OPERATIONALLY DEFINING THE ASSUMPTION OF INDEPENDENCE AND CHOOSING THE APPROPRIATE UNIT OF ANALYSIS

A Dissertation
for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY

Linda K. Glendening

1977

This is to certify that the

thesis entitled

Operationally Defining the Assumption of Independence and Choosing the Appropriate Unit of Analysis

presented by

Linda K. Glendening

has been accepted towards fulfillment of the requirements for

Ph.D. degree in Educational Psychology

A C Porter

Major professor

Date January 28, 1977

O-7639

ABSTRACT

OPERATIONALLY DEFINING THE ASSUMPTION OF INDEPENDENCE AND CHOOSING THE APPROPRIATE UNIT OF ANALYSIS

By

Linda K. Glendening

The assumption of independence was operationally defined as: Individual units (such as students) can be considered independent on some dimension whenever the variance of the grouped units (such as classrooms) can be predicted from the grouping size and the variance of the individual units. When this definition of independence is satisfied, the expected mean squares between and within groups are equal. Given this operational definition, two types of dependence are possible, positive and negative. Positive dependence was defined by the expected mean square between groups being larger than the expected mean square within groups. Negative dependence was defined by the expected mean square within groups being larger than the expected mean square between groups.

Both empirical and analytical methods were used to study the effect of violating the assumption of independence, where the design model was balanced and had two levels of nesting, subjects within groups and groups within treatments. Group data were independent of each other, while subjects within group data were manipulated to create different degrees and types of dependence. The simulated data were analyzed using

two ANOVA models, the "never pool" model where group was the unit of analysis and so was an always correct model and the "always pool" model with student as the unit of analysis.

First, sampling distributions using the "never pool" model and the "always pool" model were compared for independent, positively dependent, and negatively dependent conditions. Given independence of subject responses, either subject or group can be used as the unit of analysis as both the "never pool" and the "always pool" tests proved to have acceptable Type I error rates for the test of treatment effects. The "always pool" test is the preferable test, however, as it had more power than did the "never pool" test. Given positive dependence, the proper unit of analysis is the grouped unit. Using subject as the unit of analysis caused the pooled error term for the "always pool" test to be too small, and so the "always pool" test was too liberal and had spuriously high power. Given negative dependence, the correct unit of analysis is again the grouped unit. Using subject as the unit of analysis caused the pooled error term for the "always pool" F test to be too large and thus the "always pool" test was too conservative and had spuriously low power. The empirical results indicated clearly that the F test is not robust to violations of the assumption of independence, even given small degrees of positive and negative dependence.

Next, a conditional testing procedure (a "sometimes pool" model) was studied where an initial test of independence was done and then on the basis of that test a unit of analysis was chosen for the primary test of treatment effects. Sampling distributions using the "never

pool" and the "sometimes pool" models were compared for independent, positively dependent, and negatively dependent conditions. Given independence of ungrouped units, the "sometimes pool" F test had acceptable Type I error rates for the test of treatment differences, as did the "never pool" test. In addition, the powers of the "sometimes pool" test tended to be greater than the powers of the "never pool" test. Given positive dependence, the "sometimes pool" F test generally was too liberal and thus had spuriously high empirical power. And given negative dependence, the "sometimes pool" test was somewhat conservative and generally had less power than the "never pool" F test. These results suggest that, as a general rule of thumb, a preliminary test of independence should not be done to choose a unit of analysis to use in testing for treatment differences.

OPERATIONALLY DEFINING THE ASSUMPTION OF INDEPENDENCE AND CHOOSING THE APPROPRIATE UNIT OF ANALYSIS

By

Linda K. Glendening

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Personnel Services,
and Educational Psychology

1977

ACKNOWLEDGMENTS

First and foremost, I would like to acknowledge the members of my dissertation committee for their assistance. Special thanks go to my advisor and friend, Professor Andrew Porter. Working with him has strengthened my capabilities as a researcher and broadened my research interests and experiences. In addition, I would like to thank Drs. William Schmidt, Lee Shulman, and James Stapleton for their interest in my research.

I also wish to acknowledge the National Institute of Education, where I did my research. The Institute has provided me with opportunities to increase my awareness of and concern for current educational research problems.

* * * * * * * *

TABLE OF CONTENTS

		Page
LIST OF TABLES		v
Chapter		
I.	STATEMENT OF THE PROBLEM	1
II.	REVIEW OF LITERATURE	6
	Definitions of Independence	6
	Threats to Independence	11
	Selection of Analytic Units	13
	Statistical Arguments for Unit Selection	14
	Logical Arguments for Unit Selection	17
III.	AN OPERATIONAL DEFINITION OF INDEPENDENCE	21
IV.	ANALYTIC RESULTS	25
	The Effects of Dependence	25
	Classroom as Unit	30
	Student as Unit	33
	The Preliminary Test	38
V.	SIMULATION PROCEDURES	46
	Simulation Parameters	47
	Data Generation Routine	52
VI.	UNITS OF ANALYSIS: EMPIRICAL ESTIMATES OF EFFECTS	56
	Classroom as Unit	57
	Independence	57
	Positive Dependence	61
	Negative Dependence	62
	Student as Unit	63
	Independence	64
	Positive Dependence	67
	Negative Dependence	72

Chapter Page

VII. THE CONDITIONAL F TEST: EMPIRICAL ANALYSIS 75

The Two-Tailed Preliminary Test 77
Independence 78
Positive Dependence 94
Negative Dependence 100
The Upper-Tailed Preliminary Test 105
Independence 112
Positive Dependence 120
Negative Dependence 121
The Lower-Tailed Preliminary Test 123
Independence 129
Positive Dependence 136
Negative Dependence 138

VIII. SUMMARY AND CONCLUSIONS 142

Appendix

A. PRELIMINARY ANALYSIS OF SIMULATED DATA 153

B. ESTIMATED ALPHAS OF THE RESCALED F STATISTIC 162

C. POWERS OF THE CONDITIONAL F GIVEN A TWO-TAILED PRELIMINARY TEST 163

D. POWERS OF THE CONDITIONAL F GIVEN AN UPPER-TAILED PRELIMINARY TEST 168

E. POWERS OF THE CONDITIONAL F GIVEN A LOWER-TAILED PRELIMINARY TEST 173

BIBLIOGRAPHY 178

LIST OF TABLES

Table		Page
1.	Power Computations Using Groups and Individuals as Units of Analysis, Given Individuals Are Independent . . .	17
2.	Expected Mean Squares for Model A	26
3.	Expected Mean Squares for Model B	27
4.	Theoretical Intraclass Correlation Coefficients for Selected Numbers of Students and Degrees of Dependence	51
5.	Design of Study	51
6.	Empirical Type I Errors for $F = MS_T/MS_{C:T}$	58
7.	Empirical Powers for $F = MS_T/MS_{C:T}$	59
8.	Empirical Type I Errors for $F = MS_T/MS_{S:T}$	65
9.	Empirical Powers for $F = MS_T/MS_{S:T}$	66
10.	Discrepancy Between Observed and Theoretical $E(MS_{C:T})$ and $E(MS_{S:T})$	71
11.	Actual Alphas of the Conditional F Test Given a Two-Tailed Preliminary Test, $c = 2$ and $s = 12$	79
12.	Actual Alphas of the Conditional F Test Given a Two-Tailed Preliminary Test, $c = 5$ and $s = 5$	80
13.	Actual Alphas of the Conditional F Test Given a Two-Tailed Preliminary Test, $c = 5$ and $s = 12$	81
14.	Actual Alphas of the Conditional F Test Given a Two-Tailed Preliminary Test, $c = 5$ and $s = 20$	82
15.	Actual Alphas of the Conditional F Test Given a Two-Tailed Preliminary Test, $c = 10$ and $s = 12$	83

Table Page

16. Power of the Conditional F Test Minus Power of the Test F = $MS_T/MS_{C:T}$ Given a Two-Tailed Preliminary Test, c = 2 and s = 12 86

17. Power of the Conditional F Test Minus Power of the Test F = $MS_T/MS_{C:T}$ Given a Two-Tailed Preliminary Test, c = 5 and s = 5 87

18. Power of the Conditional F Test Minus Power of the Test F = $MS_T/MS_{C:T}$ Given a Two-Tailed Preliminary Test, c = 5 and s = 12 88

19. Power of the Conditional F Test Minus Power of the Test F = $MS_T/MS_{C:T}$ Given a Two-Tailed Preliminary Test, c = 5 and s = 20 89

20. Power of the Conditional F Test Minus Power of the Test F = $MS_T/MS_{C:T}$ Given a Two-Tailed Preliminary Test, c = 10 and s = 12 90

21. Actual Alphas of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 2 and s = 12 107

22. Actual Alphas of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 5 and s = 5 108

23. Actual Alphas of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 5 and s = 12 109

24. Actual Alphas of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 5 and s = 20 110

25. Actual Alphas of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 10 and s = 12 111

26. Power of the Conditional F Test Minus Power of the Test F = $MS_T/MS_{C:T}$ Given an Upper-Tailed Preliminary Test, c = 2 and s = 12 114

27. Power of the Conditional F Test Minus Power of the Test F = $MS_T/MS_{C:T}$ Given an Upper-Tailed Preliminary Test, c = 5 and s = 5 115

28. Power of the Conditional F Test Minus Power of the Test F = $MS_T/MS_{C:T}$ Given an Upper-Tailed Preliminary Test, c = 5 and s = 12 116

Table		Page
29.	Power of the Conditional F Test Minus Power of the Test F = $MS_T/MS_{C:T}$ Given an Upper-Tailed Preliminary Test, c = 5 and s = 20	117
30.	Power of the Conditional F Test Minus Power of the Test F = $MS_T/MS_{C:T}$ Given an Upper-Tailed Preliminary Test, c = 10 and s = 12	118
31.	Actual Alphas of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 2 and s = 12	124
32.	Actual Alphas of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 5 and s = 5	125
33.	Actual Alphas of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 5 and s = 12	126
34.	Actual Alphas of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 5 and s = 20	127
35.	Actual Alphas of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 10 and s = 12	128
36.	Power of the Conditional F Test Minus Power of the Test F = $MS_T/MS_{C:T}$ Given a Lower-Tailed Preliminary Test, c = 2 and s = 12	131
37.	Power of the Conditional F Test Minus Power of the Test F = $MS_T/MS_{C:T}$ Given a Lower-Tailed Preliminary Test, c = 5 and s = 5	132
38.	Power of the Conditional F Test Minus Power of the Test F = $MS_T/MS_{C:T}$ Given a Lower-Tailed Preliminary Test, c = 5 and s = 12	133
39.	Power of the Conditional F Test Minus Power of the Test F = $MS_T/MS_{C:T}$ Given a Lower-Tailed Preliminary Test, c = 5 and s = 20	134
40.	Power of the Conditional F Test Minus Power of the Test F = $MS_T/MS_{C:T}$ Given a Lower-Tailed Preliminary Test, c = 10 and s = 12	135
A-1.	Distribution of Sample Classroom Means and Student Observations	153
A-2.	Moments for Student within Treatment Type Data when c = 2 and s = 12	154

Table Page

A-3. Moments for Student within Treatment Type Data when c = 5 and s = 5 154

A-4. Moments for Student within Treatment Type Data when c = 5 and s = 12 155

A-5. Moments for Student within Treatment Type Data when c = 5 and s = 20 155

A-6. Moments for Student within Treatment Type Data when c = 10 and s = 12 156

A-7. Three Distributional Statistics for Standardized Mean Square Values, Given c = 2 and s = 12 157

A-8. Three Distributional Statistics for Standardized Mean Square Values, Given c = 5 and s = 5 158

A-9. Three Distributional Statistics for Standardized Mean Square Values, Given c = 5 and s = 12 159

A-10. Three Distributional Statistics for Standardized Mean Square Values, Given c = 5 and s = 20 160

A-11. Three Distributional Statistics for Standardized Mean Square Values, Given c = 10 and s = 12 161

B-1. Estimated Type I Errors for $F = MS_T/MS_{S:T}$ Using a Rescaled F Statistic 162

C-1. Power of the Conditional F Test Given a Two-Tailed Preliminary Test, c = 2 and s = 12 163

C-2. Power of the Conditional F Test Given a Two-Tailed Preliminary Test, c = 5 and s = 5 164

C-3. Power of the Conditional F Test Given a Two-Tailed Preliminary Test, c = 5 and s = 12 165

C-4. Power of the Conditional F Test Given a Two-Tailed Preliminary Test, c = 5 and s = 20 166

C-5. Power of the Conditional F Test Given a Two-Tailed Preliminary Test, c = 10 and s = 12 167

Table Page

D-1. Power of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 2 and s = 12 168

D-2. Power of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 5 and s = 5 169

D-3. Power of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 5 and s = 12 170

D-4. Power of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 5 and s = 20 171

D-5. Power of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 10 and s = 12 172

E-1. Power of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 2 and s = 12 173

E-2. Power of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 5 and s = 5 174

E-3. Power of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 5 and s = 12 175

E-4. Power of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 5 and s = 20 176

E-5. Power of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 10 and s = 12 177

CHAPTER I

STATEMENT OF THE PROBLEM

At least three standard assumptions are necessary whenever parametric hypotheses are tested or confidence intervals constructed. These three "minimum" assumptions cut across models, e.g., correlational models and experimental models, and across research designs, e.g., crossed and nested designs. There is the assumption of equal variances or homoscedasticity, the assumption of normality and the assumption of independence. Since assumptions are rarely, if ever, exactly met in real world situations, researchers need to be sensitive to departures from the assumptions underlying the model to be used and they need to be aware of the consequences of such departures. Thus when making assumptions, it seems necessary to consider two questions in particular. First, what happens when the assumptions are violated? Second, if it is important that an assumption be satisfied, how can one tell when it has been satisfied?

Failure to meet the assumptions of a model may affect both the significance level of a test and the sensitivity of a test (Cochran & Cox, 1957, p. 91). It is well known that under certain circumstances the analysis of variance is robust to violations of the assumptions concerning homoscedasticity and normality (Glass, Peckham & Sanders, 1972). However violations of the assumption of independence are less

well understood and may substantially affect the validity of any confidence statements based in part on that assumption and made regarding the hypothesized effects. In comparison to research efforts studying the effects of violating the homoscedasticity and normality assumptions, very little research effort has systematically dealt with the effects of using correlated units of analysis.

In cases where the analytic model is not robust to violating either the homoscedasticity assumption or the normality assumption, tests exist (e.g., Levene's [1960] test for equal variances and, if n is reasonably large, the chi-square test for normality of data) that can be used for making a decision about whether the assumption was violated. The question remains as to whether or not the validity of the independence assumption can also be tested.

The major intent of this study was to propose a definition of independence that was both conceptually meaningful within the typical educational paradigm and operationally measurable. Secondly, this study addressed the effects on the sampling distribution of the F statistic and the effects on parameter estimates when correlated units of analysis were assumed to be independent of each other. More specifically, the probability of Type I errors, the power of the test and the biasedness of parameter estimates were examined when different degrees and types of dependencies were present within the research model. And thirdly, this study considered distributional problems with using a conditional testing procedure which included a preliminary test of independence and either one of two subsequent primary tests of treatment effects.

In particular, the probability of Type I errors and the power of the conditional F test (where the error term for the primary test of treatment effects was selected on the basis of the results of a preliminary test of independence) were of interest, both for conditions where the preliminary test should fail to reject and for conditions where the preliminary test should reject that individual observational units were independent. The two distributional problems were studied both analytically (to determine the existence and direction of effects) and empirically (to estimate the magnitude of effects). The general design model used for studying these problems was a balanced, hierarchically-nested analysis of variance model, having only one outcome measure per subject.

Throughout this paper the effects of violating the assumption of independence will be discussed within the context of the general educational setting, where students are reacting to stimuli within a group atmosphere and where treatments or programs are usually given to the entire classroom. Within each classroom situation, it seems that different classroom or grouping components can make the responses of individual students (observational units) dependent upon each other to some degree. Some classroom components that can affect the relationships between students' responses are:

- classroom environment effect;
- teacher or instructional effect; and
- classmate effect.

The first problem is knowing and measuring how dependence operates within a particular classroom. A second problem involves designing statistical models which minimize the possibility of dependence due to one or more of the classroom components. Clearly the answer to the first problem should guide solutions to the second.

As one research effort to empirically study the effect of a common learning environment on the achievement of students, Steck (1966) randomly assigned thirty 7th graders to receive mathematical instruction in a group and another thirty to receive the instruction on a one-to-one basis. The presentation of the lesson was made by tape recorder to ensure similarity of treatment. In both situations, students were encouraged to ask questions concerning the lesson before they took a test to measure the extent to which they had mastered the material. Steck found that the variance of the scores of students who received the presentation as a group was significantly smaller than the variance of the scores of students who received the lesson individually. Steck's results suggest that in this particular study either common classroom experience decreased the variability of student responses and/or individualized instruction increased the variability of student responses. Clearly, however, this one study is not enough to conclude that the opposite condition (where classroom experience actually increases student variation and/or individualized instruction decreases student variation) does not occur.

Just how classroom components affect student responses may depend on such things as the student population. For example, the responses

of kindergarteners within a classroom may tend to be more related to each other than the responses of 12th graders within a classroom. The type of program instruction might also make a difference. For example, a classroom discussion probably tends to make students' responses more related to each other than does a classroom lecture. Independence (or dependence) is not an all or nothing situation. Rather it is a matter of degree. At one extreme, observational units can be completely dependent upon each other. In this situation, N observational units are no more informative than one observational unit. At the other extreme, observational units can be completely independent (at least in theory). In this situation, N observational units give N pieces of nonoverlapping information. In-between these two extremes is a complete continuum of dependency. It is this in-between area that gives practicing statisticians headaches.

CHAPTER II

REVIEW OF LITERATURE

Definitions of Independence

Understanding the assumption of independence is prerequisite to studying the consequences of violating that assumption. Thus this section contains a review of the varying definitions of independence contained in texts and papers dealing with statistical and research design issues. The conclusion is that, for the most part, these statistical sources have been too theoretical and mystical to inform practice on the assumption of independence itself and the consequences of violating that assumption. Below are examples of how different statisticians formally define independence. By themselves, these definitions seem inadequate as they are conflicting in the suggested causes of dependency and deficient in the assessment of dependency.

> The experimental errors must all be independent. That is, the probability that the error of any observation has a particular value must not depend on the values of the errors for other observations. (Cochran, 1947)

> It is also assumed that the ε_{ij}'s are independent, both within each treatment level and across all treatment levels. If subjects are randomly assigned to treatment levels, the value of ε_{ij} for any observation can be assumed independent of the values of ε_{ij} for other observations. (Kirk, 1968, p. 52)

Before looking at additional definitions of independence, a distinction needs to be made between independence and random assignment. (Note that Kirk's definition equated the two.) The condition of random assignment in a study is not synonymous with the condition of independence of observations. In theory, independence can happen without random assignment. Students judgmentally assigned to any treatment condition can react to that treatment individually or independently of others assigned to that same treatment. On the other hand, it can and does happen that units randomly assigned to treatment conditions do not, in fact, receive the treatments independently and thus the assigned units are not independent of each other. One example of this would be where children are randomly assigned to treatments but then all children assigned to any one treatment condition are treated as a group, e.g., they may all receive the treatment from the same teacher. Random assignment can only be counted upon to control disturbing or confounding variables that are present at the start of the study. Random assignment cannot control variables that are introduced, maybe only for convenience sake, into the experiment by the experimenter's manipulations, e.g., having one teacher instruct all students assigned to one treatment level.

A third definition of independence has been given by Draper and Smith (1966, p. 17). In discussing the general regression model, $Y_i = \beta_o + \beta_1 X + \varepsilon_i$, for $i = 1, 2, \ldots, n$, they make the following assumption: "ε_i and ε_j are uncorrelated, $i \neq j$, so that $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$. Thus . . . Y_i and Y_j, $i \neq j$, are uncorrelated."

They then conclude that, given the above assumption and an additional assumption that ε_i is a normally distributed random variable, "ε_i, ε_j are not only uncorrelated but necessarily independent."

Draper and Smith's definition, in particular, seems especially hard to conceptualize. The difficulty lies in finding a way to measure the covariance between subjects when there is only one outcome score per subject. Within the single sample paradigm and under the condition that subjects are measured before selection, the cov (ε_i, ε_j), for $i \neq j$, should always equal zero whenever subjects have been randomly selected from an infinitely large target population. When subjects within a sample are more homogeneous on the dependent variable than subjects in the population, the cov (ε_i, ε_j) will be greater than zero. The opposite condition of the cov (ε_i, ε_j) being less than zero occurs whenever subjects within a sample are more heterogeneous on same outcome measure than subjects in the population.

Draper and Smith have suggested that one way observations would be independent of each other, before any treatment has taken place, is if the subjects are randomly selected from a normally distributed population. The condition of random selection from a normal population before treatment is not synonymous, however, with the condition of independence of observations after treatment. First, there is no reason that independence need be a function of the population being normally distributed. Theoretically, observations from a non-normal population can be independent of each other. Secondly, random selection does not insure independence of observations after a treatment has been

effected. Random selection is concerned with the external validity of the experiment and cannot control variables that are later introduced into the study. For example, the situation where subjects randomly chosen were not, in fact, treated independently could disturb any condition of independence that was due to subjects being randomly selected.

Cox (1958, p. 15) defines independence as the condition where "the observation on one unit is unaffected by the treatments applied to other units." He then goes on to say (1958, p. 19) that independence

> is the requirement that the observation on one unit should be unaffected by the particular assignment of treatments to the other units, i.e., that there is no "interference" between different units. In many experiments the different units are physically distinct and the assumption is automatically satisfied. If, however, the same object is used as a unit several times, or if different units are in physical contact, difficulties can arise.

Cox's definition of independence suggests that physical contact causes dependence. He states that if units are physically distinct the assumption of independence has been met. Within the educational setting, however, defining what "physically distinct" means is as difficult as defining what independence means. In looking over Cox's definition of independence, one must remember that Cox was mainly concerned with agricultural experiments. In these types of experiments, it happens to be much easier to control the interaction between different units. With educational studies, this is not usually the case. For example, the educational researcher cannot control what happens to his subjects outside the school environment.

A fifth definition of independence concerns equating the experimental unit and the unit of analysis. If the statistical analysis of an experiment is to yield a valid confidence statement about the chances of drawing false conclusions from the data, the analytic unit should coincide with the experimental unit (Glass & Stanley, 1970; Peckham, Glass & Hopkins, 1969a, 1969b; Porter, 1972). This suggests that another definition of independence is that the condition of independence is satisfied when the unit of analysis is identical to the experimental unit. A unit of analysis refers to the smallest observational unit (or data point) which in the data analysis is to be considered distinct from other observational units.

Cox (1958, p. 2) has given the following definition of experimental unit: "The formal definition of an experimental unit is that it corresponds to the smallest division of the experimental material such that any two units may receive different treatments in the actual experiment." Cox then goes on to say that it is very "desirable" that experimental units also respond independently of one another. Peckham et al. (1969a, p. 341) have defined the experimental unit as:

> The experimental units are the smallest divisions of the collection of the experimental subjects which have been randomly assigned to the different conditions in the experiment and which have responded independently of each other for the duration of the experiment or which, if allowed to interact during the experimental period, have had the influence of all extraneous variables controlled through randomization.

Most of the above definitions of the condition of independence are not complete in that they are written only within the context of

experimental situations. Definitions of independence should include correlational situations as well as experimental situations as the assumption of independence is made in correlational studies as well as in experimental studies. It is also interesting to note that some of the definitions of independence are procedural and imply cause, i.e., Cox's and Kirk's. Other definitions are given only in terms of outcomes or effects, i.e., Cochran's and Draper and Smith's.

Threats to Independence

Peckham et al. (1969a, 1969b) suggest two different ways that group influence can exert a dependency among units within a group on one or more dimensions. The first of these is an additive effect, which can raise or lower the group mean by tending to raise or lower the score of each unit within the group by a constant amount. This additive effect influences the variability of classroom means, but at the same time does nothing to the variability of students' scores within the classroom. In other words, an additive type of dependency would disturb the predictability of the between class variance from the within class variance by affecting the former. The second type of group influence is what Peckham et al. call a proportional effect which leaves the mean performance of a group unchanged but has a marked effect on the variability of responses within the group. This type of dependence also disturbs the predictability of the between class variance from the within class variance, but this time by affecting the latter.

These two general types of group influence defined by Peckham et al. can be broken into four conditions of dependence which systematically affect either the variation of students within classrooms or the variation of students between classrooms:

- An additive type of dependence which decreases the variation between classes,
- An additive type of dependence which increases the variation between classes,
- A proportional type of dependence which decreases the variation within classrooms, and
- A proportional type of dependence which increases the variation within classrooms.

In real world situations it is possible that any one of the four dependency conditions suggested might occur simply by nonrandom assignment of students to classrooms.

A decrease in the between class variation could occur in a study where judgmental assignment of teachers to intact classrooms has taken place. A situation where principals have assigned particularly effective teachers to difficult classes and average teachers to the more motivated classes could have the effect of decreasing the between classroom variation perhaps without changing the within classroom variation. On the other hand, an increase of the between class variation could be a function of classroom composition systematically affecting teachers' attitudes toward teaching. Some teachers may become extremely ineffective when assigned to a particularly difficult class. This could have the effect of decreasing their students' achievement by a constant amount. Other teachers when assigned to a particularly motivated

class may enjoy their working conditions so much that the achievement level of all their students within the class is increased by a constant amount, apart from the effect of the treatment being investigated. A situation occurring such as this could have the effect of increasing the between classroom variability without affecting the within classroom variability.

A decrease of variation within classes could be due, for example, to teachers working hard to increase the achievement level of disadvantaged students while at the same time ignoring the achievement potential of the brighter students. A situation such as this could decrease the within classroom variability without similarly affecting the between class variability. An increase of within class variation could be a function of teachers paying more attention to the more capable students in the class while ignoring the slower students. A situation such as this could increase the within classroom variance without changing the between classroom variance.

Selection of Analytic Units

In order that any definition of independence be operational, it must be able to help the researcher choose between alternative units of analysis and/or at least realize the consequences of a wrong choice of unit. Thus in this section, past theoretical research will be reviewed that suggests how using a unit of analysis inappropriate to the research design can affect the results of a study. Empirical research will be presented that suggests how statistical power can

differ given two distinct units of analysis, both being appropriate to the research design. The studies reviewed suggest that the test of the hypothesis of no treatment effects may be affected by the choice of unit of analysis. The magnitude of treatment effects (α_i's), however, should not depend upon the choice of the unit of analysis. Logical arguments will also be reviewed that discuss whether or not research questions dealing with educational programs should focus on the individual student. That is, are educational research questions involving the individual student functional given the present educational system?

Statistical Arguments for Unit Selection

Scheffé (1959) has discussed the effects of violating the assumption of independence when observations are serially correlated. He considered the single group case where the random variables Y_i and Y_{i+1}, for $i = 1, 2, \ldots, n-1$, had a serial correlation equal to ρ and all other pairs of observations had a serial correlation equal to zero. Scheffé found that the effect of serial correlation can be serious on inferences about means. As the serial correlation went from 0 to -0.4, the test of the hypothesis became very conservative. As the serial correlation went from 0 to +0.4, the test of the hypothesis became very liberal. Measurements that are either close in time or close in space can be serially correlated. Scheffé's results seem applicable in educational research when successive measurements have been taken on each experimental unit.

Cochran (1947), on the other hand, considered a simple group comparison case where every pair of observational units within a treatment level had a simple correlation of ρ. With correlated (ρ) units of analysis, the error of the treatment total ($e_1 + e_2 + \ldots + e_n$) should have a variance equal to $n\sigma^2 + n(n-1)\rho\sigma^2$, rather than $n\sigma^2$ which would be the variance of the treatment total when errors are uncorrelated. Consequently, with correlated observations, the true variance of the treatment mean is $[\sigma^2 + (n-1)\ \rho\sigma^2]/n$. However the variance of the treatment mean is estimated by calculating the sum of squared deviations within each treatment level and then pooling across treatment levels. The variance of this treatment mean is equal to $\sigma^2\ (1-\rho)/n$. Therefore Cochran concludes that when observations are positively correlated, the true variance of the treatment means is underestimated. When observations are negatively correlated, the true variance of the treatment means is overestimated. This suggests that Cochran's conclusions are consistent with Scheffé's in that when observations are positively correlated, the actual alpha level for the test of no treatment effect will be larger than the nominal alpha level, indicating that the test is too liberal; and when observations are negatively correlated, the actual alpha level for the test of the null hypothesis will be smaller than the nominal alpha level, indicating that the test is too conservative. Cochran handled the problem of dependent units of analysis within the same context as Draper and Smith's definition of independence, correlating pairs of units of analysis on one outcome measure.

Lissitz and Chardos (1975) empirically verified Cochran's theoretical analysis, where every pair of subjects had a simple correlation of ρ, and replicated Scheffé's analysis, where subjects were serially correlated both positively and negatively. They also extended Cochran's case to data which were negatively correlated ($-\rho$), although they failed to explain what this negative correlation meant. Their empirical analysis showed that positively dependent data, as defined by Cochran, for both $\rho = .2$ and $\rho = .4$, made the t-test too liberal a statistic; while negatively correlated data, $\rho = -.2$ and $\rho = -.4$, made the t-test too conservative. The same general conclusions were found with the positive and negative serially correlated data except the results were not as extreme in any of the cases.

Probably the one most persuasive argument used by educational researchers in selecting the individual within a group, rather than the group itself, as their unit of analysis is the well-ingrained notion that studies with few observations tend to have little power for detecting treatment differences. Peckham et al. (1969a, 1969b) considered the hypothetical case of having 200 subjects randomly assigned to one of eight groups and groups randomly assigned to one of two treatment conditions. Using Kirk's (1968, p. 107) formula for estimating the noncentrality parameter, ϕ, under the condition of independence of individual observations, they computed power estimates using both the group as the unit and the individual as the unit (Table 1). Table 1 shows that, under these conditions, only for small treatment effects

Table 1

Power Computations Using Groups and Individuals as Units of Analysis, Given Individuals Are Independent

Analysis unit	d.f. (error)	Power ($\alpha = .05$) Treatment Effect ($\mu_1 - \mu_2$ in sigma units)		
		.25	.50	.75
Individuals	198	.42	.94	.991
Groups	6	.25	.82	.987

is the power using the group as the analysis unit much less than the power using the individual as the analysis unit. Under the conditions hypothesized by Peckham et al. no systematic differences occur between subjects across groups within treatment levels and therefore the expected mean square for groups within treatments equals the expected mean square for subjects within group/treatment combinations. Because these two expected mean squares are equal the two F tests, one using group as the unit of analysis and the other using individual as the unit of analysis, have identical noncentrality parameters (ϕ) and thus the difference in the power of the two F tests is totally a function of the difference in degrees of freedom associated with each F test.

Logical Arguments for Unit Selection

Working within the realm of education, a question that needs to be asked is, Can the responses of individual students ever legitimately

be assumed to be independent of each other? Or, in other words, is student ever the appropriate unit of analysis? Within traditional education, the classroom is most often the functional treatment unit. That is, most often it occurs that all students within the same classroom receive the same basic instructional treatment. For example, it usually is the case that all children within a classroom learn math instructed by the same method. Rarely does it happen that, within one classroom, math is taught to some students using one method and to other students using yet another method. Wiley (1970) has stated that "if the object of evaluation is a typical classroom instructional program where the instruction is received simultaneously by all students in the class, then the appropriate vehicle (or sampling unit) is the class and not the individual pupil."

Peckham et al. (1969a, 1969b), Raths (1967), and Wiley (1970) have stated that if, however, treatments are presented in the form of individualized instructional techniques, such as programmed learning texts, dependency between students is unlikely to occur since the students should be working through the program on their own. Thus in programmed instructional experiments such as this, students would be the proper units of analysis. On the other hand, Haney (1974) has suggested that, even if students are given individualized instruction within the same classroom area, students' responses may not be independent of each other. Even though student A is working in his own carrel, he may find out at recess or after school that student B is moving along on his work a lot faster than he, student A, is moving.

This could, for example, make student A hurry through his work too fast and thus the activities of student B would be having a negative effect on student A's learning, even if an individualized instructional setting.

Some people argue that using classroom means as units of analysis rather than individual student observations deprives them of stratifying on student variables of interest and hence the researcher loses all possibilities of finding aptitude/treatment interactions (ATI's). In other words, using classroom as the unit of analysis prevents researchers from investigating treatment by student characteristic interactions. This need not be the case. Repeated measures designs (Winer, 1962) do allow researchers to test for aptitude/treatment interactions while still using the classroom as the unit of analysis. Porter (1972), however, has suggested that from a program evaluation perspective, these types of interactions (ATI's) may often not be relevant to educators. In talking about classroom-oriented Follow Through approaches, Porter states:

> If one approach works better with black children and another approach works better with white children, then what are the implications for integrated classrooms? Should both approaches be used in an integrated classroom? Such a decision would not be based on data from the evaluation since the interaction was observed for situations where children were in classrooms receiving only one approach. It seems more appropriate to investigate treatment interactions with variables defined on classroom composition, such as percent of white children in the class.

The definitions of independence and the research studies on the selection of appropriate analytic units discussed thus far are not sufficient to inform practice on the assumption of independence and consequences of violating that assumption. What is needed is an operational definition that is measurable. As long as the condition of independence is conceived to be an "ideal" property, undefinable operationally and hence unmeasurable directly, researchers will have trouble not only validating the assumption itself but also agreeing among themselves as to what the appropriate unit of analysis should be for specific research studies. An operational definition of independence can be surmised by returning to what was said earlier about threats to the assumption of independence.

CHAPTER III

AN OPERATIONAL DEFINITION OF INDEPENDENCE

Peckham et al. (1969a, 1969b) suggest two general threats to independence. The first threat (an additive effect) affects the variation of the group or aggregate variable. The second threat (a proportional effect) affects the variation of individual units within the aggregate variable. And given random assignment of subjects to groups, these two variabilities are related. That is, given random assignment of subjects to groups, the variance of the group means will equal the variance of the population of subjects divided by the number of subjects per group. Translating this into an educational example, if students (S) are randomly assigned to classrooms (C) and classrooms randomly nested within treatment levels (T) and students remain independent throughout the ongoing treatments, then the variance of the random sampling distribution of classroom means (σ^2_C) should equal the variance of the student population within classes and treatments ($\sigma^2_{S:CT}$) divided by the number of students per classroom(s).

The two threats to independence and the general rule described above for relating between group and within group variations provide the basis for the following proposed operational definition of independence. It is proposed that independence be operationally defined as the condition that σ^2_C equals $\sigma^2_{S:CT}/s$, or equivalently the condition

that the expected mean square of the grouping variable, $E(MS_{C:T})$, equals the expected mean square of individual units within the grouping variable, $E(MS_{S:CT})$. More generally, this definition states that individuals or disaggregated units can be treated as independent units whenever the variance of aggregate units is predictable from the grouping size and the variance of the disaggregate units.

