*Journal of Educational Psychology*
1975, Vol. 67, No. 6, 717-724

# Between-Class and Within-Class Effects in a Reported Aptitude × Treatment Interaction: Reanalysis of a Study by G. L. Anderson

Lee J. Cronbach and Noreen Webb
*Stanford University*

In this article the authors reanalyzed a study by G. L. Anderson. He reported finding an interaction of drill and meaningful methods of arithmetic instruction with student ability and achievement. Drill was superior for "overachievers" and meaningful instruction for "underachievers" in 18 fourth-grade classrooms. Pretest measures were the Minnesota School Ability Test and the Compass Survey Test. In a reanalysis to separate between-class and within-class components of the outcome on aptitude regression, the Aptitude × Treatment interaction finding disappeared. An apparent interaction in the between-class analysis was dismissed as unreliable. No interaction was found within classes. The importance of separating class, individual, and individual-within-class interaction effects is discussed.

Anderson, in a Minnesota doctoral dissertation completed in 1941 under the direction of T. R. McConnell, compared two styles of instruction in arithmetic (Anderson, 1941, 1949). Anderson's study has attracted interest even in recent years (e.g., Cronbach, 1966) as an impressive example of a finding of Aptitude × Treatment interaction. Anderson had a large sample (some 400 cases) and collected data on numerous outcomes of a full year of instruction. The study, taken as a whole, seemed to show that a drill method was superior for "overachievers," persons with good past achievement but relatively poor scores on a group test of general ability. In contrast, when the instruction emphasized meanings behind the processes, "underachievers" tended to do better than they did in the drill treatment (Cronbach, 1966).

We set out to trace the sources of that finding, using the full data reported in Anderson's appendix (Anderson, 1941). We were particularly interested in separating the between-class and within-class components of the effect. Aptitude × Treatment interaction effects may arise from the individual's response to the treatment (individual effect); in a laboratory study where persons are treated one at a time, effects are always interpreted in this way. When persons are treated in groups, however, at least two other causal explanations must be entertained. Some process may affect the group as a unit. For example, when the mean aptitude of the class is high, the teacher may crowd more material into the course. Consequently, the class as a whole may learn more; or it may, on the average, suffer from the fast pace. The third possibility is comparative effects within a group. Thus, if one method provides special opportunities or rewards for whoever is ablest within the class, the experience of a student with an IQ of 110 depends on whether the mean of his class is 100 or 120.

The comparative effect influences the within-class regression of outcome on aptitude. The class-level process influences the between-class regression. The strictly individual effect influences both regressions. No method has been found for separating the three kinds of effects in a classroom experiment, but one can contrast the between- and within-class regressions. If a strong interaction appears, for example, between classes and not within classes, a group level effect is the simplest explanation. But in interpretation one must be cautious, as two effects can work against each other to reduce a slope.

Anderson's analysis followed a tradition in instructional research that has persisted

This work was performed under a grant from the Spencer Foundation.

Requests for reprints should be sent to Lee J. Cronbach, School of Education, Stanford University, Stanford, California 94305.

(with few exceptions) to this day: he pooled all students within a treatment, disregarding class membership. He interpreted his results as if they reflected individual Aptitude × Treatment interaction effects. We have reanalyzed his data taking class membership into account, to demonstrate the methodology for those who plan future Aptitude × Treatment interaction studies and to learn more about the causal basis of Anderson's effect.

The reanalysis, we regret to say, overturns Anderson's conclusion. There is no evidence of Aptitude × Treatment interaction *within* classes. The regression onto Anderson's pretests within the drill classes was not appreciably different from the regression within the meaning classes. As for the between-class regressions, no conclusion could be reached. The number of classes was too small to give accurate information on the between-class regressions. Worse, an accident in the selection of classes produced a collinearity that made multivariate analysis untrustworthy. This distortion produced the impression of a significant effect in Anderson's pooled analysis. In our reworking of the data, the Aptitude × Treatment interaction effect vanished.

This article documents the importance of separating the two kinds of regression effects, between class and within class, and illustrates the kind of report that results. Such an analysis can be expected to produce important positive results at times. The rationale underlying our procedures is spelled out elsewhere (Cronbach & Snow, in press, Chapter 4). A more technical manuscript on the statistical model is in preparation.

