DOCUMENT RESUME

ED 099 306 SP 008 480

AUTHOR Cronbach, Lee J.

TITLE Consistency and Stability of Interaction Effects from Classroom to Classroom: Pilot Studies.

INSTITUTION Srencer Foundation, Chicago, Ill.

PUB DATE 30 Sep 74

NOTE 21p.

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE

DESCRIPTORS Analysis of Variance; Classroom Research; *Educational Research; Multiple Regression Analysis; Research Design; *Research Methodology; Research Problems; *Statistical Analysis

ABSTRACT

This interim report is concerned with the analysis of educational experiments and quasiexperiments where alternative teaching methods are applied to intact classes or where alternative programs are set up in samples of schools or communities. Generally, such studies have used the class, the school, or the community as the unit of sampling with the consequence that investigators have based the number of degrees of freedom in their statistical calculations on the number of classes. However, the author believes that a further distinction must be made: that is, between effects that operate at the class level and effects that operate within the class. The variance in an outcome measure can be divided into between-class and within-class components. The between-class variance can in turn be divided into variance predicted from the aptitude level of the class and a residual. The within-class variance can be subdivided into the effect predicted from the within-class regression equation and a residual. The regression equation with each single class can then be computed and the residual reduced still further. The author has developed methods for completing such analysis and applies them to previous studies to show the effect of such analysis on the outcomes of studies. (HMD)

Consistency and Stability of Interaction Effects From Classroom to Classroom: Pilot Studies

Lee J. Cronbach
Stanford University

Report to Spencer Foundation

September 30, 1974

Administrative

This project was undertaken as a "tooling-up" year, to define questions more precisely and to develop techniques that could be applied to the reanalysis of large-scale educational studies. The work has progressed as far as was expected during the year, although the specific questions that commanded attention have shifted somewhat. A grant to continue the work was approved by the Spencer Foundation, and this document is an interim report even though for administrative purposes, the original grant is considered to be terminated.

Joseph G. Deken, a doctoral student in the Department of Statistics, played a major role in the work reported. Noreen Webb, a doctoral student in educational psychology and measurement, joined the project in September and will continue through the next year. She helped in the preparation of this report.

Overview

This project is concerned with the analysis of educational experiments and quasiexperiments where alternative teaching methods are applied to intact classes, or where alternative programs are set up in samples of schools or communities. In such a study the class (or school or community) is the "unit of sampling". Recognition of that fact by investigators typically has had just one consequence: some of them have based the number of degrees of freedom in their statistical calculations on the number of classes. Beyond this, we argue, a conceptual distinction must be made: between effects that operate "at the class level" and effects that operate within the class. This becomes critically important in examining the effects of a treatment on different types of pupils.

The variance in an outcome measure (within a treatment) can be divided into between-class and within-class components. The between-class variance can in turn be divided into variance predicted from the aptitude* level of the class and a residual. In the process, the slope of the outcome-on-aptitude regression between classes is examined. If the slopes in the two treatments differ, Aptitude × Treatment interaction operates between classes; then one might pursue a policy of assigning high-aptitude classes to one treatment while applying the second treatment in low-aptitude classes.

The within-class variance (within either treatment) can be similarly subdivided into the effect predicted from the pooled-within-class regression equation, and a residual. Then the regression equation within each single class can be computed, and the residual reduced still further. The dependence of outcome on aptitude within the class may reflect how individual ability responds to the treatment or it may reflect social-psychological effects (e.g., response to competition). Differences in slope from one class to another (within a treatment) may be a consequence of ascertainable differences in the teachers' practices.

In the course of our research to date, we have identified some important decisions the data analyst needs to make in separating such components as the preceding paragraph describes. We have adopted one set of operations (grounded in SPSS computer programs), and have applied them to a significant study by G. L. Anderson, to demonstrate the kind of results expected. The new analysis undermines Anderson's conclusion that has

*Aptitude is a formal term applying to any individual characteristic at pretest. In some studies demographic characteristics are critical.

stood since 1941. This is unfortunate, but it demonstrates the value of the new procedure.

Judgment is required to choose between the alternatives open to the data analyst. While the choices we made in the Anderson analysis are, we believe, appropriate for that study, we intend to spell out the rationale underlying each choice, so that future investigators can make choices appropriate to their own studies. For example, we have chosen to regard classes as randomly sampled from a population, and to regard students within classes as "fixed" in the statistical sense. Under some circumstances an investigator might prefer to regard both as random, or to regard classes as fixed and pupils as random.

