3I + e$ (Individual-level).

Or, sometimes, despite the fact that the relationship of interest is one posited to exist among individuals, we find class-level analyses as depicted in equation (10):

(10) $\bar{O} = \bar{b}_3\bar{I} + \bar{e}$ (class-level).

Except under very special circumstances (e.g., groups randomly formed or grouped formed on the basis of I), the appropriate model for hierarchical data or intact classrooms includes the grouping variable G in an individual-level analysis:

(11) $O = b_2G + b_{3w}I + e$.

That is, if the researcher were interested in the relationship of student learning as measured by an achievement test to student entering aptitude or to students' family background, then b_{3w} (the pooled within-class regression coefficient) is deemed preferable to b_3 (the individual coefficient) and certainly to $\bar{b}_3$ (the between-groups coefficients).

The reason for preferring the within-class coefficient to the individual coefficient is that equation (11) is more correctly specified, having better accounted for the factors affecting O than does equation (9). And, the one benefit from calculating the between-groups coefficient from model (10) is that we have evidence that equation (9) is misspecified, and at least G should be incorporated when $b_3 - \bar{b}_3 \neq 0$ (Burstein, 1975a,b; Cronbach, 1976; Hannan, Young and Nielsen, 1975).

It is worth reiterating that the nature of the question is the primary determiner of the appropriate choice of analysis model. We might have asked a question requiring between-class analysis as described by model (10) (e.g., Does the teacher's training ($\bar{I}$) affect the amount of class time spent on drill activities ($\bar{O}$); both variables measured at the class level). Or, when both outcome and input were determined prior to assignment to classrooms (e.g., relationship of student's family background (I) to their entering ability (O)), the individual-level regression model (Equation (9)) is preferable to either model (10) or model (11).

The latter point about the preference for individual-level model is a tricky one. Suppose that some form of tracking was employed to assign students to classrooms. The result would probably be that G is correlated with both aptitude and background. If this were the case, then the classes would be internally homogeneous with respect to O and I and most of the variation would lie between classes. Under these circumstances, $\bar{b}_3$ is an inflated estimate of the desired relationship (b_3 in this case) and the pooled within-class coefficient (b_{3w}) is likely to be an underestimate.

One last point about the regression case. Most of the work by my colleagues and me (Burstein, 1974, 1975a,b, 1976; Burstein and Hannan, 1975; Burstein and Knapp, 1975; Burstein and Smith, 1975; Hannan, 1971;

Hannan and Burstein, 1974; Hannan, Freeman and Meyer, 1976; Hannan, Young, and Nielsen, 1975) has focused on the difference between the between-groups regression coefficient ($\bar{b}_3$) and the individual-level regression coefficient (b_3). Cronbach (1976; see also Cronbach and Webb (1975).) recommends that a two-step analysis be carried out (between-groups, $\bar{b}_3$, and within-groups, b_{3w}). We are in agreement with his general recommendation and below we discuss our strategy for carrying them out (see section on Multilevel Analysis).

It is important to point out, however, that the guidelines we have previously proposed for determining when the between-group coefficient yields poor estimates of the individual-level coefficient also identify cases where between-groups coefficients yield poor estimates of the pooled within-groups coefficient. This occurs because of the inextricable linkages among b_3, $\bar{b}_3$ and b_{3w}. After all, covariances and variances at the individual level can be decomposed into their within-group and between-group components so that guidance regarding the relationship of $\bar{b}_3$ to b_3 also provides guidance for the relation of $\bar{b}_3$ to b_{3w}.

Multilevel Analysis. In previous papers (Burstein, 1975a, 1976; Burstein and Knapp, 1975; Burstein and Smith, 1975), I have suggested that we begin to utilize multilevel designs in regression-based analyses of school effects.[4] In such analyses, each variable is measured and analyzed at the lowest level at which the observations on the measure tend to vary independently. Cronbach's presentation (1976) urges essentially the same procedure.

The two examples of multilevel analyses which are most frequently cited are a study by Rock, Baird, and Linn (1972) of the interaction of student aptitude and college characteristics and Keesling and Wiley's (1974) reanalysis of a subset of the Coleman data. The basic analysis

steps in each method are discussed below in the context of a two-level school effects problem.

(A) Rock-Baird and Linn (1972) --

(1) Calculate within-class regression of outcome (O) on input (I),

(2) cluster classrooms on the basis of their parameters (α, β, plus mean predictor score for the class) of within-class regressions,

(3) generate discriminant functions to test for statistical distinctions among the clusters of classes, and

(4) identify classroom-level variables that discriminate among the clusters at the classroom-level.

(B) Keesling and Wiley --

(1) Perform within-class regressions of O on I (they used the common pooled within-class coefficient in this step, presumably for simplicity),

(2) aggregate pupil's predicted scores to the class level ($\bar{\hat{O}}$), and

(3) in a between-class analysis, regress $\bar{O}$ on $\bar{\hat{O}}$ and class-level input variables.

From my perspective, each method has certain merits and certain drawbacks. The Keesling and Wiley approach provides effect perameters more nearly mirroring the structural form of school effects than the Rock, Baird and Linn approach or the usual single-level analysis models. Yet, the appropriate algorithm for actually performing the analysis suggested by Keesling and Wiley is unclear. Should pooled within-class coefficients be used in step 1 or are the individual within-class coefficients more appropriate?

How can we best aggregate predicted outcomes for entry into the between-class analyses: Does this approach provide adequate adjustment?

The above questions about the Keesling and Wiley approach are important, but my most serious concern is that their approach fails to adequately reflect the effects of between-class differences in slopes. For example, Figure 1 depicts outcome-on-input regressions for hypothetical classrooms. I would expect Keesling and Wiley's strategy to be able to distinguish among the performances in classrooms (A) through (C) (and (F) and (E), for that matter) quite easily and be able to separate the effects in (A)—(C) from those in (D), (E) and (F). But is their technique sensitive to the case where the slopes are quite different with common means on the outcome and input variables (comparing (D) with (E))? If not, some improvements are needed in the Keesling and Wiley approach.

Theoretically, the approach suggested by Rock, Baird and Linn should place classrooms (D) and (E), and perhaps all classrooms in Figure 1, into separate clusters. In practice, I am not so sure that this will be the case. Currently available clustering algorithms like those due to Ward (1963) and to Johnson (1967) are dependent on the choice of characteristics on which to base the formation of clusters. Also, there would likely be stability problems in some clusterings that make the discrimination among clusters all the more difficult.

