American Educational Research Journal
Spring 1982, Vol. 19, No. 1, Pp. 5–18

The Unit of Analysis: Group Means Versus Individual Observations

KENNETH D. HOPKINS
University of Colorado

This paper shows that the common recommendation to use group means when there may be nonindependence among observational units is unnecessary, unduly restrictive, impoverishes the analysis, and limits the questions that can be addressed in a study. When random factors are properly identified and included in the analysis, the results (Fs and critical Fs) are identical in balanced ANOVA designs, irrespective of whether group means or individual observations are employed. The use of individual observations also allows the exploration of other interesting questions pertaining to interaction and generalizability. In addition, the pooling strategy can be considered. Thus, the question of the proper experimental unit or unit of analysis for treatment effects is answered directly, correctly, and implicitly when the proper statistical model is employed.

INTRODUCTION

As early as 1940, Lindquist recognized a problem that characterizes many educational and behavioral experiments—the use of individuals as the statistical unit of analysis when the treatment is applied to a class or to a group. When the treatment is not administered individually to each subject, the statistical assumption of "independence of error" can be violated if individual scores (rather than class means) are used as the unit of analysis.

BACKGROUND

In his review of Lindquist's (1940) classic test *Statistical Analysis in Educational Research*, McNemar (1940) observed:

The author is indebted to Gene V Glass, Rick Kroc, Lynn Sherretz, and an AERJ reviewer for their helpful suggestions on earlier versions of this paper.

> We next raise a puzzling question for which we have no definite answer. . . . [T]he analysis of variance technique is applied to an educational-methods experiment involving five schools and three methods, with twenty pupils in each of fifteen classes. The analysis is carried through on the basis of the fifteen class means in such a way that neither the number of pupils nor the pupil variation enters into the analysis . . . . The reviewer suspects that something is wrong with a test of significance that does not involve the variation of the individuals upon which the means are based. We are unable to locate the fallacy here, if there be such . . . . We are not arguing that the author is wrong in arguing that intact groups are the proper sampling units, but rather that the case is not convincingly stated.[1] (p. 747)

The confusion expressed by McNemar is still evident, although several efforts have shed light on the relevant issues (Addelman, 1970; Barcikowski, 1980; Burstein, undated; Campbell & Stanley, 1963; Fletcher, 1968; Glass & Stanley, 1970, pp. 501–508; Kempthorne, 1952, p. 163; Lindquist, 1953; Lumsdaine, 1963; Page, 1965; Peckham, Glass, & Hopkins, 1969; Raths, 1967; Steck, 1966; Wright, 1969). Cronbach (1976, p. 5.22) has stated, "Significance tests based on individual-level analysis are unacceptable when classes are the unit of sampling."

Glass and Stanley (1970), Page (1975), and others have noted that the independence assumption can be violated when the subjects receive the treatment as a group even if individuals are randomly assigned to groups. In such situations, the type 1 error probability statements can be seriously underestimated. Glass and Stanley (1970) stated:

> Educational researchers are especially prone to making the error of analyzing data in terms of units other than the legitimate unit. . . . The researcher has two alternatives, though he is seldomly aware of the second one: (1) he can run a potentially illegitimate analysis of the experiment by using the 'pupil' as the unit of statistical analysis, or (2) he can run a legitimate analysis on the means of the classrooms, in which case he is almost certain to obtain statistically nonsignificant results. (p. 507)

---

[1] For purposes of this paper, certain facets of the illustrative studies that are not relevant to the points being made were ignored, such as the elimination of factors that would make the examples unnecessarily complex. The elimination of these distractions has no effect on the *F* ratios or conclusions of the studies. The conclusion of the studies, and their internal validity, are irrelevant to the purpose of this paper. Be studies weak or strong, the common effects evaluated when group means and individual scores are used as the unit of analysis will be associated with identical observed and critical Fs with balanced designs. With unbalanced designs the results will not be identical, but will differ depending on the extent to which the sources of variation in the design are nonorthogonal, and the particular analysis strategy employed (e.g., unweighted means, hierarchical, "classical" fitting constants or "saturated" regression solution).