Our scheme leaves us in a position to test regression and interaction effects for significance. But we have become increasingly convinced that statistical inference ought to emphasize confidence intervals rather than tests of the null hypothesis.

We have studied Potthoff's (1954) extension of the Johnson-Neyman method. The method does appear to be suitable, though with the data collected in typical educational studies the confidence limits for effects turn out to be very wide. I.e., the typical investigation gives only skimpy information on the strength of effects in the population. By way of demonstration, we have applied Potthoff's method to a study by Austin Bond. Since confidence intervals for regression effects are difficult to comprehend unless they are displayed visually, we are making use of the computer to plot confidence intervals with one or two predictors. We have some further work to do to get displays that are adequately clear.

The method applies to the formation of confidence intervals for each within-treatment regression, as well as to the interactions Potthoff considered. Especially in quasiexperiments, this is likely to be more appropriate than to focus on the interaction directly.

From the point of view of the Foundation, the most important matter to report is that work has proceeded according to plan, that the tentative findings have confirmed the importance of work along these lines, and that we expect highly useful results from the successor grant, which will extend our pilot work into a more or less definitive report and will demonstrate the impor ance of analysis between- and within-classes. We shall elaborate in the following sections, primarily to give colleagues who see the report a more concrete idea of the direction our work is taking.

The work appears to have implications for a kind of question not considered in our proposal. Few issues regarding methodology in educational research are more significant than the legitimacy of covariance adjustments in quasiexperiments. Many of the controversies over, for example, the Westinghouse evaluation of Headstart had to do with this issue. The distinction between within-class and between-class regressions is not considered in the traditional analysis of covariance. We suspect that many analyses have inappropriately calculated regressions for purposes of adjustment by pooling individuals from all classes within treatments. While work on this extension of our thinking is not scheduled, we mention it here to indicate that our program of investigation has wider significance than our proposals have indicated. As a matter of fact, a current request for proposals from the Office of Education, for the Follow Through Planned Variation quasiexperiment, asks explicitly for analysis of both main effects and interactions at the between and within site levels. So our results will be ready none too soon!

Publications

Cronbach and Snow had worked from 1966 on studies on individual differences in response to instruction. Their monograph Aptitudes and Instructional Methods is nearing completion and will be sent to press (Irvington Press, New York) within a few weeks. It has been possible to insert into the manuscript a brief account of the work on this project, including specifically the findings and mathematical formulations summarized below. Since the monograph is expected to influence subsequent investigations, this grant has served to increase significantly the methodological soundness and the probable benefit from the Cronbach-Snow program of work.

A paper by Cronbach should also be mentioned. His invited address to the American Psychological Association in September, 1974 will be published in the February, 1975, American Psychologist under the title "Beyond the two disciplines of scientific psychology". That paper does not derive directly from the grant, yet it addresses the large issue to which the research is relevant: What are the limits of interpretation of educational research studies as conventionally conceived and designed? The technical work under this and the successor grant will converge with the philosophical argument in the 1974 paper, to raise sharp questions about the appropriate function of social-science inquiries pointed toward social policy. The contribution emerging from the grant is the evidence that studies of reasonable size probably cannot pin down quantitative effects definitively.

The Bond study reanalyzed

In 1940 Austin Bond published a doctoral dissertation of unusual methodological quality, much influenced by Helen Walker. He taught genetics in college by two procedures, one traditional and the other with constant attention to social implications. Using the Johnson-Neyman method of statistical analysis, he reported significant differences on many outcomes, but showed that the effects shifted from strong to negligible or even to reverse effects, depending on the characteristics of the student at the start of the course. This data set was used for our application of Potthoff's extension of the Johnson-Neyman method. (The study, as reported, does not lend itself to a separation of class and individual effects as only two classes were involved.)

Figure 1 displays confidence limits for one of Bond's outcomes. His analysis established regions of significance. These regions appear in our figure in the horizontal plane through the origin, bounded by an hyperbola. Only for persons in the outer corners could Bond conclude that the treatments differed significantly. The three-dimensional display given by Potthoff's method augments this information with an estimate of the possible range of the treatment effect at each point in the space defined by the pretests. Also, the curved surface defines the limiting positions of the regression plane. Any plane that does not slice into the surface might be the regression plane that describes the treatment difference in the population. This is a powerful method of taking into account the uncertainties introduced by quasicollinearity, a problem of considerable