The intraclass correlation coefficient (ρI), which measures the extent to which observations within the same group tend to be homogeneous relative to observations across different groups (Kirk, 1968, pp. 126-127), can also be used to define independence. Computationally the intraclass correlation coefficient equals:

$$\rho I = \frac{E(MS_{C:T}) - E(MS_{S:CT})}{E(MS_{C:T}) + (s-1)\ E(MS_{S:CT})} = \frac{s\sigma^2_C - \sigma^2_{S:CT}}{s\sigma^2_C + (s-1)\ \sigma^2_{S:CT}}$$

This formula indicates that whenever σ^2_C equals $\sigma^2_{S:CT}/s$, or equivalently whenever the $E(MS_{C:T})$ equals the $E(MS_{S:CT})$, the ρI will equal zero. Thus an intraclass correlation coefficient equal to zero also operationally defines independence.

Given these three equivalent indicators of independence, operationally there are only two distinguishable types of dependence. Positive dependence is that condition where σ^2_C is greater than $\sigma^2_{S:CT}/s$. Equivalently, positive dependence occurs whenever the ratio of the $E(MS_{C:T})$ to the $E(MS_{S:CT})$ is greater than one or whenever the ρI is greater than zero. Positive dependence is maximal when scores within each group are identical (i.e., when the $\sigma^2_{S:CT}$ equals zero) and the

scores differ only from group to group (i.e., when the σ^2_C is greater than zero). Negative dependence is that condition where σ^2_C is less than $\sigma^2_{S:CT}/s$. Equivalently, negative dependence occurs whenever the ratio of the $E(MS_{C:T})$ to the $E(MS_{S:CT})$ is less than one or whenever the ρI is less than zero. Negative dependence is maximal when group scores are identical (i.e., when the σ^2_C equals zero) and when scores within each group differ (i.e., when the $\sigma^2_{S:CT}$ is greater than zero).

Either type of dependency (positive or negative) can result from any of the following:

a. an additive effect,

b. a proportional effect, and

c. nonrandom assignment which will most probably result in either an additive or a proportional effect.

It happens that factors which would by themselves cause dependency can occur simultaneously in an experiment so as to counterbalance each other leaving the σ^2_C equal to the $\sigma^2_{S:CT}/s$. For example, if interactions between students caused both the between class variance and the within class variance to decrease, positive dependency would not show up at least so long as the $E(MS_{C:T})$ remained equal to the $E(MS_{S:CT})$. These situations, however, really do not matter; as will be seen later, the only thing that upsets the random sampling distribution of the F statistic of treatment effects when individual students are the analytic unit is the situation where σ^2_C does not equal $\sigma^2_{S:CT}/s$, given that the normality and homoscedasticity assumptions hold.

Again it should be mentioned that random assignment of students to classrooms and treatments is not synonymous with independence. However

the condition of nonrandom assignment of students to classrooms almost certainly flags the condition of dependence between students, That is, under nonrandom assignment of students to classrooms, it seems unlikely that the operational definition of independence of students will hold.

Importantly, the proposed operational definition of independence seems to be a useful one within the typical educational framework of hierarchical designs (i.e., designs which have students nested within classrooms, classrooms nested within schools, etc.). Further, this definition captures the usual definitions of independence at least insofar as variations from them affect the nature of the data. Finally, this definition has the advantage of being readily estimable.

CHAPTER IV

ANALYTIC RESULTS

Armed with an operational definition of independence which is readily measurable, it is useful to reconsider the consequence of using correlated analytic units. As stated previously, failure to have independent units of analysis may bias parameter estimates, and this in turn may alter both the actual significance level and power of a test.

The Effects of Dependence

The effects of violating the assumption of independence were studied within the context of a balanced, hierarchically-nested design, which had students nested within classrooms and classrooms nested within program or treatment levels. There was only one outcome measure per student, observations between classrooms were independent of each other and observations on classes, students within classes and students within treatments were normally distributed and had homoscedastic variances.

Data fitting the general educational model described above was analyzed using two different analysis of variance models. Model A (Table 2) was defined as:

$$Y_{ijk} = \mu + \alpha_i + b_{j(i)} + e_{ijk}$$

Table 2

Expected Mean Squares for Model A

$$Y_{ijk} = u + \alpha_i + b_{j(i)} + e_{ijk}$$

Sources of variation	d.f.	E(MS)[a]
Treatment (T)	t-1	$\sigma^2_{S:CT} + s\sigma^2_{C:T} + sc\sigma^2_T$
Classroom:T (C:T)	(c-1)t	$\sigma^2_{S:CT} + s\sigma^2_{C:T}$
Student:CT (S:CT)	(s-1)ct	$\sigma^2_{S:CT}$
Total	sct-1	

[a] $\sigma^2_T = \Sigma\alpha^2_i/(t-1)$.

where α_i represents the effect of treatment i, $b_{j(i)}$ represents the effect of classroom j within treatment i and e_{ijk} represents the error of the kth student observation within the jth classroom and the ith treatment. Model B (Table 3) was defined as:

$$Y_{ik} = \mu + \alpha_i + e_{ik}$$

where α_i again represents the effect of treatment i and e_{ik} represents the error of the kth student observation within the ith treatment. Both models regard student as random. Model B differs from Model A in that Model B contains no classroom component. Classroom is the analytic unit for Model A, while student is the analytic unit for Model B. The model having classroom as the unit of analysis (Model A) is also called the "never pool" model. The model with student as the

Table 3

Expected Mean Squares for Model B

$Y_{ik} = \mu + \alpha_i + e_{ik}$		
Sources of variation	d.f.	E(MS)[a]
Treatment (T)	t-1	$\sigma^2_{S:T} + sc\sigma^2_T$
Student:T (S:T)	(sc-1)t	$\sigma^2_{S:T}$
Total	sct-1	

[a] $\sigma^2_T = \Sigma\alpha^2_i/(t-1)$.

unit of analysis (Model B) is called the "always pool" model as its expected mean square error term, $E(MS_{S:T})$, is a pooled or weighted sum of the expected mean square between classrooms, $E(MS_{C:T})$, and the expected mean square within classrooms, $E(MS_{S:CT})$. Any hierarchically-nested data which fits Model A can also be analyzed using Model B.

The expected mean square tables for Model A and Model B, which were defined using the Millman and Glass (1967) rules of thumb, indicate that each model can test the hypothesis that there are no treatment differences. For the "never pool" model, the test is $F = MS_T/MS_{C:T}$, while for the "always pool" model, the test is $F = MS_T/MS_{S:T}$. Regardless of whether students are independent or not or whether there is a treatment effect or not, the $E(MS_T)$ for the "never pool" model equals the $E(MS_T)$ for the "always pool" model. Because the two F tests mentioned above have computationally identical numerators, problems

caused by having correlated units of analysis only become evident in comparing the denominators of the two F statistics.

If students themselves are operationally independent on the dependent variable and there are no true treatment effects ($\sigma^2_T = 0$), then all expected mean squares in Models A and B estimate $\sigma^2_{S:CT}$. One operational definition of independence on the dependent variable discussed above equated the $E(MS_{C:T})$ to the $E(MS_{S:CT})$. So, for the case of independence of student responses within classes, the following holds for Model A:

$$E(MS_{S:CT}) = E(MS_{C:T})$$

$$E(MS_{S:CT}) = \sigma^2_{S:CT} + s\sigma^2_{C:T}$$

$$E(MS_{S:CT}) = \sigma^2_{S:CT}$$

$$\text{and } \therefore \sigma^2_{C:T} = 0$$

Thus the $\sigma^2_{C:T}$ component in the $E(MS_{C:T})$ formula can also be used to define independence. That is, whenever the $\sigma^2_{C:T}$ component equals zero, observations on students can be considered independent observations. That the $E(MS_{S:T})$ in Model B estimates $\sigma^2_{S:CT}$ when given independent student observations can best be seen by looking at the sources of variation that combine to form the $SS_{S:T}$:

$$SS_{S:T} = SS_{S:CT} + SS_{C:T}$$

$$E(SS_{S:T}) = E(SS_{S:CT} + SS_{C:T})$$

$$E(SS_{S:T}) = E(SS_{S:CT}) + E(SS_{C:T})$$

$$(sc-1)t\ E(MS_{S:T}) = (s-1)ct\ E(MS_{S:CT}) + (c-1)t\ E(MS_{C:T})$$

$$E(MS_{S:T}) = [sc\sigma^2_{S:CT} - \sigma^2_{S:CT} + sc\sigma^2_{C:T} - s\sigma^2_{C:T}]/(sc-1)$$

$$= \sigma^2_{S:CT} + \frac{(c-1)s}{(sc-1)}\sigma^2_{C:T}$$

and if $\sigma^2_{C:T} = 0$

$$E(MS_{S:T}) = \sigma^2_{S:CT}$$

As stated previously, both Model A and Model B regard the student as a random variable. However, it also makes conceptual sense to think of students as fixed. This comes from thinking of individual classrooms as being unique and defined only by the particular students in the class. Given this is the case, in Model A (Table 2), classrooms would remain random and students would be considered fixed. And if so, the $E(MS_T)$ and the $E(MS_{C:T})$ in Model A would no longer contain the variance component $\sigma^2_{S:CT}$. Thus with subjects fixed, the $E(MS_{C:T})$ would not be dependent on the size of the $\sigma^2_{S:CT}$ term, as it is when subjects are considered random. Each would vary independently, which explains how the $E(MS_{C:T})$ can be smaller than the $E(MS_{S:CT})$, the operational definition of negative dependence. What this all suggests is that negative dependence seems possible only when student represents a fixed independent variable, while positive dependence is possible with either fixed or random subjects.

Classroom as Unit

Independence. Whenever classroom observations are operationally independent of each other, normally distributed and homoscedastic and there are no treatment effects, the test statistic $F = MS_T/MS_{C:T}$ in Model A will have a central F distribution with (t-1) and (c-1)t degrees of freedom. This is true, in fact, regardless of what the distribution of students within classes looks like or regardless of whether or not student observations within classrooms are independent.

Given independence of student responses within treatments, varying the number of students per class and/or the number of classes per treatment should have no affect on the actual significance level of the F test $MS_T/MS_{C:T}$. On the other hand, it is predictable that increasing the number of students per class and/or increasing the number of classes per treatment should increase the noncentrality parameter, defined as $\lambda = 1 + sc\sigma_T^2/(\sigma_{S:CT}^2 + s\sigma_{C:T}^2)$, and thus increase the power of the test $F = MS_T/MS_{C:T}$. Given independence of students, i.e., $\sigma_{C:T}^2 = 0$, one can predict, by looking at the noncentrality parameter, that increasing the number of classes per treatment or the number of students per class should identically inflate the $E(MS_T)$ and at the same time have no affect on the $E(MS_{C:T})$. However, increasing the number of classes per treatment should increase the power of the test $F = MS_T/MS_{C:T}$ to greater than that gotten by increasing the number of students per classroom by the same amount. This is due to the fact that increasing the number of students per class has no affect on the degrees of freedom error when classroom is being used as the unit of analysis. Whereas increasing

the number of classes per treatment does increase the degrees of freedom error.

Positive dependence. Positive dependence between student observations has been defined as the condition where the $E(MS_{C:T})$ is greater than the $E(MS_{S:CT})$, or concurrently the condition where the σ^2_C is greater than $\sigma^2_{S:CT}/s$. Positive dependence can occur either because a proportional type of dependency has decreased the within classroom variation or because an additive type of dependency has increased the between classroom variation. The positive dependency condition is identical to the case where the cov $(\varepsilon_i,\varepsilon_j)$, or similarly the intra-class correlation coefficient, would be greater than zero, under Draper and Smith's definition of independence. It is also identical to the instances where Cochran and Scheffé speak of responses of subjects being positively correlated.

When positive dependency occurs, under the condition of no treatment effects, $F = MS_T/MS_{C:T}$ has a central F distribution with (t-1) and (c-1)t degrees of freedom, given that classroom observations within treatment levels are independent, homoscedastic and distributed normally. This is true because the $\sigma^2_{C:T}$ component, which is greater than zero when student responses are positively correlated, affects the $E(MS_T)$ and the $E(MS_{C:T})$ in the "never pool" model (Model A) similarly, given there are no treatment effects. Increasing the number of student responses within each class and/or increasing the number of classes per treatment should have no affect on the actual significance level of the test $F = MS_T/MS_{C:T}$.

The power of $F = MS_T/MS_{C:T}$ and the method of manipulating the degree of positive dependence (how big $\sigma^2_{C:T}$ is) are confounded. If degree of positive dependence is defined by increasing the $E(MS_{C:T})$, just such an increase will decrease power. However, if degree of positive dependence is defined by decreasing the $E(MS_{S:CT})$, degree of positive dependence will not affect the power of $F = MS_T/MS_{C:T}$. Degree of positive dependence is not confounded with significance level and will in no way affect the significance level of the F statistic using class as the unit of analysis.

Negative dependence. Negative dependence between student responses has been defined as the condition where the $E(MS_{C:T})$ is less than the $E(MS_{S:CT})$, or concurrently the condition where the σ^2_C is less than the $\sigma^2_{S:CT}/s$. Negative dependence can occur either because a proportional type of dependency has increased the within classroom variation or because an additive type of dependency has decreased the between classroom variation. Again, when dealing with real world data, these two conditions of dependency are indistinguishable and can be considered as one. This particular type of dependency is identical both to the case where the cov $(\varepsilon_i, \varepsilon_j)$, or similarly the intraclass correlation coefficient, is less than zero, under Draper and Smith's definition of independence, and to the instances where Cochran and Scheffé speak of subjects being negatively correlated.

As with positive student dependency, the amount of negative student dependency should have no affect on the distribution of the statistic $F = MS_T/MS_{C:T}$, given the condition of no treatment effects,

and neither should increasing the number of students per class, the latter because the test $F = MS_T/MS_{C:T}$ uses classrooms as the unit of analysis rather than students. In addition, given negative dependence, increasing the number of classes per treatment should also not affect the actual significance level of this test of no treatment effects as, under the central case, increasing the number of classes per treatment should have no effect on the parameters $E(MS_T)$ and $E(MS_{C:T})$. As with positive dependence, when the test of the hypothesis is $F = MS_T/MS_{C:T}$, the method of manipulating degree of negative dependence is confounded with affect on the power, but not the significance level, of $F = MS_T/MS_{C:T}$.

Student as Unit

The expected mean square formulas in Table 2 (the "never pool" model) indicate that the $E(MS_{C:T})$ is the appropriate error term in testing for treatment effects. This is true whether or not student responses within classes are independent of each other (and, in fact, whether they be considered as fixed or random). Therefore, in order to check the validity of the test $F = MS_T/MS_{S:T}$, which has student as the unit of analysis, when students are not independent of each other, the $E(MS_{S:T})$ in Model B (Table 3) can be compared to the $E(MS_{C:T})$ in Model A (Table 2). If these two expected means squares are not equal, then the $E(MS_{S:T})$ has to be biased in some way. Much of the subsequent discussion will be made treating students as random, but the conclusions would hold even had students been considered fixed.

Independence. Given independence of student responses, $F = MS_T/MS_{S:T}$ would be an appropriate test of treatment effects for, as indicated earlier, the $MS_{S:T}$ and the $MS_{C:T}$ then estimate the same parameter ($\sigma^2_{S:CT}$). So, given independence of student responses within treatments, varying the number of students per class and/or the number of classes per treatment should have no affect on the actual significance level of the F test $MS_T/MS_{S:T}$. Whenever students' responses are operationally independent, there are no treatment effects and the assumptions of normality and homoscedasticity hold for observations on students within treatment levels, the test statistic $F = MS_T/MS_{S:T}$ will have a central F distribution with (t-1) and (sc-1)t degrees of freedom. On the other hand, it is predictable, by looking at the expected mean square formula for the $E(MS_T)$ in Model B (Table 3), that increasing the number of students per class and/or increasing the number of classes per treatment should increase the power of the test $F = MS_T/MS_{S:T}$. Under the independence condition, the two F tests, $F = MS_T/MS_{C:T}$ and $F = MS_T/MS_{S:T}$, should differ only in their power, with $F = MS_T/MS_{S:T}$ being the more powerful test as it has more degrees of freedom error.

Positive dependence. The effect of positive student dependency on the dependent variable can be seen by comparing the formulas for the $E(MS_{C:T})$ and the $E(MS_{S:T})$.

$$E(MS_{C:T}) = \sigma^2_{S:CT} + s\sigma^2_{C:T}$$

$$E(MS_{S:T}) = \sigma^2_{S:CT} + \frac{(c-1)}{(sc-1)} s\sigma^2_{C:T}.$$

Positive dependence deflates the $E(MS_{S:T})$ to less than the $E(MS_{C:T})$ as when the $E(MS_{C:T})$ is greater than the $E(MS_{S:CT})$, $\sigma^2_{C:T}$ will be greater than zero. However, $\sigma^2_{C:T}$ carries less weight in the $E(MS_{S:T})$ formula than it does in the $E(MS_{C:T})$ formula. Thus it can be seen that for all values of c and s, the $E(MS_{S:T})$ should be smaller than the $E(MS_{C:T})$ and therefore $F = MS_T/MS_{S:T}$ should be too liberal a test. So, when the $E(MS_{C:T})$ is greater than the $E(MS_{S:CT})$ and there is no treatment effect, $F = MS_T/MS_{S:T}$ should not be distributed as a central F. Rather $F = MS_T/MS_{S:T}$ should have an F distribution which is spread out and lies to the right of the distribution for the same F statistic under the condition of students being independent. That the positive dependency condition should result in too liberal a test statistic, when student is the unit of analysis, is consistent with what Scheffé (1959) and Cochran (1947) concluded when observations are positively correlated.

Given positive dependence, the degree of liberalness of the test $F = MS_T/MS_{S:T}$ is monotonically related to the degree of positive dependence. That is, as the $\sigma^2_{C:T}$ increases beyond zero, the discrepancy between the $E(MS_{C:T})$ and the $E(MS_{S:T})$ increases and thus the degree of liberalness increases. Given any one positive dependence level, this liberalness should be reduced as c increases and increased as s increases. Because of the general liberalness of the $F = MS_T/MS_{S:T}$ test, given positive dependence, the power of that test should be spuriously high.

Negative dependence. What happens to the magnitude of the $E(MS_{S:T})$, which designates student as the unit of analysis, when

there is negative dependence, i.e., when the $E(MS_{C:T})$ is less than the $E(MS_{S:CT})$? The $E(MS_{S:T})$ is inflated to greater than the $E(MS_{C:T})$. The following relationships show why:

$$SS_{S:T} = SS_{S:CT} + SS_{C:T}$$

$$(sc-1)t\ E(MS_{S:T}) = (s-1)ct\ E(MS_{S:CT}) + (c-1)t\ E(MS_{C:T})$$

$$E(MS_{S:T}) = \frac{(s-1)c}{(sc-1)} E(MS_{S:CT}) + \frac{(c-1)}{(sc-1)} E(MS_{C:T})$$

Whenever the $E(MS_{S:CT})$ is at all larger than the $E(MS_{C:T})$, the $E(MS_{S:T})$ is larger than the $E(MS_{C:T})$, since for s greater than or equal to 2, (s-1)c is greater than (c-1). Therefore, the test $F = MS_T/MS_{S:T}$ under this dependency condition should give too conservative a test. This indicates that whenever the $E(MS_{C:T})$ is less than the $E(MS_{S:CT})$ and there is no treatment effect, $F = MS_T/MS_{S:T}$ should not be distributed as a central F with (t-1) and (sc-1)t degrees of freedom. Rather $F = MS_T/MS_{S:T}$ should have an F distribution which is compressed and lies to the left of the distribution for the same F statistic under the situation where the $E(MS_{S:CT})$ is equal to the $E(MS_{C:T})$, or concurrently where the $E(MS_{S:T})$ equals the $E(MS_{C:T})$. That the negative dependency condition should result in a too conservative a test statistic, when the unit of analysis is the student, is consistent with what Scheffé (1959) and Cochran (1947) concluded when observations are negatively correlated.

Given negative dependency, the degree of conservativeness of the test $F = MS_T/MS_{S:T}$ is monotonically related to the degree of negative

dependence. That is, as the degree of negative dependence increases (as the σ^2_C becomes smaller and smaller relative to the ratio $\sigma^2_{S:CT}/s$) the discrepancy between the $E(MS_{C:T})$ and the $E(MS_{S:T})$ increases and thus the degree of conservativeness should increase. As for positive dependence, given any one level of negative dependence, this conservativeness should be reduced as c is increased and increased as s is increased.

The power of the test $F = MS_T/MS_{S:T}$ should spuriously be reduced as the conservativeness is increased. However both an increase in c and an increase in s inflates the $E(MS_T)$ and gives the error term of students within treatments more degrees of freedom. Thus both increasing c and s should increase the power of the test $F = MS_T/MS_{S:T}$. Whether $F = MS_T/MS_{S:T}$ will end up having more power than $F = MS_T/MS_{C:T}$ depends on how conservative $F = MS_T/MS_{S:T}$ is and how many additional degrees of freedom having students as the unit of analysis, rather than classroom, gives.

In summary, failure to have independent units of analysis, as operationally defined in this study, biases the parameter estimate of the students within treatment error term, $E(MS_{S:T})$. And this bias, in turn, influences the empirical alpha and power of the test $F = MS_T/MS_{S:T}$, which has student as the unit of analysis. The magnitude of this bias, given different degrees of both positive and negative dependence, will be studied later using simulated data.

The Preliminary Test

There are available at least two tests of treatment effects given an hierarchically-nested design with students nested within classrooms and classrooms nested within treatments. These two tests may conveniently be referred to as the "never pool" and the "always pool" procedures. The "never pool" test, $F = MS_T/MS_{C:T}$, uses classroom observations, which in this study are independent of each other, homoscedastic and normally distributed, as the units of analysis. Even though dependence of student responses does not affect the actual significance level of this test, always using the aggregate variable as the unit of analysis restricts the degrees of freedom error which, in turn, limits the power of this test of treatment effects. This few degrees of freedom problem motivates the need for another test or testing procedure with possibly greater power. The second and so-called "always pool" test, $F = MS_T/MS_{S:T}$, uses disaggregate or individual student observations, which within treatments are homoscedastic, normally distributed but not necessarily independent of each other, as the units of analysis. This particular analysis model has more degrees of freedom error, but, on the other hand, it has been shown analytically that any dependence between student responses adversely affects the significance level of the test of treatment effects. Thus it seems that there is no one "best" choice of unit of analysis for all hierarchical designs. Rather, the best testing procedure would entail using student as the analytic unit when student responses are

independent of each other, and for all other circumstances using classroom as the analytic unit. This suggests a need for some sort of conditional testing procedure in which the unit of analysis to be used in the primary test of treatment effects is determined by a preliminary test of whether or not disaggregate units are operationally independent of each other.

When there is a question as to the validity of an assumption within an experiment, a preliminary test of significance can be used to support or reject the validity of that assumption. The procedure followed is to view the assumption as a hypothesis which is testable. If this hypothesis that the assumption is true is rejected, one takes action as if the assumption were false. On the other hand, if this hypothesis that the assumption is true is not rejected, one takes action as if the assumption were, in fact, true. This sort of assumption testing procedure is sometimes used by researchers in analysis of covariance models for testing the equality of regression slopes within groups. Another example of its use includes testing the equality of variances in analysis of variance models when groups are of different sizes.

If the researcher has some a priori notion that his individual observations are independent, Peckham et al. (1969a, 1969b) and Poynor (1974) recommend using this preliminary test of the assumption to choose the unit of analysis for the primary test of treatment differences. The two-staged procedure takes the following form. The researcher begins with Model A (Table 2) and examines the null hypothesis that

the variation between classes is no different than the variation within classes, H_o: $E(MS_{C:T})$ equals $E(MS_{S:CT})$. If the hypothesis of this preliminary test is not rejected, the researcher adopts Model B (Table 3) and pools the two mean squares, $MS_{C:T}$ and $MS_{S:CT}$. Using the pooled $MS_{C:T}$ and $MS_{S:CT}$ as the error term in a test of treatment differences is identical to using individual observations as the units of analysis. On the other hand, if the hypothesis of the preliminary test, H_o: $E(MS_{C:T})$ equals $E(MS_{S:CT})$, is rejected, the researcher retains Model A in testing for treatment differences and by doing so selects the classroom as the appropriate unit of analysis. This type of conditional testing procedure can be claimed a success if the actual alpha level of the primary or conditional test remains equal to the theoretical alpha and if the power of the conditional or final F test is greater than the power of the unconditional, always correct $F = MS_T/MS_{C:T}$ test.

Actually in describing this conditional testing procedure for choosing an appropriate unit of analysis, Peckham et al. considered that this procedure would only detect dependencies of the additive type, where a constant is added to or subtracted from an entire class. Furthermore, they, as well as Poynor, considered only the possibility of the expected mean square between classes, $E(MS_{C:T})$, being larger than the expected mean square within classes, $E(MS_{S:CT})$. These two limiting considerations are needlessly restrictive. First, as previously indicated in this paper, the additive and proportional types of dependency which Peckham et al. define are in reality indistinguishable. And second, the preliminary F test can actually be rejected for

one of two reasons. The $E(MS_{C:T})$ can be greater than the $E(MS_{S:CT})$, which will happen when the σ^2_C is greater than $\sigma^2_{S:CT}/s$. This signifies a positive dependency condition. On the other hand, theoretically the $E(MS_{C:T})$ can also be less than the $E(MS_{S:CT})$, which will happen when the σ^2_C is less than $\sigma^2_{S:CT}/s$. This signifies a negative dependency condition. Thus this preliminary F test should be a two-tailed test rather than the usual one-tailed F test which Peckham et al. and Poynor recommend.

Determining that individual observations rather than group observations should be the correct unit of analysis based on the initial test of independence is, in fact, a questionable analysis procedure. If the primary test of no treatment differences is based on the results of the preliminary test of independence, then the F test for no treatment effects is a conditional test and not a regular F test. This makes the test statistic have an unknown conditional distribution (Kirk, 1968). The conditional F test statistic need not be distributed either as the regular or "always pool" F statistic with (t-1) and (sc-1)t degrees of freedom, nor as the regular or "never pool" F statistic with (t-1) and (c-1)t degrees of freedom.

Paull (1950) has investigated the distributional properties of the conditional or so-called "sometimes pool" F statistic where the $MS_{C:T}$ and the $MS_{S:CT}$ form a pooled error term in testing for treatment effects only when the $E(MS_{C:T})$ is significantly greater then the $E(MS_{S:CT})$, which is the same limited condition that Peckham et al. and Poynor talked about. Paull found that under the currently proposed

condition of operational independence, where the $E(MS_{C:T})$ equals the $E(MS_{S:CT})$, the preliminary F test, designed to choose the appropriate unit of analysis to use for the primary F test of treatment effects, is effective in making the power of the conditional test greater than the power of the "never pool" or $F = MS_T/MS_{C:T}$ test. But given the general condition of positive dependence, Paull found that the conditional test was more liberal and less powerful than the unconditional test $F = MS_T/MS_{C:T}$. It should be noted that Paull compared the powers of the "sometimes pool" and "never pool" tests at equal empirical alpha levels. That is, he did not confound power with the liberalness of the "sometimes pool" or conditional F test. As the ratio of the $E(MS_{C:T})$ to the $E(MS_{S:CT})$ increased from equal to one, Paull found that the observed alpha level of the "sometimes pool" or conditional test increased to a maximum and then decreased slowly to being equal to the nominal alpha level. This occurs because there is usually little power to find very small degrees of dependence, or very small differences between the $E(MS_{C:T})$ and the $E(MS_{S:CT})$. This means that given very small degrees of dependence, the preliminary test will most often signal the individual, rather than the group, as the appropriate unit of analysis. And using the individual as the analytic unit will make the primary test of treatment effects too liberal a test. As the degree of positive dependence increases though, the researcher rejects the null hypothesis of operational independence more often. This dictates using classroom as the unit of analysis more often in the test for treatment differences. This, in turn, suggests that as the

degree of positive dependence increases, the distributional properties of the "sometimes pool" or conditional F test become more and more similar to the distributional properties of the "never pool" or unconditional F = $MS_T/MS_{C:T}$ test. From this, one can conclude that for great amounts of dependency, the conditional test is just fine because it simply becomes an unconditional, "never pool" F test, F = $MS_T/MS_{C:T}$.

Paull also found that the number of classes per treatment and the number of students per class clearly affected the magnitude of the distributional differences between the "sometimes pool" and the "never pool" F tests, given positive dependence. Under the condition of positive dependence, a large number of classes per treatment is desirable in two respects. First, as c increases the preliminary test becomes more powerful and correctly identifies classrooms as the appropriate unit more often. And second, when pooling of mean square error terms is prescribed, the pooled mean square $MS_{S:T}$ is weighted in favor of the valid and correct mean square error, $MS_{C:T}$. As the number of students per class, s, increases, the preliminary test F = $MS_{C:T}/MS_{S:CT}$ again becomes more powerful and thus correctly signals the classroom as the proper unit of analysis more often. However, counterbalancing this positive effect of increasing s when given positive dependence is the fact that increasing s gives more weight to the wrong error term, $MS_{S:CT}$, which is smaller than the valid error term, $MS_{C:T}$. Thus the effect on the primary F test of increasing the number of individuals per group is due to a combination of two factors and most importantly

depends on how much larger than the $E(MS_{S:CT})$ the $E(MS_{C:T})$ is. Lastly, Paull considered the effect of increasing the nominal alpha level of the preliminary F test and found that, given positive dependence, the magnitude of the undesirable property (liberalness) of the conditional F test was reduced somewhat with just such an increase. There was, however, a critical alpha level above which increasing the alpha level of the preliminary F test resulted in the conditional F test becoming more liberal. In Paull's example, this critical alpha value was very large, around 0.77.

Paull finally comes up with recommending the following rule as when and when not to pool the two mean square error terms, $MS_{C:T}$ and $MS_{S:CT}$. The rule entails pooling the two mean square error terms only if their ratio is less than $2F_{50}$, where F_{50} is the 50% point of the F distribution with (c-1) and (s-1)ct degrees of freedom. Paull claims that this pooling decision rule is one which tends to "stabilize the disturbances" between the distributions of the two statistics $F = MS_T/MS_{C:T}$ and $F = MS_T/MS_{S:T}$, given "intermediate" conditions of positive dependence, while still taking advantage of a considerable portion of the possible gain in power of pooling, given very low levels of positive dependence. The present author, however, questions this "rule of thumb." Ideally, the researcher wants most not to reject the null preliminary hypothesis of operational independence, as when this null hypothesis is not rejected the degrees of freedom error and mean squares between classes ($MS_{C:T}$) and within classes ($MS_{S:CT}$) can be pooled. The error the researcher needs to guard most against is a

Type II error (β), or not rejecting the null hypothesis when it is, in fact, false. One way to decrease the probability of a Type II error is to increase the probability of a Type I error (α). However, doing as Paull recommends and taking twice the critical value given a large alpha of .50 has the same effect as selecting a small alpha level in the first place.

Again, as mentioned above, within a two level, hierarchically-nested design with individual units nested within groups and groups nested within treatments, Paull studied the distributional properties of the conditional F test given the usual one- and upper-tailed only preliminary F test. These distributional properties were studied only under the condition of positive dependence. It has been noted, however, that this preliminary test can be rejected for one of two reasons (the occurrence of positive dependence and the occurrence of negative dependence) and thus instead should be a two-tailed F test. Or, if negative dependence is suspected, a one- and lower-tailed only preliminary F test would seem an appropriate possibility. On the other hand, there is no reason to believe that the presence and direction of the distributional effects found by Paull by changing the four parameters--the number of individuals per group, the number of groups per treatment, the degree of dependence and the nominal alpha level of the preliminary F test--should differ given negative dependence and/or a two-tailed preliminary F test. It is predictable, however, that given these other conditions, the magnitude of these effects should change. The estimates of the magnitude of these distributional effects will be studied later using simulated data.

CHAPTER V

SIMULATION PROCEDURES

The present investigation has addressed and discussed two specific questions. First, what is the effect of using correlated units of analysis? That is, what happens to parameter estimates and the probability of Type I and II errors when the assumption of independence is violated? And second, what is the effect of using a preliminary test of independence to choose the unit of analysis for the primary test of treatment effects? Thus far the two questions have been presented and discussed analytically. As yet, though, no attempt has been made to describe the actual magnitude of effects, whose presence and direction were predicted in the preceding analytic chapter, and which is the whole purpose of the simulation study. Thus the simulation study will demonstrate the size of the distributional effects for situations held to be common in educational settings.

The procedures employed to empirically study the magnitude of distributional effects will now be discussed. First, the description of the design parameters will be given. Second, the data generating routine will be described along with a presentation of tests performed on the generation routine.

Simulation Parameters

As stated previously, the general research design considered in the present study was a balanced, hierarchically-nested design. There were two levels to the nesting. Individuals were nested within groups and groups were nested within treatments. The design assumed there to be one outcome measure per subject. Data such as this can be analyzed using one of two analysis of variance models, which were described in the previous chapter and presented in Tables 2 and 3.

For this simulation study, the number of treatment groups, t, was held constant at two. Both the number of classes per treatment, c, and the number of students per class, s, were allowed to vary so that possible trends in the sampling distributions of the F statistic could be investigated as these two parameters increased. Three values of classes per treatment (2, 5, and 10) were selected. Two classes per treatment is the minimum allowable number of classes per treatment such that the treatment effects can be kept unconfounded from the classroom effects. Ten classes per treatment was chosen as the upper limit as in practice most educational studies do not employ more than ten classes or groups per treatment condition. The sample size of students ranged from five observations per classroom to twelve and twenty. The two extreme values of subjects per class were chosen because five subjects per group is relevant for small group studies and 20 subjects per group comes relatively close to the average number of students per classroom in elementary school settings.

Two competing methods for defining and manipulating levels of positive and negative dependence among units were considered, both

of which seem equally valid. One method uses the ratio of the $E(MS_{C:T})$ over the $E(MS_{S:CT})$. This method allows the intraclass correlation coefficient (ρI) to vary as the number of students per class varies (i.e., the ρI naturally gets smaller as the number of students per class increases). The intraclass correlation coefficient could also have been used to define and manipulate levels of dependence. This alternative method would entail keeping the ρI at a constant value for each level of dependence regardless of how many classes there were per treatment or students there were per class, but would force either the $E(MS_{C:T})$ or the $E(MS_{S:CT})$ to vary as the number of students per class varied. The first method described above for defining and manipulating degrees of dependence, altering the relationship between the $E(MS_{C:T})$ and the $E(MS_{S:CT})$ by varying the $E(MS_{C:T})$, was selected for several reasons. First, this method keeps the conceptual population of students the same. That is, randomly deleting or adding students to any classroom does not affect the within class variability. Second, previous research (Paull, 1950) has used the ratio of the expected mean squares between groups to the expected mean squares within groups in order to study the effects of preliminary tests for pooling mean squares.