Page (1975) observed:

> Classrooms are potentially very rich in research variables: variables of student, teacher, curriculum, environment, variables measured and yet unmeasured, conjectured or yet unimagined. And the interactions of such variables are, theoretically at least, many times more numerous. Yet most researchers have little understanding of how classrooms may be analyzed. Many treat each student as an independent subject, but in doing so produce findings which are not replicable by others. More rigorous investigators are apt to suppress most of the richness within the classroom by using class means. Is there a way to overcome this dilemma, to have both rigor and richness? (p. 339)

Page pointed out that the use of class means (as the observational unit) makes it impossible to evaluate, statistically, interaction hypotheses between treatments and learner characteristics (e.g., ability, sex, ethnicity). These interactions speak directly to important questions regarding external validity.

Researchers have failed to recognize that if the proper ANOVA model is explicated and employed, the "problem" of the appropriate unit of analysis disappears. Furthermore, it will be shown that the use of class means does not necessarily ensure that the relevant independence assumption has been met. The author did not locate any study in which the question of the proper unit of analysis and the related independence assumptions were evaluated in the context of the available alternative statistical models. Several different statistical models will be compared in this paper, along with an illustration in which the same data are used. Initially, three models will be compared using data from a study (DeRosia, 1980) in which two methods of instruction are contrasted, with three teachers nested within each method; twenty-five students are nested within each teacher.

Model A

The single-factor fixed effects general linear model for the score for the ith student ($i = 1, \ldots n$) nested (":") in the mth ($m = 1, \ldots, M$) method group is

$$X_{im} = \mu + \alpha_m + \epsilon_{i:m} \qquad \text{(A)}$$

The assumptions of normality, homogeneity of variance, and independence pertain to the ϵ values, viz: $\epsilon_{i:m} \sim NID\ (0, \sigma_\epsilon^2)$ (i.e., within each of the M methods, the ϵ values are normally and independently distributed and have a mean of 0, and common variance, σ_ϵ^2). The expected mean squares, E(MS), for Model A are given in panel *a* of Table I; the right-hand portion of panel *a* gives the results of the analysis of the DeRosia data using Model A. Analyses such as this that use scores from individual students have been widely criticized since Lindquist (1940). For example, Kempthorne (1961) observed:

TABLE I

Expected Mean Squares for a Balanced ANOVA Design (and Illustrative Analyses) in which n Students are Nested within T Teachers, which are Nested Within M Methods, using Three Models

SV	*v*	*E(MS)*	Example		
			v	*MS*	*F*
a)	[Model A: $X_{im} = \mu + \alpha_m + \epsilon_{i:m}$]				
Methods (*M*)	*M*-1	$\sigma^2_{s:m} + n\,\sigma^2_m$	1	2814.6	7.69*
Students nested *within* method (*S*:*M*)	*M* (*n* − 1)	$\sigma^2_{s:m}$	148	366.2	
b)	[Model B: $X_{tm} = \mu + \alpha_m + \beta_{t:m}$]				
Methods (*M*)	*M*-1	$\sigma^2_{t:m} + T\sigma^2_m$	1	112.6	1.99
Teachers nested *within* methods (*T*:*M*)	*M*(*T*-1)	$\sigma^2_{t:m}$	4	56.5	
c)	[Model C: $X_{itm} = \mu + \alpha_m + \beta_{t:m} + \epsilon_{i:tm}$]				
Methods (*M*)	*M*-1	$\sigma^2_{s:tm} + n\sigma^2_{t:m} + nT\sigma^2_m$	1	2814.5	1.99
Teachers nested *within* methods (*T*:*M*)	*M*(*T*-1)	$\sigma^2_{s:tm} + n\sigma^2_{t:m}$	4	1412.7	4.19*
Students nested *within* teachers and methods (*S*:*TM*)	*MT*(*n* − 1)	$\sigma^2_{s:tm}$	144	337.1	

* $p < .01$.