## Method

Anderson started with 18 fourth-grade arithmetic classes whose teachers had agreed to participate in his study. Each of the 18 teachers answered questions regarding his usual practices and philosophy. On the basis of these responses, Anderson assigned the teacher to whichever method seemed to suit the teacher best. One method (D) emphasized practice in computational procedures and the other (M) emphasized explanation and intuitive understanding. The teacher developed daily lessons, within the guidelines for the style to which he was assigned.

Pupils were tested in September (1939) and again in May (1940) on the Compass Survey Test in Arithmetic, the van Wagenen Analytic Scales of Attainment (ASA), and certain other achievement tests. One further test was given in midyear. The Minnesota School Ability Test (MSAT) was given as a pretest only.

Anderson analyzed each subtest of an outcome measure as a separate dependent variable. Treatment means differed little. In the sample, method D usually worked slightly better than M on the average. Anderson used two predictors at a time: either MSAT with the Compass pretest (hereafter denoted by PRECOM), or MSAT with the ASA test. Entering these three scores ($Y$ = outcome, $X_1$ = MSAT, $X_2$ = PRECOM or ASA pretest) in a Johnson-Neyman analysis, Anderson established the linear equation describing the difference between treatments in expected outcome as a function of $X_1$ and $X_2$. In the Johnson-Neyman approach, any significant difference between treatments is reported as a region of significance in the $X_1$, $X_2$ plane, a region bounded by a conic section. If the bound is a hyperbola, Aptitude × Treatment interaction is strongly indicated. An elliptical region does not enable one to decide between the hypotheses of interaction and no interaction.

Anderson reduced the original set of 18 classes by discarding 3 classes of exceptional makeup or small size. Since a ceiling effect lowered posttest-on-pretest regression slopes in the ablest groups, Anderson made two analyses, one on able classes (114 cases in all) and one on less able classes (234 cases). Few significant effects appeared in the abler subsample. We now know (Cronbach & Snow, in press) that samples of 100 cases per treatment or less lack power for testing Aptitude × Treatment interaction hypotheses, hence the negative result means little.

Among eight analyses of the less able subsamples using PRECOM as $X_2$, Anderson found a region of significance in seven; he accepted the null hypothesis in the eighth analysis. Regions of significance appeared in two out of four analyses with ASA as $X_2$. In seven of the plots of significant effects, the region of significance was a hyperbola. Typically, in that corner of the plot where standing on $X_2$ was higher than $X_1$ (overachievers at pretest), the drill method was superior to the meaning method. And in the opposite corner the reverse was true (see Figure 1). This result seemed to make sense psychologically.

Five regions of significance appeared in the analysis of high-ability classes. Two regions were described by hyperbolas (though one branch of each hyperbola was outside the range of cases). With each of these dependent variables, the meaning method was advantageous for the most able students and the D method was better for the less superior students.

### *Procedure in the Reanalysis*

In our analysis we chose not to separate classes of high and low ability; to reduce loss of information, we retained two of the three classes Anderson