Two general types of dependence were studied. Positive dependence was defined as occurring whenever the ratio of the $E(MS_{C:T})$ to the $E(MS_{S:CT})$ was greater than one (similarly the ρI was positive). Negative dependence was defined as occurring whenever the ratio of the $E(MS_{C:T})$ to the $E(MS_{S:CT})$ was less than one (similarly the ρI was

negative). Within each type of dependence, two degrees or levels of dependence were studied. The degree of positive and negative dependence was varied in order to study probable trends of disturbance in sampling distributions of F tests of treatment effects, given different degrees of dependence. Along with the condition of independence the four defined conditions of dependence were:

- Independence

$$E(MS_{C:T})/E(MS_{S:CT}) = 1$$

- Positive dependence

$$E(MS_{C:T})/E(MS_{S:CT}) = 2$$

$$E(MS_{C:T})/E(MS_{S:CT}) = 3$$

- Negative dependence

$$E(MS_{C:T})/E(MS_{S:CT}) = .50$$

$$E(MS_{C:T})/E(MS_{S:CT}) = .33$$

The choice of particular degrees of positive and negative dependence was somewhat arbitrary. However, an attempt was made to investigate degrees of dependence which typically occur within educational research studies. Smith (1974) suggested that, for elementary school children, classroom variance usually accounts for anywhere between 20 and 50% of the student variation within treatment levels for achievement measures such as reading and arithmetic. For studies conducted within a "tight" regional area, the classroom variation most likely would account for approximately 20% of the student variation; while for studies conducted nationwide the classroom variation most likely would account for about 50% of the student variation.

Haney (1974), working with data from 14 Philadelphia schools in the Follow Through study, found that students (ungrouped) had a variance of 137 on the Metropolitan Achievement Test total mathematics score. When Haney randomly formed groups of eight students (which was, on the average, the number of students per actual group), the variance between random groups was 25, while the variance between actual groups (or classrooms) was 52, which was twice the size of the variance using random groupings. Haney's data seem consistent with Smith's suggested limits in that when using the intraclass correlation coefficient to calculate the percentage of student variance attributable to classroom differences, the random groups accounted for approximately 6% of the student variation, while the actual classrooms accounted for approximately 28%. However, in calculating the variances, Haney did not take into account Follow Through, non-Follow Through differences. If he had, his intraclass correlation coefficients would likely have been somewhat smaller.

The actual degrees of positive dependence chosen reflect fairly well Haney's data, using sample sizes which seem in the range of common usage. It is also desirable that for different numbers of students, the intraclass correlation coefficient remain relatively small, as it is with small to intermediate degrees of dependence that effects of dependence and effects of the preliminary test of independence seem most nebulose. As Table 4 indicates, the intraclass correlation coefficients used in the present study were in the small to intermediate range.

Table 4

Theoretical Intraclass Correlation Coefficients for Selected Numbers of Students and Degrees of Dependence

	$E(MS_{C:T})/E(MS_{S:CT})$				
	.33	.50	1	2	3
s = 5	-.153	-.111	.000	.167	.286
s = 12	-.059	-.044	.000	.077	.143
s = 20	-.034	-.026	.000	.048	.091

Table 5 indicates all possible combinations of the four parameters, the number of students per class, the number of classes per treatment, the type of dependence, and the degree of dependence, included in this simulation study. An "†" marks the cells actually used in this study.

Table 5

Design of Study

		$E(MS_{C:T})/E(MS_{S:CT})$				
c	s	.33	.50	1	2	3
2	5	†	†	†	†	†
	12					
	20					
5	5	†	†	†	†	†
	12	†	†	†	†	†
	20	†	†	†	†	†
10	5	†	†	†	†	†
	12					
	20					

For studying statistical power, the noncentral case was created by adding 0.4 within class standard deviation units to each student's observation in one treatment group. This value of 0.4 was determined by approximating power under condition that the $E(MS_{C:T})$ equalled the $E(MS_{S:CT})$ for the selected numbers of classes per treatment and students per class. The size of the effects was selected so as to give theoretical power values within the desired moderate range, given independence of analytic units.

Data Generation Routine

The generation of unit normal variates involved two steps. First, pseudo-random variates were obtained by calling subroutine RANDU (IBM, 1970, p. 77). This subroutine uses the power residue method to generate uniform random variables. Second, the GAUSS subroutine (IBM, 1970, p. 77) took 12 RANDU variates and used the Central Limit Theorem to rescale and normalize the uniform variates to be distributed as N(0,1) variates. Each time the generation program was run, which was once for each marked row in Table 5, the seed or initial random number was changed to insure independence among the resulting F distributions. The number of iterations per selected row in Table 5 was 1000.

Four basic steps were used to create the dependent variable Y_{ijk} with a known degree of dependence, where k indexes a student observation within classroom j and treatment i. First, s number of N(0,1) variates (Y'_{ijk}) were summed and averaged to get a classroom mean ($\bar{Y}'_{ij.}$). Second, within each classroom, the classroom mean was subtracted from each

individual observation ($Y'_{ijk} - \bar{Y}'_{ij.}$). Third, the class means were adjusted by the square root of the dependence level, $E(MS_{C:T})/E(MS_{S:CT})$. And finally, the adjusted means were added back to their respective set of deviations. Thus the dependent variable was calculated as

$$Y_{ijk} = Y'_{ijk} - \bar{Y}'_{ij.} + \left\{ \bar{Y}'_{ij.} \sqrt{E(MS_{C:T})/E(MS_{S:CT})} \right\} .$$

There was special concern that, after adjusting the normal random variates within classrooms for dependency, the student observations, within treatments but across classrooms, remain normally distributed. This is necessary when Model B (ignoring classrooms) is being considered. Theoretically the variates within treatments, adjusted for dependency, should be normally distributed as each adjusted variate is a linear combination of two variates independently distributed as normal variables (Graybill, 1961, pp. 56-57).

Chi-square tests were run to see if example distributions of class means and individual observations, for both independent and defined dependent situations, would approximate the normal distribution, which theoretically they should. The abscissa of the normal distribution was divided into 12 sections. A sample of 10,000 observations, adjusted to fit both the independent and four defined dependent conditions, and 2000 classroom means (there were five observations for each class) were generated and the number of cases falling into each of the 12 defined intervals was counted (Table A-1 in Appendix A). None of the six χ^2 tests of fit were rejected at the .10 level, which suggests

good approximation to the normal distribution for both class means and individual observations.

In addition, the mean, variance, skewness and kurtosis for student within treatment type data were calculated for each marked cell in Table 5. These distributional statistics are displayed in Appendix A (Tables A-2 through A-6). In all cases the four distributional statistics were visually in close agreement to their known parameters. For data constructed under both dependent and independent conditions, the means for the adjusted student within treatment observations should equal zero for each treatment, given the central case. Given the non-central case, the means for student within one treatment condition should equal zero and within the other treatment condition 0.4. The expected variances of the student within treatment data can be calculated from the following:

$$\mathrm{Var}(Y_{ijk}) = \mathrm{Var}(Y'_{ijk}) + (\theta-1)^2\mathrm{Var}(\bar{Y}'_{ij.}) + 2(\theta-1)\ \mathrm{Cov}(Y'_{ijk}, \bar{Y}'_{ij.}),$$

$$\text{where } \theta = \sqrt{E(MS_{C:T})/E(MS_{S:CT})}$$

$$\mathrm{Var}(Y_{ijk}) = 1 + 2(\theta-1)/s + (\theta-1)^2/s, \text{ as the}$$

$$\mathrm{Cov}(Y'_{ijk}, \bar{Y}'_{ij.}) = 1/s.$$

As one example, if $s = 12$ and $\theta = \sqrt{3}$, the expected variance of the adjusted observations within one treatment level is 1.167. Given these two parameters, the empirical variances, one for each treatment level, are very close to this predicted value (see Tables A-2, A-4,

and A-6). The remaining empirical variances for the simulated data, for all combinations of s and θ, were also close in value to their respective expected values. And finally, the empirical skewness and kurtosis estimates for the simulated student within treatment type data were also very close to their expected values of zero for all simulated runs.

Distributional properties of four observed mean squares adjusted by their expected values were also examined (Tables A-7 through A-11 in Appendix A). Each of these four standardized mean squares, MS_T, $MS_{C:T}$, $MS_{S:CT}$, and $MS_{S:T}$, should be distributed as chi-square variables with a mean equal to its degrees of freedom, a variance equal to twice its degrees of freedom and a skewness equal to the square root of eight divided by its degrees of freedom (Glass & Stanley, 1970, pp. 231-232). Under the condition of no treatment effect, the mean, variance and skewness of these standardized mean squares across 1000 samples were visually close to their respective known chi-square parameters. All three sets of above analysis suggest that the data generation routine was in proper working order.

CHAPTER VI

UNITS OF ANALYSIS: EMPIRICAL ESTIMATES OF EFFECTS

Chapter IV dealt analytically with how dependence between disaggregate units affects the sampling distribution of two F statistics, one using the aggregate unit as the unit of analysis ($F = MS_T/MS_{C:T}$) and the other using the disaggregate unit as the unit of analysis ($F = MS_T/MS_{S:T}$). The present chapter demonstrates empirically the size of the effects hypothesized in Chapter IV for situations held to be common in educational research. The variables of interest in both the analytic and the empirical investigations were number of subjects per group, number of groups per treatment, type of dependence, and degree of dependence. Any combination of levels of the above variables represents a sampling distribution which could have been generated. An attempt was made to generate sampling distributions for a subset of the totality which would afford as much information as possible about the effects of the above mentioned variables on the sampling distributions of the two F statistics. The subset of variable levels chosen to study is represented by the three-dimensional matrix in Table 5.

Classroom as Unit

Model A or the "never pool" model (Table 2) is the analytic model of concern when classroom is designated as the unit of analysis in testing for treatment effects. With classroom as the unit of analysis, the test of treatment effects is $F = MS_T/MS_{C:T}$.

Independence

Whenever student responses within and between classrooms were operationally independent of each other, homoscedastic and normally distributed between classrooms and given the data were contrived such that there were no treatment effects, the test statistic $F = MS_T/MS_{C:T}$ was always distributed as a central F with (t-1) and (c-1)t degrees of freedom. Under this set of conditions, the observed alpha level of the test $F = MS_T/MS_{C:T}$ consistently agreed to within 1.96 standard deviation units, $\sqrt{\alpha(1-\alpha)/1000}$, with the nominal alpha levels (Table 6). Across the five different combinations of s and c, the mean observed alpha levels equalled .008 for $\alpha = .01$, .023 for $\alpha = .025$, .050 for $\alpha = .05$, .100 for $\alpha = .10$, and .240 for $\alpha = .25$. As one can see, these averaged observed alpha levels are relative close to their nominal counterparts. Table 6 also indicates that the number of students per class and the number of classes per treatment had no affect on the actual significance level of the test $F = MS_T/MS_{C:T}$. Another view of the effect on the actual alpha levels of increasing s and increasing c can be gotten by calculating mean empirical alpha levels across the five nominal alpha levels. Doing this and keeping c constant at 5 gave mean empirical

Table 6

Empirical Type I Errors for $F = MS_T/MS_{C:T}$

c	s	d.f. error	Nominal alpha .010	.025	.050	.100	.250	Mean alpha
2	12	(2)	.010[a]	.028[a]	.049[a]	.090[a]	.262[a]	.088
5	5	(8)	.011[a]	.025[a]	.061[a]	.107[a]	.244[a]	.090
5	12	(8)	.007[a]	.018[a]	.046[a]	.095[a]	.226[a]	.078
5	20	(8)	.007[a]	.028[a]	.049[a]	.100[a]	.248[a]	.086
10	12	(18)	.006[a]	.017[a]	.049[a]	.094[a]	.231[a]	.079
Mean alpha			.008	.023	.050	.100	.240	

[a]Empirical alpha is within 1.96 standard errors of the nominal alpha.

alpha levels equalling .090 for s = 5, .078 for s = 12, and .086 for s = 20. Averaging across the five nominal alpha levels and keeping s constant at 12 gave mean empirical alpha levels equal to .088 for c = 2, .078 for c = 5, and .079 for c = 10. The standard by which each of the mean empirical alpha levels, as s and c were varied, should be judged is the mean of the five nominal alpha levels, which equals .087. As expected, this standard mean is in close agreement to those found when s and c were varied.

As predicted analytically, given independence of student responses, i.e., $E(MS_{C:T})/E(MS_{S:CT})$ equalled one, and noncentral conditions, both increasing the number of students per class and increasing the number of classes per treatment increased the power of the test $F = MS_T/MS_{C:T}$ (Table 7). Averaging across the five nominal

Table 7

Empirical Powers for $F = MS_T/MS_{C:T}$

$E(MS_{C:T})/E(MS_{S:CT})$	c	s	d.f. error	Nominal alpha .010	.025	.050	.100	.250	Mean power
.33	2	12	(2)	.074	.149	.297	.499	.793	.362
	5	5	(8)	.265	.436	.594	.760	.896	.590
	5	12	(8)	.661	.839	.916	.966	.997	.876
	5	20	(8)	.896	.967	.990	.998	.999	.970
	10	12	(18)	.986	.996	.999	1.000	1.000	.996
	Mean power			.576	.677	.759	.845	.937	
.50	2	12	(2)	.055	.124	.217	.389	.708	.299
	5	5	(8)	.162	.287	.420	.592	.813	.455
	5	12	(8)	.450	.651	.794	.889	.967	.750
	5	20	(8)	.721	.861	.937	.978	.997	.899
	10	12	(18)	.920	.965	.984	.995	1.000	.973
	Mean power			.462	.578	.670	.769	.897	
1	2	12	(2)	.036	.086	.153	.261	.533	.214
	5	5	(8)	.078	.149	.227	.359	.621	.287
	5	12	(8)	.193	.322	.478	.657	.843	.499
	5	20	(8)	.381	.552	.699	.822	.938	.678
	10	12	(18)	.621	.773	.852	.915	.967	.826
	Mean power			.262	.376	.482	.603	.780	
2	2	12	(2)	.023	.057	.103	.194	.406	.157
	5	5	(8)	.038	.085	.141	.229	.447	.188
	5	12	(8)	.075	.160	.261	.399	.638	.307
	5	20	(8)	.164	.292	.429	.569	.781	.447
	10	12	(18)	.268	.411	.562	.705	.857	.561
	Mean power			.114	.201	.299	.419	.626	
3	2	12	(2)	.021	.047	.086	.167	.353	.135
	5	5	(8)	.024	.069	.107	.191	.383	.155
	5	12	(8)	.053	.110	.176	.299	.524	.232
	5	20	(8)	.094	.195	.302	.448	.666	.341
	10	12	(18)	.173	.275	.386	.539	.765	.428
	Mean power			.075	.138	.211	.329	.538	

alpha levels but keeping c constant at 5 gave mean power values equallying .287 for s = 5, .499 for s = 12, and .678 for s = 20. Averaging across the same nominal alpha levels but keeping s constant at 12 gave mean power values equalling .214 for c = 2, .499 for c = 5, and .826 for c = 10. Table 7 also shows that increasing the number of classrooms per treatment had a more positive effect on increasing the power of $F = MS_T/MS_{C:T}$ than did increasing the number of students per classroom. This is illustrated in Table 7 by comparing the difference in estimated powers as c is increased, but keeping s constant at 12, from c equals 5 to c equals 10 (a difference of five classrooms per treatment) to the difference in estimated powers as s is increased, but keeping c constant at 5, from s equals 12 to s equals 20 (a difference of eight students per classroom). As the nominal alpha level was increased from .01 to .25, an increase of five classes per treatment increased the estimated power from .193 to .621, which is an increase in power of 221% to an increase in power from .843 to .967, which is an increase of 14.7%. Correspondingly, an increase of eight students increased the estimated power from .193 to .381, which is an increase in power of only 97.4%, to an increase in power from .843 to .938, which is an increase of only 11.3%. This same sort of general relationship between estimated powers as c and s were increased by relative amounts held up across all five nominal alpha levels examined (.01, .025, .05, .10, and .25).

Positive Dependence

Positive dependence between student responses within classrooms was defined as the condition where the $E(MS_{C:T})$ was greater than the $E(MS_{S:CT})$. For the simulation study, the $E(MS_{S:CT})$ always equalled one and the degree of positive dependence was defined by manipulating the value of the $E(MS_{C:T})$. The degree of positive dependence was said to increase as the value of the $E(MS_{C:T})$ increased above one. In particular, two degrees of positive dependence were studied. They were $E(MS_{C:T})/E(MS_{S:CT})$ equal to 2 and $E(MS_{C:T})/E(MS_{S:CT})$ equal to 3.

Whenever student responses within classrooms were operationally dependent upon each other in a positive manner, but between classroom observations were independent, homoscedastic and normally distributed, and given no treatment effects, the test statistic $F = MS_T/MS_{C:T}$ again had a central F distribution with (t-1) and (c-1)t degrees of freedom. And as predicted in the earlier analytic work, neither the number of students per class nor the number of classes per treatment nor the existence and/or degree of positive dependence had any effect on the actual significance level of the test $F = MS_T/MS_{C:T}$. The observed alpha values, given both degrees of positive dependence between student responses, for the five nominal alpha levels are identical to those when given independence between student responses and are displayed in Table 6 for prespecified values of s and c.

The estimated powers for $F = MS_T/MS_{C:T}$ (the "never pool" test) for the two simulated positive dependence conditions are shown at the bottom of Table 7. The estimated statistical powers when using class

as the unit of analysis decreased as the degree of positive dependence increased from $E(MS_{C:T}/E(MS_{S:CT}) = 2$ to $E(MS_{C:T})/E(MS_{S:CT}) = 3$. This inverse relationship, however, is purely a function of the degree of dependence being defined by manipulating the $E(MS_{C:T})$ and keeping the $E(MS_{S:CT})$ constant. The relationship between degree of positive dependence and magnitude of the estimated power values would have equalled zero if the degree of dependence had instead been defined by altering the $E(MS_{S:CT})$ or if the noncentral case had been created by adding 0.4 of the between class variance, rather than 0.4 of the within class variance, to the adjusted random variates of one treatment level. In other words, because of the way the data base in the noncentral case was built, power and degree of dependency are confounded when classroom is the unit of analysis. The effect of increasing s and/or c would have increased the power of $F = MS_T/MS_{C:T}$ no matter how the degree of dependence and the noncentral case had been defined. Similarly, the number of students per class and the number of classes per treatment would have had no effect on the actual significance level of $F = MS_T/MS_{C:T}$ no matter how degree of dependence had been defined.

Negative Dependence

Negative dependence between student responses within classes was defined as the condition where the $E(MS_{C:T})$ was less than the $E(MS_{S:CT})$ or concurrently where the variance between classrooms was less than that predicted had equal numbers of students been randomly assigned to classrooms. Two degrees of negative dependence were studied. They

were $E(MS_{C:T})/E(MS_{S:CT})$ equal to .5 and $E(MS_{C:T})/E(MS_{S:CT})$ equal to .33.

The amount of negative student dependency, an increase in students per class and an increase in classes per treatment all had no effect on the distribution of the test statistic $F = MS_T/MS_{C:T}$, given the central case and given that observations between classes were independent, homoscedastic and normally distributed. When classroom is the unit of analysis, the observed alpha values, given negative dependence within classrooms, for the five nominal alpha levels are identical to those when given independence and are displayed in Table 6 for prespecified values of s and c.

The empirical powers for $F = MS_T/MS_{C:T}$ for the two simulated negative dependence conditions are shown at the top of Table 7. As with the case of positive dependency, the noncentrality parameter, and thus power, and the degree of negative dependence are confounded when classroom is the unit of analysis. Once again, however, increasing s and/or c would have increased the power of $F = MS_T/MS_{C:T}$ no matter how degree of negative dependence and the noncentral case had been defined.

Student as Unit

Model B or the "always pool" model (Table 3) is the analytic model of concern when student is used as the unit of analysis in testing for treatment effects. With student as the analytic unit, the test of treatment effects is $F = MS_T/MS_{S:T}$.

Independence

Empirical Type I errors for the $F = MS_T/MS_{S:T}$ test given independence of student responses within treatments are given in Table 8. All estimated Type I errors, across all five combinations of s and c and across all five nominal alpha levels, were within 1.96 standard errors of the nominal alphas. Across the five combinations of s and c, the mean observed alpha levels equalled .010 for $\alpha = .01$, .024 for $\alpha = .025$, .051 for $\alpha = .05$, .098 for $\alpha = .10$, and .241 for $\alpha = .25$. The mean observed alpha levels appear very close to their respective nominal values. Table 8 also indicates that increasing the number of students per class and/or increasing the number of classes per treatment had no effect on the probability of Type I errors for $F = MS_T/MS_{S:T}$. Averaging across the five nominal alpha levels and keeping c constant at 5 gave mean observed alpha levels of .082 for $s = 5$, .081 for $s = 12$, and .092 for $s = 20$. Averaging across the nominal alpha values and keeping s constant at 12 gave mean observed alpha levels of .087 for $c = 2$, .081 for $c = 5$, and .082 for $c = 10$. All of the above mean empirical alpha levels are in close agreement to their expected mean empirical alpha level which equals .087. Thus, as expected, given that students within treatments were independent, homoscedastic and normally distributed and given there were no treatment effects, $F = MS_T/MS_{S:T}$ was distributed as a central F having (t-1) and (sc-1)t degrees of freedom.

Empirical statistical powers for $F = MS_T/MS_{S:T}$ for the five analysis of variance designs are given in Table 9. As expected, the

Table 8

Empirical Type I Errors for $F = MS_T/MS_{S:T}$

$E(MS_{C:T})/E(MS_{S:CT})$	c	s	d.f. error	Nominal alpha .010	.025	.050	.100	.250	Mean alpha
	2	12	(46)	.000	.001	.001	.008	.059	.014
	5	5	(48)	.000	.000	.004	.011	.058	.015
.33	5	12	(118)	.000	.000	.001	.005	.049	.011
	5	20	(198)	.000	.000	.001	.005	.050	.011
	10	12	(238)	.000	.000	.000	.004	.053	.011
Mean alpha				.000	.000	.001	.007	.054	
	2	12	(46)	.001	.002	.009	.029	.100	.028
	5	5	(48)	.000	.007	.012	.023	.114	.031
.50	5	12	(118)	.000	.003	.007	.018	.108	.027
	5	20	(198)	.000	.002	.008	.021	.122	.031
	10	12	(238)	.000	.000	.004	.021	.108	.027
Mean alpha				.000	.003	.008	.022	.110	
	2	12	(46)	.011[a]	.032[a]	.058[a]	.091[a]	.242[a]	.087
	5	5	(48)	.012[a]	.022[a]	.048[a]	.097[a]	.231[a]	.082
1	5	12	(118)	.008[a]	.020[a]	.046[a]	.094[a]	.237[a]	.081
	5	20	(198)	.014[a]	.024[a]	.050[a]	.111[a]	.260[a]	.092
	10	12	(238)	.007[a]	.022[a]	.052[a]	.096[a]	.235[a]	.082
Mean alpha				.010	.024	.051	.098	.241	
	2	12	(46)	.064	.093	.146	.226	.383	.182
	5	5	(48)	.044	.081	.123	.195	.351	.159
2	5	12	(118)	.051	.096	.139	.210	.395	.178
	5	20	(198)	.067	.116	.158	.242	.405	.198
	10	12	(238)	.056	.093	.144	.215	.356	.173
Mean alpha				.056	.096	.142	.218	.378	
	2	12	(46)	.102	.165	.224	.312	.457	.252
	5	5	(48)	.085	.122	.180	.252	.417	.211
3	5	12	(118)	.106	.153	.211	.298	.465	.247
	5	20	(198)	.126	.179	.249	.324	.475	.271
	10	12	(238)	.108	.158	.210	.290	.451	.243
Mean alpha				.105	.155	.215	.295	.453	

[a]Empirical alpha is within 1.96 standard errors of the nominal alpha.

Table 9

Empirical Powers for $F = MS_T/MS_{S:T}$

$E(MS_{C:T})/E(MS_{S:CT})$	c	s	d.f. error	Nominal alpha .010	.025	.050	.100	.250	Mean power
	2	12	(46)	.028	.087	.177	.342	.685	.264
	5	5	(48)	.031	.105	.200	.397	.727	.292
.33	5	12	(118)	.282	.485	.694	.855	.962	.656
	5	20	(198)	.685	.857	.929	.980	.998	.890
	10	12	(238)	.854	.942	.976	.996	1.000	.954
Mean power				.376	.495	.595	.714	.874	
	2	12	(46)	.053	.126	.211	.376	.656	.284
	5	5	(48)	.057	.139	.229	.406	.689	.304
.50	5	12	(118)	.321	.481	.646	.800	.937	.635
	5	20	(198)	.644	.815	.897	.948	.991	.859
	10	12	(238)	.805	.898	.957	.981	.998	.928
Mean power				.374	.492	.588	.702	.854	
	2	12	(46)	.114	.190	.286	.409	.607	.321
	5	5	(48)	.102	.186	.274	.402	.645	.322
1	5	12	(118)	.346	.470	.590	.742	.869	.600
	5	20	(198)	.604	.730	.810	.881	.947	.794
	10	12	(238)	.720	.826	.883	.931	.971	.866
Mean power				.377	.480	.569	.669	.808	
	2	12	(46)	.185	.262	.345	.437	.612	.368
	5	5	(48)	.149	.221	.301	.400	.616	.337
2	5	12	(118)	.357	.454	.544	.641	.782	.556
	5	20	(198)	.563	.670	.733	.798	.884	.730
	10	12	(238)	.642	.730	.795	.859	.920	.789
Mean power				.379	.467	.544	.627	.763	
	2	12	(46)	.226	.312	.390	.455	.621	.401
	5	5	(48)	.171	.236	.320	.420	.617	.353
3	5	12	(118)	.367	.442	.526	.610	.752	.539
	5	20	(198)	.537	.623	.703	.753	.846	.692
	10	12	(238)	.589	.687	.744	.807	.879	.741
Mean power				.378	.460	.537	.609	.743	

table shows that, given $E(MS_{C:T})/E(MS_{S:CT})$ equals one and the noncentral condition, increasing the number of subjects per class and increasing the number of classes per treatment increased the power of the test $F = MS_T/MS_{S:T}$. Averaging across the five nominal alpha levels and keeping c constant at 5 gave mean power values equal to .322 for s = 5, .600 for s = 12, and .794 for s = 20. In similar fashion, averaging over the nominal alpha levels but keeping s constant at 12 gave mean power values equalling .321 for c = 2, .600 for c = 5, and .866 for c = 10. Each one of these mean power values exceeds its respective mean power value when class, rather than student, is used as the unit of analysis. In fact, given independence of student responses, all of the 25 powers shown in Table 9 are larger than the 25 corresponding powers displayed in Table 7, which indicates that, as expected under independence, $F = MS_T/MS_{S:T}$ was always more powerful a test than $F = MS_T/MS_{C:T}$. This increase in power can be accredited to an increase in degrees of freedom error. (Comparing any two respective rows in Tables 7 and 9, given $E(MS_{C:T})/E(MS_{S:CT})$ equals 1, is similar to Table 1 from Peckham et al.)

Positive Dependence

The analytic analysis (Chapter IV) of the effect of positive dependence given student as the unit of analysis suggested that the test $F = MS_T/MS_{S:T}$ would result in a too liberal test statistic. Determining just how liberal that test statistic would be given certain parameters can be estimated using either of two methods: (a) Monte

Carlo analysis and (b) rescaling the biased F statistic and then using the tabled F values to measure the degree of liberalness. Both methods of estimation were used in this study and their comparable results will be discussed in this section, beginning with the Monte Carlo analysis.

The effects of positive dependence between disaggregate units on the actual significance levels of the test $F = MS_T/MS_{S:T}$, measured by the Monte Carlo method, are reported at the bottom of Table 8. None of the observed alpha levels are within 1.96 standard errors of the nominal alpha levels. All observed alpha levels are larger than their theoretical complements, which signals liberalness of the test statistic. This indicates that, given positive dependence and no treatment effects, $F = MS_T/MS_{S:T}$ is not distributed as a central F, but in fact is distributed as an F distribution which is located to the right of the central F distribution found when given independence of student responses and the same F statistic. This is in complete agreement with the work of Scheffé (1959) and Cochran (1947) and with analytic work presented in Chapter IV.

As a second and alternative way to determine the extent of the liberalness, the F statistic $MS_T/MS_{S:T}$ can be adjusted such that it has a central F distribution under the null hypothesis. Doing this would allow using the F table to estimate the degree of effect positive dependence has on the alpha level. Given the null, the $E(MS_T)$ equals the $E(MS_{C:T})$ and thus the two ratios of expected values $E(MS_{C:T})/E(MS_{S:T})$ and $E(MS_T)/E(MS_{S:T})$ are equivalent and greater than one, given positive dependence. If, however, the F is rescaled by a

constant $1/\eta$, where η equals $E(MS_{C:T})/E(MS_{S:T})$, the two ratios of expected values given above equal one, from which it follows that the rescaled F, F_o, will have a central F distribution. This rescaled F equals $(MS_T/\eta)/MS_{S:T}$. The actual alpha level then equals the $P(F_o > d/\eta)$, where d equals the critical F value at any nominal alpha, given (t-1) and (sc-1)t degrees of freedom.

The estimated Type I errors, given positive dependence, obtained using the rescaled F statistic are reported at the bottom of Table B-1 in Appendix B. These estimated values closely match those found in the Monte Carlo study (Table 8). (Ninety percent of the matched alpha values from the two empirical analyses were within 1.96 standard errors of each other.) Because the empirical alpha levels of both techniques were similar both in absolute value and trend, the remaining analysis of the empirical effects on the alpha level of increasing s, increasing c, and increasing degree of positive dependence will be discussed and illustrated using only the simulated or Monte Carlo data found in Table 8.

As c and degree of positive dependence were held constant and the actual alpha levels were averaged across the five nominal alpha levels, an increase in s was directly related to an increase in liberalness. For example, at $E(MS_{C:T})/E(MS_{S:CT})$ equal to 2 and c equal to 5, averaging across the five nominal alpha levels gave mean observed alpha levels equal to .159 for s = 5, .178 for s = 12, and .198 for s = 20 (Table 8). This direct relationship between liberalness and increasing s, given positive dependence, occurs because as s is increased the discrepancy between the $E(MS_{C:T})$ and the $E(MS_{S:T})$ is increased

(Table 10). For example, given $E(MS_{C:T})/E(MS_{S:CT})$ equal to 2 and c equal to 5, this observed discrepancy goes from .81 to .94 to .97 as s equals 5, 12, and 20, respectively. On the other hand, as s and the degree of positive dependence were held constant and the actual alpha levels were averaged across the nominal alpha values, an increase in c was indirectly related to the degree of liberalness. For example, at $E(MS_{C:T})/E(MS_{S:CT})$ equal to 2 and s = 12, the mean observed alpha levels equalled .182 for c = 2, .178 for c = 5, and .173 for c = 10. This indirect relationship occurs because as c is increased the discrepancy between the $E(MS_{C:T})$ and the $E(MS_{S:T})$ decreases. For example, given $E(MS_{C:T})/E(MS_{S:CT})$ equal to 2 and s equal to 12, this observed discrepancy equals .96 for c = 2, .94 for c = 5, and .92 for c = 10.

Table 8 also shows that the degree of liberalness is monotonically related to the degree of positive dependence. That is, as $E(MS_{C:T})/E(MS_{S:CT})$ increased from 2 to 3, the liberalness of the F test using student as the unit of analysis also increased. At $E(MS_{C:T})/E(MS_{S:CT})$ equal to 2 averaging across the five combinations of s and c gave mean observed alpha levels equal to .056 for α = .01, .096 for α = .025, .142 for α = .05, .218 for α = .10, and .378 for α = .25. These five mean values are all smaller than their matches, given $E(MS_{C:T})/E(MS_{S:CT})$ equal to 3, which, respectively, equal .105, .155, .215, .295, and .453.

Because of the general liberalness of the test $F = MS_T/MS_{S:T}$, under all observed cases of s, c, and degree of positive dependence, the powers shown at the bottom of Table 9 will not be discussed as these powers are all spuriously large.

Table 10

Discrepancy Between Observed and Theoretical $E(MS_{C:T})$ and $E(MS_{S:T})$

		$E(MS_{C:T})/E(MS_{S:CT})$				
c	s	.33	.50	1	2	3
2	12	-.638[a]	-.478	.000	.957	1.913
		(-.638)[b]	(-.478)	(.000)	(.957)	(1.913)
5	5	-.560	-.423	-.013	.808	1.629
		(-.556)	(-.417)	(.000)	(.833)	(1.667)
5	12	-.634	-.476	-.004	.939	1.883
		(-.622)	(-.466)	(.000)	(.932)	(1.864)
5	20	-.645	-.483	.002	.974	1.946
		(-.640)	(-.480)	(.000)	(.960)	(1.919)
10	12	-.624	-.470	-.008	.915	1.839
		(-.616)	(-.462)	(.000)	(.924)	(1.849)

[a]Observed $E(MS_{C:T})$ minus Observed $E(MS_{S:T})$.

[b]Theoretical $E(MS_{C:T})$ minus Theoretical $E(MS_{S:T})$.

Negative Dependence

The effects of negative dependence between disaggregate units on the actual significance levels of the test $F = MS_T/MS_{S:T}$, estimated using Monte Carlo procedures, are reported at the top of Table 8. None of the empirical alpha levels were within 1.96 standard errors of the nominal alpha levels. All empirical alpha levels were smaller than their nominal counterparts, which means the test statistics were too conservative. This indicates that, given that σ^2_C was less than $\sigma^2_{S:CT}/s$ and there were no treatment effects, $F = MS_T/MS_{S:T}$ was not distributed as a central F but instead had an F distribution which was located to the left of the central F distribution found when given the same F statistic and independence of individual units. This finding concurs with the analytic work of Scheffé (1959) and Cochran (1947) and also with the analytic work presented in Chapter IV.

The rescaled F statistic was also used to estimate the magnitude of effects given negative dependence and prespecified parameters. The results of this analysis are reported at the top of Table B-1 in Appendix B. Once again the estimated alpha values in Table B-1 closely match, both in absolute value and trend, those reported in Table 8. (Ninety-eight percent of the matched alpha levels from the two empirical analyses were within 1.96 standard errors of each other.) Because of this, the effects on the alpha level of increasing s, increasing c, and increasing the level of negative dependence will be discussed and illustrated using only the simulated data found in Table 8.