> If all experimental units receiving each particular treatment receive it together, as for instance by all being taught in one way by one instructor, the only conclusion about any treatment difference observed is that it is attributable to the way of teaching or the instructor or partly due to each. (p. 123)

In other words, an analysis that includes only two sources of variation, methods and students-within-methods (Model A), would correctly assess the inferential question pertaining to methods only in the unlikely circumstances in which (1) there are no teacher effects, and (2) each student's performance is independent of the particular set of students in his class.

Model B

If instead of using student scores as the observational unit, class means for the T teachers ($t = 1, \ldots, T$) are employed, the model becomes

$$X_{tm} = \mu + \alpha_m + \beta_{t:m}, \tag{B}$$

where $\beta_{t:m} \sim NID\ (0, \sigma^2_\beta)$ (NID = normally distributed and independent). This is the model advocated by Lindquist (1940), Campbell and Stanley (1963), etc. for studies in which the treatment is group-oriented and hence

can result in nonindependence among the students' scores. The E(MS) values for Model B along with the analysis of DeRosia data using Model B is found in panel *b* of Table I. Notice that the "highly significant" ($p < .01$) methods effect that was found when Model A (panel *a*) was used disappears ($p > .20$) when Model B is employed.

But the analysis using the class mean as the observational unit (Model B) does not ensure that the important independence assumption has been met; to test the methods effect, Model B exchanges one independence assumption (among teachers) for another (among students). Model B is not the best response to the analysis problem; Models A and B ignore the important distinction between the experimental unit and the observational unit. As Addelman (1970) noted,

> The experimental unit is that entity that is allocated to a treatment 'independently' of other entities. It may contain several observational units. (p. 1,095)

Clearly the experimental unit could either be individual students or teachers (classes). Even if teachers are the experimental unit, students' scores may (and should) serve as the observational unit. The preferred model for the analysis (Model C) contains terms for all the available sources of variation in the experiment—in this example methods (M), teachers within methods (T:M), and students within teachers (S:TM). As Addelman (1970) observed,

> When there are several observational units per experimental unit, both the experimental unit error [$\beta_{t:m}$] and the observational unit error [$\epsilon_{i:tm}$] should be included in the model. Since both types of errors include variability due to factors unknown to or beyond the control of the experimenter, neither should be deleted from the model at the whim of the experimenter or statistician. (pp. 1,097–1,098)

Model C

Model C incorporates components both for teachers (or classes) and students; both are viewed as random effects because the desired inference is to teachers "like these" as well as to students "like these." The linear model for the design in which n students are nested within T teachers which, in term, are nested within M methods is

$$X_{itm} = \mu + \alpha_m + \beta_{t:m} + \epsilon_{i:tm}. \qquad \text{(C)}$$

In Model C there are two sets of assumptions, i.e., $\epsilon_{i:tm} \sim NID\ (0, \sigma_\epsilon^2)$ and $\beta_{t:m} \sim NID\ (0, \sigma_\beta^2)$.

In other words, since teachers are properly viewed as a random effect, a second "layer" of assumptions is required in the desired universe of inference. The second set of assumptions is rarely recognized or considered

in practice. The assumptions of homogeneity of variance and normality for $\beta_{t:m}$ and $\epsilon_{i:tm}$ can be tested using the common statistical tests for these purposes, but are less important than the independence assumptions, especially with balanced designs (Glass, Peckham & Sanders, 1972). The independence assumptions must often be evaluated logically. If scores from individual students are used as observational units and the data are analyzed using Model C (students are nested within teachers who are nested within methods), and if classes are appropriately designated as a random factor, the expected mean squares for the effects are given in panel *c* of Table I. The analysis of the sample data using Model C is given in the right-hand portion of panel *c* in Table I.