Insert Figure 1 here

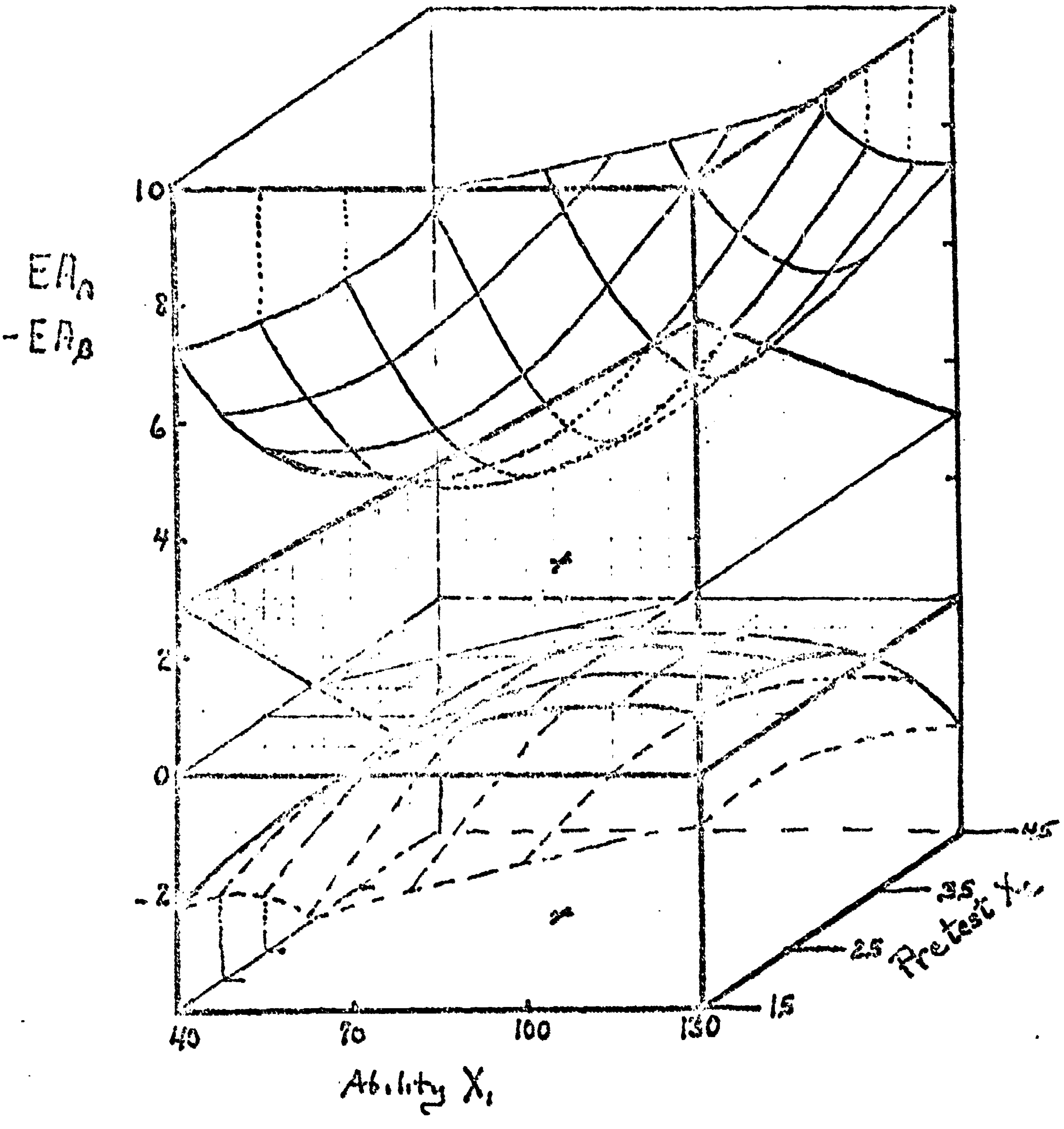

Figure 1. Confiuence limits for an interaction effects in Bond's Bond's study of genetics. EA is a test on Evaluation of Authorities.

importance in interpreting multiple regressions.

The analytic model

Our mathematical model can only be sketched here. First we hypothesize that, within a treatment, the outcome Y depends on three kinds of effect.

1. Individual effect of aptitude X_{∞}. Equals the expected Y for persons with a particular level of X_{∞}, averaged over all the classes in the population.

2. Class effect. Equals the Y value for the class, after adjusting for the individual effect of the members. Includes any experiences common to the group (e.g., teacher excellence, or high group morale). This effect may be a function of the class mean on X.

3. Individual-within-class effect. Described by the regression of Y or X within the class, adjusted for Effect 1. Reflects any special distribution of opportunity, encouragement, or the like within the class (e.g., teacher slows pace to fit the weaker students).