The theoretical and observed discrepancies (Table 10) between the $E(MS_{C:T})$ and the $E(MS_{S:T})$ for each level of negative dependence indicate that as s increases $F = MS_T/MS_{S:T}$ should become more conservative because the $E(MS_{S:T})$ becomes increasingly larger than the $E(MS_{C:T})$ as s increases. Table 10 indicates that the opposite should occur as c is increased. Neither one of these two expectations appeared in the simulated data. It may have been that the observed alpha values were just too close to zero and the three different levels of number of students and classes were just not different enough to bring out the expected trends.

Table 8 also shows that the degree of conservativeness is monotonically related to degree of negative dependence. For example, at $E(MS_{C:T})/E(MS_{S:CT})$ equal to .5, averaging across the five combinations of s and c gave mean observed alpha levels equal to .000 for $\alpha = .01$, .003 for $\alpha = .025$, .008 for $\alpha = .05$, .022 for $\alpha = .10$, and .110 for $\alpha = .25$; while at $E(MS_{C:T})/E(MS_{S:CT})$ equal to .33, averaging across the combinations of s and c gave mean observed alpha levels equal to .000 for $\alpha = .01$, .000 for $\alpha = .025$, .001 for $\alpha = .05$, .007 for $\alpha = .10$, and .054 for $\alpha = .25$.

The empirical powers for $F = MS_T/MS_{S:T}$ for the two simulated negative dependence conditions are shown at the top of Table 9. As expected, the estimated power values increased both as the number of students per class increased and as the number of classes per treatment increased. For example, given the least degree of negative dependence, i.e., $E(MS_{C:T})/E(MS_{S:CT})$ equal to .5, averaging across the five nominal

alpha levels and keeping c constant at 5 gave mean power values equalling .304 for s = 5, .635 for s = 12, and .859 for s = 20. Given that same degree of negative dependence, averaging across the nominal alpha levels and keeping s constant at 12 gave mean power values equal to .284 for c = 2, .635 for c = 5, and .928 for c = 10.

By itself a conservative test statistic should have spuriously less power. Thus the power of the test $F = MS_T/MS_{S:T}$ should be reduced as the negative dependence is increased from $E(MS_{C:T})/E(MS_{S:CT})$ equals .5 to $E(MS_{C:T})/E(MS_{S:CT})$ equals .33. For small degrees of freedom error, i.e., (sc-1)t equals 46 and 48, increasing negative dependence did decrease the power of $F = MS_T/MS_{S:T}$. However, for large degrees of freedom error, across most nominal alpha levels the reverse occurred. Of greatest significance to the practitioner, however, $F = MS_T/MS_{S:T}$ had, in all cases but one, ($E[MS_{C:T}]/E[MS_{S:CT}] = .5$, c = 2, s = 12, and $\alpha = .025$), less power than the $F = MS_T/MS_{C:T}$ test. Thus, in this simulation situation increasing the degrees of freedom error for the F statistic by using student, rather than class, as the unit of analysis did not compensate for the fact that using student made the test of no treatment effects too conservative a test.

CHAPTER VII

THE CONDITIONAL F TEST: EMPIRICAL ANALYSIS

The most desirable situation in testing hypotheses is, of course, both a small probability of a Type I error (α) and a small probability of a Type II error (β). Table 7 of the preceding chapter showed that the probability of a Type II error, given the always correct $F = MS_T/MS_{C:T}$ test, the most commonly used alpha level of .05 and the operational definition of independence, was relatively high for four of the five simulated combinations of s and c. For c = 2 and s = 12, β equalled .847; for c = 5 and s = 5, β equalled .773; for c = 5 and s = 12, β equalled .522; for c = 5 and s = 20, β equalled .301, and for c = 10 and s = 12, β equalled .148. Clearly, it would be nice to improve on these rather high probabilities, if possible, without increasing the probability of Type I errors. It was shown earlier, both analytically and empirically (Table 9), that using F tests with disaggregate units as the units of analysis, $F = MS_T/MS_{S:T}$, would reduce the probability of Type II errors, given independence, by increasing the degrees of freedom error. However, if the individual data values were positively dependent upon each other, which appears to be quite common in ordinary classroom situations, then using the test $F = MS_T/MS_{S:T}$ increased the probability of Type I errors. On the other hand, if observations within groups were negatively dependent, using the test

$F = MS_T/MS_{S:T}$ decreased the probability of Type I errors but at the same time the natural increase in degrees of freedom error was not enough to offset the increase of the probability of Type II errors caused by this spurious decrease in the probability of Type I errors. All in all, given simulated conditions common to educational data, it seemed "best" to use classroom as the unit of analysis given dependence (either positive or negative) between student responses and student as the unit of analysis when student responses were independent of each other. Herein lies the motivation for performing a preliminary test of independence. That is, the sole purpose of this preliminary test is to choose the appropriate unit of analysis for the primary test of treatment effects.

The problem with using this operational test of independence, $F = MS_{C:T}/MS_{S:CT}$, to select a unit of analysis for the primary test is that this procedure makes the primary test of no treatment effects have a conditional F distribution. Of interest then is the difference between the distribution of a conditional F test statistic (also called the "sometimes pool" test statistic) and the distribution of the appropriate unconditional and always correct F statistic from the "never pool" model, $F = MS_T/MS_{C:T}$. Variables which were examined to see how they affected this difference included the number of students per class, the number of classes per treatment, the type and degree of dependence, the alpha level of the preliminary test, and the alpha level of the primary test. The effects of each one of the above mentioned variables on the distributional properties of the conditional F test were empirically studied within the content of three different preliminary F

tests. The three preliminary F tests included: (a) a two-tailed preliminary F test, (b) the usual, upper-tailed only preliminary F test, and (c) a lower-tailed only preliminary F test.

In order to claim this two-stage testing procedure a success, the observed alpha level of the conditional F test should be close to the nominal alpha level at which the researcher thinks he is working and the procedure should have greater power than the always correct, unconditional test $F = MS_T/MS_{C:T}$. Simulated data, identical to those used for looking at the effects of correlated units of analysis, were used to examine both empirical probabilities of Type I errors and empirical powers of the conditional F tests. Based on the results of one of the preliminary tests, either Model A (Table 2), $F = MS_T/MS_{C:T}$, or Model B (Table 3), $F = MS_T/MS_{S:T}$, was designated as the appropriate model to use in testing the primary hypothesis of no treatment effects. The actual alpha level for the conditional F test was defined by $(n_A\alpha_A + n_B\alpha_B)/(n_A + n_B)$, where n_A and n_B equalled the number of preliminary F tests rejected and not rejected, respectively, and α_A and α_B equalled the actual alpha levels for the primary tests of no treatment effects analyzed by Models A and B, respectively.

The Two-Tailed Preliminary Test

The two-tailed preliminary F test tested the hypothesis that the $E(MS_{C:T})$ equalled the $E(MS_{S:CT})$, or equivalently that ρI equalled zero. The effects of the two-tailed preliminary test were examined at five different preliminary test alpha levels (i.e., .02, .05, .10, .20, and

.50). Actual conditional test alpha levels, given the two-tailed preliminary test, are shown in Tables 11 through 15. Corresponding differences between empirical powers of the conditional test and the unconditional, always correct test $F = MS_T/MS_{C:T}$ are shown in Tables 16 through 20. Appendix C (Tables C-1 through C-5) contains the actual statistical powers of the conditional F test, given the two-tailed preliminary test. Each separate table describes the effect on the conditional F test alpha level or power of varying the type and degree of dependence, the two-tailed preliminary test alpha level and the conditional test alpha level for one specific combination of s and c. Examining the effects of s and c requires between table comparisons. In this study each combination of s and c will be referred to as a "design." Thus, this study includes five designs, c = 2 and s = 12, c = 5 and s = 5, c = 5 and s = 12, c = 5 and s = 20, and c = 10 and s = 12.

Independence

Independence is that condition where the variance of the aggregate units is predictable given the variance of the disaggregate units and the grouping size. Operationally speaking, within the context of this study, independence occurs whenever the ratio of $E(MS_{C:T})$ over $E(MS_{S:CT})$ equals 1. Given this situation, ideally the two-tailed preliminary test should not reject its null hypothesis, H_o: $E(MS_{C:T})$ equals $E(MS_{S:CT})$, designating the disaggregate unit (students) as the appropriate unit of analysis in testing for treatment effects.

Table 11

Actual Alphas of the Conditional F Test Given a Two-Tailed Preliminary Test, c = 2 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean alpha
.33	.02	.007[a]	.015	.017	.025	.081	.029
	.05	.010[a]	.024[a]	.033	.051	.114	.046
	.10	.010[a]	.029[a]	.044[a]	.070	.151	.061
	.20	.010[a]	.029[a]	.048[a]	.085[a]	.200	.074
	.50	.010[a]	.029[a]	.050[a]	.093[a]	.257[a]	.088
	Mean alpha	.009	.025	.038	.065	.161	
.50	.02	.008[a]	.014	.023	.044	.116	.041
	.05	.011[a]	.023[a]	.035	.060	.139	.054
	.10	.011[a]	.026[a]	.045[a]	.075	.161	.064
	.20	.011[a]	.029[a]	.052[a]	.091[a]	.200	.077
	.50	.010[a]	.028[a]	.052[a]	.098[a]	.254[a]	.088
	Mean alpha	.010	.024	.041	.074	.174	
1	.02	.016[a]	.042	.070	.102[a]	.251[a]	.096
	.05	.018	.045	.072	.104[a]	.250[a]	.098
	.10	.019	.048	.078	.113[a]	.260[a]	.104
	.20	.019	.048	.085	.122	.268[a]	.108
	.50	.016[a]	.042	.079	.126	.289	.110
	Mean alpha	.018	.045	.077	.113	.264	
2	.02	.065	.090	.138	.210	.357	.172
	.05	.066	.093	.138	.201	.340	.168
	.10	.066	.091	.134	.193	.325	.162
	.20	.055	.082	.123	.181	.309	.150
	.50	.035	.062	.098	.152	.280	.125
	Mean alpha	.057	.084	.126	.187	.322	
3	.02	.093	.144	.188	.261	.383	.214
	.05	.085	.136	.178	.244	.357	.200
	.10	.079	.125	.161	.219	.329	.183
	.20	.067	.108	.139	.196	.306	.163
	.50	.044	.070	.100	.143	.274	.126
	Mean alpha	.074	.117	.153	.213	.330	

[a]Actual alpha is within 1.96 standard errors of the nominal alpha.

Table 12

Actual Alphas of the Conditional F Test Given a Two-Tailed Preliminary Test, c = 5 and s = 5

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha		Conditional test nominal alpha .010	.025	.050	.100	.250	Mean alpha
	.33	.02	.008[a]	.019[a]	.044[a]	.070	.151	.058
		.05	.009[a]	.020[a]	.049[a]	.076	.192	.069
		.10	.010[a]	.022[a]	.055[a]	.098[a]	.209	.079
		.20	.011[a]	.024[a]	.059[a]	.107[a]	.230[a]	.086
		.50	.011[a]	.025[a]	.061[a]	.108[a]	.245[a]	.090
	Mean alpha		.010	.022	.054	.092	.205	
	.50	.02	.006[a]	.018[a]	.032	.053	.152	.052
		.05	.008[a]	.024[a]	.045[a]	.070	.180	.065
		.10	.008[a]	.026[a]	.053[a]	.084[a]	.201	.074
		.20	.010[a]	.026[a]	.057[a]	.098[a]	.224[a]	.083
		.50	.011[a]	.026[a]	.062[a]	.108[a]	.241[a]	.090
	Mean alpha		.009	.024	.050	.083	.200	
	1	.02	.013[a]	.024[a]	.053[a]	.100[a]	.235[a]	.085
		.05	.016[a]	.029[a]	.058[a]	.106[a]	.234[a]	.089
		.10	.016[a]	.031[a]	.060[a]	.108[a]	.237[a]	.090
		.20	.018	.030[a]	.064	.118[a]	.244[a]	.095
		.50	.014[a]	.031[a]	.069	.118[a]	.247[a]	.096
	Mean alpha		.015	.029	.061	.110	.239	
	2	.02	.039	.073	.112	.171	.308	.141
		.05	.036	.067	.104	.155	.289	.130
		.10	.033	.060	.095	.144	.278	.122
		.20	.026	.054	.091	.133	.263[a]	.113
		.50	.020	.040	.074	.119	.248[a]	.110
	Mean alpha		.031	.059	.095	.144	.277	
	3	.02	.056	.082	.119	.170	.319	.149
		.05	.047	.073	.109	.159	.299	.137
		.10	.038	.059	.095	.147	.285	.125
		.20	.031	.046	.079	.127	.274[a]	.111
		.50	.019	.033	.071	.115[a]	.253[a]	.098
	Mean alpha		.038	.059	.095	.144	.286	

[a]Actual alpha is within 1.96 standard errors of the nominal alpha.

Table 13

Actual Alphas of the Conditional F Test Given a Two-Tailed Preliminary Test, c = 5 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha		Conditional test nominal alpha .010	.025	.050	.100	.250	Mean alpha
	.33	.02	.006[a]	.015	.030	.049	.120	.044
		.05	.006[a]	.015	.036	.071	.155	.057
		.10	.007[a]	.017[a]	.040[a]	.080	.190	.067
		.20	.007[a]	.018[a]	.045[a]	.091[a]	.212	.075
		.50	.007[a]	.018[a]	.046[a]	.095[a]	.224[a]	.078
	Mean alpha		.007	.017	.039	.077	.180	
	.50	.02	.003	.008	.021	.042	.136	.042
		.05	.005[a]	.016[a]	.032	.058	.157	.054
		.10	.006[a]	.017[a]	.036	.070	.175	.061
		.20	.006[a]	.017[a]	.042[a]	.086[a]	.198	.070
		.50	.007[a]	.018[a]	.046[a]	.096[a]	.222	.078
	Mean alpha		.005	.015	.035	.070	.178	
	1	.02	.009[a]	.021[a]	.047[a]	.095[a]	.235[a]	.081
		.05	.009[a]	.021[a]	.047[a]	.099[a]	.235[a]	.082
		.10	.010[a]	.023[a]	.054[a]	.106[a]	.235[a]	.086
		.20	.011[a]	.025[a]	.057[a]	.110[a]	.237[a]	.088
		.50	.011[a]	.027[a]	.056[a]	.117[a]	.234[a]	.089
	Mean alpha		.010	.023	.052	.105	.235	
	2	.02	.040	.081	.111	.167	.323	.144
		.05	.033	.067	.095	.148	.298	.128
		.10	.028	.058	.084	.132	.274[a]	.115
		.20	.024	.050	.074	.115[a]	.253[a]	.103
		.50	.014[a]	.029[a]	.055[a]	.101[a]	.237[a]	.087
	Mean alpha		.057	.057	.084	.133	.277	
	3	.02	.055	.076	.108	.157	.298	.139
		.05	.042	.056	.085	.132	.280	.119
		.10	.032	.044	.072	.120	.265[a]	.107
		.20	.023	.030[a]	.058[a]	.109[a]	.244[a]	.093
		.50	.012[a]	.023[a]	.053[a]	.099[a]	.231[a]	.084
	Mean alpha		.033	.046	.061	.123	.264	

[a]Actual alpha is within 1.96 standard errors of the nominal alpha.

Table 14

Actual Alphas of the Conditional F Test Given a Two-Tailed Preliminary Test, c = 5 and s = 20

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean alpha
.33	.02	.006[a]	.019[a]	.030	.051	.126	.046
	.05	.007[a]	.024[a]	.040[a]	.077	.168	.063
	.10	.007[a]	.026[a]	.045[a]	.089[a]	.209	.075
	.20	.007[a]	.028[a]	.049[a]	.097[a]	.233[a]	.083
	.50	.007[a]	.028[a]	.049[a]	.100[a]	.248[a]	.086
	Mean alpha	.007	.025	.043	.083	.197	
.50	.02	.005[a]	.012	.020	.037	.143	.043
	.05	.005[a]	.018[a]	.029	.054	.168	.055
	.10	.007[a]	.023[a]	.039[a]	.069	.192	.066
	.20	.007[a]	.025[a]	.044[a]	.089[a]	.211	.075
	.50	.007[a]	.028[a]	.049[a]	.100[a]	.245[a]	.086
	Mean alpha	.006	.021	.036	.070	.192	
1	.02	.014[a]	.025[a]	.050[a]	.110[a]	.260[a]	.092
	.05	.016[a]	.028[a]	.054[a]	.111[a]	.260[a]	.094
	.10	.016[a]	.030[a]	.056[a]	.112[a]	.262[a]	.095
	.20	.016[a]	.031[a]	.054[a]	.110[a]	.260[a]	.094
	.50	.015[a]	.035	.057[a]	.107[a]	.259[a]	.095
	Mean alpha	.015	.030	.054	.110	.260	
2	.02	.052	.092	.120	.193	.328	.157
	.05	.048	.085	.111	.175	.305	.145
	.10	.040	.075	.102	.157	.290	.133
	.20	.032	.063	.084	.129	.274[a]	.116
	.50	.021	.039	.060[a]	.112[a]	.252[a]	.097
	Mean alpha	.039	.071	.095	.153	.290	
3	.02	.064	.089	.123	.170	.303	.150
	.05	.050	.069	.098	.144	.286	.129
	.10	.032	.052	.082	.136	.266[a]	.114
	.20	.024	.043	.067	.118[a]	.257[a]	.102
	.50	.010[a]	.031[a]	.054[a]	.103[a]	.250[a]	.090
	Mean alpha	.036	.057	.085	.134	.272	

[a]Actual alpha is within 1.96 standard errors of the nominal alpha.

Table 15

Actual Alphas of the Conditional F Test Given a Two-Tailed Preliminary Test, c = 10 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean alpha
.33	.02	.006[a]	.016[a]	.044[a]	.078	.200	.069
	.05	.006[a]	.017[a]	.046[a]	.088[a]	.217	.075
	.10	.006[a]	.017[a]	.048[a]	.092[a]	.225[a]	.078
	.20	.006[a]	.017[a]	.049[a]	.094[a]	.230[a]	.079
	.50	.006[a]	.017[a]	.049[a]	.094[a]	.231[a]	.079
	Mean alpha	.006	.017	.047	.089	.221	
.50	.02	.003	.009	.023	.049	.160	.049
	.05	.006[a]	.014	.037[a]	.070	.191	.064
	.10	.006[a]	.016[a]	.044[a]	.079	.210	.071
	.20	.006[a]	.017[a]	.047[a]	.085[a]	.225[a]	.076
	.50	.006[a]	.017[a]	.049[a]	.094[a]	.230[a]	.079
	Mean alpha	.005	.015	.040	.075	.203	
1	.02	.007[a]	.023[a]	.054[a]	.099[a]	.237[a]	.084
	.05	.007[a]	.023[a]	.054[a]	.099[a]	.238[a]	.084
	.10	.007[a]	.023[a]	.054[a]	.098[a]	.238[a]	.084
	.20	.006[a]	.021[a]	.056[a]	.097[a]	.235[a]	.083
	.50	.006[a]	.022[a]	.053[a]	.095[a]	.232[a]	.082
	Mean alpha	.007	.022	.054	.098	.236	
2	.02	.032	.053	.087	.152	.277	.120
	.05	.024	.043	.071	.132	.262[a]	.106
	.10	.015[a]	.033[a]	.063[a]	.117[a]	.254[a]	.096
	.20	.011[a]	.026[a]	.054[a]	.104[a]	.240[a]	.087
	.50	.007[a]	.021[a]	.051[a]	.096[a]	.235[a]	.082
	Mean alpha	.018	.035	.065	.120	.254	
3	.02	.019	.031[a]	.064	.116[a]	.248[a]	.096
	.05	.014[a]	.026[a]	.056[a]	.102[a]	.243[a]	.088
	.10	.011[a]	.021[a]	.053[a]	.099[a]	.241[a]	.085
	.20	.007[a]	.018[a]	.050[a]	.097[a]	.235[a]	.081
	.50	.006[a]	.017[a]	.049[a]	.094[a]	.231[a]	.079
	Mean alpha	.011	.023	.054	.102	.240	

[a]Actual alpha is within 1.96 standard errors of the nominal alpha.

Actual alpha levels. It was expected that, given independence of student data and no treatment effects, the empirical and nominal alpha levels for all the conditional F tests would be equal. Generally the simulated data verified this expectation. Excluding all situations where c equalled 2 and s equalled 12, 96% of the remaining 100 observed alpha levels (Tables 12 through 15), given $E(MS_{C:T})/E(MS_{S:CT}) = 1$, were within 1.96 standard errors of the nominal alpha levels. However, given all situations where c equalled 2 and s equalled 12 (Table 11), only 9 of the 25 observed alpha values (36%) were within 1.96 standard errors of the nominal alpha levels. The remaining 16 observed alpha levels were too liberal. That is, their probabilities of a Type I error were consistently too large. These 16 liberal observed alpha levels were concentrated at the lower conditional test nominal alpha levels (i.e., .01, .025, and .05). Given independence, the actual alpha levels of the conditional F tests, averaged across the five preliminary test alpha levels and the four designs c = 5 and s = 5, c = 5 and s = 12, c = 5 and s = 20, and c = 10 and s = 12, increased from .012 to .026 to .055 as the nominal alpha levels increased from .01 to .025 to .05. At those same three conditional test nominal alpha levels, however, the actual alpha levels of the conditional tests for c = 2 and s = 12, averaged across the five preliminary test alpha levels, increased from .018 to .045 to .077. Because there seemed to be no reasonable explanation for the liberalness that dominated when c equalled 2 and s equalled 12, a second simulation run was done for that particular design. The results of this run deviated even more from the expected,

given independence, as 22 of the 25 (88%) actual alpha values were too liberal.

Estimated powers. Given that $E(MS_{C:T})/E(MS_{S:CT})$ equalled one, the statistical powers of the conditional F tests were, almost without exception, greater than the powers of their respective "never pool," unconditional $F = MS_T/MS_{C:T}$ tests (Tables 16 through 20). Across the five designs and the five preliminary alpha levels, as the five nominal alpha levels increased from .01 to .25, the average difference between the conditional test powers and the "never pool" test powers decreased from .102 to .091 to .076 to .057 to .021. Comparing the estimated power values of the conditional F tests (Appendix C) with comparable power values of the unconditional $F = MS_T/MS_{C:T}$ tests (Table 7) shows that this decrease in discrepancy is probably due to the fact that the average powers of the "never pool" tests are rather high given an alpha level of .25 and thus it is harder for the "sometimes pool" tests to improve on that already "high" power. This is especially so given the two designs c = 5 and s = 20 and c = 10 and s = 12. While this negative relationship held up across the five designs or combinations of s and c, it did not hold up within each combination of s and c. Consider the design c = 2 and s = 12 (Table 16). Averaged across the five preliminary test alpha levels, the observed power differences for this one design equalled .085, .115, .142, .146, and .060 as their respective nominal alphas increased from .01 to .25.

Table 16

Power of the Conditional F Test Minus Power of the Test $F = MS_T/MS_{C:T}$ Given a Two-Tailed Preliminary Test, c = 2 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean power dif.[a]
.33	.02	-.018	-.034	-.097	-.137	-.099	-.077
	.05	.014	.006	-.057	-.108	-.087	-.046
	.10	.019	.034	-.012	-.065	-.060	-.017
	.20	.020	.055	.055	.006	-.027	.022
	.50	.013	.039	.067	.075	.016	.042
	Mean power dif.[a]	.010	.020	-.009	-.046	-.051	
.50	.02	.019	.019	.011	.001	-.046	.001
	.05	.036	.045	.033	.019	-.036	.019
	.10	.046	.069	.060	.039	-.024	.038
	.20	.046	.089	.094	.088	.006	.065
	.50	.028	.064	.093	.114	.021	.082
	Mean power dif.	.035	.057	.058	.052	-.016	
1	.02	.089	.115	.143	.155	.077	.116
	.05	.090	.116	.142	.152	.068	.114
	.10	.092	.122	.150	.156	.065	.117
	.20	.090	.127	.154	.156	.051	.116
	.50	.062	.093	.120	.113	.038	.085
	Mean power dif.	.085	.115	.142	.146	.060	
2	.02	.152	.191	.223	.218	.159	.189
	.05	.145	.180	.210	.202	.139	.175
	.10	.134	.166	.191	.175	.110	.155
	.20	.114	.149	.167	.147	.082	.132
	.50	.068	.088	.101	.070	.029	.071
	Mean power dif.	.123	.155	.178	.162	.104	
3	.02	.171	.221	.250	.218	.166	.205
	.05	.154	.196	.219	.183	.136	.178
	.10	.137	.175	.194	.157	.108	.154
	.20	.114	.144	.146	.105	.058	.113
	.50	.059	.078	.073	.043	.028	.056
	Mean power dif.	.127	.163	.176	.141	.099	

[a]Mean power differences.

Table 17

Power of the Conditional F Test Minus Power of the Test $F = MS_T/MS_{C:T}$ Given a Two-Tailed Preliminary Test, c = 5 and s = 5

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean power dif.[a]
.33	.02	-.101	-.177	-.239	-.228	-.099	-.169
	.05	-.045	-.094	-.150	-.153	-.068	-.102
	.10	-.016	-.044	-.083	-.092	-.054	-.058
	.20	-.004	-.018	-.034	-.042	-.025	-.025
	.50	.000	.004	.002	.001	-.003	.001
	Mean power dif.[a]	-.033	-.066	-.101	-.103	-.050	
.50	.02	-.060	-.096	-.141	-.141	-.092	-.106
	.05	-.031	-.057	-.099	-.104	-.074	-.073
	.10	-.013	-.025	-.060	-.072	-.052	-.044
	.20	.005	.008	-.012	-.026	-.032	-.011
	.50	.012	.022	.011	.008	-.002	.010
	Mean power dif.	-.017	-.030	-.060	-.067	-.050	
1	.02	.029	.041	.051	.044	.022	.037
	.05	.030	.043	.052	.042	.024	.038
	.10	.032	.043	.051	.040	.028	.039
	.20	.039	.041	.043	.035	.022	.036
	.50	.025	.025	.025	.024	.007	.021
	Mean power dif.	.031	.039	.044	.037	.021	
2	.02	.090	.101	.106	.108	.102	.101
	.05	.076	.078	.074	.077	.075	.076
	.10	.066	.071	.066	.062	.057	.064
	.20	.053	.050	.043	.044	.039	.046
	.50	.024	.022	.017	.020	.008	.018
	Mean power dif.	.062	.064	.061	.062	.056	
3	.02	.083	.083	.101	.093	.082	.088
	.05	.069	.063	.080	.072	.055	.068
	.10	.050	.046	.060	.052	.036	.049
	.20	.033	.032	.038	.022	.025	.030
	.50	.019	.011	.014	.003	.006	.011
	Mean power dif.	.051	.047	.059	.048	.041	

[a]Mean power differences.

Table 18

Power of the Conditional F Test Minus Power of the Test $F = MS_T/MS_{C:T}$ Given a Two-Tailed Preliminary Test, c = 5 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean power dif.[a]
.33	.02	-.218	-.240	-.150	-.079	-.026	-.143
	.05	-.135	-.158	-.099	-.052	-.015	-.092
	.10	-.050	-.086	-.061	-.033	-.011	-.048
	.20	-.003	-.026	-.023	-.011	-.005	-.014
	.50	.018	.007	.000	-.002	.000	.005
	Mean power dif.[a]	-.078	-.101	-.067	-.035	-.011	
.50	.02	-.085	-.127	-.120	-.073	-.028	-.087
	.05	-.051	-.090	-.083	-.056	-.021	-.060
	.10	-.002	-.046	-.049	-.039	-.015	-.030
	.20	.033	-.003	-.010	-.020	-.006	-.001
	.50	.045	.041	.018	.003	.002	.022
	Mean power dif.	-.012	-.045	-.049	-.037	-.014	
1	.02	.148	.141	.105	.062	.026	.096
	.05	.140	.130	.096	.058	.025	.090
	.10	.141	.130	.092	.054	.018	.087
	.20	.135	.119	.082	.044	.016	.079
	.50	.090	.073	.052	.032	.006	.051
	Mean power dif.	.131	.119	.085	.050	.018	
2	.02	.192	.180	.161	.130	.078	.148
	.05	.153	.142	.122	.100	.050	.113
	.10	.119	.105	.090	.081	.031	.085
	.20	.081	.067	.045	.045	.016	.051
	.50	.034	.030	.016	.018	.004	.020
	Mean power dif.	.116	.105	.087	.075	.036	
3	.02	.114	.107	.100	.081	.061	.093
	.05	.075	.074	.071	.056	.040	.063
	.10	.056	.057	.054	.037	.027	.046
	.20	.030	.032	.031	.016	.013	.024
	.50	.014	.010	.010	.006	.004	.009
	Mean power dif.	.058	.056	.053	.039	.029	

[a]Mean power differences.

Table 19

Power of the Conditional F Test Minus Power of the Test $F = MS_T/MS_{C:T}$ Given a Two-Tailed Preliminary Test, $c = 5$ and $s = 20$

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean power dif.[a]
.33	.02	-.143	-.072	-.043	-.016	-.001	-.055
	.05	-.090	-.049	-.029	-.012	-.001	-.036
	.10	-.041	-.027	-.015	-.005	.000	-.018
	.20	-.004	-.009	-.008	-.001	.001	-.004
	.50	.020	.004	.002	.000	.001	.005
	Mean power dif.[a]	-.052	-.031	-.019	-.007	.000	
.50	.02	-.050	-.034	-.035	-.028	-.006	-.031
	.05	-.027	-.016	-.023	-.023	-.006	-.019
	.10	.003	.002	-.011	-.014	-.005	-.005
	.20	.039	.025	.001	-.007	-.004	.011
	.50	.059	.037	.013	.000	.000	.022
	Mean power dif.	.005	.003	-.011	-.014	-.004	
1	.02	.218	.173	.106	.056	.007	.112
	.05	.207	.161	.095	.053	.007	.105
	.10	.195	.152	.087	.049	.005	.098
	.20	.178	.134	.083	.040	.001	.087
	.50	.113	.080	.041	.015	-.001	.050
	Mean power dif.	.182	.140	.082	.043	.004	
2	.02	.259	.235	.173	.116	.047	.166
	.05	.209	.187	.132	.087	.033	.130
	.10	.168	.141	.088	.058	.020	.095
	.20	.115	.095	.059	.035	.011	.063
	.50	.043	.037	.018	.006	.004	.022
	Mean power dif.	.159	.139	.094	.060	.023	
3	.02	.158	.137	.123	.073	.039	.106
	.05	.101	.086	.080	.043	.026	.067
	.10	.073	.061	.058	.026	.015	.047
	.20	.049	.031	.031	.012	.008	.026
	.50	.014	.008	.011	.002	.003	.008
	Mean power dif.	.079	.065	.061	.031	.018	

[a]Mean power differences.

Table 20

Power of the Conditional F Test Minus Power of the Test $F = MS_T/MS_{C:T}$ Given a Two-Tailed Preliminary Test, c = 10 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean power dif.[a]
.33	.02	-.033	-.015	-.007	-.001	.000	-.011
	.05	-.014	-.005	-.003	-.001	.000	-.005
	.10	-.004	-.003	-.002	-.001	.000	-.002
	.20	-.001	.000	.000	.000	.000	.000
	.50	.001	.000	.000	.000	.000	.000
	Mean power dif.	-.004	-.005	-.002	-.001	.000	
.50	.02	-.070	-.043	-.021	-.011	-.002	-.029
	.05	-.047	-.025	-.013	-.008	.000	-.019
	.10	-.025	-.015	-.007	-.006	.000	-.011
	.20	-.009	-.003	-.002	-.004	.000	-.004
	.50	.005	.000	.000	.000	.000	.001
	Mean power dif.	-.029	-.017	-.017	-.006	.000	
1	.02	.100	.054	.031	.015	.003	.041
	.05	.095	.051	.028	.013	.003	.038
	.10	.092	.048	.028	.011	.002	.036
	.20	.079	.038	.023	.011	.001	.030
	.50	.047	.021	.010	.006	-.001	.017
	Mean power dif.	.083	.042	.024	.011	.002	
2	.02	.165	.131	.090	.047	.019	.090
	.05	.109	.081	.053	.026	.013	.056
	.10	.070	.046	.033	.017	.008	.035
	.20	.044	.021	.015	.009	.002	.018
	.50	.013	.004	.004	.001	.000	.004
	Mean power dif.	.080	.057	.039	.020	.008	
3	.02	.040	.035	.020	.020	.006	.024
	.05	.022	.019	.010	.012	.003	.013
	.10	.006	.007	.005	.008	.002	.006
	.20	.003	.004	.003	.004	.001	.003
	.50	.001	.000	.000	.001	.000	.000
	Mean power dif.	.014	.013	.008	.009	.002	

[a]Mean power differences.

As the number of observations per class increased (compare across Tables 17, 18, and 19), the discrepancy between the power of the conditional test and the power of the unconditional $F = MS_T/MS_{C:T}$ test was expected to increase. The rationale for this expectation follows. Given independence of student responses, $E(MS_{C:T})/E(MS_{S:CT}) = 1$, pooling should be prescribed all the time. If there were only one student per class, the power of the conditional test and the power of $F = MS_T/MS_{C:T}$ should be identical. As the number of students increases, however, the power of the conditional test and the power of the test $F = MS_T/MS_{C:T}$ should become more discrepant, with the power of the conditional test being greater as it would have more degrees of freedom error. Basically the simulated data upheld this prediction, especially given the more stringent nominal conditional test alpha levels (.01 and .025). Given a conditional test nominal alpha of .01 and averaging across the five preliminary test alpha levels gave average differences between the power of the conditional F test and the respective power of the test $F = MS_T/MS_{C:T}$ of .031 for $c = 5$ and $s = 5$, .131 for $c = 5$ and $s = 12$, and .182 for $c = 5$ and $s = 20$. As the number of students per class increased, the empirical powers of the unconditional test $F = MS_T/MS_{C:T}$ became rather high (Table 7) given the larger nominal alpha values. Thus it became more difficult to detect this expected difference in powers between the conditional test procedure and the unconditional test $F = MS_T/MS_{C:T}$ as s increased at the higher nominal conditional test alpha levels.

On the other hand, with the exception of the design c = 2 and s = 12 at the smallest alpha levels, as the number of classes increased (compare across Tables 16, 18, and 20), the discrepancies between the powers of the conditional F test and the unconditional test $F = MS_T/MS_{C:T}$ tended to decrease. For example, for the conditional test alpha of .25, averaging the discrepancies across the five preliminary test alpha levels gave power differences of .060 for the design c = 2 and s = 12, .018 for the design c = 5 and s = 12, and .002 for the design c = 10 and s = 12. This trend was expected as "sometimes pooling" should be more advantageous for increasing power than "never pooling" when only a few classrooms per treatment have been sampled. The increased degrees of freedom brought on by pooling has a larger effect when the "never pool" test has relatively few degrees of freedom error than when it has many degrees of freedom error. The powers of the conditional test for the design c = 2 and s = 12 turned out very curiously as it was the one design where the discrepancies between the conditional test power and the $F = MS_T/MS_{C:T}$ test power did not fit the predicted trend as c was varied for nominal conditional test alphas of .01 and .025. At those two alpha levels, the estimated powers of the conditional tests were spuriously high because the actual alpha levels were too liberal. Thus, one would have expected the average discrepancies between powers of the conditional test and powers of the test $F = MS_T/MS_{C:T}$ to also be spuriously large. However, the opposite occurred.