Moreover, treating the resulting clusters as groups in a discriminant analysis, as Rock Baird, and Linn do, discards any metric differences existing among the clusters and thereby eliminates the possiblity of describing school effects in structural terms. (I raised the same concerns about the analysis reported by ETS from Phase II of the Beginning Teacher Evaluation Study (McDonald, et al., 1975). The use of discriminant groups

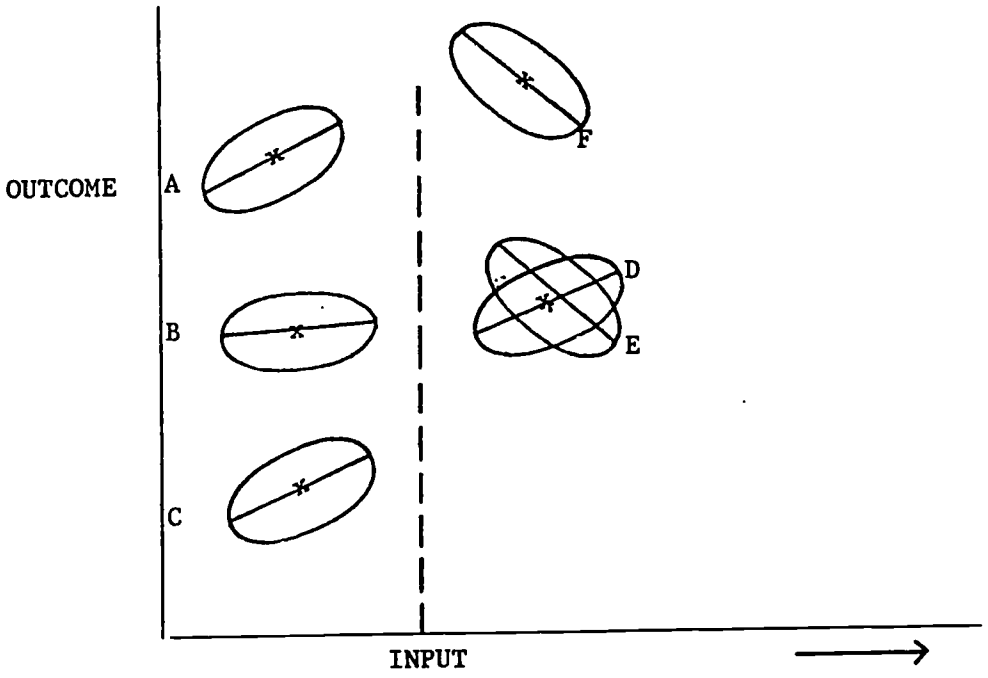

Figure 1. Regression of Outcome on Input for Different Hypothetical Classrooms.

results in some cost in generalizability of findings that should be avoided.

The problem then is that though we have candidates for performing our multilevel analysis, each has drawbacks. I think that it would be worthwhile to explore the following alternative:

(C) An Alternative Multilevel Analysis Strategy. --

(1) Perform within-class regressions (not pooled) of outcomes on input, and

(2) use the parameters (α, β, perhaps SEE) from the within-class regressions as "outcomes" in a between-class analysis.

This alternative strategy combines certain features of approaches by Keesling and Wiley and by Rock, Baird and Linn. The techniques should be able to treat the classrooms depicted in Figure 1 as "different" and provide effect estimates in structural terms. In fact, using the within-class parameter estimates at outcomes should lead to more sensitive interpretations of effects and clearer policy implications from findings. For if one needs to decide whether classrooms have behaved in a compensatory fashion and, if so, at what levels of input, the proposed strategy has merit.

Necessary Caveats

It is perhaps appropriate to end the paper with some caveats about what I feel are potentially the best analytical methods for handling hierarchical data. The proposed alternative strategy for multilevel analysis has never been tried out in the form described here to my knowledge[5], much less in comparison with other approaches. Furthermore, we have considered only analysis procedures in the two-level case when it is obvious that a great deal of educational effects research involves at least a third level (the school).

The main point is that we have to move in this direction if we have any hope of avoiding the painful, tedious, unparsimonious alternative of looking at effects one class (or maybe even one person) at a time. If the current rate of improvement in dealing with units of analysis problems continues, I should be able to close on a more optimistic note this time next year.

FOOTNOTES

[1]The exchange among Bloom, Gagné and Wiley as recorded in Wittrock and Wiley (1970) also predates our work, but addresses the problem from a perspective not reflected in the work from the other social sciences.

[2]It seems to me that the differences in the independence of treatment and response for between-class and within-class allocation to treatment or control is a matter of degree rather than existance (of independence), and for the moment, we will treat both experimental setups in the same fashion.

[3]I will use b_1, b_2, b_3 and e in all subsequent models to represent the same parameters and variables at the individual level though their values might change. $\bar{b}_1$, $\bar{b}_2$, $\bar{b}_3$ and $\bar{e}$ represent their corresponding group value. A w subscript will be used to denote within-class coefficient, e.g., b_{3w}.

[4]The same analysis carried out at two or more levels does not qualify as "multilevel analysis" by the present definition. Thus the analyses of covariance at the student, class, school and sponsor levels in the Project Follow Through studies (Abt Associates, 1973, Emrick, Sorensen and Stearns, 1973; see Haney (1974) for discussion.) are not considered to be multilevel.

[5]Apparently, Baker and Snow (1972) explore a related procedure to assess teacher differences in student aptitude-achievement relationships.

REFERENCES

Abt Associates. National Evaluation of Project Follow Through 1971-72. Cambridge, Mass.: Abt Associates, 1973.

Blalock, H.M. Causal Inferences in Nonexperimental Research. New York: W.W. Norton and Company, 1964. (Also, Chapel Hill, N.C.: University of North Carolina Press.)

Burstein, L. "Issues Concerning Inferences from Grouped Observations." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April, 1974.

Burstein, L. "Data Aggregation in Educational Research: Applications." Technical Report No. 1, Consortium on Methodology for Aggregating Data in Educational Research. Milwaukee, Wisc: Vasquez Associates, Ltd., March, 1975a. (Also, paper presented at the Annual Meeting of the Educational Research Association, Washington, D.C., April, 1975.)

Burstein, L. "The Use of Data from Groups for Inferences about Individuals in Educational Research." Unpublished doctoral dissertation, Stanford University, December, 1975b. (Also, Technical Report No. 7, Consortium on Methodology for Aggregating Data in Educational Research. Milwaukee, Wisc.: Vasquez Associates, Ltd., December, 1975.)

Burstein, L. and Knapp, T.R. "The Unit of Analysis in Educational Research." Technical Report No. 4, Consortium on Methodology for Aggregating Data in Educational Research. Milwaukee, Wisc.: Vasquez Associates, Ltd., July, 1975.

Burstein, L. and Smith, I.D. "Choosing the Unit for Investigating School Effects: IEA Study of Science Education in Australia." Technical Report No. 6, Consortium on Methodology for Aggregating Data in Educational Research. Milwaukee, Wisc.: Vasquez Associates, Ltd., November, 1975.

Burstein, L. and Hannan, M.T. Progress Report NIE-C-74-0123: Consortium on Methodology for Aggregating Data in Educational Research. Milwaukee, Wisc.: Vasquez Associates, Ltd., July, 1975.

Burstein, L. "The Choice of Unit of Analysis in the Investigation of School Effects: IEA in New Zealand." New Zealand J. of Educational Studies, 1976. (Also, Technical Report No. 5, Consortium on Methodology for Aggregating Data in Educational Research. Milwaukee, Wisc.: Vasquez Associates, Ltd., October, 1975).

Baker, Katherine D. and Snow, Richad D. "Teacher Differences as Reflected in Student Aptitude-Achievement Relationships." Stanford, CA: Memorandum No. 85, Stanford Center for Research and Development in Teaching, 1972.

Burks, Barbara. "Statistical Hazards in Nature-Nurture Investigations." 27th Yearbook of the National Society for the Study of Education, Part 1, 1928: 9-33.