Now each of these effects may differ from one treatment to another. Hence there are three kinds of "Aptitude × Treatment interaction". having different theoretical and practical implications. This has not previously been recognized.

Analyzing data within a treatment, one obtains a between-classes slope and a within-class slope for the class. Thus one has three "unknowns" (three effect sizes) and only two observations (slopes). One can regard the between slope as a composite of Effects 1 and 2, the within slope as a composite of 2 and 3.

The sampling model we emphasize at present regards the set of classes as randomly sampled from the population of classes receiving the same treatment. This allows us to consider quasiexperiments in which nonrandom causes determine what treatment a particular class gets. (Even true experiments, if conducted on a large scale, are likely to be reduced to quasiexperiments when some classes depart from the plan.) We treat pupils as fixed within classes. In future work we shall trace the possibilities, limitations, and procedures associated with alternative models.

We have to date operated as if the pupil's X score is fixed. The work needs to be extended to consider measurement error (regarding X as a random observation from a collection of X observations on the fixed pupil). The resulting "correction for attenuation" takes a different form in studies with nested data than it does in the usual individual study.

Ignoring measurement error and analyzing solely within a treatment, we consider the outcome score Y_{pc} of person p in class c to be made up of components:

(1)	$Y_{pc} = \beta_o$	General mean
	$+ \beta_1 \bar{X}_c$	Between-class regression
	$+ \beta_c C$	Class-level residual; C is a dummy variable.
	$+ \beta_2 X_{2p}$	Expected (over classes) within-class regression. $X_{2p} = X_p - \bar{X}_c$
	$+ (\beta_{2c} - \beta_2) X_{2p}$	Specific within-class regression
	$+ \delta$	Residual

The first three terms add to $\bar{Y}_c$. Hence we make a decomposition:

(2)	$SS(Y_{pc})$	$= SS(\bar{Y}_c)$	$+$	$SS(Y_{pc} - \bar{Y}_c)$
	Total	Between classes		Within classes

Any parameter which is a function of the class means (μ_x, or $\sigma^2(Y)$, or β_1, etc.) can be defined taking class size into account as a weight, or weighting classes equally. The value of one or more parameters will shift between