Although the trend was not perfect across the five conditional test nominal alpha levels and across the five combinations of s and c,

the simulated data generally showed the discrepancies between the power of the "sometimes pool," conditional test and the power of the "never pool" test to decrease with an increase in the nominal alpha level of the preliminary test. This too was predictable as the distributions of the "sometimes pool" test and the "never pool" test become more similar as the nominal preliminary test alpha level increases. As the alpha level of the preliminary test increases to .50, the power of the preliminary test, $F = MS_{C:T}/MS_{S:CT}$, increases and thus pooling of degrees of freedom and mean squares between and within classrooms are prescribed less often, which makes the conditional F test more of a "never pool" test. A good example of this indirect relationship between the power difference between the conditional and unconditional test and alpha level of the preliminary test is evident when the nominal alpha level of the conditional test equals .25 and the design is c = 2 and s = 12 (Table 16). Given these three prespecified parameters, the discrepancies between the power of the conditional test and the $F = MS_T/MS_{C:T}$ test equal .077, .068, .065, .051, and .038 for preliminary test alpha levels of .02, .05, .10, .20, and .50.

Given independence of individual responses within and between groups, one might also wonder how the power of the conditional, "sometimes pool" test compared to the power of the "always pool" $F = MS_T/MS_{S:T}$ test. These two powers can be compared by looking at the $E(MS_{C:T})/E(MS_{S:CT}) = 1$ sections in Appendix C (Tables C-1 through C-5) and Table 9. One would expect the power of the "always pool" test to always exceed the power of the "sometimes pool" test. While that was

usually the case, it was not always the case. For example, given the design c = 2 and s = 12 (Table C-1), 17 of the 25 (68%) conditional test powers exceeded the "always pool" $F = MS_T/MS_{S:T}$ test powers. A little over two-thirds of these "exceptions," given this design, coincided with alpha levels which were too liberal, which would explain this result. However, as one example of a curious and unexplained result, given a preliminary alpha of .02, a conditional alpha of .01, c = 2 and s = 12, the power of the conditional test equalled .125 (Table C-1), while the power of the "always pool" test only equalled .114 (Table 9) even though the actual alpha level of this conditional test was within 1.96 standard errors of the nominal value. The other design that had several of these surprising and unexplanable findings was c = 5 and s = 5 (Table C-2). Here 14 of the 25 (56%) conditional test powers exceeded the $F = MS_T/MS_{S:T}$ test powers. And for this particular design, in all cases but one, these "exceptions" occurred even though the conditional test alphas were not too liberal.

Positive Dependence

Positive dependence is that condition where the variance of the aggregate units exceeds that predicted given random assignment of individual units to groups, the variance of the disaggregate units and the grouping size. Given positive dependence, the two-tailed preliminary test should reject its null hypothesis, H_o: $E(MS_{C:T})$ equals $E(MS_{S:CT})$, designating the aggregate unit (classrooms) as the appropriate unit of analysis in testing for treatment effects.

Actual alpha levels. Because positive dependence is defined by the $E(MS_{C:T})$ being greater than the $E(MS_{S:CT})$, it was expected that given no treatment effects the observed probabilities of Type I errors for the conditional or "sometimes pool" F distribution would be too large. While the simulations generally showed this (see bottom sections of Tables 11 through 15), they also empirically showed that all four factors--the value of s, the value of c, the nominal alpha level of the preliminary test, and the nominal alpha level of the conditional test--and their interactions affected whether or not there was any liberalness in the conditional distribution and, if so, the degree of that liberalness.

While one generally expected the observed alphas to be too liberal, given positive dependence, that liberalness decreased as the nominal alpha level of the preliminary F test increased from .02 to .50 and the nominal alpha level of the conditional F test increased from .01 to .25. This was so for both defined degrees of positive dependence, $E(MS_{C:T})/E(MS_{S:CT})$ equal to 2 and $E(MS_{C:T})/E(MS_{S:CT})$ equal to 3. For example, Table 12 shows that, given $E(MS_{C:T})/E(MS_{S:CT})$ equal to 3, c = 5 and s = 5, if the preliminary test alpha level equals .20, four of the five conditional test actual alphas were too liberal; if, however, the alpha of the preliminary test is increased to .50, only three of the five conditional test alphas were too liberal. Given the same set of conditions, but letting the conditional test nominal alpha remain constant at .10, four of the five actual alphas were too liberal; when, on the other hand, the nominal alpha level of the

conditional test was increased to .25, only three of the five actual conditional test alpha levels were too liberal. That an increase in the nominal level of the preliminary test should cause a decrease in the liberalness of the conditional F test was expected, as when the alpha level of the preliminary test is increased, the primary test of treatment effects becomes less of a conditional test. That the liberalness, given positive dependence, tended to disappear as the nominal alpha level of the conditional F test increased suggests that the tail of the conditional F distribution, given no treatment effects, was too thick in comparison to the tail of the distribution of $F = MS_T/MS_{C:T}$ at the extreme alpha levels, such as .01. The conditional distribution had a much closer fit to the central F distribution for (t-1) and (c-1)t degrees of freedom at the large alpha levels, such as .25.

Generally there was a trend for the fit of the observed alpha levels of the conditional F tests to improve as the number of classes increased. This improvement was more evident given the greater degree of positive dependence, $E(MS_{C:T})/E(MS_{S:CT}) = 3$, than given the lesser degree of positive dependence, $E(MS_{C:T})/E(MS_{S:CT}) = 2$. For example, given that $E(MS_{C:T})/E(MS_{S:CT})$ equals 3 for $c = 2$ and $s = 12$, 100% of the actual alphas of the conditional test were too liberal; for $c = 5$ and $s = 12$, 60% were too liberal; and for $c = 10$ and $s = 12$, only 8% were too liberal. Given that $E(MS_{C:T})/E(MS_{S:CT})$ equals 2, for $c = 2$ and $s = 12$, 100% of the actual alphas were too liberal; for $c = 5$ and $s = 12$, 68% were too liberal; and for $c = 10$ and $s = 12$, 36% were too

liberal. This trend, of the actual alpha levels of the conditional F tests becoming less liberal as c increased, was predictable as when the number of classes increased the discrepancy between the $E(MS_{C:T})$ and the $E(MS_{S:T})$ decreased (Table 10). As c increases, pooling should be prescribed less often as the preliminary test becomes more powerful. And when pooling is prescribed, the pooled mean square error is weighted in favor of the proper error term, $E(MS_{C:T})$. Both factors are contributing toward a decrease in the bias of the conditional test error term and thus less disagreement between the actual and nominal alpha values of the conditional F test.

How increasing the number of students per classroom affected the actual alpha levels of the conditional F test was less clear. As s increases, pooling of error terms should be prescribed less often, causing the conditional distribution to become more similar to the distribution of the F statistic using classrooms as the unit of analysis. But when pooling is prescribed, the pooled error term is weighted toward the improper, too small error term, $E(MS_{S:CT})$, causing the conditional F test to be too liberal. The effect of these two competing factors on the actual alpha level of the conditional test clearly depends on the combined value of s and the ratio of the $E(MS_{C:T})$ to the $E(MS_{S:CT})$. A comparison across Tables 12, 13, and 14 shows that as s was increased in the simulations from 5 to 12 to 20, no simple trend on the actual alpha level of the conditional F tests showed up. For c = 5 and s = 5, 92% and 88% of the actual alpha levels were too liberal given $E(MS_{C:T})/E(MS_{S:CT})$ equal to 2 and 3, respectively;

for c = 5 and s = 12, 68% and 60% of the actual alphas were too liberal given $E(MS_{C:T})/E(MS_{S:CT})$ equal to 2 and 3, respectively; and for c = 5 and s = 20, 84% and 68% of the actual alpha values were too large given $E(MS_{C:T})/MS_{S:CT})$ equal to 2 and 3, respectively.

When comparing the empirical alpha levels for both degrees of positive dependence, $E(MS_{C:T})/E(MS_{S:CT})$ equal to 2 and $E(MS_{C:T})/E(MS_{S:CT})$ equal to 3, the conditional or "sometimes pool" F tests generally appeared more liberal given the lesser degree of positive dependence, $E(MS_{C:T})/E(MS_{S:CT})$ equal 2, than given the higher degree of positive dependence. For example, given c = 5 and s = 20 (Table 14), 84% of the conditional tests' actual alphas were too liberal when $E(MS_{C:T})/E(MS_{S:CT})$ equalled 2; while only 68% of the actual alphas were too liberal when $E(MS_{C:T})/E(MS_{S:CT})$ equalled 3. An exception to this trend appeared with the design c = 2 and s = 12 (Table 11). Given c = 2 and s = 12, all (100%) of the actual alpha values were too liberal for both degrees of positive dependence. And for this design, given a specific preliminary test nominal alpha value and a specific conditional test nominal alpha value, all 25 actual conditional test alpha values given the condition $E(MS_{C:T})/E(MS_{S:CT})$ equal to 3 were more liberal than their matched values given the condition $E(MS_{C:T})/E(MS_{S:CT})$ equal to 2. For example, given this one design, if $E(MS_{C:T})/E(MS_{S:CT})$ equals 2, the nominal preliminary test alpha level equals .10 and the nominal conditional test alpha level equals .05, then the actual conditional test alpha value equals .134; while for those same conditions, except letting $E(MS_{C:T})/E(MS_{S:CT})$ equal 3, the nominal conditional test alpha level

equals .161. This is exactly opposite to what one would expect and to what actually occurred in the other four designs. When it is true that the $E(MS_{C:T})/E(MS_{S:CT})$ equals 3, the researcher rejects the null hypothesis of independence more often than when it is true that the $E(MS_{C:T})/E(MS_{S:CT})$ equals 2. This dictates using $MS_{C:T}$ more often as the error term for the test of treatment effects. This, in turn, suggests that the conditional F distribution, given $E(MS_{C:T})/E(MS_{S:CT})$ equal 3, is closer to the F distribution of the unconditional $F = MS_T/MS_{C:T}$ test than is the conditional F distribution, given $E(MS_{C:T})/E(MS_{S:CT})$ equals 2. Thus the observed alpha values of the conditional test, given the greater degree of positive dependence, should be less liberal than the observed alpha values of the conditional test, given the lesser degree of positive dependence.

Estimated powers. Without exception, the statistical powers of the conditional F test of treatment effects were more powerful than the powers of the unconditional test $F = MS_T/MS_{C:T}$ (see bottom sections of Tables 16 through 20 and Appendix C). This was expected because of the general liberalness of the conditional F test, given positive dependence. In this study, the liberalness of a test statistic and its power are completely confounded. Because of this confounding, statistical powers, given positive student dependence, were not examined to any great extent since a liberal alpha is generally considered a "no-no." One can, however, compare powers of the conditional F tests to powers of the $F = MS_T/MS_{C:T}$ tests for simulated designs which did not have alpha levels that were too large. For example, given $E(MS_{C:T})/E(MS_{S:CT})$ equal to 3, c = 10

and s = 12, only 8% of the actual alpha levels of the conditional F tests were too liberal (Table 15). The power differences between the "sometimes pool" test and the "never pool" test for this one design are very small, but positive (Tables 20 and C-5).

Negative Dependence

Negative dependence is that condition where the variance of the aggregate units is smaller than that predicted given random assignment of individual units to groups, the variance of the individual units and the grouping size. As with positive dependence situations, the two-tailed preliminary test should reject its null hypothesis, H_o: $E(MS_{C:T})$ equals $E(MS_{S:CT})$, given $E(MS_{C:T})/E(MS_{S:CT})$ is less than one, designating the aggregate unit (classrooms) as the appropriate unit of analysis in testing for treatment differences.

Actual alpha levels. Given negative dependence, where the $E(MS_{C:T})$ is less than the $E(MS_{S:CT})$, and a two-tailed preliminary test, one would expect the observed probabilities of Type I errors for the conditional F tests of treatment effects to be too small. However, for the two specific degrees of negative dependence in this simulation study, many (66.4% in total) observed alpha levels of the conditional or "sometimes pool" F tests were within 1.96 standard errors of the theoretical or nominal alpha values (see top sections of Tables 11 through 15).

The conservativeness that did appear in the data decreased as the nominal alpha level of the preliminary tests increased from .02 to .50 and the nominal alpha level of the conditional F tests decreased from

.25 to .01. This held true for both simulated degrees of dependence, $E(MS_{C:T})/E(MS_{S:CT})$ equal to .50 and $E(MS_{C:T})/E(MS_{S:CT})$ equal to .33. Table 11 shows that, given $E(MS_{C:T})/E(MS_{S:CT})$ equal .33, c = 2 and s = 12, if the preliminary test alpha level equals .02, one out of five conditional test actual alphas fell within a 95% confidence interval of the appropriate nominal value; if, however, the preliminary test alpha level equals .05, two of the five conditional test actual alphas were within the 95% confidence interval. Given the same set of conditions, if the conditional test nominal alpha equals .025, four of the five actual alphas fell within the 95% confidence interval of their respective nominal values; if, on the other hand, the conditional test nominal alpha equals .01, all five actual alphas were within the 95% confidence interval. That the conservativeness of the conditional F tests decreased as the nominal alpha level of the preliminary test increased directly compares to the similar results found given positive dependence and liberalness of the conditional F test alpha level. That the conservativeness, given negative dependence, tended to dissipate as the nominal alpha level of the conditional F test decreased suggests that the conditional F distribution, given no treatment effects, was too thin in comparison to the distribution of the $F = MS_T/MS_{C:T}$ test at the larger alpha levels, such as .25. On the other hand, the conditional F distribution had a much closer fit to the central F distribution for (t-1) and c-1)t degrees of freedom given the extreme alpha levels, such as .01.

It was expected that the fit of the observed alpha level of the conditional F test would improve as the number of classes increased from c equal 2 to c equal 5 and 10. As in the case of positive dependence, this improvement was expected because when the number of classes increases the discrepancy between the $E(MS_{C:T})$ and the $E(MS_{S:T})$ decreases (Table 10). Once more it was expected that more improvement would take place given the greater degree of negative dependence, $E(MS_{C:T})/E(MS_{S:CT}) = .33$, than given the lesser degree of negative dependence, $E(MS_{C:T})/E(MS_{S:CT}) = .50$. The simulated data situations did not empirically verify very well these expected improvements in the fit of the conditional test actual alpha levels as c was increased. For example, given that $E(MS_{C:T})/E(MS_{S:CT})$ equals .33, for c = 2 and s = 12, 40% of the actual alpha values of the conditional test were too conservative; for c = 5 and s = 12, 44% were too conservative; and for c = 10 and s = 12, 12% were too conservative. Given that $E(MS_{C:T})/E(MS_{S:CT})$ equalled .50, for c = 2 and s = 12, 40% of the observed alpha values were too conservative; for c = 5 and s = 12, 52% were too conservative; and for c = 10 and s = 12, 40% were too conservative. That the expected trend achieved by increasing c did not appear in the simulated data was probably due somewhat to the fact that many of the actual alpha levels of the conditional F test were not statistically different from their theoretical values.

As in the case with positive dependency, comparisons within the two negative dependency conditions but across Tables 12, 13, and 14 showed no simple trend on how increasing the number of students per

class affected the actual alpha values of the conditional F test. For example, given $E(MS_{C:T})/E(MS_{S:CT})$ equal to .33 and c = 5, as s was increased from 5 to 12 to 20, the percentage of actual alpha levels of the conditional test that were too conservative equalled 20%, 44%, and 24%, respectively. Given $E(MS_{C:T})/E(MS_{S:CT})$ equal to .50 and c = 5, again as s was increased from 5 to 12 to 20, the percentage of conservative actual alpha levels of the conditional F tests equalled 24%, 52%, and 40%, respectively.

When comparing the observed alpha levels for both defined degrees of negative dependency, the conditional or "sometimes pool" F test generally appeared more conservative given the lesser degree of negative dependence, $E(MS_{C:T})/E(MS_{S:CT}) = .50$, than given the larger degree of negative dependence, $E(MS_{C:T})/E(MS_{S:CT}) = .33$. For example, given c = 10 and s = 12 (Table 15), 40% of the conditional test alpha levels were too conservative when $E(MS_{C:T})/E(MS_{S:CT})$ equalled .50; while only 12% of the actual alphas were too conservative when $E(MS_{C:T})/E(MS_{S:CT})$ equalled .33. One design deviated from this predicted finding. Given c = 2 and s = 12, 40% of the actual alpha values were too conservative for both degrees of negative dependence. And for this one design, given a specific preliminary test nominal alpha value and a large conditional test nominal alpha value, a majority of the actual conditional test alpha values, given $E(MS_{C:T})/E(MS_{S:CT})$ equal to .33, were more conservative than their respective alpha values given $E(MS_{C:T})/E(MS_{S:CT})$ equal to .50. This one design also deviated from the general trend in the simulations done given positive dependence.

Estimated powers. It was hoped that, since the fit of the empirical alpha levels of the conditional test were fairly good and in the conservative direction which is more acceptable than a test being too liberal, use of the conditional test would increase power, relative to the unconditional test $F = MS_T/MS_{C:T}$. For all five designs (see top sections of Tables 16 through 20 and Appendix C), only when the nominal alpha level of the preliminary test was very large (preferably .50) and the nominal alpha level of the conditional test was small (.01 or .025) did the powers of the conditional or "sometimes pool" test tend to be greater than the powers of the "never pool" $F = MS_T/MS_{C:T}$ test. Given these two conditions, however, the differences between the estimated powers of the two F tests tended to be on the small side. For example, given $E(MS_{C:T})/E(MS_{S:CT})$ equal to .33, a preliminary test alpha value equal to .50, a conditional test alpha value equal to .01, c = 5 and s = 12, the estimated power of the conditional test was .679 (Table C-3); while the estimated power of the unconditional $F = MS_T/MS_{C:T}$ test was .661 (Table 18), a .018 difference in favor of the conditional test. C = 2 and s = 12 was the one design where the power of the conditional test was in a majority of cases, at each of the two defined negative dependence degrees, larger than the power of the $F = MS_T/MS_{C:T}$ test (Table 16). Given c = 2 and s = 12, across all alphas, 52% and 88% of the conditional test powers were greater than the unconditional test powers, given $E(MS_{C:T})/E(MS_{S:CT})$ equal to .33 and .50, respectively. This same design had a majority of its actual alpha levels of the conditional test to within 1.96

standard errors of the nominal alpha at the same large preliminary test nominal alphas and small conditional test nominal alpha levels (Table 11).

Comparing the powers of the conditional and unconditional F tests, given negative dependence levels prescribed in this simulation study, one would have to conclude that generally using a two-tailed preliminary test to choose a "correct" unit of analysis lowers the powers of the F test rather than raises them, as is the desired case. For example, given $E(MS_{C:T})/E(MS_{S:CT})$ equal to .50, c = 5 and s = 5 (Table 17), 76% of the "never pool" test powers exceeded the "sometimes pool" test powers. And given $E(MS_{C:T})/E(MS_{S:CT})$ equal to .33, c = 10 and s = 12 (Table 20), 52% of the "never pool" test powers were larger than the "sometimes pool" test powers; and 44% of the time, the two estimated powers were equal.

The Upper-Tailed Preliminary Test

Paull (1950), Peckham et al. (1969a, 1969b) and Poynor (1974) considered only two possible situations. One, student responses within classrooms could be independent of each other, defined by the $E(MS_{C:T})$ equalling the $E(MS_{S:CT})$; or two, student responses within classrooms could be positively dependent upon each other, defined by the $E(MS_{C:T})$ being greater than the $E(MS_{S:CT})$. And because they considered only the one alternative to independence, they suggested that the choice of unit of analysis could be determined by using an upper-tailed only preliminary F test, rather than the two-tailed preliminary test

discussed in the previous section. This upper-tailed preliminary F test tests the null hypothesis that the $E(MS_{C:T})$ is less than or equal to the $E(MS_{S:CT})$ or equivalently that ρI is less than or equal to zero. Using this upper-tailed only preliminary F test suggests that the negative dependency situation, where the $E(MS_{C:T})$ is less than the $E(MS_{S:CT})$, could never occur or that this situation is no more interesting than a zero difference between the between class and within class expected mean squares. The latter is clearly not the case as was shown in Chapter VI, the chapter which discussed the empirical results of correlated units.

In this section the effects of using an upper-tailed preliminary test to choose an analytic unit for the conditional test are studied for five different preliminary test alpha levels (i.e., .01, .025, .05, .10, and .25). Actual conditional test alpha levels, given the upper-tailed preliminary F test, are shown in Tables 21 through 25. Corresponding differences between estimated powers of the conditional F test and the unconditional, always correct $F = MS_T/MS_{C:T}$ test are shown in Tables 26 through 30. Appendix D (Tables D-1 through D-5) contains the empirical powers of the conditional test, given the upper-tailed preliminary test. Each separate table describes the effect on the conditional F tests' actual alpha or power of varying the type and degree of dependence, the alpha level of the upper-tailed preliminary test and the alpha level of the conditional test for one specific combination of s and c.

Table 21

Actual Alphas of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 2 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha		Conditional test nominal alpha .010	.025	.050	.100	.250	Mean alpha
	.33	.01	.000	.001	.001	.008	.059	.014
		.025	.000	.001	.001	.008	.059	.014
		.05	.000	.001	.001	.008	.059	.014
		.10	.000	.001	.001	.008	.058	.014
		.25	.000	.001	.001	.008	.057	.013
	Mean alpha		.000	.001	.001	.008	.058	
	.50	.01	.001	.002	.009	.029	.100	.028
		.025	.001	.002	.009	.029	.100	.028
		.05	.001	.002	.009	.029	.099	.028
		.10	.001	.002	.009	.029	.098	.028
		.25	.000	.001	.007	.026	.090	.025
	Mean alpha		.001	.002	.009	.028	.097	
	1	.01	.011[a]	.032[a]	.058[a]	.090[a]	.240[a]	.086
		.025	.011[a]	.032[a]	.057[a]	.089[a]	.235[a]	.085
		.05	.009[a]	.030[a]	.055[a]	.085[a]	.227[a]	.081
		.10	.009[a]	.029[a]	.054[a]	.082[a]	.217	.078
		.25	.008[a]	.026[a]	.049[a]	.073	.190	.069
	Mean alpha		.010	.030	.055	.084	.222	
	2	.01	.061	.084	.132	.205	.351	.167
		.025	.060	.082	.126	.190	.328	.157
		.05	.059	.079	.121	.181	.310	.150
		.10	.049	.067	.104	.158	.282	.132
		.25	.035	.049	.074	.117[a]	.237[a]	.102
	Mean alpha		.053	.072	.111	.170	.302	
	3	.01	.090	.141	.185	.258	.379	.211
		.025	.081	.128	.169	.235	.348	.192
		.05	.074	.117	.151	.208	.317	.173
		.10	.062	.099	.128	.184	.289	.152
		.25	.044	.062	.085	.124	.247[a]	.112
	Mean alpha		.070	.109	.144	.202	.316	

[a]Actual alpha is within 1.96 standard errors of the nominal alpha.

Table 22

Actual Alphas of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 5 and s = 5

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean alpha
.33	.10	.000	.000	.004	.011	.058	.015
	.025	.000	.000	.004	.011	.058	.015
	.05	.000	.000	.004	.011	.058	.015
	.10	.000	.000	.004	.011	.058	.015
	.25	.000	.000	.004	.011	.058	.015
	Mean alpha	.000	.000	.004	.011	.058	
.50	.01	.000	.007	.012	.023	.114	.031
	.025	.000	.007	.012	.023	.114	.031
	.05	.000	.007	.012	.023	.114	.031
	.10	.000	.007	.012	.023	.113	.031
	.25	.000	.007	.011	.022	.113	.031
	Mean alpha	.000	.007	.012	.023	.114	
1	.01	.011[a]	.021[a]	.047[a]	.095[a]	.230[a]	.081
	.025	.011[a]	.021[a]	.047[a]	.095[a]	.227[a]	.080
	.05	.011[a]	.021[a]	.046[a]	.093[a]	.222	.079
	.10	.011[a]	.020[a]	.044[a]	.091[a]	.217	.077
	.25	.007[a]	.015	.039[a]	.083[a]	.203	.069
	Mean alpha	.010	.020	.045	.091	.220	
2	.01	.039	.072	.111	.170	.308	.140
	.025	.035	.065	.101	.153	.289	.129
	.05	.031	.058	.091	.141	.277	.120
	.10	.024	.051	.085	.128	.262[a]	.110
	.25	.017	.035	.066	.112[a]	.244[a]	.095
	Mean alpha	.029	.056	.091	.141	.276	
3	.01	.056	.082	.119	.170	.319	.149
	.025	.047	.072	.108	.158	.299	.137
	.05	.038	.058	.094	.146	.285	.124
	.10	.030	.044	.077	.125	.274[a]	.110
	.25	.018	.030[a]	.067	.113[a]	.252[a]	.096
	Mean alpha	.038	.057	.093	.142	.286	

[a]Actual alpha is within 1.96 standard errors of the nominal alpha.

Table 23

Actual Alphas of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 5 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean alpha
.33	.01	.000	.000	.001	.005	.049	.011
	.025	.000	.000	.001	.005	.049	.011
	.05	.000	.000	.001	.005	.049	.011
	.10	.000	.000	.001	.005	.049	.011
	.25	.000	.000	.001	.005	.049	.011
	Mean alpha	.000	.000	.001	.005	.049	
.50	.01	.000	.003	.007	.018	.108	.027
	.025	.000	.003	.007	.018	.108	.027
	.05	.000	.003	.007	.018	.108	.027
	.10	.000	.003	.007	.018	.107	.027
	.25	.000	.003	.007	.018	.105	.027
	Mean alpha	.000	.003	.007	.018	.107	
1	.01	.008[a]	.020[a]	.045[a]	.093[a]	.234[a]	.080
	.025	.008[a]	.020[a]	.045[a]	.093[a]	.228[a]	.079
	.05	.008[a]	.020[a]	.045[a]	.093[a]	.222	.078
	.10	.008[a]	.019[a]	.044[a]	.090[a]	.215	.075
	.25	.007[a]	.015	.038[a]	.081	.196	.067
	Mean alpha	.008	.019	.043	.090	.219	
2	.01	.040	.081	.111	.167	.323	.144
	.025	.033	.067	.095	.148	.298	.128
	.05	.028	.058	.084	.132	.274[a]	.115
	.10	.024	.050	.071	.112[a]	.253[a]	.102
	.25	.014[a]	.029[a]	.049[a]	.095[a]	.234[a]	.084
	Mean alpha	.028	.057	.082	.131	.276	
3	.01	.055	.076	.108	.157	.298	.139
	.025	.042	.056	.085	.132	.280	.119
	.05	.032	.044	.072	.120	.265[a]	.107
	.10	.023	.030[a]	.058[a]	.109[a]	.244[a]	.093
	.25	.012[a]	.023[a]	.051[a]	.099[a]	.231[a]	.083
	Mean alpha	.033	.046	.075	.123	.264	

[a]Actual alpha is within 1.96 standard errors of the nominal alpha.

Table 24

Actual Alphas of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 5 and s = 20

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha		Conditional test nominal alpha .010	.025	.050	.100	.250	Mean alpha
	.33	.01	.000	.000	.001	.006	.050	.011
		.025	.000	.000	.001	.006	.050	.011
		.05	.000	.000	.001	.006	.050	.011
		.10	.000	.000	.001	.006	.050	.011
		.25	.000	.000	.001	.006	.050	.011
	Mean alpha		.000	.000	.001	.006	.050	
	.50	.01	.000	.002	.008	.021	.122	.031
		.025	.000	.002	.008	.021	.122	.031
		.05	.000	.002	.008	.021	.122	.031
		.10	.000	.002	.007	.020	.121	.030
		.25	.000	.002	.007	.020	.120	.030
	Mean alpha		.000	.002	.008	.021	.121	
	1	.01	.013[a]	.023[a]	.048[a]	.108[a]	.257[a]	.090
		.025	.012[a]	.022[a]	.047[a]	.107[a]	.255[a]	.089
		.05	.011[a]	.021[a]	.046[a]	.105[a]	.252[a]	.087
		.10	.011[a]	.021[a]	.043[a]	.099[a]	.243[a]	.083
		.25	.010[a]	.018[a]	.036	.088[a]	.227[a]	.076
	Mean alpha		.011	.021	.044	.101	.247	
	2	.01	.051	.091	.119	.192	.328	.156
		.025	.047	.084	.110	.174	.305	.144
		.05	.039	.073	.100	.155	.289	.131
		.10	.031	.060	.082	.127	.273[a]	.115
		.25	.017	.036	.057[a]	.109[a]	.249[a]	.094
	Mean alpha		.037	.069	.094	.151	.289	
	3	.01	.063	.088	.122	.170	.303	.149
		.025	.049	.068	.097	.144	.286	.129
		.05	.031	.051	.081	.136	.266[a]	.113
		.10	.023	.042	.066	.118[a]	.256[a]	.101
		.25	.009[a]	.029[a]	.052[a]	.103[a]	.249[a]	.088
	Mean alpha		.035	.056	.084	.134	.272	

[a]Actual alpha is within 1.96 standard errors of the nominal alpha.

Table 25

Actual Alphas of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 10 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha		Conditional test nominal alpha .010	.025	.050	.100	.250	Mean alpha
	.33	.01	.000	.000	.000	.004	.053	.011
		.025	.000	.000	.000	.004	.053	.011
		.05	.000	.000	.000	.004	.053	.011
		.10	.000	.000	.000	.004	.053	.011
		.25	.000	.000	.000	.004	.053	.011
	Mean alpha		.000	.000	.000	.004	.053	
	.50	.01	.000	.000	.004	.021	.108	.027
		.025	.000	.000	.004	.021	.108	.027
		.05	.000	.000	.004	.021	.108	.027
		.10	.000	.000	.004	.021	.108	.027
		.25	.000	.000	.004	.021	.108	.027
	Mean alpha		.000	.000	.004	.021	.108	
	1	.01	.007[a]	.022[a]	.052[a]	.096[a]	.234[a]	.082
		.025	.007[a]	.022[a]	.052[a]	.095[a]	.233[a]	.082
		.05	.007[a]	.021[a]	.051[a]	.093[a]	.230[a]	.080
		.10	.006[a]	.018[a]	.048[a]	.088[a]	.225[a]	.077
		.25	.006[a]	.016[a]	.040[a]	.082[a]	.210	.071
	Mean alpha		.007	.020	.049	.091	.226	
	2	.01	.032	.053	.087	.152	.277	.120
		.025	.024	.043	.071	.132	.262[a]	.106
		.05	.015[a]	.033[a]	.063[a]	.117[a]	.254[a]	.096
		.10	.011[a]	.026[a]	.054[a]	.103[a]	.240[a]	.087
		.25	.007[a]	.020[a]	.050[a]	.095[a]	.233[a]	.081
	Mean alpha		.018	.035	.065	.120	.253	
	3	.01	.019	.031[a]	.064	.116[a]	.248[a]	.096
		.025	.014[a]	.026[a]	.056[a]	.102[a]	.243[a]	.088
		.05	.011[a]	.021[a]	.053[a]	.099[a]	.241[a]	.085
		.10	.007[a]	.018[a]	.050[a]	.097[a]	.235[a]	.081
		.25	.006[a]	.017[a]	.049[a]	.094[a]	.231[a]	.079
	Mean alpha		.011	.023	.054	.102	.240	

[a]Actual alpha is within 1.96 standard errors of the nominal alpha.

Independence

Ideally, the upper-tailed preliminary test should designate the disaggregate unit (student) as the appropriate unit of analysis to use in testing the primary or conditional null hypothesis of no treatment differences.

Actual alpha levels. Comparisons of actual alpha levels to their respective nominal values were excellent as all four parameters, c, s, the preliminary test alpha level and the conditional test alpha level, were varied (Tables 21 through 25). Given $E(MS_{C:T})/E(MS_{S:CT})$ equal one and an upper-tailed preliminary F test, the observed alphas for the conditional or "sometimes pool" F test of treatment effects were 89.6% of the time within 1.96 standard errors of the theoretical alpha levels. All conditional test alphas (10.4%) that were not in agreement with their respective nominal values were too conservative. Unlike the situation found given independence and the two-tailed preliminary test, the design c = 2 and s = 12 was no exception to the rule as, given c = 2 and s = 12, 88% of the actual alphas were within 1.96 standard errors of the nominal alphas.

Estimated powers. Given $E(MS_{C:T})/E(MS_{S:CT})$ equal one, the estimated statistical powers of the conditional or "sometimes pool" F tests were, in most cases (91.2% of the time), greater than the estimated powers of the unconditional or "never pool" tests of treatment effect, $F = MS_T/MS_{C:T}$ (Tables 26 through 30). Paull (1950), in studying the distributional properties of the conditional F test, given independence and an upper-tailed preliminary F test, also found that

the upper-tailed only preliminary test was effective in making the power of the "sometimes pool" test greater than the power of the "never pool" test. Given $E(MS_{C:T})/E(MS_{S:CT})$ equal one, all trends in power discrepancies between the conditional and unconditional F tests due to varying the four parameters, c, s, the preliminary test alpha level, and the conditional test alpha level, mirrored those found given the two-tailed preliminary test of independence.

Across all five designs and all five preliminary test alpha levels, as the five conditional test alpha levels increased from .01 to .25, the average difference between the "sometimes pool" and the "never pool" test powers went from .089 to .076 to .059 to .042 to .012. This indirect relationship did not, however, hold up within two of the five combinations of c and s (i.e., c = 2 and s = 12; c = 5 and s = 5). For example, consider the design c = 5 and s = 5 (Table 27). Averaged across the five preliminary test alpha levels, the estimated power differences equalled .017, .023, .026, .017, and .007 as their counterpart nominal values increased from .01 to .25.