It is apparent from Table I that for balanced designs, the *F*-ratio for the *methods effect (with Model B or Model C) is the same whether class means or individual observations are used,* since the methods mean square would be divided by the teachers within methods means square ($T{:}M$) in both instances. Even though the methods mean squares will differ (by a factor of n), the *F*-ratios for treatment will be *identical* in the two analyses given in Table I, and these *F* values will have identical degrees of freedom and critical F's.

Independence Assumption. There has been much confusion regarding the assumption of independence in ANOVA designs. As in Model C, there is often more than one independence assumption to be considered in a given model. In Model A, the NID assumptions pertain to the student scores (within teacher and method). In Model B the NID assumptions pertain to class means (within method); in Model C the NID assumptions pertain to both class means (within method) and student scores (within teacher). If a model has a third random factor, the NID assumptions would also pertain to it. *The critical independence question is the independence assumption pertaining to the effect being tested.* The lack of independence among students within classes does not necessarily affect the independence among classes within treatment, and nonindependence among classes does not necessarily result in violation of the independence assumption among schools. Each independence assumption of the model must be evaluated separately. Thus, in the Lindquist example described earlier (in the McNemar [1940] quote), pupils are nested within teachers, who are nested within schools. The desired universe would require students, teachers, and schools to be random factors, and the levels within each would be assumed to be independent. Dependency among teachers (within schools) and/or among students (within teachers) does not necessarily result in a lack of independence among schools (which is all that would be required to test the method effect). For example, consider the analysis of a second dependent variable from DeRosia's study shown in Table II.

Notice that the "conservative" analysis using class means (Model B) would yield a significant method effect ($F = 40.8$, $p < .01$), whereas the preferred

TABLE II

An Illustration of Data in which Class Means are not Independent

SV	ν	*MS*
Methods (M)	1	106.0
Teachers nested within method ($T{:}M$)	4	2.6
Students within teachers and methods ($S{:}TM$)	144	21.0

analysis (Model C) would not. Notice the mean square for the $T{:}M$ effect is much smaller than the $S{:}TM$ effect, yet Model C shows that the E(MS) for $T{:}M$ includes the $S{:}TM$ effect (see Table I). Why would the class means differ by less than would be expected "by chance"? (If a nondirectional F-test had been employed, F would have a value of 8.1 and the null hypothesis could have been rejected at $\alpha = .05$.) Several explanations are possible; each illustrates a violation of the independence assumption for the β values. The models assume that the teachers in each method are a random sample of teachers from the population who implement the method independently. Team-teaching, common planning, exchange of incidental materials, or activities could cause class means to differ by less than would be expected. Likewise, if in the assignment of students to classes within each method, a careful effort is made to keep the classes equivalent in intelligence, etc., the class means can be more nearly equal than would be expected from a random assignment of pupils to classes, and hence the mean square associated with $T{:}M$ could be less than for $S{:}TM$.

Such dependency among class means does not necessarily affect the variance among pupils within classes, i.e., pupil scores can be independent even if class means are not, and vice-versa. The analysis using Model C (but not Model B) would have suggested that the independence assumption regarding the β values is untenable and therefore, the generalizability of the method effect lacks credibility.

Mixed-model Designs

Just as in hierarchical designs, in balanced mixed-model ANOVA designs the F-ratio for the fixed effect is unchanged by the unit of analysis decision when the proper model is employed; this is illustrated in the methods-by-teachers design in Table III.