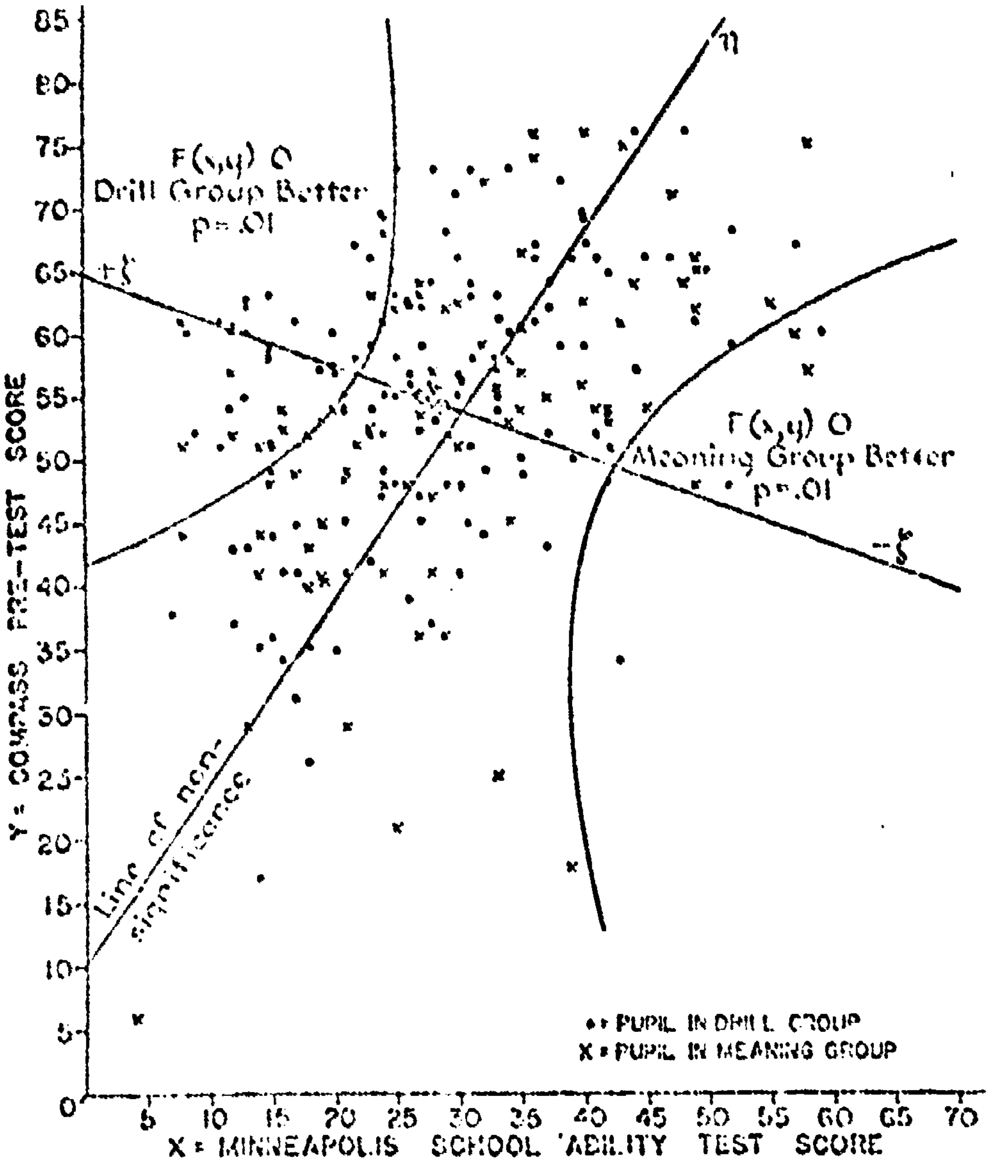

Figure 1. Comparison of the Low Subgroups on the Learning Exercise

Figure 2. Anderson's presentation of an interaction finding.

the two calculations, unless the $\bar{\bar{X}}$, $\bar{\bar{Y}}$ distribution is the same for large and small classes. We have chosen to estimate parameters in the weighted way, which weights pupils equally. This has not been the usual practice in educational research. The arguments for and against this formulation are to appear in our technical report.

The between-class sum of squares is now decomposed

$$(3)\qquad SS(\bar{Y}_c) = \underset{\substack{\text{Aptitude}\\ \text{effect}}}{SS_1\text{Regr}} + \underset{\substack{\text{Class}\\ \text{effect}}}{SS_1\text{Residual}}$$

These two sums of squares come out at the first step of a generalized regression analysis with (1) as the model. For that analysis, one must carry every student (not every class) as a case.

The within-class decomposition requires two more steps of regression.

$$(4)\qquad SS(\bar{\bar{Y}}_{pc}-\bar{Y}_c) = \underset{\substack{\text{General within-}\\ \text{class aptitude}\\ \text{effect}}}{SS_2\text{Regr}} + \underset{\substack{\text{Specific within-}\\ \text{class aptitude}\\ \text{effect}}}{(SS_3\text{Regr}-SS_2\text{Regr})} + \underset{\text{Unpredicted}}{SS_3\text{Residual}}$$

The strength of various effects is now estimated by the proportion of SS(Y) in each of the five segments.

The model generalizes to more than one aptitude. We find that for results to be interpreted it is necesary to convert the aptitude to dimensions that are close to orthogonal (in the pooled-within-treatments distribution).

The next question has to do with inferences about the size of effects in the population. Under the sampling model emphasized above, one can look at confidence intervals on the effects at the class level (the general mean and the between-class slope), followin; the Potthoff rationale. Almost invariably, the confidence intervals will be quite wide because data are available on only a few classes. Hence it will rarely be possible to draw definite conclusions about population parameters in instructional experiments of a practical size. As we see it, the use of the number of pupils as the basis for calculating confidence intervals or significance tests requires a special justification that may not often apply. We will spell out the grounds for such justification in next year's work.

If one is prepared to regard assignment of classes to treatments as random, the generalized regression analysis can be used to test the significance of aptitude-treatment interactions or the Potthoff method can be applied to the interaction.

The sketch above speaks of a single predictor X which is decomposed into between and within-class components. All the statistical arguments developed apply to multivariate prediction, as in the Anderson study.