As the number of individual units per group increased from s equals 5 to s equal 12 and 20 (compare Tables 27, 28, and 29), the discrepancies between the powers of the "sometimes pool" test and the powers of the "never pool" test tended to increase. For example, given independence, an upper-tailed preliminary test, a conditional test nominal alpha of .01 and a c equal to 5, averaging over the five preliminary test alpha levels gave average power differences between the "sometimes pool" and "never pool" tests of .017 for s = 5, .119 for s = 12

Table 26

Power of the Conditional F Test Minus Power of the Test $F = MS_T/MS_{C:T}$ Given an Upper-Tailed Preliminary Test, c = 2 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean power dif.[a]
	.01	-.046	-.062	-.120	-.157	-.108	-.099
	.025	-.046	-.062	-.120	-.157	-.108	-.099
.33	.05	-.046	-.062	-.120	-.157	-.108	-.099
	.10	-.046	-.062	-.120	-.157	-.110	-.099
	.25	-.046	-.063	-.123	-.164	-.120	-.103
	Mean power dif.[a]	-.046	-.062	-.121	-.158	-.111	
	.01	-.002	.002	-.006	-.013	-.052	-.014
	.025	-.002	.002	-.006	-.013	-.052	-.014
.50	.05	-.002	.002	-.006	-.013	-.054	-.015
	.10	-.002	.001	-.008	-.016	-.058	-.017
	.25	-.006	-.008	-.022	-.038	-.088	-.032
	Mean power dif.	-.003	.000	-.010	-.019	-.061	
	.01	.077	.103	.132	.146	.070	.106
	.025	.075	.099	.125	.137	.058	.099
1	.05	.071	.092	.117	.127	.045	.090
	.10	.069	.086	.107	.113	.021	.079
	.25	.054	.058	.065	.053	-.024	.041
	Mean power dif.	.069	.088	.109	.115	.034	
	.01	.147	.185	.217	.212	.155	.183
	.025	.137	.169	.199	.192	.132	.166
2	.05	.126	.153	.177	.161	.100	.143
	.10	.109	.128	.147	.124	.065	.115
	.25	.073	.076	.078	.040	-.001	.053
	Mean power dif.	.118	.142	.164	.146	.090	
	.01	.166	.216	.246	.214	.164	.201
	.025	.147	.190	.213	.176	.131	.171
3	.05	.129	.167	.186	.146	.100	.146
	.10	.106	.133	.135	.091	.046	.102
	.25	.063	.075	.064	.024	.006	.046
	Mean power dif.	.122	.156	.169	.130	.089	

[a]Mean power differences.

Table 27

Power of the Conditional F Test Minus Power of the Test $F = MS_T/MS_{C:T}$ Given an Upper-Tailed Preliminary Test, c = 5 and s = 5

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean power dif.[a]
.33	.01	-.234	-.331	-.394	-.363	-.169	-.298
	.025	-.234	-.331	-.394	-.363	-.169	-.298
	.05	-.234	-.331	-.394	-.363	-.169	-.298
	.10	-.234	-.331	-.394	-.363	-.169	-.298
	.25	-.234	-.331	-.394	-.363	-.169	-.298
	Mean power dif.[a]	-.234	-.331	-.394	-.363	-.169	
.50	.01	-.105	-.148	-.191	-.186	-.124	-.151
	.025	-.105	-.148	-.191	-.186	-.124	-.151
	.05	-.105	-.148	-.191	-.186	-.124	-.151
	.10	-.105	-.148	-.191	-.186	-.125	-.151
	.25	-.105	-.148	-.195	-.193	-.127	-.154
	Mean power dif.	-.105	-.148	-.192	-.187	-.125	
1	.01	.023	.037	.046	.040	.020	.033
	.025	.023	.037	.043	.035	.019	.031
	.05	.021	.030	.035	.027	.015	.026
	.10	.017	.019	.018	.007	.003	.013
	.25	.000	-.008	-.010	-.024	-.022	-.013
	Mean power dif.	.017	.023	.026	.017	.007	
2	.01	.089	.101	.105	.107	.102	.101
	.025	.075	.077	.072	.076	.075	.075
	.05	.064	.069	.063	.061	.056	.063
	.10	.051	.048	.040	.042	.038	.044
	.25	.022	.020	.012	.015	.006	.015
	Mean power dif.	.060	.063	.058	.060	.055	
3	.01	.083	.083	.099	.093	.082	.088
	.025	.068	.063	.078	.072	.055	.067
	.05	.049	.046	.058	.051	.036	.048
	.10	.032	.031	.035	.021	.025	.029
	.25	.018	.009	.011	.002	.006	.009
	Mean power dif.	.050	.046	.056	.048	.041	

[a]Mean power differences.

Table 28

Power of the Conditional F Test Minus Power of the Test $F = MS_T/MS_{C:T}$ Given an Upper-Tailed Preliminary Test, $c = 5$ and $s = 12$

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean power dif.[a]
.33	.01	-.379	-.354	-.222	-.111	-.035	-.220
	.025	-.379	-.354	-.222	-.111	-.035	-.220
	.05	-.379	-.354	-.222	-.111	-.035	-.220
	.10	-.379	-.354	-.222	-.111	-.035	-.220
	.25	-.379	-.354	-.222	-.111	-.035	-.220
	Mean power dif.[a]	-.379	-.354	-.222	-.111	-.035	
.50	.01	-.138	-.170	-.148	-.089	-.030	-.115
	.025	-.138	-.170	-.148	-.089	-.030	-.115
	.05	-.138	-.170	-.148	-.089	-.030	-.115
	.10	-.138	-.170	-.148	-.089	-.031	-.115
	.25	-.146	-.177	-.153	-.091	-.032	-.120
	Mean power dif.	-.140	-.171	-.149	-.089	-.031	
1	.01	.148	.141	.105	.061	.025	.096
	.025	.138	.128	.092	.051	.022	.086
	.05	.130	.117	.081	.041	.014	.077
	.10	.113	.096	.058	.023	.010	.060
	.25	.064	.042	.008	-.010	-.003	.020
	Mean power dif.	.119	.105	.069	.033	.014	
2	.01	.192	.180	.161	.130	.078	.148
	.025	.153	.142	.122	.100	.050	.113
	.05	.119	.105	.090	.081	.031	.085
	.10	.081	.067	.045	.045	.015	.051
	.25	.034	.029	.012	.018	.000	.019
	Mean power dif.	.116	.105	.086	.075	.035	
3	.01	.114	.107	.100	.081	.061	.093
	.025	.075	.074	.071	.056	.040	.063
	.05	.056	.057	.054	.037	.027	.046
	.10	.030	.032	.031	.016	.013	.024
	.25	.014	.010	.010	.006	.004	.009
	Mean power dif.	.058	.056	.053	.039	.029	

[a]Mean power differences.

Table 29

Power of the Conditional F Test Minus Power of the Test $F = MS_T/MS_{C:T}$ Given an Upper-Tailed Preliminary Test, c = 5 and s = 20

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean power dif.[a]
.33	.01	-.211	-.110	-.061	-.018	-.001	-.080
	.025	-.211	-.110	-.061	-.018	-.001	-.080
	.05	-.211	-.110	-.061	-.018	-.001	-.080
	.10	-.211	-.110	-.061	-.018	-.001	-.080
	.25	-.211	-.110	-.061	-.018	-.001	-.080
	Mean power dif.[a]	-.211	-.110	-.061	-.018	-.001	
.50	.01	-.077	-.046	-.040	-.030	-.006	-.040
	.025	-.077	-.046	-.040	-.030	-.006	-.040
	.05	-.077	-.046	-.040	-.030	-.006	-.040
	.10	-.078	-.046	-.040	-.031	-.006	-.040
	.25	-.083	-.051	-.042	-.032	-.006	-.043
	Mean power dif.	-.078	-.047	-.040	-.031	-.006	
1	.01	.217	.172	.106	.056	.007	.112
	.025	.203	.157	.092	.049	.006	.101
	.05	.189	.143	.080	.044	.004	.092
	.10	.163	.119	.067	.032	-.001	.076
	.25	.091	.060	.020	.002	-.006	.033
	Mean power dif.	.173	.130	.073	.037	.002	
2	.01	.259	.235	.173	.116	.047	.166
	.025	.208	.187	.132	.087	.033	.129
	.05	.167	.140	.087	.057	.020	.094
	.10	.114	.094	.058	.034	.010	.062
	.25	.041	.033	.016	.005	.002	.019
	Mean power dif.	.158	.138	.093	.060	.022	
3	.01	.158	.137	.123	.073	.039	.106
	.025	.101	.086	.080	.043	.026	.067
	.05	.073	.061	.058	.026	.015	.047
	.10	.048	.031	.031	.012	.008	.026
	.25	.013	.008	.011	.001	.003	.007
	Mean power dif.	.079	.065	.061	.031	.018	

[a]Mean power differences.

Table 30

Power of the Conditional F Test Minus Power of the Test $F = MS_T/MS_{C:T}$ Given an Upper-Tailed Preliminary Test, c = 10 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean power dif.[a]
.33	.01	-.132	-.054	-.023	-.004	.000	-.043
	.025	-.132	-.054	-.023	-.004	.000	-.043
	.05	-.132	-.054	-.023	-.004	.000	-.043
	.10	-.132	-.054	-.023	-.004	.000	-.043
	.25	-.132	-.054	-.023	-.004	.000	-.043
	Mean power dif.[a]	-.132	-.054	-.023	-.004	.000	
.50	.01	-.115	-.067	-.027	-.014	-.002	-.045
	.025	-.115	-.067	-.027	-.014	-.002	-.045
	.05	-.115	-.067	-.027	-.014	-.002	-.045
	.10	-.115	-.067	-.027	-.014	-.002	-.045
	.25	-.115	-.069	-.028	-.014	-.002	-.046
	Mean power dif.	-.115	-.067	-.027	-.014	-.002	
1	.01	.096	.050	.028	.014	.003	.038
	.025	.088	.045	.024	.012	.003	.034
	.05	.080	.040	.023	.010	.002	.031
	.10	.062	.028	.018	.007	.001	.023
	.25	.019	.000	.000	-.001	-.002	.003
	Mean power dif.	.069	.033	.019	.008	.001	
2	.01	.165	.131	.090	.047	.019	.090
	.025	.109	.081	.053	.026	.013	.056
	.05	.070	.046	.033	.017	.008	.035
	.10	.044	.021	.015	.008	.002	.018
	.25	.013	.004	.002	.000	.000	.004
	Mean power dif.	.080	.057	.039	.020	.008	
3	.01	.040	.035	.020	.020	.006	.024
	.025	.022	.019	.010	.012	.003	.013
	.05	.006	.007	.005	.008	.002	.006
	.10	.003	.004	.003	.004	.001	.003
	.25	.001	.000	.000	.001	.000	.000
	Mean power dif.	.014	.013	.008	.009	.002	

[a]Mean power differences.

and .173 for s = 20. This trend was also found when the preliminary test was a two-tailed test. Only given a conditional test alpha of .25 did the trend fail to hold. This was most likely due to a ceiling effect occurring given the large alpha value.

With the exception of the very small conditional test alpha levels of .01 and .025 for the design c = 2 and s = 12, the discrepancies between the powers of the conditional test and the $F = MS_T/MS_{C:T}$ test tended to decrease with an increase in the number of classes per treatment (compare Tables 26, 28, and 30). For example, given independence, a conditional test alpha of .25 and s equal to 12, averaging the power discrepancies across the five preliminary alpha values gave discrepancies of .034 for c = 2, .014 for c = 5, and .001 for c = 10. That the expected trend did not hold at the small conditional test alpha levels for the design c = 2 and s = 12 also occurred when the preliminary test was a two-tailed test.

The effect of increasing the alpha level of the upper-tailed preliminary test on the power discrepancies between the "sometimes pool" and "never pool" F tests also copied the trend found given the two-tailed preliminary test. That is, there was an indirect relationship between changing the alpha level of the preliminary test and the effects that it had on the power differences between the conditional and unconditional tests. For example, given independence, a conditional test alpha level of .25, c = 2 and s = 12 (Table 26), the discrepancies between the power of the conditional test and the $F = MS_T/MS_{C:T}$ test equalled .070, .058, .045, .021, and -.024 for preliminary test alpha levels of .01, .025, .05, .10, and .25, respectively.

In comparison to the strange result that occurred when comparing the powers of the "sometimes pool" test, given a two-tailed preliminary F test, and the "always pool" test, the expected always occurred when the preliminary test was an upper-tailed only test. That is, as expected, the power of the "always pool" test (Table 9) always exceeded the power of the "sometimes pool" test, given an upper-tailed preliminary test and $E(MS_{C:T})/E(MS_{S:CT})$ equal one (Tables D-1 through D-5).

Positive Dependence

Ideally the upper-tailed preliminary test should reject its null hypothesis, designating the aggregate unit (classroom) as the appropriate analytic unit in looking for treatment effects. Paull and Peckham et al. referred to this particular situation, having the $E(MS_{C:T})/E(MS_{S:CT})$ be greater than one and an upper-tailed only preliminary F test, when they discussed the effects of using a preliminary testing procedure to choose the unit of analysis.

Actual alpha levels. Given the $E(MS_{C:T})$ was greater than the $E(MS_{S:CT})$ and the preliminary test was an upper-tailed only F test done at the α level, the actual alphas of the conditional or "sometimes pool" F test were essentially duplicates of the actual alphas of the conditional F test given positive dependence and a two-tailed preliminary test done at the 2α level, which generally were too liberal. For example, compare Tables 12 and 22. If $E(MS_{C:T})/E(MS_{S:CT})$ equal 2, the alpha level of the two-tailed preliminary test equal .02 and the alpha level of the upper-tailed preliminary test equal .01, the

absolute differences between the two sets of actual conditional test alphas equalled .000, .001, .001, .001, and .000 as the nominal conditional test alpha went from .01 to .25. Because the consequences of varying the five principal parameters, s, c, the preliminary test alpha, the conditional test alpha, and the degree of positive dependence, imitated (both in size and direction) those found when studying the effects of using a two-tailed preliminary test on the actual conditional test alphas, no more will be said about this situation.

Estimated powers. As expected, given the $E(MS_{C:T})$ was greater than the $E(MS_{S:CT})$, the power of the conditional test following an upper-tailed preliminary test done at α equalled the power of the conditional test given a two-tailed preliminary test done at 2α (compare Tables C-1 through C-5 with Tables D-1 through D-5). Because of the general liberalness of the "sometimes pool" test, though, given positive dependence and an upper-tailed preliminary test, studying the power of that "sometimes pool" test is rather uninteresting.

Negative Dependence

Given the $E(MS_{C:T})$ is less than the $E(MS_{S:CT})$, the upper-tailed only preliminary F test should designate the disaggregate unit (student) as the appropriate unit of analysis in testing the primary or conditional null hypothesis of no treatment effects.

Actual alpha levels. Given negative student dependence and an upper-tailed preliminary test, all observed alpha values of the conditional test were too conservative (see top sections of Tables 21

through 25). In fact, if one compares Tables 21 through 25 with Table 8, which recorded the actual alphas of the "always pool" $F = MS_T/MS_{S:T}$ test, one finds that the "sometimes pool" alphas, at both levels of negative dependence, are just as conservative as the "always pool" alphas. For example, given $E(MS_{C:T})/E(MS_{S:CT})$ equal .33, c = 2 and s = 12, the actual alphas of the conditional test averaged across the five preliminary test alpha levels equal .000, .001, .001, .008, and .058 as the conditional test nominal alphas increase from .01 to .25 (Table 21), while the actual alphas of $F = MS_T/MS_{S:T}$ equal .000, .001, .001, .008, and .059, respectively (Table 8).

The one distinct feature of the "sometimes pool" test is its preliminary test used for choosing the so-called appropriate unit of analysis for the primary test of treatment differences. However, given $E(MS_{C:T})/E(MS_{S:CT})$ is less than one and the preliminary test is an upper-tailed only test, the alpha level of the preliminary test had virtually no affect on the actual alpha level of the conditional or "sometimes pool" test. For example, given $E(MS_{C:T})/E(MS_{S:CT})$ equal .33 and c = 2 and s = 12 (Table 21), the actual alpha levels averaged over the five conditional test nominal alpha levels equalled .014, .014, .014, .014, and .013 as the nominal alpha levels of the preliminary test ranged from .01 to .25. Because the actual alphas and the effects of s, c, and the nominal alpha level of the conditional test are exactly identical to those found when studying the empirical effects on the alpha levels of always using student as the unit of analysis, given negative dependence, no more will be said about this situation.

Estimated powers. Because the actual alpha values of the "sometimes pool" tests behaved exactly as expected and exactly as did the actual alpha values of the "always pool" test, one could expect the powers of the "sometimes pool" tests (see top sections of Tables 26 through 30) and the "always pool" tests (Table 9) to mimic each other also. This is exactly what happened. Once again, the only difference between the two tests is that a preliminary test of independence preceeds the "sometimes pool" test. The effect of increasing the nominal alpha of the preliminary test had no affect whatsoever on the power of the "sometimes pool" test. Because the "sometimes pool" test, based on the upper-tailed preliminary test, and the "always pool" test are essentially the same, given negative dependence, no more will be said about the power of the former.

The Lower-Tailed Preliminary Test

The lower-tailed preliminary test tests the null hypothesis that the $E(MS_{C:T})$ is equal to or greater than the $E(MS_{S:CT})$ or equivalently that ρI is equal to or greater than zero. This test cannot detect positive dependence situations.

Observed conditional test alpha values, given the lower-tailed only preliminary F test, are reported in Tables 31 through 35. Estimated power discrepancies between the "sometimes pool" or conditional test and the "never pool" or $F = MS_T/MS_{C:T}$ test are given in Tables 36 through 40. Appendix E (Tables E-1 through E-5) contains the estimated powers of the conditional test, given the lower-tailed

Table 31

Actual Alphas of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 2 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean alpha
.33	.01	.007[a]	.015	.017	.025	.081	.029
	.025	.010[a]	.024[a]	.033	.051	.114	.046
	.05	.010[a]	.029[a]	.044[a]	.070	.151	.061
	.10	.010[a]	.029[a]	.048[a]	.085[a]	.201	.075
	.25	.010[a]	.029[a]	.050[a]	.093[a]	.259[a]	.088
	Mean alpha	.009	.025	.038	.065	.161	
.50	.01	.008[a]	.014	.023	.044	.116	.041
	.025	.011[a]	.023[a]	.035	.060	.139	.054
	.05	.011[a]	.026[a]	.045[a]	.075	.162	.064
	.10	.011[a]	.029[a]	.052[a]	.091[a]	.202	.077
	.25	.011[a]	.029[a]	.054[a]	.101[a]	.264[a]	.092
	Mean alpha	.010	.024	.042	.074	.177	
1	.01	.016[a]	.042	.070	.103[a]	.253[a]	.097
	.025	.018	.045	.073	.106[a]	.257[a]	.100
	.05	.021	.050	.081	.119	.275[a]	.109
	.10	.021	.051	.089	.131	.293	.117
	.25	.019	.048	.088	.144	.341	.128
	Mean alpha	.019	.047	.080	.121	.284	
2	.01	.068	.099	.152	.231	.389	.188
	.025	.070	.104	.158	.237	.395	.193
	.05	.071	.105	.159	.238	.398	.194
	.10	.070	.108	.165	.249	.410	.200
	.25	.064	.106	.170	.261	.426	.205
	Mean alpha	.069	.104	.161	.243	.404	
3	.01	.105	.168	.227	.315	.461	.255
	.025	.106	.173	.233	.321	.466	.260
	.05	.107	.173	.234	.323	.469	.261
	.10	.107	.174	.235	.324	.474	.263
	.25	.102	.173	.239	.331	.484	.266
	Mean alpha	.105	.172	.234	.323	.471	

[a]Actual alpha is within 1.96 standard errors of the nominal alpha.

Table 32

Actual Alphas of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 5 and s = 5

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean alpha
.33	.01	.008[a]	.019[a]	.044[a]	.070	.151	.057
	.025	.009[a]	.020[a]	.049[a]	.086[a]	.192	.071
	.05	.010[a]	.022[a]	.055[a]	.098[a]	.209	.079
	.10	.011[a]	.024[a]	.059[a]	.107[a]	.230[a]	.086
	.25	.011[a]	.025[a]	.061[a]	.108[a]	.245[a]	.090
Mean alpha		.010	.022	.054	.094	.205	
.50	.01	.006[a]	.018[a]	.032	.053	.152	.052
	.025	.008[a]	.024[a]	.045[a]	.070	.180	.065
	.05	.008[a]	.026[a]	.053[a]	.084[a]	.201	.074
	.10	.010[a]	.026[a]	.057[a]	.098[a]	.225[a]	.083
	.25	.011[a]	.026[a]	.063[a]	.109[a]	.242[a]	.088
Mean alpha		.009	.024	.050	.083	.200	
1	.01	.014[a]	.025[a]	.054[a]	.102[a]	.236[a]	.086
	.025	.017	.030[a]	.059[a]	.108[a]	.238[a]	.090
	.05	.017	.032[a]	.062[a]	.112[a]	.246[a]	.094
	.10	.019	.032[a]	.068	.124	.258[a]	.100
	.25	.019	.038	.078	.132	.275[a]	.108
Mean alpha		.017	.031	.064	.116	.251	
2	.01	.044	.082	.124	.196	.351	.159
	.025	.045	.083	.126	.197	.351	.160
	.05	.046	.083	.127	.198	.352	.161
	.10	.046	.084	.129	.200	.352	.162
	.25	.047	.086	.131	.202	.355	.164
Mean alpha		.046	.084	.127	.199	.352	
3	.01	.085	.122	.180	.252	.417	.211
	.025	.085	.123	.181	.253	.417	.212
	.05	.085	.123	.181	.253	.417	.212
	.10	.086	.124	.182	.254	.417	.213
	.25	.086	.125	.184	.254	.418	.213
Mean alpha		.085	.123	.182	.353	.417	

[a]Actual alpha is within 1.96 standard errors of the nominal alpha.

Table 33

Actual Alphas of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 5 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean alpha
.33	.01	.006[a]	.015	.030	.049	.120	.044
	.025	.006[a]	.015	.036	.071	.155	.057
	.05	.007[a]	.017[a]	.040[a]	.080	.190	.067
	.10	.007[a]	.018[a]	.045[a]	.091[a]	.212	.075
	.25	.007[a]	.018[a]	.046[a]	.095[a]	.224[a]	.078
Mean alpha		.007	.017	.039	.077	.180	
.50	.01	.003	.008	.021	.042	.136	.042
	.025	.005[a]	.016[a]	.032	.058	.157	.054
	.05	.006[a]	.017[a]	.036	.070	.175	.061
	.10	.006[a]	.017[a]	.042[a]	.086[a]	.199	.070
	.25	.007[a]	.018[a]	.046[a]	.096[a]	.224[a]	.078
Mean alpha		.005	.015	.035	.070	.178	
1	.01	.009[a]	.021[a]	.048[a]	.096[a]	.238[a]	.082
	.025	.009[a]	.021[a]	.050[a]	.100[a]	.244[a]	.085
	.05	.010[a]	.023[a]	.055[a]	.107[a]	.250[a]	.089
	.10	.011[a]	.026[a]	.059[a]	.114[a]	.259[a]	.094
	.25	.012[a]	.032[a]	.064	.120	.275[a]	.101
Mean alpha		.010	.025	.055	.107	.253	
2	.01	.051	.096	.139	.210	.395	.178
	.025	.051	.096	.139	.210	.395	.178
	.05	.051	.096	.139	.210	.395	.178
	.10	.051	.096	.142	.213	.395	.179
	.25	.051	.096	.145	.216	.398	.181
Mean alpha		.051	.096	.141	.212	.396	
3	.01	.106	.153	.211	.298	.465	.247
	.025	.106	.153	.211	.298	.465	.247
	.05	.106	.153	.211	.298	.465	.247
	.10	.106	.153	.211	.298	.465	.247
	.25	.106	.153	.213	.298	.465	.247
Mean alpha		.106	.153	.211	.298	.465	

[a]Actual alpha is within 1.96 standard errors of the nominal alpha.

Table 34

Actual Alphas of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 5 and s = 20

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean alpha
.33	.01	.006[a]	.019[a]	.030	.051	.126	.046
	.025	.007[a]	.024[a]	.040[a]	.077	.168	.063
	.05	.007[a]	.026[a]	.045[a]	.089[a]	.209	.075
	.10	.007[a]	.028[a]	.049[a]	.097[a]	.233[a]	.083
	.25	.007[a]	.028[a]	.049[a]	.100[a]	.248[a]	.086
Mean alpha		.007	.025	.043	.083	.197	
.50	.01	.005[a]	.012	.020	.037	.143	.043
	.025	.005[a]	.018[a]	.029	.054	.168	.055
	.05	.007[a]	.023[a]	.039[a]	.069	.192	.066
	.10	.007[a]	.025[a]	.045[a]	.090[a]	.212	.076
	.25	.007[a]	.028[a]	.050[a]	.101[a]	.247[a]	.087
Mean alpha		.006	.021	.037	.070	.192	
1	.01	.015[a]	.026[a]	.052[a]	.113[a]	.263[a]	.094
	.025	.018	.030[a]	.057[a]	.115[a]	.265[a]	.097
	.05	.019	.033[a]	.060[a]	.118[a]	.270[a]	.100
	.10	.019	.034[a]	.061[a]	.122	.277	.103
	.25	.019	.041	.071	.130	.292	.111
Mean alpha		.018	.033	.060	.120	.273	
2	.01	.068	.117	.159	.243	.405	.198
	.025	.068	.117	.159	.243	.405	.198
	.05	.068	.118	.160	.244	.406	.199
	.10	.068	.119	.160	.244	.406	.199
	.25	.071	.119	.161	.245	.408	.201
Mean alpha		.069	.118	.160	.244	.406	
3	.01	.127	.180	.250	.324	.475	.271
	.025	.127	.180	.250	.324	.475	.271
	.05	.127	.180	.250	.324	.475	.271
	.10	.127	.180	.250	.324	.476	.271
	.25	.127	.181	.251	.324	.476	.272
Mean alpha		.127	.180	.250	.324	.475	

[a]Actual alpha is within 1.96 standard errors of the nominal alpha.

Table 35

Actual Alphas of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 10 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha		Conditional test nominal alpha .010	.025	.050	.100	.250	Mean alpha
	.33	.01	.006[a]	.016[a]	.044[a]	.078	.200	.069
		.025	.006[a]	.017[a]	.046[a]	.088[a]	.217	.075
		.05	.006[a]	.017[a]	.048[a]	.092[a]	.225[a]	.078
		.10	.006[a]	.017[a]	.049[a]	.094[a]	.230[a]	.079
		.25	.006[a]	.017[a]	.049[a]	.094[a]	.231[a]	.079
	Mean alpha		.006	.017	.047	.089	.221	
	.50	.01	.003	.009	.023	.049	.160	.049
		.025	.006[a]	.014	.037[a]	.070	.191	.064
		.05	.006[a]	.016[a]	.044[a]	.079	.210	.071
		.10	.006[a]	.017[a]	.047[a]	.085[a]	.225[a]	.076
		.25	.006[a]	.017[a]	.049[a]	.094[a]	.230[a]	.079
	Mean alpha		.005	.015	.040	.075	.203	
	1	.01	.007[a]	.023[a]	.054[a]	.099[a]	.238[a]	.084
		.025	.007[a]	.023[a]	.054[a]	.100[a]	.240[a]	.085
		.05	.007[a]	.024[a]	.055[a]	.101[a]	.243[a]	.086
		.10	.007[a]	.025[a]	.060[a]	.105[a]	.245[a]	.088
		.25	.007[a]	.028[a]	.065	.109[a]	.257[a]	.093
	Mean alpha		.007	.025	.058	.103	.245	
	2	.01	.056	.093	.144	.215	.356	.173
		.025	.056	.093	.144	.215	.356	.173
		.05	.056	.093	.144	.215	.356	.173
		.10	.056	.093	.144	.216	.356	.173
		.25	.056	.094	.145	.216	.358	.174
	Mean alpha		.056	.093	.144	.215	.356	
	3	.01	.108	.158	.210	.290	.451	.243
		.025	.108	.158	.210	.290	.451	.243
		.05	.108	.158	.210	.290	.451	.243
		.10	.108	.158	.210	.290	.451	.243
		.25	.108	.158	.210	.290	.451	.243
	Mean alpha		.108	.158	.210	.290	.451	

[a]Actual alpha is within 1.96 standard errors of the nominal alpha.

preliminary test. Each table is read exactly as were the comparable tables for the two-tailed and upper-tailed preliminary tests.

Independence

Ideally, the lower-tailed preliminary test should point to the disaggregate unit (student) as the appropriate unit of analysis to use in testing for treatment differences.

Actual alpha levels. Sixty-seven percent of the actual alpha values of the conditional tests were within 1.96 standard errors of their nominal values, given independence and a lower-tailed preliminary test (Tables 31 through 35). All actual conditional test alphas that were not in agreement with their respective nominal counterparts were too liberal. As in the situation with independence and the two-tailed preliminary F test, the design c = 2 and s = 12 deviated from the expected. Given independence and the lower-tailed only preliminary test, only 24% of the actual alpha values for c = 2 and s = 12 (Table 31) were within 1.96 standard errors of their respective nominal values.

Estimated powers. Given $E(MS_{C:T})/E(MS_{S:CT})$ equal one, the estimated powers of the conditional tests were, in all cases, greater than the estimated powers of the $F = MS_T/MS_{C:T}$ tests (Tables 36 through 40). In addition, the effects of varying c, s, and the conditional test alpha level followed those found when the preliminary test was either a two-tailed or upper-tailed only test.

Across all designs and all preliminary test alpha levels, as the conditional test alpha level increased from .01 to .25, the power discrepancies between the "sometimes pool" and "never pool" tests decreased

from .128 to .119 to .103 to .082 to .037. But once again, this indirect relationship did not hold up within the two designs c = 2 and s = 12 and c = 5 and s = 5 (Tables 36 and 37). For example, given c = 5 and s = 5, averaging across the five preliminary alpha levels gave power discrepancies equalling .038, .053, .065, .063, and .038 as the conditional test alpha level was monotonically increased.

As was the case given the upper-tailed and two-tailed preliminary tests, a ceiling effect at the higher conditional test alpha levels prevented the power differences between the "sometimes pool" and "never pool" tests from increasing as the number of students per classroom was increased (compare Tables 37, 38, and 39). For example, given independence, a conditional test alpha of .25 and a c equal to 5, averaging across the five preliminary alpha levels the power differences equalled .038 for s = 5, .031 for s = 12, and .011 for s = 20.

It was expected that, given independence, increasing the number of classes per treatment should decrease the discrepancies between the power of the conditional test and the power of $F = MS_T/MS_{C:T}$ (compare Tables 36, 38, and 40). With the exception of the two conditional test alpha levels of .01 and .025, given c = 2 and s = 12, this trend did appear in the simulated data. For example, given $E(MS_{C:T})/E(MS_{S:CT})$ equal one, a lower-tailed preliminary test, a conditional test alpha of .25 and an s equal to 12, averaging the power differences across the preliminary test alpha levels gave power differences of .100 for c = 2, .031 for c = 5, and .004 for c = 10. The value for c = 10 (.004) is most probably spuriously low because of the extremely large powers of the $F = MS_T/MS_{C:T}$ test.

Table 36

Power of the Conditional F Test Minus Power of the Test $F = MS_T/MS_{C:T}$ Given a Lower-Tailed Preliminary Test, c = 2 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean power dif.[a]
.33	.01	-.018	-.034	-.097	-.137	-.099	-.077
	.025	.014	.006	-.057	-.108	-.087	-.046
	.05	.019	.034	-.012	-.065	-.060	-.017
	.10	.020	.055	.055	.006	-.025	.022
	.25	.013	.040	.070	.082	.028	.047
	Mean power dif.[a]	.010	.020	-.008	-.044	-.049	
.50	.01	.019	.019	.011	.001	-.046	.001
	.025	.036	.045	.033	.019	-.036	.019
	.05	.046	.069	.060	.039	-.022	.038
	.10	.046	.090	.096	.091	.012	.067
	.25	.032	.074	.109	.139	.056	.082
	Mean power dif.	.036	.059	.062	.058	-.007	
1	.01	.090	.116	.144	.157	.081	.118
	.025	.093	.121	.150	.163	.084	.122
	.05	.099	.134	.166	.177	.094	.134
	.10	.099	.145	.180	.191	.104	.144
	.25	.086	.139	.188	.208	.136	.151
	Mean power dif.	.093	.131	.166	.179	.100	
2	.01	.167	.211	.248	.249	.210	.217
	.025	.170	.216	.253	.253	.213	.221
	.05	.170	.218	.256	.257	.216	.223
	.10	.167	.226	.262	.266	.223	.229
	.25	.157	.217	.265	.273	.236	.230
	Mean power dif.	.166	.218	.257	.260	.220	
3	.01	.210	.270	.308	.292	.270	.270
	.025	.212	.271	.310	.295	.273	.272
	.05	.213	.273	.312	.299	.276	.275
	.10	.213	.276	.315	.302	.280	.277
	.25	.201	.268	.313	.307	.290	.276
	Mean power dif.	.210	.272	.312	.299	.278	

[a]Mean power differences.

Table 37

Power of the Conditional F Test Minus Power of the Test $F = MS_T/MS_{C:T}$ Given a Lower-Tailed Preliminary Test, c = 5 and s = 5

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean power dif.[a]
.33	.01	-.101	-.177	-.239	-.228	-.099	-.169
	.025	-.045	-.094	-.150	-.153	-.068	-.102
	.05	-.016	-.044	-.083	-.092	-.054	-.058
	.10	-.004	-.018	-.034	-.042	-.025	-.025
	.25	.000	.004	.002	.001	-.003	.001
	Mean power dif.[a]	-.033	-.066	-.101	-.103	-.050	
.50	.01	-.060	-.096	-.149	-.141	-.092	-.108
	.025	-.031	-.057	-.099	-.104	-.074	-.073
	.05	-.013	-.025	-.060	-.072	-.052	-.044
	.10	.005	.008	-.012	-.026	-.031	-.011
	.25	.012	.022	.015	.015	.001	.013
	Mean power dif.	-.017	-.030	-.061	-.066	-.050	
1	.01	.030	.041	.052	.047	.026	.039
	.025	.031	.043	.056	.050	.029	.042
	.05	.035	.050	.063	.056	.037	.048
	.10	.046	.059	.072	.071	.043	.058
	.25	.049	.070	.082	.091	.053	.069
	Mean power dif.	.038	.053	.065	.063	.038	
2	.01	.112	.136	.161	.172	.169	.150
	.025	.112	.137	.162	.172	.169	.150
	.05	.113	.138	.163	.172	.170	.154
	.10	.113	.138	.163	.173	.170	.151
	.25	.113	.138	.165	.176	.171	.153
	Mean power dif.	.113	.137	.163	.173	.170	
3	.01	.147	.167	.211	.229	.234	.198
	.025	.148	.167	.211	.229	.234	.198
	.05	.148	.167	.211	.230	.234	.198
	.10	.148	.168	.212	.230	.234	.198
	.25	.148	.169	.212	.230	.234	.199
	Mean power dif.	.148	.168	.211	.230	.234	

[a]Mean power differences.