An Example In an experimental study conducted to assess the effects of the use (E) vs. nonuse (C) of hand calculators in a remedial college mathematics course (Koop, 1978), thirty students in each of three instructors' classes were randomly assigned to the E or C group. If group means are employed as the unit of analysis (Model D in Table III), there are only six observations, whereas there would be 90 observations in the student-level

TABLE III

Expected Mean Squares for a Balanced ANOVA Design (and Illustrative Analyses) in which T Teachers Cross M Methods when Class Means (Panel A) and n Student Scores (Panel B) are used as the Observational Unit

SV	ν	$E(MS)$	Example		
			ν	*MS*	*F*
a)	[Model D: $X_{tm} = \mu + \alpha_m + \beta_t + \alpha\beta_{tm}$]				
Methods (*M*)	$M - 1$	$\sigma^2_{tm} + T\sigma^2_m$	1	40.07	17.27
Teachers (*T*)	$T - 1$	$M\sigma^2_t$	2	3.80	...
MT	$(M - 1)(T - 1)$	σ^2_{tm}	2	2.32	
b)	[Model E: $X_{itm} = \mu + \alpha_m + \beta_t + \alpha\beta_{tm} + \epsilon_{i:tm}$]				
Methods (*M*)	$M - 1$	$\sigma^2_{s:tm} + n\sigma^2_{tm} + nT\sigma^2_m$	1	601.08	17.27
Teachers (*T*)	$T - 1$	$\sigma^2_{s:tm} + nM\sigma^2_t$	2	57.00	1.43
MT	$(M - 1)(T - 1)$	$\sigma^2_{s:tm} + n\sigma^2_{tm}$	2	34.81	.87
Students nested *within* teachers and methods (*S*:*TM*)	$MT(n - 1)$	$\sigma^2_{s:tm}$	84	39.98	

analysis (Model E). The results of the analyses are given as the Example in panels *a* and *b* of Table III when group means (Model D) and student scores (Model E) are used as the unit of analysis, respectively.

Note that *if the proper ANOVA model is employed,* the question of the proper unit of analysis is taken care of implicitly. When the proper ANOVA model is used, although the analyses are identical as far as the method effect is concerned, the analyses using individual students in Model E are preferred because the hypothesis concerning the method-by-teacher interaction can be evaluated. In addition, by retaining individual scores in the analysis, the researcher can consider incorporating personological variables into the design so that interactions of these factors with treatment effects can be evaluated. These interactions speak directly to critical generalizability questions.

Models in which students are used as the observational unit (like Models *C* and *E*) have the advantage of allowing the researcher to empirically test the statistical model to see if it might be simplified, that is, the pooling option can be considered. Notice in Table III that in spite of the large *F*-ratio (17.27) for the method effect, the null hypothesis for the treatment effect cannot be rejected (at $\alpha = .05$) because the error mean square has only two degrees of freedom, and hence the critical *F* is very large ($_{.95}F_{1,2} = 18.5$). If Model E can be simplified by finding that $\sigma^2_{\alpha\beta} = 0$ is tenable when tested with good power, the null hypothesis for the method effect becomes less tenable.

The hypothesis for the method-by-teacher interaction ($\sigma^2_{tm} = 0$) in Model E (and the teachers-within-method, $H_o{:}\sigma^2_{t:m} = 0$, in Model C) will often be tested with good power since the error term (the variance among students-within-teacher-and-method, $\sigma^2_{s:tm}$) will frequently have many degrees of freedom.

If there is a lack of independence among pupils' scores, the variance for the teacher-by-method interaction in Model E will ordinarily be larger than the error mean square ($\sigma^2_{s:tm}$), thereby nullifying the pooling option (see Glass & Stanley, 1970, pp. 501–507).

In Table I with Model C, the model should not be simplified since the F-ratio for the $T{:}M$ effect is large (4.19, $p < .05$). But in Table III with the example using Model E, the F-ratio for the MT interaction is less than 1.0, hence many researchers following Winer (1971), Kirk (1968), Myers (1979), or Green and Tukey (1960) would simplify the model. Given that $H_o{:}\sigma^2_{tm} = 0$ is tenable and tested with good power, the model can be simplified, and the σ^2_{tm} component can be deleted in Model E and wherever it appears in the E(MS) expressions for the various effects. The pooled estimate of $\sigma^2_{s:tm}$ in the example in Model E is 39.86 and the F-ratio for the method effect would become 15.08 ($p < .01$).