Cochran, W.G. "Some Consequences when the Assumptions for the Analysis of Variance are Satisfied." Biometrics, 1947, 3: 22-38.

Cramer, J.S. "Efficient Grouping, Regression, and Correlation in Engle Curve Analysis." J. of Amer. Statistical Assoc., 1964: 233-249.

Cronbach, L.J. "Methods of Aggregation and Choice of Units of Analysis." Paper presented at the May 12th Group meeting, Tampa, Florida, March, 1976.

Cronbach, L.J. and Webb, N. "Between-Class and Within-Class Effects in a Reported Aptitude x Treatment Interaction: Reanalysis of a Study by G. L. Anderson." J. of Educational Psych., 1975, 67: 717-724.

Emrick, J.A., Sorensen, P.H. and Stearns, M.S. Interim Evaluation of the National Follow Through Program 1969-71, A Technical Report. Menlo Park, CA: Stanford Research Institute, February, 1973.

Feige, E.L. and Watts, H.W. "An Investigation of the Consequences of Partial Aggregation of Micro-Economic Data." Econometrica, 1972, 40: 343-360.

Fennessey, James. "General Linear Model: Ecological Correlation." Amer. J. Soc., 1968, 74: 20-21.

Glass, G.V. and Stanley, J.C. Statistical Methods in Education and Psychology. Englewood Cliffs, NJ: Prentice-Hall, 1970.

Glendening, Linda and Porter, Andrew C. "The Effects of Correlated Units of Analysis: Violating the Assumption of Independence." A working paper: Measurement and Methodology Program, NIE, November, 1974.

Haney, W. "Units of Analysis Issues in the Evaluation of Project Follow Through." Unpublished Report. Cambridge, Mass.: Huron Institute, 1974.

Hannan, M.T., Young, A.A., and Nielsen, F. "Specification Bias Analysis of the Effects of Grouping of Observations in Multiple Regression Models." Paper presented at the Annual Meeting of the American Educational Research Association, Washington, D.C., April, 1975.

Hannan, M.T., Freeman, J.H. and Meyer, J.W. "Specification of Models for Organizational Effectiveness, Comment on Bidwell and Kasarda, ASR February, 1975." Amer. Sociological Rev., 1976, 41: 136-143.

Johnson, S.C. "Hierarchical Clustering Schemes." Psychometrika, 1967, 32: 241-254.

Keesling, J.W. and Wiley, D.E. "Regression Models of Hierarchical Data. Paper presented at Annual Meeting of the Psychometric Society, 1974.

Kline, F.G., Kent, K., and Davis, D. "problems in Causal Analysis of Aggregate Data with Applications to Political Instability." In: J.V. Gillespie, and Nesvold, B. (Eds.) Macroquantitative Analysis. Beverly Hills, CA: Sage Publications, 1971: 251-279.

Madansky, A. "The Fitting of Straight Lines when Both Variables are Subject to Error." Amer. Stat. Assoc. J., 1959: 173-205.

McDonald, F.J., Elias, P., Stone, M., Wheeler, P., Lambert, N., Calfree, R., Sandoval, J., Ekstrom, R. and Lookheed, M. Final Report on Phase II Beginning Teacher Evaluation Study. Prepared for the California Commission for Teacher Licensing, Sacramento, CA. Princeton, N.J.: Educational Testing Service, 1975.

McNemar, Quinn. "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages." Psychological Bulletin, 1940, 37: 747.

Peckham, P.D., Glass, G.V. and Hopkins, K.D. "The Experimental Unit in Statistical Analysis: Comparative Experiments with Intact Groups (No. 28)." Boulder, Colo.: University of Colorado, Laboratory of Educational Research, March, 1969. (b)

Poynor, Hugh. "Selecting Units of Analysis" In: Borich, Gary D. (Ed.), Evaluating Educational Programs and Products. Englewood Cliffs, N.J.: Educational Technology Publication, 1974, Chapter 15.

Prais, S.J. and Aitchison,J. "The Grouping of Observations in Regression Analysis." Revue of International Stat. Inst., , 22: 1-22.

Raths, J. "The Appropriate Experimental Unit." Educational Leadership, 1967: 263-266.

Robinson, W.S. "Ecological Correlations and the Behavior of Individuals." Amer. Sociological Review, 1950: 351-357.

Rock, Donald A., Baird, Leonard L. and Linn, Robert, L. "Interaction between College Effects and Students' Aptitudes." Amer. Educ. Res. J. , 1972, 10: 149-161.

Thorndike, E.L. "On the Fallacy of Inputing and Correlations Found for Groups to the Individuals or Smaller Groups Composing Them. Amer. J. of Psych., 1939, 52: 122-124.

Walker, Helen M. "A Note on the Correlation of Averages." J. of Educ. and Psych., 1928, 19: 636-642.

Ward, J.H., Jr. "Hierarchical Grouping to Optimize an Objective Function." J. Amer. Stat. Assoc., 1963, 58: 236-244.

Werts, C.E. and Linn, R.L. "A Regression Model for Compositional Effects." Unpublished paper. Princeton, N.J.: Educational Testing Service, 1969.

Werts,C.E and Linn, R.L. "Considerations when Making Inferences within the Analysis of Covariance Model." Educ. and Psych. Measurement, 1971, 31: 407-416.

Wiley, D.E. "Design and Analysis of Evaluation Studies." In: Wittrock, M.D. and Wiley, D.E. (Eds.), The Evaluation of Instruction: Issues and Problems. New York: Holt, Rinehart and Winston, 1970.

Wittrock, M.D. and Wiley D.E. The Evaluation of Instruction: Issues and Problems. New York: Holt, Rinehart and Winston, 1970.

Lindquist, E.F. Statistical Analysis in Educational Research. New York: Houghton Mifflin, 1940.

Hannan, M. Aggregation and Disaggregation in Sociology. Lexington, Mass.: Lexington Books, 1971.

APPENDIX A: Comparison of the Werts-Linn Approach with Hannan-Burstein and Feige-Watts Approaches for Assessing the Difference between Grouped and Ungrouped Estimators of Regression Coefficients.

In my 1975 AERA presentation (Burstein, 1975a), two techniques (Feige and Watts, 1972; Hannan and Burstein, 1974) for assessing differences between grouped and ungrouped coefficients in the single-regressor case were discussed. These techniques can be used to determine whether observations grouped according to a specific variable Z (called "A" in Hannan and Burstein and expressed in the grouping matrix $\underset{\sim}{G}$ in Feige and Watts) yield accurate estimates of the corresponding ungrouped coefficients.

Developments over the past year suggest the need for augumentation of the list of methods for assessing differences between grouped and ungrouped coefficients. At a conference sponsored by the Measurement and Methodology Division of NIE at Annapolis, Maryland, educational statisticians raised several questions regarding the F-statistics that Feige and Watts (1972) use to assess the divergence between grouped and ungrouped estimators. Burstein (1975, p. 119) describes the main concern of the statisticians about the appropriateness of the Feige-Watts F-test. Below we discuss the questions raised and provide an alternative test to the one provided by Feige and Watts.