Table 38

Power of the Conditional F Test Minus Power of the Test $F = MS_T/MS_{C:T}$ Given a Lower-Tailed Preliminary Test, c = 5 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean power dif.[a]
.33	.01	-.228	-.240	-.150	-.079	-.026	-.145
	.025	-.135	-.158	-.099	-.052	-.015	-.092
	.05	-.050	-.086	-.061	-.033	-.011	-.048
	.10	-.003	-.023	-.023	-.011	-.005	-.013
	.25	.018	.007	.000	-.002	.000	.005
Mean power dif.[a]		-.080	-.100	-.067	-.035	-.011	
.50	.01	-.085	-.127	-.120	-.073	-.028	-.087
	.025	-.051	-.090	-.083	-.056	-.021	-.060
	.05	-.002	-.046	-.049	-.039	-.015	-.030
	.10	.033	-.003	-.010	-.020	-.005	-.001
	.25	.053	.048	.023	.005	.004	.027
Mean power dif.		-.010	-.044	-.048	-.037	-.013	
1	.01	.153	.148	.112	.068	.027	.102
	.025	.155	.150	.116	.074	.029	.105
	.05	.164	.161	.123	.080	.030	.112
	.10	.175	.171	.136	.088	.032	.120
	.25	.179	.179	.156	.109	.035	.132
Mean power dif.		.165	.162	.129	.084	.031	
2	.01	.282	.294	.283	.242	.144	.249
	.025	.282	.294	.283	.242	.144	.249
	.05	.282	.294	.283	.242	.144	.249
	.10	.282	.294	.283	.242	.145	.249
	.25	.282	.294	.287	.242	.148	.251
Mean power dif.		.282	.294	.284	.242	.145	
3	.01	.314	.332	.350	.311	.228	.307
	.025	.314	.332	.350	.311	.228	.307
	.05	.314	.332	.350	.311	.228	.307
	.10	.314	.332	.350	.311	.228	.307
	.25	.314	.332	.350	.311	.228	.307
Mean power dif.		.314	.332	.350	.311	.228	

[a]Mean power differences.

Table 39

Power of the Conditional F Test Minus Power of the Test $F = MS_T/MS_{C:T}$ Given a Lower-Tailed Preliminary Test, c = 5 and s = 20

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean power dif.[a]
.33	.01	-.143	-.072	-.043	-.016	-.001	-.055
	.025	-.090	-.049	-.029	-.012	-.001	-.036
	.05	-.041	-.027	-.015	-.005	.000	-.018
	.10	-.004	-.009	-.008	-.001	.001	-.004
	.25	.020	.004	.002	.000	.001	.005
	Mean power dif.[a]	-.052	-.031	-.019	-.007	.000	
.50	.01	-.050	-.034	-.035	-.028	-.006	-.031
	.025	-.027	-.016	-.023	-.023	-.006	-.019
	.05	.003	.002	-.011	-.014	-.005	-.005
	.10	.040	.025	.001	-.006	-.004	.011
	.25	.065	.042	.015	.002	.000	.025
	Mean power dif.	.006	.004	-.011	-.014	-.004	
1	.01	.224	.179	.111	.059	.009	.116
	.025	.227	.182	.114	.063	.010	.119
	.05	.229	.187	.118	.064	.010	.122
	.10	.238	.193	.127	.067	.011	.127
	.25	.245	.198	.132	.072	.014	.132
	Mean power dif.	.233	.188	.120	.065	.011	
2	.01	.399	.378	.304	.229	.103	.283
	.025	.400	.378	.304	.229	.103	.283
	.05	.400	.379	.305	.230	.103	.283
	.10	.400	.379	.305	.230	.104	.284
	.25	.401	.382	.306	.230	.105	.285
	Mean power dif.	.400	.379	.305	.230	.104	
3	.01	.443	.428	.401	.305	.180	.351
	.025	.443	.428	.401	.305	.180	.351
	.05	.443	.428	.401	.305	.180	.351
	.10	.444	.428	.401	.305	.180	.352
	.25	.444	.428	.401	.306	.180	.352
	Mean power dif.	.443	.428	.401	.305	.180	

[a]Mean power differences.

Table 40

Power of the Conditional F Test Minus Power of the Test $F = MS_T/MS_{C:T}$ Given a Lower-Tailed Preliminary Test, c = 10 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250	Mean power dif.[a]
.33	.01	-.033	-.015	-.007	-.001	.000	-.011
	.025	-.014	-.005	-.003	-.001	.000	-.005
	.05	-.004	-.003	-.002	-.001	.000	-.002
	.10	-.001	.000	.000	.000	.000	.000
	.25	.001	.000	.000	.000	.000	.000
	Mean power dif.[a]	-.010	-.005	-.002	-.001	.000	
.50	.01	-.070	-.043	-.021	-.011	-.002	-.029
	.025	-.047	-.025	-.013	-.008	.000	-.019
	.05	-.025	-.015	-.007	-.006	.000	-.011
	.10	-.009	-.003	-.002	-.004	.000	-.004
	.25	.005	.002	.001	.000	.000	.002
	Mean power dif.	-.029	-.017	-.008	-.006	.000	
1	.01	.103	.057	.034	.017	.004	.043
	.025	.106	.059	.035	.017	.004	.062
	.05	.111	.061	.036	.017	.004	.046
	.10	.116	.063	.036	.020	.004	.048
	.25	.127	.074	.041	.023	.005	.054
	Mean power dif.	.113	.063	.036	.019	.004	
2	.01	.374	.319	.233	.154	.063	.229
	.025	.374	.319	.233	.154	.063	.229
	.05	.374	.319	.233	.154	.063	.229
	.10	.374	.319	.233	.155	.063	.229
	.25	.374	.319	.235	.155	.063	.229
	Mean power dif.	.374	.319	.233	.154	.063	
3	.01	.416	.412	.358	.268	.114	.314
	.025	.416	.412	.358	.268	.114	.314
	.05	.416	.412	.358	.268	.114	.314
	.10	.416	.412	.358	.268	.114	.314
	.25	.416	.412	.358	.268	.114	.314
	Mean power dif.	.416	.412	.358	.268	.114	

[a]Mean power differences.

The effect of increasing the nominal alpha level of the lower-tailed preliminary test on the power differences between the "sometimes pool" and "never pool" F tests were reversed from the trends found given both a two-tailed and an upper-tailed test. That is, as the preliminary test alpha level increased, there was an increase in the power differences. For example, given a conditional alpha level of .25, c = 2 and s = 12 (Table 36), the power differences, as the preliminary test alpha level increased, came to .081, .084, .094, .104, and .136, favoring the "sometimes pool" test. This reversal in trend probably was due to the fact that the alpha levels of the "sometimes pool" test leaned further and further toward being too liberal as the alpha level of the preliminary test increased.

Without exception, given independence of disaggregate units, the estimated powers of the "sometimes pool" test (Tables E-1 through E-5) were greater than the powers of the "always pool" or $F = MS_T/MS_{S:T}$ test (Table 9). The larger powers of the "sometimes pool" test could be expected when the conditional test was too liberal a test, but they also occurred when the actual alpha levels of the conditional test were satisfactory.

Positive Dependence

Given that $E(MS_{C:T})$ is greater than the $E(MS_{S:CT})$, the lower-tailed preliminary F test should not reject its null hypothesis of H_o: the $E(MS_{C:T})$ is greater than or equal to the $E(MS_{S:CT})$ and thus designate the disaggregate unit (student) as the appropriate unit of analysis for the primary test of treatment effects.

Actual alpha levels. Given positive dependence, all actual conditional test alpha levels exceeded their respective nominal values by more than 1.96 standard errors (see bottom sections of Tables 31 through 35). If one compares Tables 31 through 35 with Table 8, which documents the actual alphas of the $F = MS_T/MS_{S:T}$ and has no preliminary test associated with it, one finds that the "sometimes pool" alphas, at both levels of positive dependence, are very close in magnitude to their counterpart "always pool" alphas. This was especially so for the three designs with large (sc-1)t values, i.e., c = 5 and s = 12, c = 5 and s = 20, and c = 10 and s = 12. Given $E(MS_{C:T})/E(MS_{S:CT})$ equal 2, c = 10 and s = 12, the actual "sometimes pool" alphas, averaged across the five preliminary test alphas, equalled .056, .093, .144, .215, and .356 as the nominal "sometimes pool" test alphas increased from .01 to .25 (Table 35), while the actual "always pool" alphas also equalled .056, .093, .144, .215, and .356, respectively (Table 8).

The preliminary test is the distinguishing feature separating the "sometimes pool" F test from the "always pool" F test. And given either degree of positive dependence, the alpha level of the lower-tailed preliminary test had little to no affect on the actual alpha of the "sometimes pool" test. For c = 2 and s = 12, as the alpha level of the preliminary test increased, the alpha levels of the "sometimes pool" tests became slightly more liberal. For example, given $E(MS_{C:T})/E(MS_{S:CT})$ equal 2, a conditional test nominal alpha of .25, c = 2 and s = 12, as the five alpha levels of the preliminary test increased from .01 to .25, the five actual conditional test alphas

went from .389 to .395 to .398 to .410 to .426 (Table 31). Given the same circumstances, the actual conditional test alpha levels for designs with higher values of (sc-1)t had almost no variation as the preliminary test alpha level was varied. Because the actual alphas, given a lower-tailed preliminary test and positive dependence, closely match those found when studying the empirical alphas of the $F = MS_T/MS_{S:T}$ test, given positive dependence, no more will be said about the former.

Estimated powers. Given positive dependence and a lower-tailed preliminary test, the "sometimes pool" F test of treatment effects essentially becomes an "always pool" F test of treatment effects. Because of this, one would expect their powers to react similarly both in magnitude and in direction. This is exactly what happened. The nominal alpha of the preliminary test did have minimal effect on the power of the "sometimes pool" test for c = 2 and s = 12 though (Table E-1). This was expected, however, as this was the design for which increasing the preliminary test alpha level made the difference between the "sometimes pool" alpha and the "always pool" alpha slightly more than zero, making the former more liberal. Because the "sometimes pool" test, based on the lower-tailed preliminary test, and the "always pool" test are virtually the same, given positive dependence, no more will be said about the power of the "sometimes pool" test.

Negative Dependence

The lower-tailed preliminary test should recognize the situation where the $E(MS_{C:T})$ is less than the $E(MS_{S:CT})$ and by so doing designate

the aggregate unit (classroom) as the appropriate analytic unit in testing for treatment differences.

Actual alpha levels. Given that the $E(MS_{C:T})$ was less than the $E(MS_{S:CT})$ and the preliminary test was a lower-tailed only test done at the α level, the observed alphas of the conditional test were very nearly equal in magnitude to observed alphas of the conditional test given negative dependence and a two-tailed preliminary test done at 2α. For example, compare the observed alphas of the conditional test in Table 32 with the observed alphas in Table 12. Given $E(MS_{C:T})/E(MS_{S:CT})$ equals .50, 21 of the 25 observed alphas, given the lower-tailed preliminary test (Table 32) done at α, equal their counterpart observed alphas, given the two-tailed preliminary test done at 2α (Table 12). And four times the observed alphas of the conditional test given the lower-tailed preliminary test exceeded their matched observed alphas given the two-tailed preliminary test by only .001. In summary, given negative dependence and a lower-tailed preliminary test, the consequences of varying s, c, the preliminary test alpha, the conditional test nominal alpha, and the level of negative dependence on the observed alphas of the conditional test imitated the magnitude and pattern of effects found when studying the effects of using a two-tailed preliminary test on the observed conditional test alphas.

Estimated powers. It was expected that, given the $E(MS_{C:T})$ was less than the $E(MS_{S:CT})$, the power of the conditional test following a lower-tailed preliminary test done at α would equal the power of the conditional test given a two-tailed preliminary test done at 2α.

Under this condition, both preliminary tests have equal probabilities of rejecting the null hypothesis of the preliminary test. To see that this expectation appeared in the simulated data, compare Tables 16 through 20 with Tables 36 through 40 and/or Tables C-1 through C-5 with Tables E-1 through E-5. For example, if $E(MS_{C:T})/E(MS_{S:CT})$ equals .33, the alpha level of the two-tailed preliminary test equal .02, c = 5 and s = 5, as the nominal alpha of the conditional test increased from .01 to .25, the power differences between the "sometimes pool" test and the "never pool" test equalled -.101, -.177, -.239, -.228, and -.099 (Table 17). If the preliminary test was instead a lower-tailed only test done at alpha equal .01 (Table 37), the power discrepancies between the "sometimes pool" and "never pool" tests also equalled -.101, -.177, -.239, -.228, and -.099 as the conditional test nominal alpha was increased from .01 to .25.

Given negatively dependent data, the power of the conditional test following a two-tailed preliminary test done at α had less power than the conditional test following a lower-tailed preliminary test done at α. For example, Table C-2 empirically shows that given $E(MS_{C:T})/E(MS_{S:CT})$ equals .50, c = 5, s = 5 and the two-tailed preliminary test alpha equal .05, the power of the conditional test equalled .131, .230, .321, .488, and .739 as the nominal alpha of the conditional test increased from .01 to .25. Given the same set of conditions but instead having the lower-tailed preliminary test alpha equal .05, the power of the conditional test following the lower-tailed preliminary test equalled .149, .262, .360, .520, and .761

as the conditional alphas increased (Table E-2). Given both the lower-tailed and two-tailed tests were done at the same alpha level and the $E(MS_{C:T})$ was less than the $E(MS_{S:CT})$, the lower-tailed preliminary test would have more power to reject the null preliminary test hypothesis and thus designate class as the unit of analysis more often. Since negative dependence is defined by having the $E(MS_{C:T})$ less than the $E(MS_{S:CT})$, using class as the analytic unit decreases the error term in testing for treatment effects, which should increase the power of the conditional test. At the same time, however, using student as the analytic unit has as its advantage more degrees of freedom error, which also has the effect of increasing the power of the conditional test. Thus that the power of the conditional test following a two-tailed preliminary test done at α had less power than the conditional test following a lower-tailed preliminary test done at α was a combined function of the difference between c and s and the difference between the $E(MS_{C:T})$ and the $E(MS_{S:CT})$.

CHAPTER VIII

SUMMARY AND CONCLUSIONS

The main purpose of this study was to propose an operational definition of independence of analytic units, and to use that definition in an investigation of the effects of violating the assumption of independence. A secondary purpose was to expand upon a conditional testing procedure that had been proposed in past research and to test its validity. This conditional testing procedure included a preliminary test of independence that was used to select the unit of analysis to use in testing the primary hypothesis of no treatment differences. How the number of individual units per group, the number of groups per treatment level and the type and degree of dependence within groups affected both the validity of using correlated units of analysis and the consequences of using a conditional testing procedure were studied analytically. In addition, the size of the effects were estimated empirically.

It was proposed that independence be operationally defined as: Disaggregate or ungrouped units can be considered as independent units whenever the variance of the aggregate or grouped units can be predicted from the grouping size and the variance of the disaggregate units. This definition of independence can be translated into testing the equality of expected mean squares between and within groups. That is, given the above definition of independence, the expected mean square between

groups, $E(MS_{C:T})$, should equal the expected mean square within groups, $E(MS_{S:CT})$, for the two-level, hierarchically-nested design considered in this study, which within each treatment had subjects grouped into classrooms. Also, under independence of responses between and within groups, the intraclass correlation coefficient would equal zero.

Given the above operational definition of independence, two types of dependency are possible, positive dependence and negative dependence. Positive dependence was defined by the expected mean square between groups being larger than the expected mean square within groups, or similarly by the intraclass correlation coefficient being greater than zero. Negative dependence was defined by the expected mean square within groups being larger than the expected mean square between groups, or similarly by the intraclass correlation coefficient being less than zero. Negative dependence can occur only when subjects within groups are considered fixed, while positive dependence can occur with subjects considered as either fixed or random. Either type of dependence can be caused by an additive effect (which influences the variation between groups), a proportional effect (which influences the variation within groups), and nonrandom assignment, which most probably will result in either positive or negative dependence. Random assignment of students to classrooms does not perforce erase the possibility of positive or negative dependence occurring. It can and does happen that units randomly assigned to treatment conditions do not receive the treatments independently and thus the randomly assigned units are not independent of each other.

Given a definition of independence which is measurable, both the analytic and empirical consequences of using correlated units of analysis were considered. The empirical estimation part of the study was done to see if hypothesized effects of correlated units on alphas and powers were large enough to affect practice. In some respects, the parameter values specified for the Monte Carlo portion of this study limit generalizations of empirical effects. However, care was taken to select parameter values held to be common in educational data.

The basic research design had a balanced, two-level hierarchically-nested structure, with students nested within classrooms and classrooms nested within treatments. Two analysis of variance models were used to analyze data fitting this general design. Classroom was the analytic unit for one model, while student was the analytic unit for the other model. The model having classroom as the unit of analysis was called the "never pool" model. The model with student as the unit of analysis was called the "always pool" model as its mean square error term was a pooled or weighted sum of the mean square between classrooms and the mean square within classrooms. Parameters that were allowed to vary included the number of students per classroom, the number of classrooms per treatment and the type and degree of dependence within each classroom. Each combination of number of students per classroom (s) and number of classrooms per treatment (c) was called a design. In total, the empirical study considered five designs and within each design, data for each of 1,000 samples were altered such to create defined types and degrees of dependence. The three basic assumptions of normality,

independence, and homoscedasticity held for all simulated classroom means; however, observations within each group or classroom were controlled only to the extent that they were normally distributed and had equal variances.

Given independence of student responses, both the analytic and empirical analyses showed that either student or classroom can be used as the unit of analysis. That is, both the "never pool" test, $F = MS_T/MS_{C:T}$, and the "always pool" test, $F = MS_T/MS_{S:T}$, are appropriate tests of treatment effects. All empirical alphas for both tests were within 1.96 standard errors of their nominal values. $F = MS_T/MS_{S:T}$ is the preferable test, however, as it always had more power than did the test $F = MS_T/MS_{C:T}$. Over the five designs and the five nominal alpha levels, the power of the "always pool" test exceeded that of the "never pool" test by an average of .081. The "always pool" test's advantage in power ranged from a low of .004 ($c = 10$, $s = 12$, $\alpha = .25$) to a high of .223 ($c = 5$, $s = 20$, $\alpha = .01$). Given a seemingly appropriate design for elementary school research ($c = 5$ and $s = 20$), the power discrepancy at $\alpha = .05$ between the "never pool" test and the "always pool" test was .111, favoring the "always pool" test, which has student as the unit of analysis. Given independence, the discrepancy in power between these two ANOVA tests comes solely from the difference in degrees of freedom error for each of the F tests.

Given positive dependence, where the $E(MS_{C:T})$ was greater than the $E(MS_{S:CT})$, the proper unit of analysis is the grouped unit, classrooms. Using the disaggregate unit as the unit of analysis caused the

pooled error term, $E(MS_{S:T})$, to be biased on the small side. And this attenuation caused the "always pool" test, $F = MS_T/MS_{S:T}$, to be too liberal. For both simulated degrees of positive dependence, none of the empirical alphas were within 1.96 standard errors of their respective nominal value. This suggests that the effect of positively correlated units can result in spurious significance. In addition, the empirical magnitudes of the differences between the nominal and actual alphas were of sufficient size to suggest that having positively correlated units of analysis results in negative effects that are of meaningful and practical importance. The liberalness increased as the number of students per class increased and as the degree of positive dependence increased, i.e., the ratio of the $E(MS_{C:T})$ over the $E(MS_{S:CT})$ increased above one. The liberalness decreased as the number of classes per treatment increased. Because of the general liberalness of the "always pool" test, given positive dependence, the empirical power of that test was spuriously high.

Given negative dependence, where the $E(MS_{C:T})$ was less than the $E(MS_{S:CT})$, the correct unit of analysis is once again the grouped unit. Using the ungrouped unit as the unit of analysis caused the pooled error term, $E(MS_{S:T})$, to be biased on the high side. That is, the "always pool" test, $F = MS_T/MS_{S:T}$, was too conservative a test. For both simulated degrees of negative dependence, none of the empirical alphas were within 1.96 standard errors of their respective nominal alpha. This suggests that the effect of negatively correlated units of analysis can result in spurious lack of significance. Here too,

the actual magnitudes of the discrepancies between the nominal and empirical alpha values were of such size to indicate that having negatively correlated units results in negative effects that are of meaningful importance. While the analytic analysis suggested that increasing the number of students per class should increase the conservativeness and increasing the number of classes per treatment should decrease any conservativeness, the empirical analysis found no clear trends. This lack of trend may have been due, in part, to a "floor effect." Decreasing the ratio of the $E(MS_{C:T})$ over the $E(MS_{S:CT})$ to below one (which is the same thing as increasing negative dependence within the data) did, however, increase the conservativeness of the "always pool" F statistic, as expected. The conservativeness spuriously reduced the empirical power of the "always pool" test to such an extent that, in all simulated cases, but one, the advantage of using the "always pool" test in the first place (more degrees of freedom error) was cancelled out.

What do the above analytic and empirical results suggest for the practitioner? They suggest that when dealing with educational data, in almost all cases, the grouped unit, such as classrooms, should be the unit of analysis. If, however, the data do happen to be independent of each other, it is clearly advantageous to use the individual unit as the unit of analysis. The results indicated quite convincingly that the F test is not robust to violations of the assumption of independence, even for small degrees of dependence. This conclusion should be kept in mind both when designing and analyzing experiments

as well as when interpreting the results from studies which have not followed the advice from this investigation. Closely tied to this, of course, is a real need to empirically determine the types and degrees of dependence most common in real world, educational data.

One might ask next, Is it feasible to first do an initial test of independence, and then on the basis of that test choose a unit of analysis for the primary test of treatment effects? This study showed clearly, both analytically and empirically, that the answer to this question is no. Two criteria were used to judge the adequacy of the conditional testing procedure, which is also called the "sometimes pool" test. First, the empirical alpha should be close to its respective nominal value. And second, the power of the conditional or "sometimes pool" test should be greater than the power of the always correct, "never pool" test, $F = MS_T/MS_{C:T}$.

Actually three different preliminary tests of independence were considered (one at a time) under the "sometimes pool" procedure. The first was a two-tailed preliminary test which tested for the inequality of the mean square between and the mean square within classroom error terms. The second, an upper-tailed preliminary test, tested whether the mean square between classes was larger than the mean square within classes. The third preliminary test, a lower-tailed test, tested for the condition that the mean square within classrooms was larger than the mean square between classrooms.

Only when independence of student responses occurred did the conditional testing procedure turn out to be very effective. For

that condition and given the two-tailed, the upper-tailed only and the lower-tailed only preliminary tests, 80% of the empirical conditional test alphas were within 1.96 standard errors of the nominal alphas and 96.5% of the powers of the conditional test were greater than comparable powers of the "never pool" test, $F = MS_T/MS_{C:T}$. Across the three types of preliminary tests and across all five designs, the difference between the power of the "sometimes pool" F test and the "never pool" F test averaged .073, favoring the "sometimes pool" test. Given any of the three preliminary F tests, the discrepancy between the powers of the "sometimes pool" test and the powers of the "never pool" test decreased as the number of groups per treatment increased and as the alpha levels of the "sometimes pool" and "never pool" test increased. On the other hand, the discrepancy increased as the number of students per group increased. For both the two-tailed and upper-tailed preliminary test situations, the discrepancy between the powers of the "sometimes pool" and the "never pool" tests decreased as the alpha level of the preliminary test increased. However, this trend (which was expected) was reversed when the lower-tailed preliminary test was used, most likely because the observed alpha levels of the "sometimes pool" test tended to become too liberal as the preliminary test alpha level increased.

Given positive dependence, the conditional F test generally turned out to be too liberal a test and thus had spuriously high empirical power. Given either the two-tailed or upper-tailed preliminary tests, the liberalness of the "sometimes pool" statistic decreased as the

number of groups per treatment increased, as the degree of positive dependence increased, as the alpha level of the preliminary test increased and as the alpha level of the "sometimes pool" test increased. Increasing the number of students per group had no simple effect on the liberalness of the "sometimes pool" test. Across the five designs considered and given the two-tailed preliminary F test of independence, 22% and 35.2% of the "sometimes pool" or conditional F tests were, however, robust to the occurrence of positive dependence, where $E(MS_{C:T})/E(MS_{S:CT})$ equalled 2 and 3, respectively. For those simulated circumstances, where the "sometimes pool" test did empirically appear to be robust to the occurrence of positively correlated analytic units, the difference between the empirical powers of the "sometimes pool" and the "never pool" tests averaged only .014, favoring the "sometimes pool" test. Given positive dependence, the magnitude of effects for the conditional test, following either a two-tailed preliminary test done at 2α or an upper-tailed preliminary test done at α, were comparable. On the other hand, the "sometimes pool" F test statistic, given positive dependence and a lower-tailed preliminary test, was distributed as the $F = MS_T/MS_{S:T}$ test statistic as varying the alpha level of the preliminary test had negligible affect on the "sometimes pool" test statistic.

Given negative dependence, the conditional test was somewhat conservative and generally had less power than the "never pool" $F = MS_T/MS_{C:T}$ test. Given either the two-tailed or lower-tailed preliminary tests, the conservativeness of the "sometimes pool"

statistic decreased as the degree of negative dependence increased, as the alpha level of the preliminary test increased and as the alpha level of the "sometimes pool" test decreased. And also as expected, increasing the number of students per group again had no simple effect on the conservativeness of the "sometimes pool" test. The analytic analysis suggested that increasing the number of groups per treatment should decrease the conservativeness of the "sometimes pool" test statistic, but the simulated data were weak in confirming this expected trend. Across the designs considered and given the two-tailed preliminary F test, 60.8% and 72% of the "sometimes pool" or conditional F tests were empirically robust to the occurrence of negative dependence, where $E(MS_{C:T})/E(MS_{S:CT})$ equalled .50 and .33, respectively. For those specific empirically robust instances, the differences between the empirical powers of the "sometimes pool" and the "never pool" tests averaged -.010, favoring the "never pool" test. Given the $E(MS_{C:T})$ was less than the $E(MS_{S:CT})$, the magnitude of effects for the conditional test, following either a two-tailed preliminary test done at 2α or a lower-tailed preliminary test done at α, were comparable. The "sometimes pool" F test statistic, given negative dependence and an upper-tailed preliminary test of independence, was distributed like the "always pool" F test statistic as varying the alpha levels of the preliminary test had essentially no affect on the "sometimes pool" test statistic.

In summary, this study shows that, in an hierarchically-nested design with one outcome measure per subject, as a general rule of

thumb any preliminary test of independence should not be used to choose a unit of analysis to test for treatment differences. That is, the decision of what analytic unit to use should not be based on the test $F = MS_{C:T}/MS_{S:CT}$. If individual units are independent, the "always pool" $F = MS_{T}/MS_{S:T}$ test is best. If individual data are not independent, the "never pool" $F = MS_{T}/MS_{C:T}$ test is best. However, the problem is the researcher is never actually in the position of knowing before the analysis stage whether or not responses of subjects nested within groups are, in fact, independent responses. And this in and of itself suggests that the researcher should in general be using the grouped unit as his unit of analysis, at least in educational research where dependence most probably is the rule rather than the exception.

APPENDICES

APPENDIX A

PRELIMINARY ANALYSIS OF SIMULATED DATA

Table A-1

Distribution of Sample Classroom Means and Student Observations

Classroom Means (N = 2,000)		Student Observations (N = 10,000)					
			Observed				
			$E(MS_{C:T})/E(MS_{S:CT})$				
Expected	Observed	Expected	.33	.50	1	2	3
22	28	110	105	109	113	119	120
24	19	120	130	128	138	135	126
44	40	220	239	226	201	185	201
104	92	520	471	486	527	551	547
290	301	1,450	1,443	1,439	1,420	1,441	1,444
516	508	2,580	2,610	2,603	2,587	2,547	2,553
516	532	2,580	2,579	2,583	2,591	2,605	2,571
290	310	1,450	1,481	1,451	1,442	1,428	1,441
104	99	520	501	527	536	541	550
44	33	220	219	233	237	234	231
24	17	120	131	121	109	117	118
22	21	110	91	94	99	97	97
χ^2	11.92		13.43	6.42	9.16	14.43	8.50

Table A-2

Moments for Student within Treatment Type Data when c = 2 and s = 12

$E(MS_{C:T})/$ $E(MS_{S:CT})$	Central					Noncentral
	Treatment	Mean	Variance	Skew	Kurtosis	Mean
.33	1	.0049	.9456	-.0111	.0728	.4049
.33	2	-.0043	.9510	.0206	.0079	
.50	1	.0060	.9589	-.0106	.0732	.4060
.50	2	-.0052	.9652	.0211	.0076	
1	1	.0085	.9988	-.0097	.0724	.4085
1	2	-.0074	1.0076	.0213	.0069	
2	1	.0121	1.0785	-.0094	.0683	.4121
2	2	-.0104	1.0924	.0193	.0061	
3	1	.0148	1.1583	-.0010	.0640	.4148
3	2	-.0128	1.1772	.0162	.0056	

Table A-3

Moments for Student within Treatment Type Data when c = 5 and s = 5

$E(MS_{C:T})/$ $E(MS_{S:CT})$	Central					Noncentral
	Treatment	Mean	Variance	Skew	Kurtosis	Mean
.33	1	-.0019	.8708	.0002	-.0053	.3981
.33	2	-.0044	.8800	.0038	-.0201	
.50	1	-.0024	.9044	.0035	.0023	.3976
.50	2	-.0053	.9129	.0023	-.0202	
1	1	-.0033	1.0052	.0093	.0178	.3967
1	2	-.0075	1.0115	-.0001	-.0204	
2	1	-.0047	1.2067	.0133	.0320	.3953
2	2	-.0107	1.2087	-.0018	-.0208	
3	1	-.0058	1.4082	.0139	.0367	.3942
3	2	-.0131	1.4059	-.0020	-.0212	

Table A-4

Moments for Student within Treatment Type Data when c = 5 and s = 12

$E(MS_{C:T})/$ $E(MS_{S:CT})$	Central					Noncentral
	Treatment	Mean	Variance	Skew	Kurtosis	Mean
.33	1	-.0003	.9466	-.0040	-.0083	.3997
.33	2	-.0025	.9375	-.0046	.0053	
.50	1	-.0003	.9605	-.0029	-.0075	.3997
.50	2	-.0031	.9515	-.0055	.0064	
1	1	-.0004	1.0023	.0000	-.0063	.3996
1	2	-.0044	.9935	-.0076	.0081	
2	1	-.0006	1.0858	.0046	-.0061	.3994
2	2	-.0062	1.0774	-.0108	.0086	
3	1	-.0008	1.1694	.0084	-.0069	.3992
3	2	-.0076	1.1614	-.0133	.0072	

Table A-5

Moments for Student within Treatment Type Data when c = 5 and s = 20

$E(MS_{C:T})/$ $E(MS_{S:CT})$	Central					Noncentral
	Treatment	Mean	Variance	Skew	Kurtosis	Mean
.33	1	.0003	.9649	.0138	.0075	.4003
.33	2	-.0010	.9577	.0067	-.0104	
.50	1	.0004	.9734	.0138	.0078	.4004
.50	2	-.0013	.9658	.0067	-.0101	
1	1	.0006	.9988	.0137	.0083	.4006
1	2	-.0018	.9900	.0065	-.0095	
2	1	.0008	1.0497	.0136	.0085	.4008
2	2	-.0026	1.0385	.0062	-.0087	
3	1	.0010	1.1007	.0135	.0086	.4010
3	2	-.0031	1.0870	.0058	-.0083	

Table A-6

Moments for Student within Treatment Type Data when c = 10 and s = 12

$E(MS_{C:T})/$ $E(MS_{S:CT})$	Central					Noncentral
	Treatment	Mean	Variance	Skew	Kurtosis	Mean
.33	1	.0006	.9490	.0017	-.0048	.4006
.33	2	-.0018	.9492	.0110	.0093	
.50	1	.0007	.9628	.0019	-.0050	.4007
.50	2	-.0022	.9632	.0117	.0113	
1	1	.0010	1.0044	.0026	-.0053	.4010
1	2	-.0031	1.0051	.0130	.0153	
2	1	.0014	1.0874	.0037	-.0053	.4014
2	2	-.0043	1.0889	.0143	.0191	
3	1	.0017	1.1705	.0047	-.0052	.4017
3	2	-.0053	1.1728	.0147	.0206	

Table A-7

Three Distributional Statistics[a] for Standardized Mean Square Values, Given c = 2 and s = 12

		Mean		Variance		Skew	
Central case:							
MS_T	Mean[a]	1.00	(1.00)[b]	2.30	(2.00)	3.00	(2.83)
$MS_{C:T}$	Mean	2.00	(2.00)	3.80	(4.00)	1.90	(2.00)
$MS_{S:CT}$	Mean	44.00	(44.00)	93.00	(88.00)	0.45	(0.43)
$MS_{S:T}$	Mean	46.00	(46.00)	97.20	(92.00)	0.47	(0.42)
Noncentral case:							
MS_T	Mean	1.06		1.28		1.86	
$MS_{C:T}$	Mean	2.00		3.80		1.90	
$MS_{S:CT}$	Mean	44.00		93.00		0.45	
$MS_{S:T}$	Mean	46.00		97.20		0.47	

[a]The values for the distributional statistics were averaged over the four conditions of dependence and the independence condition.

[b]Parenthesized values are the theoretical distributional properties.

Table A-8

Three Distributional Statistics[a] for Standardized Mean Square Values, Given c = 5 and s = 5

		Mean		Variance		Skew	
Central case:							
MS_T	Mean[a]	0.96	(1.00)[b]	2.10	(2.00)	3.10	(2.83)
$MS_{C:T}$	Mean	7.90	(8.00)	17.00	(16.00)	0.93	(1.00)
$MS_{S:CT}$	Mean	40.00	(40.00)	85.00	(80.00)	0.42	(0.45)
$MS_{S:T}$	Mean	48.00	(48.00)	109.80	(96.00)	0.43	(0.41)
Noncentral case:							
MS_T	Mean	1.00		0.98		1.76	
$MS_{C:T}$	Mean	7.90		17.00		0.93	
$MS_{S:CT}$	Mean	40.00		85.00		0.42	
$MS_{S:T}$	Mean	48.00		109.80		0.43	

[a]The values for the distributional statistics were averaged over the four conditions of dependence and the independence condition.

[b]Parenthesized values are the theoretical distributional properties.

Table A-9

Three Distributional Statistics[a] for Standardized Mean Square Values, Given c = 5 and s = 12

		Mean	Variance	Skew
Central case:				
MS_T	Mean[a]	0.96 (1.00)[b]	1.80 (2.00)	2.80 (2.83)
$MS_{C:T}$	Mean	8.10 (8.00)	15.00 (16.00)	0.94 (1.00)
$MS_{S:CT}$	Mean	110.00 (110.00)	210.00 (220.00)	0.28 (0.27)
$MS_{S:T}$	Mean	120.00 (118.00)	234.00 (236.00)	0.32 (0.26)
Noncentral case:				
MS_T	Mean	1.00	0.68	1.32
$MS_{C:T}$	Mean	8.10	15.00	0.94
$MS_{S:CT}$	Mean	110.00	210.00	0.28
$MS_{S:T}$	Mean	120.00	234.00	0.32

[a]The values for the distributional statistics were averaged over the four conditions of dependence and the independence condition.