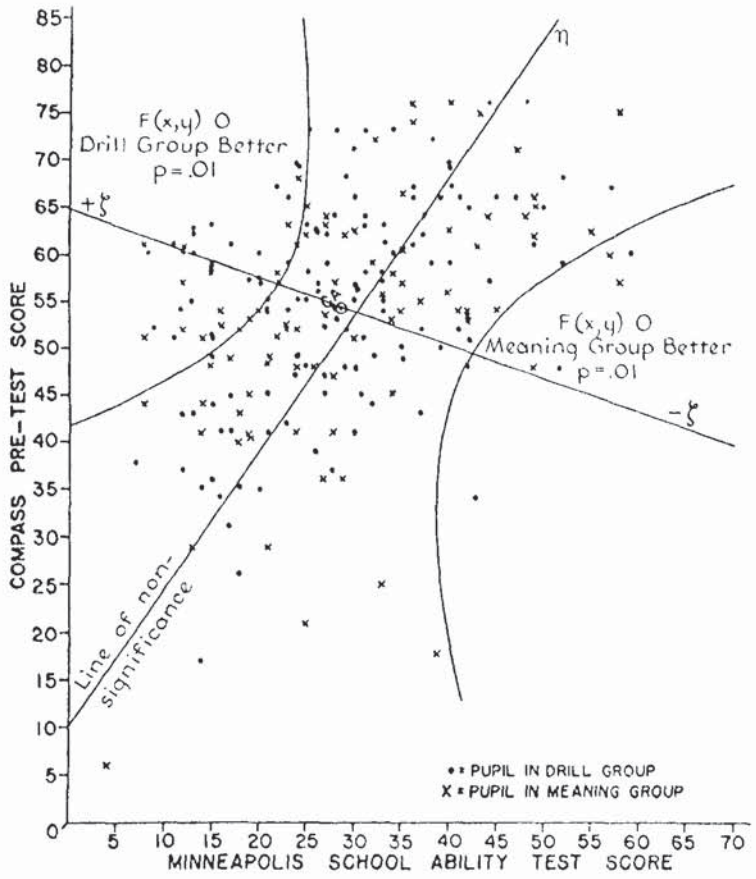


FIGURE 1. An illustrative region of significance for an interaction finding (after Anderson, 1949, p. 54; reproduced by permission).

had eliminated. We dropped any individual having several missing scores; but where only a few scores for an individual were missing, we entered regression estimates of the scores.[1] We therefore had the same 434 cases in the analysis for each set of variables. (Anderson's data set had been reduced to 348 cases. The discrepancy between Anderson's conclusion and ours did not result from our retaining more cases.)

We used the 15 subtests in turn as dependent variables, including three subtests that Anderson had chosen not to analyze statistically because the D and M distribution of scores differed. To obtain an overall picture of any consistent trend, we formed a composite dependent variable (ZACH) by combining subtests of the Compass and ASA posttests. Since the dependent variables were strongly correlated, any effect of practical importance ought to appear in ZACH.

[1] Regression estimates replaced less than 2% of the data.

We did not use the ASA pretest as an aptitude variable. It correlated strongly and about equally with PRECOM and MSAT. ASA would have added little to prediction and nothing to explanation. To get orthogonal predictors we converted MSAT to ABIL, a partial variate defined as MSAT − .042 PRECOM. Although ABIL correlated zero with PRECOM over all cases pooled, within a class or treatment the correlation was not necessarily zero. We rescaled the two predictors and all dependent variables to a mean of zero and a standard deviation of 100 over all cases pooled.

A general conclusion about the effects of methods of instruction would be a generalization over classes; therefore, the basis for degrees of freedom was the number of D and M classes, nine and eight, respectively. The small number of classes made the confidence intervals for all between-class regression lines or planes very wide. We considered students to be fixed within classes in making statistical inferences.

Our calculations were made by a regression program rather than the Johnson-Neyman pro-

TABLE 1

REGRESSION COEFFICIENTS AND CORRELATIONS RELATING ZACH TO PRECOM AND ABIL

| Treatment | Univariate | | | | Bivariate | | Multiple Correlation |
|---|---|---|---|---|---|---|---|
| | Coefficient[a] | | Correlation | | Coefficient[a] | | |
| | PRECOM | ABIL | PRECOM | ABIL | PRECOM | ABIL | |
| | | | Between-class analysis[b] | | | | |
| Drill | .74 | −.47 | .88 | −.20 | .75 | .06 | .88 |
| Meaning | .47 | .22 | .48 | .31 | .70 | .42 | .73 |
| | | | Within-class analysis[c] | | | | |
| Drill | .73 | .39 | .70 | .39 | .75 | .41 | .81 |
| Meaning | .71 | .51 | .70 | .52 | .65 | .42 | .82 |

*Note.* The variables ZACH, PRECOM, and ABIL were formed by standardizing the total distribution of a composite achievement measure, the Compass pretest, and the Minnesota School Ability Test, respectively.

[a] Unstandardized regression coefficient.

[b] Class means weighted by number of students per class.

[c] Classes pooled.

cedure. Regression analysis within each treatment provides a more direct and more complete description of effects in the sample. Had we found a believable Aptitude × Treatment interaction effect we could have evaluated its significance or set confidence limits on it in the manner of Potthoff (1964).[2] We made separate computer runs for the between-class and within-class analyses. In the computer operation, we entered for each individual the mean of his class on each original variable and the deviation of his scores from that mean. Between-class calculations thus weighted each class mean by the class size. Multiple-regression procedures were carried out stepwise with PRECOM entered before ABIL. (*F* ratios given by the computer were ignored, since the standard program took students as the unit of sampling.)