Recently, Firebaugh (personal communication) suggested that a regression model described earlier by Werts and Linn (1971) which includes both the regressor, X_{ij}, and the "compositional effect", $\bar{X}_{.j}$, (the mean on the regressor for the jth group) is more suitable for assessing the differences of grouped and ungrouped regression coefficients. Preliminary findings from our empirical analyses indicate that the Werts-Linn model has potential in the single-regressor case. As our examples demonstrate, the bias from grouping can be more accurately predicted from their model (Werts and Linn, 1971) than from the "Structural-Equations" model (Hannan and Burstein, 1974) recommended by Hannan and Burstein under a

variety of conditions. Further comparison of W-L with the Feige-Watts (F-W) F-statistic yielded the same incongruities evidenced in the comparison of H-B with F-W. Below we first outline the W-L approach and restate the equations for the structural equations (H-B) and for the original and alternative Feige-Watts models. Data from a large midwestern university are used in an empirical illustration of the three models. Finally, we point out some potential advantages and disadvantages of each approach and offer directions for future comparative research.

It is important to remember that we are discussing the comparative merits of estimating β_{YX} from the ungrouped model

$$(A.1) \qquad Y_{ij} = \alpha + \beta_{YX}X_{ij} + u_{ij}$$

from grouped data. The fact that another model may be appropriate (as when the single regression model is misspecified) is not considered here though it is addressed in Hannan and Burstein, in Firebaugh, in Feige and Watts and elsewhere.

Model Discription

Werts and Linn. The original basis for the model suggested by Firebaugh is a ETS technical report on regression analysis for compositional effects by Werts and Linn (1969) and their subsequent paper on making inferences within the Analysis of Covariance model (Werts and Linn, 1971). The model Werts and Linn discuss (1971, pp. 407-08) is

(Note: We use Z_j where they use A_j and u_{ij} where they use e_{ij})

$$(A.2) \qquad Y_{ij} = Z_j + B_w X_{ij} + u_{ij}$$

where Z_j = the Y-intercept of the Y-on-X regression for group $j(=\bar{Y}_j - B_w\bar{X}_j)$.

B_w = pooled within-group regression slope

and

u_{ij} = usual error term independent of Z_j and X_{ij} .

When the standard ANCOVA model is applied to the analysis of compositional effects, the equation is written as (our notation)

$$(A.3) \qquad Y_{ij} = \alpha' + \beta_{YX\cdot\bar{X}} X_{ij} + \beta_{Y\bar{X}\cdot X} \bar{X}_{\cdot j} + u_{ij}$$

where $\beta_{YX\cdot\bar{X}}$ = the pooled within-group regression coefficient and

$\beta_{Y\bar{X}\cdot X}$ = difference of the pooled within-group coefficient from the between-group regression coefficient $(\beta_{\bar{Y}\bar{X}} - \beta_w)$,

Also, according to Werts and Linn (1971, pp. 414),

$$\beta_{Y\bar{X}\cdot X} = \beta_{Z\bar{X}} \ .$$

$\beta_{Y\bar{X}\cdot X}$ is called the "compositional" effect and it represents the effects of group composition after holding constant individual influences on outcome. Werts and Linn point out that (a) the analysis of "compositional" effects corresponds to the ANCOVA model in which treatments are not independent of the covariate, (b) the slope $\beta_{Y\bar{X}\cdot X}$ $(=\beta_{Z\bar{X}})$ represents the net influence of composition and (c) the "compositional" effect is part of the "treatment" effect in the ANCOVA model.

Though these earlier treatments do not make the point explicit, subsequent communication with Firebaugh and with Linn indicate that $\beta_{Y\bar{X}\cdot X} = 0$ in equation A.3 is a necessary and sufficient condition for the estimator of $\beta_{\bar{Y}\bar{X}}$ from the model

$$(A.4) \qquad \bar{Y}_{\cdot j} = \alpha_{\bar{Y}\bar{X}} + \beta_{\bar{Y}\bar{X}} \bar{X}_{\cdot j} + \bar{u}_{\cdot j}$$

to be a consistent estimator of β_{YX} from equation A.1. That is, the grouped estimator, $B_{\bar{Y}\bar{X}}$ of $\beta_{\bar{Y}\bar{X}}$ will be an inconsistent estimator of β_{YX} (the ungrouped parameter) only when $\beta_{Y\bar{X}\cdot X} \neq 0$. Or, bias occurs whenever there is a "compositional" effect.

So one need only examine the estimator of $\beta_{Y\bar{X}\cdot X}$ from the analysis to assess the consistency of the estimation from grouped observations. Note also that the metric of grouping variable is irrelevant since the analytical model for assessing differences (Equation A.3) uses the individual observations and the group mean for the regressor in its regressor set.

Hannan and Burstein. Hannan and Burstein (1974) dealt primarily with the case of an ordered grouping characteristic.[1] They proposed that the

grouping variable Z be incorporated in A.1 and its structural relations to X and Y be examined by estimating the parameters from the model:

$$\text{A.5} \qquad Y_{ij} = \alpha + \beta_{YX \cdot Z} X_{ij} + \beta_{YZ \cdot X} Z_{ij} + w_{ij} ,$$
$$X_{ij} = \lambda + \beta_{XZ} Z_{ij} + v_{ij}$$

where $Z_{ij} = \bar{Z}_{\cdot j}$ ($i = 1..., n_j$ persons in groups $j = 1, ..., m$, respectively).

One version of their bias formula is

$$\text{A.6} \qquad \theta = B_{\bar{Y}\bar{X}} - b_{YX}$$
$$= \beta_{YZ \cdot X} \beta_{XZ} \left[\frac{\sigma^2_{\bar{Z}}}{\sigma^2_{\bar{X}}} - \frac{\sigma^2_Z}{\sigma^2_X} \right]$$

from which they deduce that there is no aggregation bias when any of the following condidtions hold:

(i) Z has no effect on Y net of X: $\beta_{YZ \cdot X} = 0$.

(ii) Z has no effect on X: $\beta_{XZ} = 0$.

(iii) the ratio of the variances of Z and X between groups equals the ratio of their total variances.

Since $Z_{ij} = \bar{Z}_{\cdot j}$, $\sigma^2_Z = \sigma^2_{\bar{Z}}$, and condition (iii) becomes

(iii') $\sigma^2_X = \sigma^2_{\bar{X}}$,

and thus (A.6) can also be written as

$$\text{A.6'} \qquad \beta_{YZ \cdot X} \beta_{XZ} \sigma^2_Z \left[\frac{\sigma^2_X - \sigma^2_{\bar{X}}}{\sigma^2_X \sigma^2_{\bar{X}}} \right]$$

[Note: If $\sigma^2_X = 0$ or $\sigma^2_{\bar{X}} = 0$, the bias is indeterminant.]

In the context of the present question, we can infer from the Hannan-Burstein appraoch that the difference between grouped and ungrouped coefficients ($\beta_{\bar{Y}\bar{X}} - \beta_{YX}$) is a function of conditions (i)–(iii'). Several years of experience with models of the type described here indicate that the

relation of the grouping variable, Z, to outcome Y after fixing the regressor, X, (i.e., $\beta_{YZ \cdot X}$) is the crucial determinant of the divergence of grouped and ungrouped regression coefficients.

It is also pertinent to note that if conditions (i)—(iii') are not satisfied, the initial model (A.1) is misspecified due to a correlation between its regressor and an omitted regressor which is incorrectly incorporated in the disturbance. Thus, the reason that $\beta_{\overline{Y}\overline{X}}$ diverges from β_{YX} may be that β_{YX} was inappropriate in the first place.