[b]Parenthesized values are the theoretical distributional properties.

Table A-10

Three Distributional Statistics[a] for Standardized Mean Square Values, Given c = 5 and s = 20

		Mean		Variance		Skew	
Central case:							
MS_T	Mean[a]	1.00	(1.00)[b]	2.10	(2.00)	2.80	(2.83)
$MS_{C:T}$	Mean	8.10	(8.00)	16.00	(16.00)	0.92	(1.00)
$MS_{S:CT}$	Mean	190.00	(190.00)	370.00	(380.00)	0.21	(0.21)
$MS_{S:T}$	Mean	200.00	(198.00)	404.00	(396.00)	0.20	(0.20)
Noncentral case:							
MS_T	Mean	1.00		0.50		1.07	
$MS_{C:T}$	Mean	8.10		16.00		0.92	
$MS_{S:CT}$	Mean	190.00		370.00		0.21	
$MS_{S:T}$	Mean	200.00		404.00		0.20	

[a]The values for the distributional statistics were averaged over the four conditions of dependence and the independence condition.

[b]Parenthesized values are the theoretical distributional properties.

Table A-11

Three Distributional Statistics[a] for Standardized Mean Square Values, Given c = 10 and s = 12

		Mean	Variance	Skew
Central case:				
MS_T	Mean[a]	0.95 (1.00)[b]	1.90 (2.00)	2.50 (2.83)
$MS_{C:T}$	Mean	18.00 (18.00)	33.00 (36.00)	0.72 (0.67)
$MS_{S:CT}$	Mean	220.00 (220.00)	470.00 (440.00)	0.26 (0.19)
$MS_{S:T}$	Mean	240.00 (238.00)	532.00 (476.00)	0.26 (0.18)
Noncentral case:				
MS_T	Mean	1.00	0.42	1.01
$MS_{C:T}$	Mean	18.00	33.00	0.72
$MS_{S:CT}$	Mean	220.00	470.00	0.26
$MS_{S:T}$	Mean	240.00	532.00	0.26

[a]The values for the distributional statistics were averaged over the four conditions of dependence and the independence condition.

[b]Parenthesized values are the theoretical distributional properties.

APPENDIX B

ESTIMATED ALPHAS OF THE RESCALED F STATISTIC

Table B-1

Estimated Type I Errors for $F = MS_T/MS_{S:T}$ Using a Rescaled F Statistic

$E(MS_{C:T})/E(MS_{S:CT})$	c	s	d.f. error	.010	.025	.050	.100	.250	Mean alpha
	2	12	(46)	.000	.000	.001	.006	.052	.012
	5	5	(48)	.000	.000	.002	.008	.062	.014
.33	5	12	(118)	.000	.000	.001	.006	.052	.012
	5	20	(198)	.000	.000	.001	.005	.049	.011
	10	12	(238)	.000	.000	.001	.006	.052	.012
	Mean alpha			.000	.000	.001	.006	.053	
	2	12	(46)	.001	.002	.007	.023	.110	.029
	5	5	(48)	.001	.003	.009	.027	.121	.032
.50	5	12	(118)	.000	.002	.007	.023	.111	.029
	5	20	(198)	.000	.002	.006	.022	.108	.028
	10	12	(238)	.000	.002	.007	.023	.111	.029
	Mean alpha			.000	.002	.007	.024	.112	
	2	12	(46)	.010[a]	.025[a]	.050[a]	.100[a]	.250[a]	.087
	5	5	(48)	.010[a]	.025[a]	.050[a]	.100[a]	.250[a]	.087
1	5	12	(118)	.010[a]	.025[a]	.050[a]	.100[a]	.250[a]	.087
	5	20	(198)	.010[a]	.025[a]	.050[a]	.100[a]	.250[a]	.087
	10	12	(238)	.010[a]	.025[a]	.050[a]	.100[a]	.250[a]	.087
	Mean alpha			.010	.025	.050	.100	.250	
	2	12	(46)	.058	.101	.153	.231	.414	.191
	5	5	(48)	.046	.084	.131	.206	.388	.171
2	5	12	(118)	.058	.100	.151	.228	.409	.189
	5	20	(198)	.062	.105	.157	.235	.416	.195
	10	12	(238)	.058	.099	.150	.227	.408	.188
	Mean alpha			.056	.098	.148	.225	.407	
	2	12	(46)	.113	.170	.232	.318	.494	.265
	5	5	(48)	.080	.130	.186	.269	.450	.239
3	5	12	(118)	.110	.165	.226	.310	.486	.259
	5	20	(198)	.120	.177	.238	.333	.497	.273
	10	12	(238)	.112	.167	.227	.312	.487	.261
	Mean alpha			.107	.162	.222	.308	.483	

[a]Estimated alpha is within 1.96 standard errors of the nominal alpha.

APPENDIX C

POWERS OF THE CONDITIONAL F GIVEN A TWO-TAILED PRELIMINARY TEST

Table C-1

Power of the Conditional F Test Given a Two-Tailed Preliminary Test, c = 2 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha		Conditional test nominal alpha .010	.025	.050	.100	.250
	.33	.02	.056	.115	.200	.362	.694
		.05	.088	.155	.240	.391	.706
		.10	.093	.183	.285	.434	.733
		.20	.094	.204	.352	.505	.766
		.50	.087	.188	.364	.574	.809
	C:T[a]		.074	.149	.297	.499	.793
	.50	.02	.074	.143	.228	.390	.662
		.05	.091	.169	.250	.408	.672
		.10	.101	.193	.277	.428	.684
		.20	.101	.213	.311	.477	.714
		.50	.083	.188	.310	.503	.729
	C:T		.055	.124	.217	.389	.708
	1	.02	.125	.201	.296	.416	.610
		.05	.126	.202	.295	.413	.601
		.10	.128	.208	.303	.417	.598
		.20	.126	.213	.307	.417	.584
		.50	.098	.179	.273	.374	.571
	C:T S:T[b]		.036 .114	.086 .190	.153 .286	.261 .409	.533 .607
	2	.02	.175	.248	.326	.412	.565
		.05	.168	.237	.313	.396	.545
		.10	.157	.223	.294	.369	.516
		.20	.137	.206	.270	.341	.488
		.50	.091	.145	.204	.264	.435
	C:T		.023	.057	.103	.194	.406
	3	.02	.192	.268	.336	.385	.519
		.05	.175	.243	.305	.350	.489
		.10	.158	.222	.280	.324	.461
		.20	.135	.191	.232	.272	.411
		.50	.080	.125	.159	.210	.381
	C:T		.021	.047	.086	.167	.353

[a]Power of the "never pool" test $F = MS_T/MS_{C:T}$.

[b]Power of the "always pool" test $F = MS_T/MS_{S:T}$.

Table C-2

Power of the Conditional F Test Given a Two-Tailed Preliminary Test, c = 5 and s = 5

$E(MS_{C:T})/E(MS_{S:CT})$		Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250
		.02	.164	.259	.355	.532	.797
		.05	.220	.342	.444	.607	.828
	.33	.10	.249	.392	.511	.668	.842
		.20	.261	.418	.560	.718	.871
		.50	.265	.440	.596	.761	.893
	C:T[a]		.265	.436	.594	.760	.896
		.02	.102	.191	.279	.451	.721
		.05	.131	.230	.321	.488	.739
	.50	.10	.149	.262	.360	.520	.761
		.20	.167	.295	.408	.566	.781
		.50	.174	.309	.431	.600	.811
	C:T		.162	.287	.420	.592	.813
		.02	.107	.190	.278	.403	.643
		.05	.108	.192	.279	.401	.645
	1	.10	.110	.192	.278	.399	.649
		.20	.117	.190	.270	.394	.643
		.50	.103	.174	.252	.383	.628
	C:T		.078	.149	.227	.359	.621
	S:T[b]		.102	.186	.274	.402	.645
		.02	.128	.186	.247	.337	.549
		.05	.114	.163	.215	.306	.522
	2	.10	.104	.156	.207	.291	.504
		.20	.091	.135	.184	.273	.486
		.50	.062	.107	.158	.249	.455
	C:T		.038	.085	.141	.229	.447
		.02	.107	.152	.208	.284	.465
		.05	.093	.132	.187	.263	.438
	3	.10	.074	.115	.167	.243	.419
		.20	.057	.101	.145	.213	.408
		.50	.043	.080	.121	.194	.389
	C:T		.024	.069	.107	.191	.383

[a]Power of the "never pool" test $F = MS_T/MS_{C:T}$.

[b]Power of the "always pool" test $F = MS_T/MS_{S:T}$.

Table C-3

Power of the Conditional F Test Given a Two-Tailed Preliminary Test, c = 5 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$		Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250
		.02	.443	.599	.766	.887	.971
		.05	.526	.681	.817	.914	.982
	.33	.10	.611	.753	.855	.933	.986
		.20	.658	.813	.893	.955	.992
		.50	.679	.846	.916	.964	.997
	C:T[a]		.661	.839	.916	.966	.997
		.02	.365	.524	.674	.816	.939
		.05	.399	.561	.711	.833	.946
	.50	.10	.448	.605	.745	.850	.952
		.20	.483	.648	.784	.869	.961
		.50	.495	.692	.812	.892	.969
	C:T		.450	.651	.794	.889	.967
		.02	.341	.463	.583	.719	.869
		.05	.333	.452	.574	.715	.868
	1	.10	.334	.452	.570	.711	.861
		.20	.328	.441	.560	.701	.859
		.50	.283	.395	.530	.689	.849
	C:T		.193	.322	.478	.657	.843
	S:T[b]		.346	.470	.590	.724	.869
		.02	.267	.340	.422	.529	.716
		.05	.228	.302	.383	.499	.688
	2	.10	.194	.265	.351	.480	.669
		.20	.156	.227	.306	.444	.654
		.50	.109	.190	.277	.417	.642
	C:T		.075	.160	.261	.399	.638
		.02	.167	.217	.276	.380	.585
		.05	.128	.184	.247	.355	.564
	3	.10	.109	.167	.230	.336	.551
		.20	.083	.142	.207	.315	.537
		.50	.067	.120	.186	.305	.528
	C:T		.053	.110	.176	.299	.524

[a]Power of the "never pool" test $F = MS_T/MS_{C:T}$.

[b]Power of the "always pool" test $F = MS_T/MS_{S:T}$.

Table C-4

Power of the Conditional F Test Given a Two-Tailed Preliminary Test, c = 5 and s = 20

$E(MS_{C:T})/E(MS_{S:CT})$		Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250
	.33	.02	.753	.895	.947	.982	.998
		.05	.806	.918	.961	.986	.998
		.10	.855	.940	.975	.993	.999
		.20	.892	.958	.982	.997	1.000
		.50	.916	.971	.992	.998	1.000
	C:T[a]		.896	.967	.990	.998	.999
	.50	.02	.671	.827	.902	.950	.991
		.05	.694	.845	.914	.955	.991
		.10	.724	.863	.926	.964	.992
		.20	.760	.886	.938	.971	.993
		.50	.780	.898	.950	.978	.997
	C:T		.721	.861	.937	.978	.997
	1	.02	.599	.725	.805	.878	.945
		.05	.588	.713	.794	.875	.945
		.10	.576	.704	.786	.871	.943
		.20	.559	.686	.782	.862	.939
		.50	.494	.632	.740	.837	.937
	C:T		.381	.552	.699	.822	.938
	S:T[b]		.604	.730	.810	.881	.947
	2	.02	.423	.527	.602	.685	.828
		.05	.373	.479	.561	.656	.814
		.10	.332	.433	.517	.627	.801
		.20	.279	.387	.488	.604	.792
		.50	.207	.329	.447	.575	.785
	C:T		.164	.292	.429	.569	.781
	3	.02	.252	.332	.425	.521	.705
		.05	.195	.281	.382	.491	.692
		.10	.167	.256	.360	.474	.681
		.20	.143	.226	.333	.460	.674
		.50	.108	.203	.313	.450	.669
	C:T		.094	.195	.302	.448	.666

[a]Power of the "never pool" test $F = MS_T/MS_{C:T}$.

[b]Power of the "always pool" test $F = MS_T/MS_{S:T}$.

Table C-5

Power of the Conditional F Test Given a Two-Tailed Preliminary Test, c = 10 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250
.33	.02	.953	.981	.992	.999	1.000
	.05	.972	.991	.996	.999	1.000
	.10	.982	.993	.997	.999	1.000
	.20	.985	.996	.999	1.000	1.000
	.50	.987	.996	.999	1.000	1.000
	C:T[a]	.986	.996	.999	1.000	1.000
.50	.02	.850	.922	.963	.984	.998
	.05	.873	.940	.971	.987	1.000
	.10	.895	.950	.977	.989	1.000
	.20	.911	.962	.982	.991	1.000
	.50	.925	.965	.984	.995	1.000
	C:T	.920	.965	.984	.995	1.000
1	.02	.721	.827	.883	.930	.970
	.05	.716	.824	.880	.928	.970
	.10	.713	.821	.880	.926	.969
	.20	.700	.811	.875	.926	.968
	.50	.668	.794	.862	.921	.966
	C:T	.621	.773	.852	.915	.967
	S:T[b]	.720	.826	.883	.931	.971
2	.02	.433	.542	.652	.752	.876
	.05	.377	.492	.615	.731	.870
	.10	.338	.457	.595	.722	.865
	.20	.312	.432	.577	.714	.859
	.50	.281	.415	.566	.706	.857
	C:T	.268	.411	.562	.705	.857
3	.02	.213	.310	.406	.559	.771
	.05	.195	.294	.396	.551	.768
	.10	.179	.282	.391	.547	.767
	.20	.176	.279	.389	.543	.766
	.50	.174	.275	.386	.540	.765
	C:T	.173	.275	.386	.539	.765

[a]Power of the "never pool" test $F = MS_T/MS_{C:T}$.

[b]Power of the "always pool" test $F = MS_T/MS_{S:T}$.

APPENDIX D

POWERS OF THE CONDITIONAL F GIVEN AN UPPER-TAILED PRELIMINARY TEST

Table D-1

Power of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 2 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$		Preliminary test nominal alpha	Conditional Test nominal alpha .010	.025	.050	.100	.250
		.01	.028	.087	.177	.342	.685
		.025	.028	.087	.177	.342	.685
	.33	.05	.028	.087	.177	.342	.685
		.10	.028	.087	.177	.342	.683
		.25	.028	.086	.174	.335	.673
	C:T[a]		.074	.149	.297	.499	.793
		.01	.053	.126	.211	.376	.656
		.025	.053	.126	.211	.376	.656
	.50	.05	.053	.126	.211	.376	.654
		.10	.053	.125	.209	.373	.650
		.25	.049	.116	.195	.351	.620
	C:T		.055	.124	.217	.389	.708
		.01	.113	.189	.285	.407	.603
		.025	.111	.185	.278	.398	.591
	1	.05	.107	.178	.270	.388	.578
		.10	.105	.172	.260	.374	.554
		.25	.090	.144	.218	.314	.509
	C:T		.036	.086	.153	.261	.533
	S:T[b]		.114	.190	.286	.409	.607
		.01	.170	.242	.320	.406	.561
		.025	.160	.226	.302	.386	.538
	2	.05	.149	.210	.280	.355	.506
		.10	.132	.185	.250	.318	.471
		.25	.096	.133	.181	.234	.405
	C:T		.023	.057	.103	.194	.406
		.01	.187	.263	.332	.381	.517
		.025	.168	.237	.299	.343	.484
	3	.05	.150	.214	.272	.313	.453
		.10	.127	.180	.221	.258	.399
		.25	.084	.122	.150	.191	.359
	C:T		.021	.047	.086	.167	.353

[a]Power of the "never pool" test $F = MS_T/MS_{C:T}$.

[b]Power of the "always pool" test $F = MS_T/MS_{S:T}$.

Table D-2

Power of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 5 and s = 5

$E(MS_{C:T})/E(MS_{S:CT})$		Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250
		.01	.031	.105	.200	.397	.727
		.025	.031	.105	.200	.397	.727
	.33	.05	.031	.105	.200	.397	.727
		.10	.031	.105	.200	.397	.727
		.25	.031	.105	.200	.397	.727
	C:T[a]		.265	.436	.594	.760	.896
		.01	.057	.139	.229	.406	.689
		.025	.057	.139	.229	.406	.689
	.50	.05	.057	.139	.229	.406	.689
		.10	.057	.139	.229	.406	.688
		.25	.057	.139	.225	.399	.686
	C:T		.162	.287	.420	.592	.813
		.01	.101	.186	.273	.399	.641
		.025	.101	.186	.270	.394	.640
	1	.05	.099	.179	.262	.386	.636
		.10	.095	.168	.245	.366	.624
		.25	.078	.141	.217	.335	.599
	C:T		.078	.149	.227	.359	.621
	S:T[b]		.102	.186	.274	.402	.645
		.01	.127	.186	.246	.336	.549
		.025	.113	.162	.213	.305	.522
	2	.05	.102	.154	.204	.290	.503
		.10	.089	.133	.181	.271	.485
		.25	.060	.105	.153	.244	.453
	C:T		.038	.085	.141	.229	.447
		.01	.107	.152	.208	.284	.465
		.025	.092	.132	.187	.263	.438
	3	.05	.073	.115	.167	.242	.419
		.10	.056	.100	.144	.212	.408
		.25	.042	.078	.120	.193	.389
	C:T		.024	.069	.109	.191	.383

[a] Power of the "never pool" test $F = MS_T/MS_{C:T}$.

[b] Power of the "always pool" test $F = MS_T/MS_{S:T}$.

Table D-3

Power of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 5 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha		Conditional test nominal alpha .010	.025	.050	.100	.250
	.33	.01	.282	.485	.694	.855	.962
		.025	.282	.485	.694	.855	.962
		.05	.282	.485	.694	.855	.962
		.10	.282	.485	.694	.855	.962
		.25	.282	.485	.694	.855	.962
	C:T[a]		.661	.839	.916	.966	.997
	.50	.01	.312	.481	.646	.800	.937
		.025	.312	.481	.646	.800	.937
		.05	.312	.481	.646	.800	.937
		.10	.312	.481	.646	.800	.936
		.25	.304	.474	.641	.798	.935
	C:T		.450	.651	.794	.889	.967
	1	.01	.341	.463	.583	.718	.868
		.025	.331	.450	.570	.708	.865
		.05	.323	.439	.559	.698	.857
		.10	.306	.418	.536	.680	.853
		.25	.257	.364	.486	.647	.840
	C:T		.193	.322	.478	.657	.843
	S:T[b]		.346	.470	.590	.724	.869
	2	.01	.267	.340	.422	.529	.716
		.025	.228	.302	.383	.499	.688
		.05	.194	.265	.351	.480	.669
		.10	.156	.227	.306	.444	.653
		.25	.109	.189	.273	.417	.638
	C:T		.075	.160	.261	.399	.638
	3	.01	.167	.217	.276	.380	.585
		.025	.128	.184	.247	.355	.564
		.05	.109	.167	.230	.336	.551
		.10	.083	.142	.207	.315	.537
		.25	.067	.120	.186	.305	.528
	C:T		.053	.110	.176	.299	.524

[a]Power of the "never pool" test $F = MS_T/MS_{C:T}$.

[b]Power of the "always pool" test $F = MS_T/MS_{S:T}$.

Table D-4

Power of the Conditional F Test Given an Upper-Tqiled Preliminary Test, c = 5 and s = 20

$E(MS_{C:T})/E(MS_{S:CT})$	Preliminary test nominal alpha		Conditional test nominal alpha .010	.025	.050	.100	.250
		.01	.685	.857	.929	.980	.998
		.025	.685	.857	.929	.980	.998
	.33	.05	.685	.857	.929	.980	.998
		.10	.685	.857	.929	.980	.998
		.25	.685	.857	.929	.980	.998
	C:T[a]		.896	.967	.990	.998	.999
		.01	.644	.815	.897	.948	.991
		.025	.644	.815	.897	.948	.991
	.50	.05	.644	.815	.897	.948	.991
		.10	.643	.815	.897	.947	.991
		.25	.638	.810	.895	.946	.991
	C:T		.721	.861	.937	.978	.997
		.01	.598	.724	.805	.878	.945
		.025	.584	.709	.791	.871	.944
	1	.05	.570	.695	.779	.866	.942
		.10	.544	.671	.766	.854	.937
		.25	.472	.612	.719	.824	.932
	C:T		.381	.552	.699	.822	.938
	S:T[b]		.604	.730	.810	.881	.947
		.01	.423	.527	.602	.685	.828
		.025	.372	.479	.561	.656	.814
	2	.05	.331	.432	.516	.626	.801
		.10	.278	.386	.487	.603	.791
		.25	.205	.325	.445	.574	.783
	C:T		.164	.292	.429	.569	.781
		.01	.252	.332	.425	.521	.705
		.025	.195	.281	.382	.491	.692
	3	.05	.167	.256	.360	.474	.681
		.10	.142	.226	.333	.460	.674
		.25	.107	.203	.313	.449	.669
	C:T		.094	.195	.302	.448	.666

[a]Power of the "never pool" test $F = MS_T/MS_{C:T}$.

[b]Power of the "always pool" test $F = MS_T/MS_{S:T}$.

Table D-5

Power of the Conditional F Test Given an Upper-Tailed Preliminary Test, c = 10 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$		Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250
		.01	.854	.942	.976	.996	1.000
		.025	.854	.942	.976	.996	1.000
	.33	.05	.854	.942	.976	.996	1.000
		.10	.854	.942	.976	.996	1.000
		.25	.854	.942	.976	.996	1.000
	C:T[a]		.986	.996	.999	1.000	1.000
		.01	.805	.898	.957	.981	.998
		.025	.805	.898	.957	.981	.998
	.50	.05	.805	.898	.957	.981	.998
		.10	.805	.898	.957	.981	.998
		.25	.805	.896	.956	.981	.998
	C:T		.920	.965	.984	.995	1.000
		.01	.717	.823	.880	.929	.970
		.025	.709	.818	.876	.927	.970
	1	.05	.701	.813	.875	.925	.969
		.10	.683	.801	.870	.922	.968
		.25	.640	.773	.852	.914	.965
	C:T		.621	.773	.852	.915	.967
	S:T[b]		.720	.826	.883	.931	.971
		.01	.433	.542	.652	.752	.876
		.025	.377	.492	.615	.731	.870
	2	.05	.338	.457	.595	.722	.865
		.10	.312	.432	.577	.713	.859
		.25	.281	.415	.564	.705	.857
	C:T		.268	.411	.562	.705	.857
		.01	.213	.310	.406	.559	.771
		.025	.195	.294	.396	.551	.768
	3	.05	.179	.282	.391	.547	.767
		.10	.176	.279	.389	.543	.766
		.25	.174	.275	.386	.540	.765
	C:T		.173	.275	.386	.539	.765

[a]Power of the "never pool" test $F = MS_T/MS_{C:T}$.

[b]Power of the "always pool" test $F = MS_T/MS_{S:T}$.

APPENDIX E

POWERS OF THE CONDITIONAL F GIVEN A LOWER-TAILED PRELIMINARY TEST

Table E-1

Power of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 2 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$		Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250
		.01	.056	.115	.200	.362	.694
		.025	.088	.155	.240	.391	.706
	.33	.05	.093	.183	.285	.434	.733
		.10	.094	.204	.352	.505	.768
		.25	.087	.189	.367	.581	.821
	C:T[a]		.074	.149	.297	.499	.793
		.01	.074	.143	.228	.390	.662
		.025	.091	.169	.250	.408	.672
	.50	.05	.101	.193	.277	.428	.686
		.10	.101	.214	.313	.480	.720
		.25	.087	.198	.326	.528	.765
	C:T		.055	.124	.217	.389	.708
		.01	.126	.202	.297	.418	.614
		.025	.129	.207	.303	.424	.617
	1	.05	.135	.220	.319	.438	.627
		.10	.135	.231	.333	.452	.637
		.25	.122	.225	.341	.469	.669
	C:T		.036	.086	.153	.261	.533
	S:T[b]		.114	.190	.286	.409	.607
		.01	.190	.268	.351	.443	.616
		.025	.193	.273	.356	.447	.619
	2	.05	.193	.275	.359	.451	.622
		.10	.190	.283	.365	.460	.629
		.25	.180	.274	.368	.467	.642
	C:T		.023	.057	.103	.194	.406
		.01	.231	.317	.394	.459	.623
		.025	.233	.318	.396	.462	.626
	3	.05	.234	.320	.398	.466	.629
		.10	.234	.323	.401	.469	.633
		.25	.222	.315	.399	.474	.643
	C:T		.021	.047	.086	.167	.353

[a]Power of the "never pool" test $F = MS_T/MS_{C:T}$.

[b]Power of the "always pool" test $F = MS_T/MS_{S:T}$.

Table E-2

Power of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 5 and s = 5

$E(MS_{C:T})/E(MS_{S:CT})$		Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250
		.01	.164	.259	.355	.532	.797
		.025	.220	.342	.444	.607	.828
	.33	.05	.249	.392	.511	.668	.842
		.10	.261	.418	.560	.718	.871
		.25	.265	.440	.596	.761	.893
	C:T[a]		.265	.436	.594	.760	.896
		.01	.102	.191	.271	.451	.721
		.025	.131	.230	.321	.488	.739
	.50	.05	.149	.262	.360	.520	.761
		.10	.167	.295	.408	.566	.782
		.25	.174	.309	.435	.607	.814
	C:T		.162	.287	.420	.592	.813
		.01	.108	.190	.279	.406	.647
		.025	.109	.192	.283	.409	.650
	1	.05	.113	.199	.290	.415	.658
		.10	.124	.208	.299	.430	.664
		.25	.127	.219	.309	.450	.674
	C:T		.078	.149	.227	.359	.621
	S:T[b]		.102	.186	.274	.402	.645
		.01	.150	.221	.302	.401	.616
		.025	.150	.222	.303	.401	.616
	2	.05	.151	.223	.304	.401	.617
		.10	.151	.223	.304	.402	.617
		.25	.151	.223	.306	.405	.618
	C:T		.038	.085	.141	.229	.447
		.01	.171	.236	.320	.420	.617
		.025	.172	.236	.320	.420	.617
	3	.05	.172	.236	.320	.421	.617
		.10	.172	.237	.321	.421	.617
		.25	.172	.238	.321	.421	.617
	C:T		.024	.069	.109	.191	.383

[a]Power of the "never pool" test $F = MS_T/MS_{C:T}$.

[b]Power of the "always pool" test $F = MS_T/MS_{S:T}$.

Table E-3

Power of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 5 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$		Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.100	.050	.250
		.01	.433	.599	.766	.887	.971
		.025	.526	.681	.817	.914	.982
	.33	.05	.611	.753	.855	.933	.986
		.10	.658	.816	.893	.955	.992
		.25	.679	.846	.916	.964	.997
	C:T[a]		.661	.839	.916	.966	.997
		.01	.365	.524	.674	.816	.939
		.025	.399	.561	.711	.833	.946
	.50	.05	.448	.605	.745	.850	.952
		.10	.483	.648	.784	.869	.962
		.25	.503	.699	.817	.894	.971
	C:T		.450	.651	.794	.889	.967
		.01	.346	.470	.590	.725	.870
		.025	.348	.472	.594	.731	.872
	1	.05	.357	.483	.601	.737	.873
		.10	.368	.493	.614	.745	.875
		.25	.372	.501	.634	.766	.878
	C:T		.193	.322	.478	.657	.843
	S:T[b]		.346	.470	.590	.724	.867
		.01	.357	.454	.544	.641	.782
		.025	.357	.454	.544	.641	.782
	2	.05	.357	.454	.544	.641	.782
		.10	.357	.454	.544	.641	.783
		.25	.357	.454	.548	.641	.786
	C:T		.075	.160	.261	.399	.638
		.01	.367	.442	.526	.610	.752
		.025	.367	.442	.526	.610	.752
	3	.05	.367	.442	.526	.610	.752
		.10	.367	.442	.526	.610	.752
		.25	.367	.442	.526	.610	.752
	C:T		.053	.110	.176	.299	.524

[a] Power of the "never pool" test $F = MS_T/MS_{C:T}$.

[b] Power of the "always pool" test $F = MS_T/MS_{S:T}$.

Table E-4

Power of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 5 and s = 20

$E(MS_{C:T})/E(MS_{S:CT})$		Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250
		.01	.753	.895	.947	.982	.998
		.025	.806	.918	.961	.986	.998
	.33	.05	.855	.940	.975	.993	.999
		.10	.892	.958	.982	.997	1.000
		.25	.916	.971	.992	.998	1.000
	C:T[a]		.896	.967	.990	.998	.999
		.01	.671	.827	.902	.950	.991
		.025	.694	.845	.914	.955	.991
	.50	.05	.724	.863	.926	.964	.992
		.10	.761	.886	.938	.972	.993
		.25	.786	.903	.952	.980	.997
	C:T		.721	.861	.937	.978	.997
		.01	.605	.731	.810	.881	.947
		.025	.608	.734	.813	.885	.948
	1	.05	.610	.739	.817	.886	.948
		.10	.619	.745	.826	.889	.949
		.25	.626	.750	.831	.894	.952
	C:T		.381	.552	.699	.822	.938
	S:T[b]		.604	.730	.810	.881	.947
		.01	.563	.670	.733	.798	.884
		.025	.564	.670	.733	.798	.884
	2	.05	.564	.671	.734	.799	.884
		.10	.564	.671	.734	.799	.885
		.25	.565	.674	.735	.799	.886
	C:T		.164	.292	.429	.569	.781
		.01	.537	.623	.703	.753	.846
		.025	.537	.623	.703	.753	.846
	3	.05	.537	.623	.703	.753	.846
		.10	.538	.623	.703	.753	.846
		.25	.538	.623	.703	.754	.846
	C:T		.094	.195	.302	.448	.666

[a]Power of the "never pool" test $F = MS_T/MS_{C:T}$.

[b]Power of the "always pool" test $F = MS_T/MS_{S:T}$.

Table E-5

Power of the Conditional F Test Given a Lower-Tailed Preliminary Test, c = 10 and s = 12

$E(MS_{C:T})/E(MS_{S:CT})$		Preliminary test nominal alpha	Conditional test nominal alpha .010	.025	.050	.100	.250
		.01	.953	.981	.992	.999	1.000
		.025	.972	.991	.996	.999	1.000
	.33	.05	.982	.993	.997	.999	1.000
		.10	.985	.996	.999	1.000	1.000
		.25	.987	.996	.999	1.000	1.000
	C:T[a]		.986	.996	.999	1.000	1.000
		.01	.850	.922	.963	.984	.998
		.025	.873	.940	.971	.987	1.000
	.50	.05	.895	.950	.977	.989	1.000
		.10	.911	.962	.982	.991	1.000
		.25	.925	.967	.985	.995	1.000
	C:T		.920	.965	.984	.995	1.000
		.01	.724	.830	.886	.932	.971
		.025	.727	.832	.887	.932	.971
	1	.05	.732	.834	.888	.932	.971
		.10	.737	.836	.888	.935	.971
		.25	.748	.847	.893	.938	.972
	C:T		.621	.773	.852	.915	.967
	S:T[b]		.720	.826	.883	.931	.971
		.01	.642	.730	.795	.859	.920
		.025	.642	.730	.795	.859	.920
	2	.05	.642	.730	.795	.859	.920
		.10	.642	.730	.795	.860	.920
		.25	.642	.730	.797	.860	.920
	C:T		.268	.411	.562	.705	.857
		.01	.589	.687	.744	.807	.879
		.025	.589	.687	.744	.807	.879
	3	.05	.589	.687	.744	.807	.879
		.10	.589	.687	.744	.807	.879
		.25	.589	.687	.744	.807	.879
	C:T		.173	.275	.386	.539	.765

[a]Power of the "never pool" test $F = MS_T/MS_{C:T}$.

[b]Power of the "always pool" test $F = MS_T/MS_{S:T}$.

BIBLIOGRAPHY

BIBLIOGRAPHY

Cochran, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics, 1947, 3, 22-38.

Cochran, W. G., & Cox, G. M. Experimental designs. New York: Wiley & Sons, 1957.

Cox, D. R. Planning of experiments. New York: Wiley & Sons, 1958.

Draper, N. R., & Smith, H. Applied regression analysis. New York: Wiley & Sons, 1966.

Glass, G. V., Peckham, P. D., & Sanders, J. R. Consequences of failure to meet assumptions underlying the analysis of variance and covariance. Review of Educational Research, 1972, 42, 237-288.

Glass, G. V., & Stanley, J. C. Statistical methods in education and psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Graybill, F. A. An introduction to linear statistical models. New York: McGraw-Hill, 1961.

Haney, W. Units of analysis issues in the evaluation of Project Follow Through or There must be heresy in this some place (Contract No. OEC-0-74-0394). Cambridge, Mass.: Huron Institute, September 1974.

International Business Machines Corporation. System/360 Scientific Subroutine Package, Version III, Programmer's Manual. Program Number 360A-CM-03X, August 1970.

Kirk, R. E. Experimental design: Procedure for the behavioral sciences. Belmont, Calif.: Brooks/Cole, 1968.

Levene, H. Robust tests for equality of variances. In I. Olkin (Ed.), Contributions to probability and statistics. Stanford, Calif.: Stanford University Press, 1960, 278-292.

Lissitz, R. W., & Chardos, S. A study of the effect of the violation of the assumption of independent sampling upon the Type I error rate of the two-group t-test. Educational and Psychological Measurement, 1975, 35, 353-359.

Millman, J., & Glass, G. V. Rules of thumb for writing the ANOVA table. Journal of Educational Measurement, 1967, 4, 41-51.

Paull, A. E. On a preliminary test for pooling mean squares in the analysis of variance. Annals of Mathemtical Statistics, 1950, 21, 539-556.

Peckham, P. D., Glass, G. V., & Hopkins, K. D. The experimental unit in statistical analysis. The Journal of Special Education, 1969, 3, 337-349. (a)

Peckham, P. D., Glass, G. V., & 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)

Porter, A. C. Some design and analysis concerns for quasi-experiments such as Follow Through. Paper presented at the meeting of the American Psychological Association, Hawaii, August 1972.

Poynor, H. Selecting units of analysis. In G. Borich (Ed.), Evaluating educational programs and products. Englewood Cliffs, N.J.: Educational Technology Press, 1974.

Raths, J. The appropriate experimental unit. Educational Leadership, 1967, 263-266.

Scheffé, H. The analysis of variance. New York: Wiley & Sons, 1959.

Smith, M. Personal communication, October 21, 1974.

Steck, J. C. The independence of observations obtained in classroom research. Unpublished M.A. thesis. College Park, Md.: University of Maryland, 1966.

Wiley, D. E. Design and analysis of evaluation studies. In M. C. Wittrock & D. E. Wiley (Eds.), The evaluation of instruction: Issues and problems. New York: Holt, Rinehart & Winston, 1970.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.