## RESULTS

### *Composite Outcome Measure*

We report the findings on ZACH in full. We have a between-classes regression coefficient within each treatment, from the analysis of class means. Second, we have a "within-classes" regression coefficient within each treatment. The *X* and *Y* scores were restated as deviations from the class mean, and then all cases within the treatment were pooled in the regression analysis.[3] Third, we have the regression within each separate class.

The between-class regression coefficients appear in Table 1. The most straightforward and interpretable result is the simple regression slope onto PRECOM. This slope seems to be higher in Treatment D. When we look at the plot of class means for PRECOM (see Figure 2), however, the D and M classes obviously fit into the same distribution. The difference between the calculated slopes (.74 vs. .47) must be dismissed as a chance result, even though so large a difference in the population would probably be of practical interest. Very large studies are required to estimate between-class slopes with reasonable precision. (Firm conclusions about main effects *can* be reached in studies with a limited number of classes, because the variance of class means tends to be small and so the error variance of a sample of classes is small. Similarly, the pooled within-class slope can often be evaluated well when the number of classes is modest. But the sampling error of a regression slope based on a few class means is large.)

When we examined the regressions onto

[2] Such inference requires assumptions that are not strictly satisfied in quasi-experiments such as Anderson's, and little is known about the effect of such violations.

[3] We might have reported instead the mean of within-class slopes, with or without weighting by class size. The results would have been much the same.

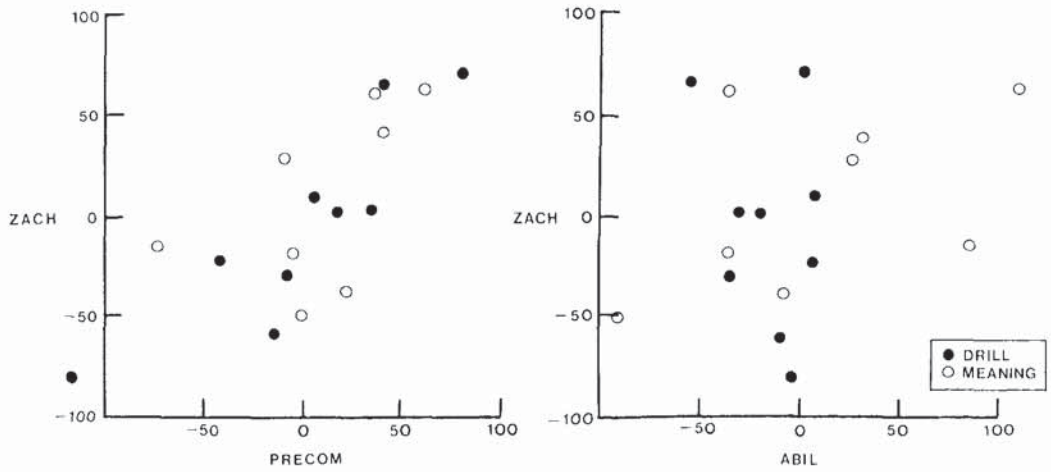


FIGURE 2. Plot of class means on posttest against pretest.

ABIL we discovered an anomaly in Anderson's sample. The joint distribution of PRECOM and MSAT means was radically different in D classes and in M classes. Across D classes, PRECOM and the original MSAT were highly correlated (.74); across M classes the two correlated .09. The discrepancy was surely an accident of sampling. The collinearity in D made the between-class variance of the partial variate ABIL very small among the D classes. ABIL could then contribute little or nothing to a between-class multiple correlation for D. The multiple regression coefficients between M classes are more meaningful.

The simple regression coefficients for ABIL differ dramatically. Classes with positive means on ABIL did little better in M than classes with negative means. In D, the classes that did best had MSAT means lower than would have been expected from PRECOM. The negative slope (−.47) of ZACH on ABIL in the drill treatment, however, is not a reliable result. If the class with means −56 and 67 were deleted (see chart for ABIL, Figure 2), the slope would be close to zero. Even among the M classes, the slope onto ABIL is not reliably positive. The D and M points could easily come from the same distribution.

We conclude that at the between-class level there is no evidence of an interaction, nor are the data adequate to justify the conclusion that interaction is absent in the population. The difference between treatments comes from the difference in regression slopes onto ABIL, a difference we cannot trust. The collinearity which prevents interpretation of the between-class slopes of ZACH onto ABIL equally prevents drawing conclusions about the other outcome variables.