Feige and Watts. The details of the Feige-Watts technique can be found in the original source (1972) and in later discussions by Burstein (1975a, 1975b). Feige and Watts developed a measure of the divergence between estimators of grouped and ungrouped coefficients, $\bar{\underset{\sim}{\beta}}$ and $\bar{\underset{\sim}{\beta}}$, in the multivariate case. They attributed the divergence to three sources -- (i) specification bias, (ii) bias introduced by grouping that is not independent of the disturbances from the structural model and (iii) sampling error induced by the loss of information through grouping.

For the sake of comparison, we will provide below a single-regressor version of their "F-Statistic" for divergence. First, however, we describe the components of their statistic as it was developed so that our recommended alterations can be better understood.

The Feige-Watts F-statistic is predicted on the following developments[2]:

(1) Under the null hypothesis that the grouped and ungrouped coefficients are the same ($\bar{\underset{\sim}{\beta}} = \underset{\sim}{\beta}$), the divergence between grouped and ungrouped estimators (B and b, respectively) has a zero mean

$$(\Delta = \underset{\sim}{b} - \underset{\sim}{B} = \underset{\sim}{0})$$

and a variance-covariance matrix

$$\left(C(\Delta) = \sigma_u^2[(\bar{\underset{\sim}{X}}'\bar{\underset{\sim}{X}})^{-1} - (\underset{\sim}{X}'\underset{\sim}{X})^{-1}]\right)$$

where $\bar{X}'\bar{X}$ and $X'X$ are the between-groups and total matrices of sum of squares and cross-products for the regressor and σ_u^2 is the variance of the disturbance.

(2) Let $\bar{e} = \bar{Y} - \bar{X}\,B$ so the $\bar{e}'\bar{e}$ is the sum of squared residuals from the between-groups regression.

(3) According to Feige and Watts,

$$Q_1 = \frac{\Delta'[(\bar{X}'\bar{X})^{-1} - (X'X)^{-1}]^{-1}}{\sigma_u^2}$$

and

$$Q_2 = \frac{\bar{e}'\bar{e}}{\sigma_u^2}$$

are distributed as χ^2 with k and m-k degrees of freedom, respectively, where m = number of groups and k is the number of regressors.

(4) Assuming correct specification and independence of $\bar{X}$ and $\bar{u}$, Feige and Watts claim that

$$F = \frac{Q_1/k}{Q_2/(m-k)} \qquad \text{(A.7)}$$

is distributed as an F-statistic with k and m-k degrees of freedom. Values of their F-statistic beyond the critical values of the F-distribution indicate differences between estimators that cannot be attributed to sampling error.

It can be shown that the single-regressor analogue of the equation (A.7) can be written as

$$F = \frac{(B_{\bar{Y}\bar{X}} - b_{YX})^2\left[\frac{1}{SS(\bar{X})} - \frac{1}{SS(X)}\right]^{-1}}{SS(\overline{res})/m-1} \qquad \text{(A.8)}$$

where $SS(\bar{X})$ = between-group sum of squares

$SS(\overline{res})$ = sum of squares for residuals from the between-groups regression

Thus if F from (A.8) were significant with 1 and m-1 d.f.[3], then the grouped estimator diverges significantly from the ungrouped estimator.

Alternative Statistic for Assessing Divergence

Several questions have been raised with regard to the adequacy of the Feige-Watts statistic. Feige and Watts' conclusions about the independence and distributions of Q_1 and Q_2 have been challenged and the inherent asymmetry in the components of the numerator and denominator have been noted (Hubert, Olkin, Rubin, Timm, personal communications). We have not yet been able to determine the viability of the above mentioned criticism with the exception that Q_1 and Q_2 can be shown to be independent.

However, there is one other point in the development of the Feige-Watts statistic that represents a clear problem. It appears that Feige and Watts have chosen an inappropriate denominator for their F-test. In the behavioral sciences the traditional form of the F-test for differences in regression models takes the form:

$$F = \frac{(R_F^2 - R_R^2)/(df_F - df_R)}{(1-R_F^2)/(N-df_F)}$$

where

R_F = squared multiple correlation for the so-called "full" model (the more inclusive model)

R_R = squared multiple correlation for the "restricted" model

and

df_F, df_R = degrees of freedom for the full and restricted models, respectively.

There is no recognizable standard for interpreting the comparison of individual-level and aggregate regression models in this fashion. Intuitively, however, it is appealing to associate the individual-level model

with the "full" model above and the aggregate with the "restricted". If this interpretation is defensible, then the residual sum of squares from the individual-level regression ($\underset{\sim}{e}'\underset{\sim}{e}$, where $\underset{\sim}{e} = \underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{b}$) would seem to be more appropriate than Feige and Watts' choice for the denominator. Thus we propose the following alternative to Feige and Watts's F-Test:

$$(A.9) \qquad F_B = \frac{(b_{YX}-B_{YX})^2\left[\frac{1}{SS(\bar{X})} - \frac{1}{SS(X)}\right]^{-1}}{SS(res)/N-1}$$

where SS(res) = sum of squared residuals from the individual-level regression models.

In the remainder of the appendix, we will add subscripts to the F-test to designate the Feige and Watts version (F_{FW}) and our suggested alternative (F_B).

Empirical Example.

Table A.1 contains a summary of the alternative models and procedures for assessing differences described above. To simplify matters even further for our illustration, we standardized all observations before grouping. This also allows us to further simplify the formulas for differences and F-tests. After standardization, the following hold:

$$\sigma_X^2 = \sigma_Y^2 = \sigma_Z^2 = \sigma_{\underline{Z}}^2 = 1$$

$$\beta_{YX} = \rho_{YX},\ \beta_{YZ} = \rho_{YZ},\ \beta_{XZ} = \rho_{XZ}$$

$$SS(X) = N-1$$

$$E_X^2 = \sigma_{\underline{X}}^2$$

The two difference formulas become:

$$(A.10) \qquad \hat{\theta}_{WL} = \hat{\beta}_{Y\bar{X}\cdot X}\ (1-\hat{\sigma}_{\underline{X}}^2) \quad \text{(Firebaugh; Werts and Linn).}$$

and

$$\text{(A.11)} \qquad \hat{\theta}_{HB} = \hat{\beta}_{YZ \cdot X}\hat{\beta}_{XZ}\left(\frac{\sigma_X^2}{\sigma_X^2}\right) \quad .$$

In the alternate form of the F-test, $\frac{1}{SS(X)}$ is replaced by $\frac{1}{N-1}$ and SS(res)/N-1 in (A.9) becomes $1-R_{YX}^2$ where R_{YX}^2 denote the squared multiple correlation from the individual-level regression analysis.

Our data set contains 2676 observations from entering freshmen at a large midwestern university (see Burstein, 1974, 1975a; 1975b for further details about the data set) on a variety of variables. Here we estimate the standardized coefficients from the regression of academic self-appraisal (SRAA) on Achievement (ACH) and Achievement (ACH) on Aptitude (SAT).