The Anderson study reanalyzed

In a Minnesota doctoral dissertation completed in 1941, G. L. Anderson compared drill methods and meaningful methods of instruction in 17 fourth-grade class classrooms. The Johnson-Neyman analysis indicated differences on several outcomes, and showed that the effects depended on

Table 1

Regression Coefficients predicting Overall Outcome in the Anderson Data

	Treatment	Multiple regression		$s^2(Y{\cdot}X)$	Simple regression		$s^2(Y{\cdot}X)$	
		b_{Abil}	b_{Precom}		b_{Abil}	b_{Precom}	Abil	Precom
Between classes	D	0.06	0.75	499	−0.47	0.74	2063	498
	M	0.43	0.70	705	0.22	0.47	1365	1153
Within classes	D	0.41	0.75	2679	0.39	0.73	6656	3980
	M	0.42	0.65	2825	0.51	0.71	6242	4369

the abilities of the student at the beginning of the year. Anderson divided the classes into two subgroups (high or low mean initial abilities), and made separate analyses in the subgroups, counting the individual student as the sampling unit. His results are typified by Figure 2, from his short published report. It looked as if (under the conditions of the experiment) the meaningful methods then being advocated by educational theorists worked for underachievers and were harmful -- relative to drill -- for overachievers. This was a kind of answer to Binet's old problem of doing something about the "bright" pupil with a poor school record. It has its modern echo in various statements about the methods that work best with middle-class and lower-class children.

Anderson did not consider class effects in his regression analyses. We analyzed the data to separate class and individual effects.

In preparation for the reanalysis we dropped some cases with missing data (reducing N to 435), made regression estimates of missing scores for others, and eliminated one small class. We did not form subgroups with different initial abilities as Anderson had done. For the analysis reported here we formed a composite dependent variable by combining the Compass and van Wagenen posttests. Anderson used the several subtests as dependent variables. As predictors we used two aptitude variables: the Compass pretest and the Minnesota School Ability test. We converted the latter to ABIL (=MSAT - 0.42 Compass) to get orthogonal predictors. (Within a class or treatment the correlation was not necessarily zero.) We rescaled the two predictors and dependent variable to mean zero and s.d. 1.00, over all cases pooled.

The basis for degrees of freedom was not the number of students, but the number of classes. The small number of degrees of freedom made the confidence intervals for all between-class regression lines or planes very wide.

Our original analysis was a multiple regression. The multiple regression coefficients (Table 1) between classes are consistent with Anderson's finding, since the overachiever is a person high on ABIL, and overachieving classes seem to do better in the Meaning treatment. The computer printout drew our attention to the fact that ABIL accounts for very little variance in the Drill group. This led us to examine the plot of class means for raw scores (Figure 3). It is evident that Anderson's treatment groups had markedly dissimilar distributions of class means. He had assigned classes to methods on the basis of the teachers' preferences, and happened -- in a small sample -- to draw a set of Drill classes in which MSAT and PRECOM means correlated highly. The collinearity in the Drill classes nearly wiped out the variance of ABIL, the partial variate. Consequently, the multiple-regression weights in the between-classes analysis for Drill are not interpretable; many other sets of weights would give nearly the same multiple correlation.

Insert Figure 3 here

While we calculated within-class multiple regressions for each class, it is more useful here to report the calculation for classes pooled. The within-class estimates represent β_2. There is obviously no difference between the within-class regression slopes.