An investigator collecting fresh data who wishes to avoid the collinearity which made Anderson's data uninterpretable would need to control assignment to treatments. Random assignment is not enough when samples are small. Hindsight suggests that Anderson could have plotted the class means at the time of the pretest. After identifying the presence of collinearity, he could have reversed the tentative assignment of about four teachers to make the D and M plots similar. It would have been better yet to pair the classes having adjacent pretest means and then to assign one member of each pair to each treatment. Overriding teacher preference, which alters the substantive hypothesis under test, might be considered good or bad.

*Within-Class Results*

The collinearity of class means in the D sample does not interfere with the within-class analysis, as the pooled within-class distributions of PRECOM and ABIL were similar within treatments. The within-class multiple regressions of the overall outcome onto PRECOM and ABIL were virtually the

same in the M and D samples (see Table 1). The within-treatment univariate regression slopes were also similar. (The difference between .39 and .51 is in the direction of Anderson's findings, but it is a very weak effect.) In the within-class analysis—which is rather powerful statistically—we see no evidence of an Aptitude × Treatment interaction.

The slope onto PRECOM in Treatment D was the same between and within classes (.74, .73). The level of the class, then, does not affect the regression slope. A student whose PRECOM score is (say) 25 can expect to achieve as well in a D class where the mean is 45 as in a D class where the mean is 5. With regard to M, if the smaller between-class slope (.47 vs. .71) were found in a larger sample of classes, it would imply that the student with a given PRECOM score tends to do better in an M class where the PRECOM mean is lower than his. This could result from a class level effect. In M classes such a class level effect would hold down the advantage of able students. An alternative explanation is that a comparative within-class effect enhances the learning of the abler members of the typical M class.

The class-by-class regression slopes varied greatly. The simple regression slopes onto PRECOM ranged from .51 to 1.07 over D classes, and from .43 to 1.12 over M classes. Likewise, the slopes onto ABIL ranged from .14 to .92 in D and from .22 to .71 in M.

The logic for testing the significance of such variations is not obvious. With persons fixed, the pertinent sources of error are those arising from measurement and from the person's experience with the treatment (absences from class, points of confusion). These errors would not be replicated if he could be treated a second time, independently. The first source can be appraised but the second cannot. One may, of course, treat pupils as random and ask whether another sample of persons forming another class would have the same regression slope. But only the teacher and treatment assignment would be constant over the two classes; the details of the group experience would not. If one puts logic aside, and simply takes the pupil as the unit of sampling, the variation is of doubtful significance. The standard error of slopes, with an $n$ of 20 per class, is about .17.

### *Size of Effects*

Table 2 indicates the size of various effects. To describe effect size, each sum of squares was divided by the number of persons, to put it on the scale of the sample variance of scores for individuals. The class effect, individual effect, and individual within-class effect of aptitudes PRECOM and ABIL together account for 72.1% and 70.9% of the ZACH variance within the D and M treatments, respectively. The class effect reflects how well the individual's score can be predicted merely from the ability of his class. The variance is around 8%–16% of the total. The individual effect reflects how much

TABLE 2

DECOMPOSITION OF SUMS OF SQUARES FOR ZACH WITH PRECOM AND ABIL AS PREDICTORS

| Source | Drill treatment | | Meaning treatment | |
|---|---|---|---|---|
| | SS/n | Percentage | SS/n | Percentage |
| Total | 9,880.28 | | 10,011.29 | |
| Between class | 2,126.60 | 21.5 | 1,491.68 | 14.9 |
| Predicted (class effect) | 1,634.45 | 16.5 | 797.06 | 8.0 |
| Residual | 492.15 | 5.0 | 694.62 | 6.9 |
| Within class | 7,753.67 | 78.5 | 8,519.61 | 85.1 |
| Predicted (individual effect) | 5,086.47 | 51.5 | 5,736.88 | 57.3 |
| Predicted (individual within-class effect) | 409.08 | 4.1 | 562.94 | 5.6 |
| Residual | 2,258.12 | 22.9 | 2,219.78 | 22.2 |

*Note.* The variables ZACH, PRECOM, and ABIL were formed by standardizing the total distribution of a composite achievement measure, the Compass pretest, and the Minnesota School Ability Test, respectively.

of the outcome can be predicted from the deviation of the individual's score from the mean of his class (using the regression slope for all classes together). This effect accounts for about half of the variance in ZACH. Specific individual within-class effects may further modify the regression. The predictive contribution of the within-class regression over and above that of the pooled regression averages about 5%. The specific within-class regression thus did not add much to the prediction, and some of that increment capitalized on chance.