The ungrouped equations are:

$$SRAA = (.529)ACH \quad [SE(b_{YX}) = .032]$$

and

$$ACH = (.829)SAT \quad [SE(b_{YX}) = .0105]$$

Tables A.2 through A.4 provide the parameter estimates from the Hannan-Burstein and the Werts-Linn approaches, the resulting expected differences and F-statistic calculated according to the Feige-Watts and alternative models for the SRAA-on-ACH regression. Tables A.5 and A.7 provide the same information for the ACH-on-SAT regression.

Results from the Empirical Analysis

(1) The Werts-Linn approach predicts differences more accurately than the Hannan-Burstein approach in 14 of 20 cases, substantially so (better by .05) in 6 cases.

(2) The Hannan-Burstein approach provides more accurate prediction in over 5 cases, in one case by more than .05.

(3) The Werts-Linn approach performs poorest for the approximation to random grouping (ID1). This is to be expected since the W-L model

capitalizes on any linear relation between group membership and the dependent variable.

(4) In general, however, W-L and H-B procedures identify the same grouping variables as yielding small differences and large differences.

(5) The grouping methods that were expected to yield small differences between grouped and ungrouped coefficients had small F-statistics by both Feige-Watts and Burstein tests.

(6) With the exception of grouping by ID1 (supposedly random grouping), large F-statistics coincided with large expected differences in every case for the alternative F-test while the Feige-Watts F-tests were not significant under conditions of large expected difference for two grouping methods.

Conclusions

There are several standards by which we can judge the relative merits of the techniques described above. The questions we might ask are:

1. How accurate are the predictions of descrepancy between grouped and ungrouped coefficients generated by the technique?
2. How easy is it to use each technique?
3. How adaptable is the technique to the nominal grouping variables we face most often in educational research?
4. What is lost or gained for each technique when we move to more complex models involving multiple regressors?
5. If, as is often the case, it is impossible to reconstruct relations at the individual level, what happens to the utility of the proposed technique?

The complications alluded to in question 5 are of a different order of magnitude than those in the other questions and this question will be considered separately. We discuss the comparative advantages and disadvantages of the WL and HB models first and then talk about the relative merits of the alternative F-tests and their utility in relation to the WL and HB methods.

Comparing W-L and H-B. Though both the Werts-Linn and Hannan-Burstein approaches identify essentially the same "good" and "poor" grouping methods, the Werts-Linn approach yields slightly more accurate predictions in the single-regressor case and their procedure for estimating differences does not have to be altered when the grouping variable is nominal (such as school). In the full information situation (both individual and grouped data accessible), it is equally easy to classify grouping variables as good or poor using either approach.

The primary disadvantage of the Werts-Linn approach is that there has to be one variable consisting of group means associated with each regressor in the model, which quickly becomes tideous when there are multiple regressors. When the grouping variable is ordered, this presents less of a problem in the structural equations approach advocated by Hannan and Burstein as only a single grouping variable is entered in the analysis.

There is as of yet no consensus regarding the best method of modeling the effects of a nominal grouping variable (Burstein (1974, 1975b) discusses some alternative strategies.), especially in the multiple-regressor case. With both approaches, we would need to generate a single variable (or some small set of variables) with metric properties that distinguish among the groups. Otherwise, the modeling of the multivariate case will be too cumbersome.

Alternative F-Tests. The utility of the alternative F-tests is not affected by either having a nominal grouping variable or having multiple-regressors. Their chief problem is that, in the final analysis, the F-tests are global tests and therefore are insensitive to differential fitting of regression parameters in the multiple-regression case. There is evidence that grouping differentially affects the estimation of regression coefficients. For example, grouping on one regressor yields a more consistent estimate of its own parameter than of the other parameters in

multiple-regression models (Burstein, 1975b; Burstein and Hannan, 1975). If this is the case, we would prefer to use a technique that will predict where the differences are largest and smallest, and both the Werts-Linn and Hannan-Burstein approaches are better suited for this task.

One further complication with using the F-tests is that of ease of calculation in the multiple-regressor case. We (Burstein and Hannan, 1975) are in the process of modifying the Feige-Watts software so that the F-statistics can be generated for a variety of data sets.

If the present examples are any indication, the F-statistics generated by our alternative test more closely approximate the results from other approaches than the Feige-Watts statistics do. Furthermore, the alternative test is somewhat easier to calculate (constant denominator over a variety of different grouping methods). However, more empirical work, and perhaps computer simulation should be carried out before making any final judgment in the comparison of alternative F-tests.

Utility in the Limited Information Case. Often in educational research, we analyze grouped data simply because the data on individuals are unavailable or unobtainable in disaggregated form. For example, schools often report only school-mean test scores and demographic characteristics to the public, and in many cases, fail to retain individual information.

This limited information situation is troublesome for all the methods discussed above. One needs to be able to estimate $\beta_{Y\bar{X}\cdot X}$ and E^2_X for the Werts-Linn approach and neither σ_{YX} nor σ^2_X, which enter the calculations, is ascertainable from grouped data alone. The same problem exists in applying the Hannan-Burstein a ach as σ_{YX} is necessary for calculating $\beta_{YZ\cdot Y}$ and σ^2_X also enters the expected difference formula.

We (Burstein, 1974, 1975a, 1975b; Burstein and Hannan, 1975) have explored the possibility of using an approximation for the expected difference formula which does not require the investigator to know σ_{YX}.

Our results to date are mixed, and we are not yet prepared to offer firm guidelines on how to proceed with one exception. If the grouping method has a stronger relationship to the outcome variable than to the regressor ($\rho_{YZ} > \rho_{XZ}$, or for the Werts-Linn approach, $E_Y^2 > E_X^2$), large differences between grouped and ungrouped coefficients are inevitable. To some degree the reverse is true -- small differences are associated with cases where $\rho_{XZ} > \rho_{YZ}$ -- especially when the difference is magnitude between ρ_{YZ} and ρ_{XZ} is large.

We experience even greater difficulties in applying either F-test in the limited information case. Estimates of individual-level regression coefficients and the covariance matrix for the regressors enter the calculation of the tests and neither is obtainable in the "grouped data only" situation. The limited usefulness of these techniques under the present circumstances is not surprising since Feige and Watts first proposed their techniques for application when individual-level data is available but must remain confidential (cf. 1972). Furthermore, economists in general have focused on applications where the existence of individual-level data is not a problem (A notable exception is Haitovsky (1966; 1968).).

Where do the caveats implied above leave us? Well, unless viable modifications of the techniques described above can be generated that are less sensitive to problems of limited information, we are left with sound alternative mathematical models with limited utility for addressing the practical problems encountered in educational research.

Perhaps this sobering analysis is for the best. Perhaps, educational record keeping and data bases will become more informative if the persons charged with the responsibility of collecting educational data are made aware of the methodological morass that results from failing to keep track of data at the individual level. For, whether one is interested in purely academic or in policy questions, there is no viable substitute for the

following guides for collecting and maintaining educational data (Burstein, 1975a, 1976; Burstein and Knapp, 1975; Burstein and Smith, 1975):

1. Measure all variables at their lowest possible level.
2. Data from individual students should be matched with data from their teachers/classrooms and characteristics of their school setting.
3. Keep track of all information at its lowest possible level with retrieval capabilities at multiple levels.

Anything less can remove the possibility of applying the "appropriate" analysis procedure to answer the questions that the policy maker and/or researcher wants answered.

Table A.1. Alternative models for assessing differences between grouped and ungrouped regression coefficients--single-regressor case.