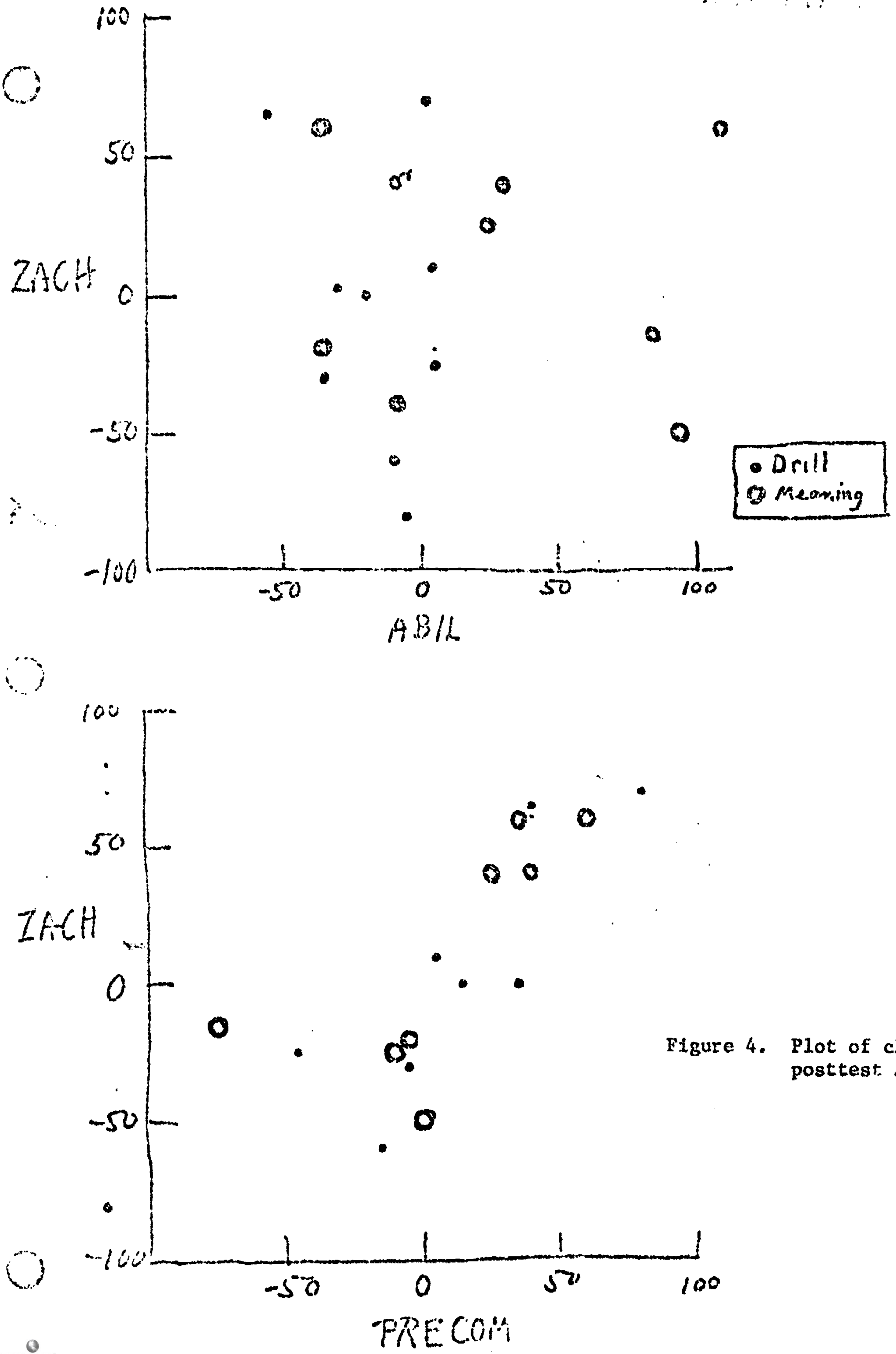

Figure 4. Plot of class means on posttest against pretests.

To reach conclusions less perturbed by collinearity at the between-classes level we calculated simple regression slopes. Again, the within-group slopes differ little. The difference between 0.39 and 0.51 for ABIL is in line with Anderson's conclusion but is a very weak effect.

The differences in the between-groups analysis are impressive, and entirely consistent with Anderson's conclusion. However, one must recognize the very large sampling error in a slope based on 8 or 9 classes. A plot of the means for ABIL, PRECOM, and ZACH (the standardized outcome) appears in Figure 4.

Insert Figure 4 here

In the chart for ABIL, the narrow range of ABIL in the Drill group should be noted. This follows from the collinearity. The slope is determined quite unreliably. In fact, if it were not for the one class at -55, 65, the slope in the Drill group would be slightly positive not negative. In the Meaning group, the slope is low, and not reliably positive. The two sets of points could easily be from the same distribution, hence we cannot say that the data support an ATI hypothesis. Yet this between-group effect must be the major source of Anderson's interaction.

In the PRECOM chart it is even more obvious that the two sets of cases form one distribution. The two leftmost cases account for the steeper slope in the Drill treatment.

Finally, it is noted that the residual variances are not so different as to suggest an important difference between treatments.

The variance decomposition (as far as we have carried it) is as follows:

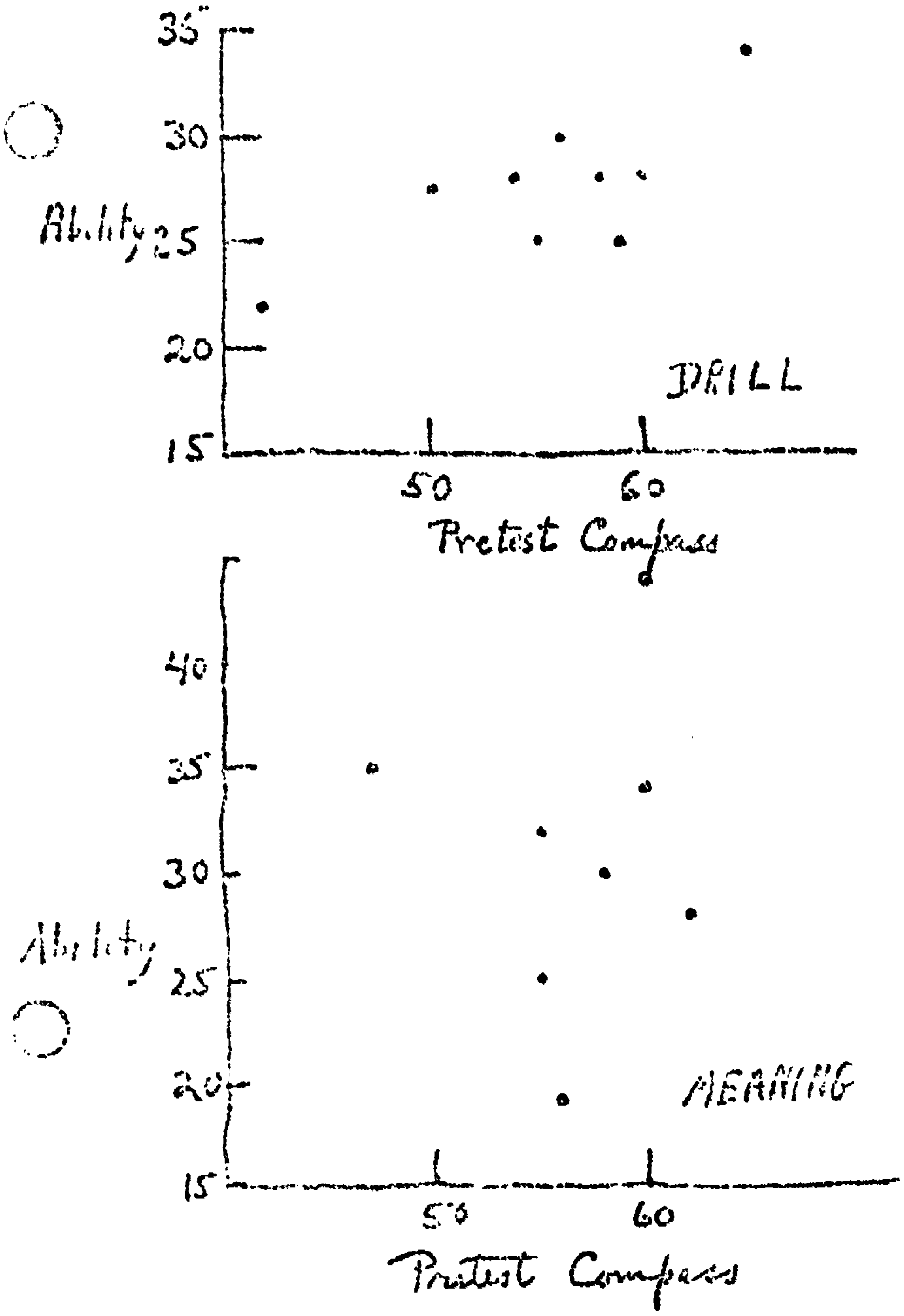

Figure 3. Plot of class means on pretests

	Drill		Meaning	
Total SS	9920	100.0%	1000	100.0%
Between classes	2134	21.5%	1498	14.9%
Predicted	1645	16.6%	793	7.9%
Residual	499	4.9%	705	7.0%
Within classes	7788	78.5%	8560	85.1%
Predicted by $\hat{\beta}_2$	5109	51.5%	5735	57.0%
Residual	2679	27.0%	2825	28.1%

It remains to comment on the specific within-class analyses. With about 20 cases in most classes, regression slopes are not to be regarded as precise, even when pupils are considered to be fixed. The slopes onto PRECOM ranged from 0.51 to 1.07 over the DRILL classes, and from 0.43 to 1.12 over the MEANING classes. The largest differences may imply that the teachers at the extremes were following quite different practices. As for ABIL, the range of within-class slopes was 0.14 to 0.92, and 0.22 to 0.71, in the respective treatments.

Conclusion

All in all, we are satisfied that our approach adds a great deal to the understanding of educational data and is likely to alter conclusions in many studies.

There are formal statistical problems to think through and explicate. Our computational methods can be made more efficient. And as we apply the method to further bodies of data in the present year we will no doubt learn more about the limits of the method and its value.

References

Anderson, G. L. Quantitative thinking as developed under connectionist and field theories of learning. In E. J. Swenson, et. al. Learning theory in school situations. University of Minnesota Studies in Education, No. 2. Minneapolis: University of Minnesota Press, 1949. Pp. 40-73.

Bond, A. D. An experiment in the teaching of genetics. Teachers College Contributions to Education. No. 797. New York: Teachers College, Columbia, 1940.

Potthoff, R. F. On the Johnson-Neyman technique and some extensions thereof. Psychometrika, 1964, 29, 241-256.