The residuals (see Table 2) are much the same in the two treatments. Abilities not measured by PRECOM and ABIL, uncontrolled experiences during the school year, and errors of observation all enter into the residual.

*Specific Outcome Measures*

Anderson located regions of significance for 9 out of 12 subtests among the less able groups. In 7 of these plots Anderson reported D superior for overachievers and M advantageous for underachievers. Among high-ability groups he found significant effects on 5 out of 12 subtests. These findings in able classes were not wholly consistent in character, but M tended to be superior to D over a large part of the pretest range.

Anderson was troubled by the ceiling effects and grossly nonnormal score distributions in his data. Nine subtests exhibited such problems. We applied exponential transformations to the subtest scores and restandardized; we also fitted quadratic regression equations. Neither of these procedures substantially improved prediction. The transformations raised both of the between-class regression slopes onto ABIL, and to a lesser extent the slopes onto PRECOM; but they had no effect on the pooled within-classes slopes. The changes do not seem to be worthy of discussion. We therefore report only the usual linear regressions for the original scores.

Our between-class regressions resembled some but not all of Anderson's individual level regressions. In the univariate analyses of several subtests, we obtained dramatic negative slopes onto ABIL, which at face value suggested that treatment D was harmful to able classes. In every instance, a single outlying class was responsible for the negative slope. If it were not for that one class, the slopes onto PRECOM and ABIL would be similar in the two treatments for most outcome measures.

No difference between treatments appeared in the within-class regressions of specific subtests. So Anderson's interactions must have been the result of regression effects between classes, which we interpret as chance effects.

Anderson found no region of significance in comparing the two methods of instruction on three subtests in low-ability groups, and on seven subtests in high-ability groups. Our within-class regressions likewise showed no differences between methods. With regard to those subtests that Anderson did not analyze statistically, our regression analyses showed no difference between the drill and meaning treatments.

## Conclusion

The Anderson data apparently do not challenge the null hypothesis. The contrast of the drill and meaning treatments did not indicate a main effect or an interaction. This is important psychologically, since the underlying issue is general and of long standing.

Can methods of instruction be found to serve better the student whose fluid ability is high relative to his crystallized ability? This question has been in the mind of the user of mental tests since the days of Binet. General mental tests are used alongside achievement tests to locate students who could benefit by a different teaching approach. This presumes that an Aptitude × Treatment interaction exists (Cronbach & Gleser, 1965, p. 144 ff.). Anderson's study originally seemed to suggest such an interaction. The reanalysis casts some doubt on the hypothesis of interactions involving the fluid/crystallized distinction. The disconfirmation is important also because Cronbach & Snow (in press) found no study in the literature supporting the Aptitude × Treatment interactions that Anderson originally had appeared to demonstrate.

This does not mean that there is clear negative evidence. Studies of interactions usually have not been powerful enough to evaluate outcome-on-aptitude regressions accurately. Using the class as the unit of analysis, even the rather large Anderson study could not set narrow confidence limits on the regression slopes. In the light of our experience with the Anderson study, we urge investigators collecting data on intact classes to examine between-group and within-group regressions separately.

## REFERENCES

Anderson, G. L. *A comparison of the outcomes of instruction under two theories of learning.* Unpublished doctoral dissertation, University of Minnesota, 1941.

Anderson, G. L. Quantitative thinking as developed under connectionist and field theories of learning. In E. J. Swenson, G. L. Anderson, & L. S. Chalmers (Eds.), *Learning theory in school situations.* University of Minnesota Studies of Education, no. 2. Minneapolis: University of Minnesota Press, 1949.

Cronbach, L. J. How can instruction be adapted to individual differences? In R. M. Gagné (Ed.), *Learning and individual differences.* Columbus, Ohio: Merrill, 1966.

Cronbach, L. J., & Gleser, G. C. *Psychological tests and personnel decisions* (2nd. ed.), Urbana. University of Illinois Press, 1965.

Cronbach, L. J., & Snow, R. E. *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington Press, in press.

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

(Received April 30, 1975)