PURPOSE: To assess differences between $B_{\bar{Y}\bar{X}}$ (between group coefficient) and b_{YX} (unbiased estimate of β_{YX} from individual-level data).

BASIC MODELS:

Werts and Linn; Firebaugh	Hannan and Burstein (Structural Equations)
$Y_{ij}=\alpha'+\beta_{YX\cdot X}\bar{Y}_{ij}+\beta_{Y\bar{X}\cdot X}\bar{X}_{\cdot j}+u_{ij}$	$Y_{ij}=\alpha+\beta_{YX\cdot Z}X_{ij}+\beta_{YZ\cdot X}Z_{ij}+w_{ij}$
with $\beta_{Y\bar{X}\cdot X}=\beta_{Z\bar{X}}$	$X_{ij}=\lambda+\beta_{XZ}Z_{ij}+v_{ij}$
from $Z_{ij}=\zeta+\beta_{Z\bar{X}}\bar{X}_{j}+q_{ij}$	with $Z_{ij}=\bar{Z}_{\cdot j}$

PREDICTED DIFFERENCE:

Werts and Linn	Hannan and Burstein
DIFFERENCE = $\hat{\theta}_{WL}=\hat{\beta}_{Y\bar{X}\cdot X}(1-E_X^2)$	DIFFERENCE = $\hat{\theta}_{HB}=\hat{\beta}_{YZ\cdot X}\hat{\beta}_{XZ}\left(\frac{\hat{\sigma}^2_{\bar{Z}}}{\hat{\sigma}^2_X}-\frac{\hat{\sigma}^2_Z}{\hat{\sigma}^2_X}\right)$
	$=\hat{\beta}_{YZ\cdot X}\hat{\beta}_{XZ}\hat{\sigma}^2_Z\left(\frac{\hat{\sigma}^2_X-\hat{\sigma}^2_{\bar{X}}}{\hat{\sigma}^2_X\hat{\sigma}^2_{\bar{X}}}\right)$
where $E_X^2=\frac{SS(\bar{X})}{SS(X)}$	since $\hat{\sigma}^2_Z=\hat{\sigma}^2_{\bar{Z}}$

F-STATISTICS

Feige-Watts	Burstein
$F=\frac{(b_{YX}-B_{\bar{Y}\bar{X}})^2\left(\frac{1}{SS(\bar{X})}\right)-\left(\frac{1}{SS(X)}\right)^{-1}}{SS(\overline{res})/m-1}$	$F=\frac{\left(b_{YX}-B_{\bar{Y}\bar{X}}\right)^2\left[\frac{1}{SS(\bar{X})}-\frac{1}{SS(X)}\right]^{-1}}{SS(res)N-1}$
with 1 and m-1 d.f.	with 1 and N-1 d.f.
where $SS(\overline{res})$=sum of squares for residuals from the between-groups regression	where SS(res)=sum of squares for residuals from the individual-level regression
m=number of groups formed.	N=total number of persons.

Table A.2. Estimates of parameters relating ACH(X) and SRAA(Y) to possible grouping variables (Z)[a].

Variable Name	Group Size (m)	Parameter Estimates				
		$\hat{\beta}_{YX \cdot Z}$	$\beta_{YZ \cdot X}$	β_{XZ}	β_{YZ}	$\hat{\sigma}_{\bar{X}}$
ID1	10	.528 (.0164)	-.011 (.0164)	-.042 (.0193)	-.033 (.0193)	.078
SAT2	13	.194 (.0282)	.406 (.0282)	.827 (.0109)	.566 (.0160)	.831
PARINC	10	.527 (.0164)	.028 (.0164)	.070 (.0193)	.064 (.0193)	.122
ACH2	10	.460 (.0896)	.070 (.0896)	.983 (.0035)	.522 (.0165)	.984
POPED	6	.519 (.0165)	.073 (.0165)	.139 (.0192)	.145 (.0191)	.150
NOBOOK	5	.511 (.0164)	.122 (.0164)	.146 (.0191)	.196 (.0190)	.148
HSPHYS	5	.515 (.0173)	.046 (.0173)	.318 (.0183)	.209 (.0189)	.365
HSMATH	5	.561 (.0187)	-.066 (.0187)	.479 (.0170)	.202 (.0189)	.489
SRAA2	5	.139 (.0099)	.819 (.0099)	.476 (.0170)	.885 (.0090)	.481
PARASP	5	.520 (.0162)	.138 (.0162)	.066 (.0193)	.172 (.0190)	.077

[a]All variables have been standardized prior to grouping so that $\sigma_Y = \sigma_X = \sigma_Z = 1$, $\beta_{XZ} = \rho_{XZ}$, and $\beta_{YZ} = \rho_{YZ}$.

[b]Numbers in parentheses are standard errors of regression coefficients.

Table A.3. Estimates of parameters relating ACH(=X) and SRAA(=Y) to group means on the regressor (=$\bar{X}$) for selected grouping variables[a].

Variable Name	Group Size (m)	Parameter Estimates				
		$\hat{\beta}_{YX\cdot\bar{X}}$	$\hat{\beta}_{Y\bar{X}\cdot X}$	$\hat{\beta}_{X\bar{X}}$[b]	$\hat{\beta}_{Y\bar{X}}$	$\hat{\sigma}_{\bar{X}}$
ID1	10	.531 (.0165)[c]	.047 (.2085)	-.387 (.2443)	-.159 (.2454)	.078
SAT2	13	.220 (.0286)	.448 (.0342)	.992 (.0129)	.666 (.0193)	.838
PARINC	10	.530 (.0166)	.028 (.1359)	1.000 (.1570)	.558 (.1585)	.122
ACH2	10	.522 (.0904)	.009 (.0920)	1.000 (.0036)	.531 (.0169)	.984
POPED	6	.522 (.0166)	.388 (.1109)	1.000 (.1275)	.911 (.1283)	.150
NOBOOK	5	.513 (.0165)	.820 (.1120)	1.000 (.1297)	1.333 (.1292)	.148
HSPHYS	5	.525 (.0177)	.047 (.0486)	1.000 (.0494)	.572 (.0522)	.365
HSMATH	5	.567 (.0188)	-.153 (.0386)	1.000 (.0345)	.414 (.0389)	.489
SRAA2	5	.134 (.0099)	1.719 (.0206)	1.000 (.0353)	1.853 (.0187)	.481
PARASP	5	.522 (.0164)	1.425 (.2126)	1.000 (.2500)	1.947 (.2489)	.077

[a] X and Y have been standardized so that $\sigma_X = \sigma_Y = 1$.

[b] $E(\beta_{X\bar{X}}) = 1$ for all regressions.

[c] Numbers in parenthesis are the standard errors of the regression coefficients.

Table A.4. Assessment of differences between grouped and ungrouped coefficients ACH -- A comparison of Werts-Linn with Hannan-Burstein predictions and alternative F-tests.[a]

Grouping Variables[b]	$B_{\bar{Y}\bar{X}}$	$SE(B_{\bar{Y}\bar{X}})$[c]	Observed Difference $B_{\bar{Y}\bar{X}} - b_{YX}$	Werts-Linn Predicted Difference[d]	Hannon-Burstein Predicted Difference[d]	Feige-Watts F-test[d]	Burstein Alternative F-test[d]
ACH2	.531	.0615	.002	.000	.002	.049	.667
PARINC	.558	.1314	.029	.027	.129	.051	.049
HSPHYS	.571	.0915	.043	.041	.095	.252	1.045
ID1	.442	.1831	-.087	.046	.075	.228	.173
HSMATH	.414	.0248	-.115	-.116	.100	27.82**	15.39***
SAT2	.671	.0670	.142	.133	.150	14.52**	166.67***
POPED	.911	.1626	.382	.380	.440	5.64	12.50***
NOBOOK	1.334	.1133	.805	.802	.800	51.76**	195.98***
SRAA2	1.853	.0631	1.324	1.321	1.295	571.66**	1962.79***
PARASP	1.946	.7339	1.417	1.417	1.519	3.75	44.73***

[a]Estimates from ungrouped data: $b_{YX} = .529$; $SE(b_{YX}) = .0032$.

[b]Ordered on the basis of size of observed difference

[c]Standard errors of between-group coefficients do not include component for bias in estimation.

[d]See Table A.1 for appropriate formulas.

**Exceeds the 95 percent critical value for F with 1 and m-1 d.f.

***Exceeds the 95 percent cricital value for F with 1 and N-1 d.f.

Table A.5. Estimates of parameters relating SAT(X) and ACH(Y) to possible grouping variables (Z)[a].

Variable Name	Group Size (m)	Parameter Estimates				
		$\hat{\beta}_{YX \cdot Z}$	$\hat{\beta}_{YZ \cdot X}$	$\hat{\beta}_{XZ}$	$\hat{\beta}_{YZ}$	$\hat{\sigma}_X$
ID1	10	.839 (.0105)	-.003 (.0105)	-.046 (.0193)	-.042 (.0193)	.069
SAT2	13	.884 (.0662)	-.042 (.0662)	.987 (.0031)	.828 (.0109)	.989
PARINC	10	.838 (.0106)	.006 (.0106)	.076 (.0193)	.070 (.0193)	.146
ACH2	10	.082 (.0061)	.916 (.0061)	.827 (.0109)	.983 (.0035)	.835
POPED	6	.838 (.0106)	.007 (.0106)	.157 (.0191)	.139 (.0192)	.169
NOBOOK	5	.844 (.0107)	-.025 (.0107)	.203 (.0189)	.146 (.0191)	.204
HSPHYS	5	.811 (.0107)	.109 (.0107)	.257 (.0187)	.318 (.0183)	.294
HSMATH	5	.765 (.0104)	.214 (.0104)	.346 (.0181)	.480 (.0170)	.349
SRAA2	5	.811 (.0123)	.054 (.0123)	.520 (.0165)	.476 (.0170)	.531
PARASP	5	.839 (.0106)	-.007 (.0106)	.087 (.0193)	.066 (.0193)	.101

[a]All variables have been standardized prior to grouping so that $\sigma_Y = \sigma_X = \sigma_Z = 1$, $\beta_{XZ} = \rho_{XZ}$, and $\beta_{YZ} = \rho_{YZ}$.

[b]Numbers in parenthesis are the standard errors of the regression coefficients.

Table A.6. Estimates of parameters relating SAT(=X) and ACH(=Y) to group means on the regressor (=$\bar{X}$) for selected grouping variables[a].

Grouping Variable	Group Size (m)	Parameter Estimates				
		$\hat{\beta}_{Y\bar{X}\cdot X}$	$\hat{\beta}_{Y\bar{X}\cdot X}$	$\hat{\beta}_{X\bar{X}}$	$\hat{\beta}_{X\bar{X}}$	$\hat{\sigma}_{\bar{X}}$
ID1	10	.839 (.0105)[c]	-.232 (.1490)	-.217 (.2738)	-.414 (.2737)	.069
SAT2	13	.875 (.0663)[c]	-.037 (.0671)	1.000 (.0031)	.838 (.0110)	.989
PARINC	10	.839 (.0106)	-.019 (.0721)	.984 (.1297)	.807 (.1302)	.148
ACH2	10	.089 (.0078)	1.080 (.0093)	1.000 (.0128)	1.168 (.0053)	.835
POPED	6	.838 (.0107)	.040 (.0634)	1.000 (.1133)	.878 (.1136)	.169
NOBOOK	5	.844 (.0107)	-.125 (.0526)	1.000 (.0927)	.719 (.0937)	.204
HSPHYS	5	.801 (.0107)	.436 (.0365)	1.000 (.0629)	1.237 (.0613)	.294
HSMATH	5	.778 (.0103)	.584 (.0296)	.878 (.0528)	1.266 (.0497)	.349
SRAA2	5	.815 (.0124)	.084 (.0234)	1.000 (.0321)	.899 (.0321)	.531
PARASP	5	.840 (.0106)	-.095 (.1051)	.999 (.1911)	.744 (.1916)	.109

[a] X and Y have been standardized so that $\sigma_X = \sigma_Y = 1$.

[b] $E(\beta_{X\bar{X}}) = 1$ for all regressions.

[c] Numbers in parenthesis are the standard errors of the regression coefficients.

Table A.7. Assessment of differences between grouped and ungrouped coefficients SAT -- A comparison of Werts-Linn with Hannan-Burstein predictions and alternative F-tests.[a]

Grouping Variables[b]	$B_{\bar{Y}\bar{X}}$	$SE(B_{\bar{Y}\bar{X}})$[c]	Observed Difference $B_{\bar{Y}\bar{X}} b_{YX}$	Werts-Linn Predicted Difference[d]	Hannan-Burstein Predicted Difference[d]	Feige-Watts F-test[d]	Burstein Alternative F-test[d]
SAT2	.838	.0190	-.001	-.001	-.001	.00	.000
PARINC	.817	.0598	-.022	-.018	.021	.14	.0986
POPED	.877	.0685	.039	.039	.038	.33	.399
SRAA2	.899	.0543	.060	.060	.072	.18	1.336
PARASP	.744	.0903	-.095	-.094	-.059	1.12	.837
NOBOOK	.718	.0372	-.121	-.120	-.174	11.10**	5.800***
ID1	1.053	.2168	.214	-.231	.029	2.47	2.002
ACH2	1.168	.0541	.329	.327	.329	197.73**	3647.420***
HSPHYS	1.237	.0422	.398	.398	.295	98.74**	136.820***
HSMATH	1.396	.0478	.557	.513	.531	149.28**	388.010***

[a]Estimates from ungrouped data: $b_{YX} = .839$; $SE(b_{YX}) = .0105$

[b]Ordered on the basis of size of observed bias.

[c]Standard errors of between group coefficients do not include component for bias in estimation.

[d]See Table A.1 for appropriate formulas.

**Exceeds the 95 percent critical value for F with 1 and m-1 d.f.

***Exceeds the 95 percent critical value for F with 1 and N-1 d.f.

APPENDIX FOOTNOTES

1. Burstein (1974, 1975b) discusses ways of handling nominal characteristics in structural equation models. Ironically, scaling by substituting the group mean on the regressor is one way of ensuring a sutiable metric for Z.

2. We have simplified the notation somewhat from earlier presentations to avoid the introduction of grouping matrices $\underset{\sim}{G}$ and $\underset{\sim}{H}$.