Note that the result of the analysis in Table III when pooling is employed is very similar to the result that would be obtained had individual observations been used (and instructors been ignored or viewed as a fixed factor) in the analysis. But this is not the type of analysis frowned on by Lindquist; there is an important difference. What Lindquist criticized was, in effect, a priori pooling—blind pooling without any statistical safeguards. If indeed there is nonindependence among individuals, a priori pooling is inappropriate and would greatly increase the probability of a Type I error.

If the teacher factor is ignored in the analysis, as in Model A in Table I, the only sources of variation represented in the analysis are methods and students, and there is de facto pooling that greatly increases the probability of spurious significance since the degrees of freedom for the error mean square for the F-test for the method effect are much too large, hence the critical F is too small. The pooling strategy in Model C (Table I) would allow one to pool only when $H_o{:}\sigma^2_{t:m} = 0$ has been shown empirically to be tenable. In Model E, pooling is legitimized only when $H_o{:}\sigma^2_{tm} = 0$ is tenable. Of course, these hypotheses must be tested with good power—which suggests that perhaps α should often be relaxed to .20 or .25, especially if the degrees of freedom for the error term are not large. Pooling is usually less undesirable than making a Type II error. When several levels of random factors are included in the design, the pooling is less apt to be needed (Scheffé, 1959, pp. 126–127). The principal point of this paper, however, does not presuppose a pooling strategy, but illustrates that when the proper statistical model is

specified with a balanced design, the choice of the observational unit employed will not affect the *F*-test for any common effects that are tested, and that the inclusions of student scores will yield more information and analysis options.

An Illustration

A study (Gehler, 1979) compared two methods of kindergarten instruction. Six kindergarten teachers were nested within each of two methods, but crossed the time-of-class (AM/PM) factor. Scores from 18 pupils per class are used in the analysis. Table IV gives results from the ANOVAs when class means (Model F, panel *a*) and when individual pupil scores (Model G, panel *b*) are used as the observational unit. Teachers and pupils are defined as random effects.

Note that, although the two analyses (panels *a* and *b* in Table IV) yield different *MS* values for common sources of variation, that they yield identical *F* values for the three common hypotheses (*M*, *A*, and *MA*). The pupil level analysis (Model G) has the advantage of providing information on two other effects: (1) differences among teachers (within methods) and (2) the generalizability of any teacher difference across the time-of-day factor. In the example, the teacher-AM/PM interaction is tested and the AM/PM mean difference was found to generalize across teachers.

TABLE IV

Results of a Balanced ANOVA Design in which n Pupils are Nested Within T Teachers who are Nested within M Methods, but Cross the AM/PM Factor, when Class Means (Panel A) and Pupil Scores (Panel B) are Used as the Observational Unit

SV	ν	$E(MS)$	*MS*	*F*
a)	[Model F: $X_{atm} = \mu + \alpha_m + \beta_{t:m} + \gamma_a + \alpha\gamma_{am} + \beta\gamma_{at:m}$]			
Methods (*M*)	1	$\sigma^2_{t:m} + 12\sigma^2_m$	104.8	.64
Teachers (*T*:*M*)	10	$\sigma^2_{t:m}$	162.9	...
AM/PM (*A*)	1	$\sigma^2_{a:tm} + 12\sigma^2_a$	96.3	10.02*
MA	1	$\sigma^2_{a:tm} + 6\sigma^2_{am}$	13.8	1.43
AT:*M*	10	$\sigma^2_{at:m}$	9.61	
b)	[Model G: $X_{iatm} = \mu + \alpha_m + \beta_{t:m} + \gamma_a + \alpha\gamma_{am} + \beta\gamma_{at:m} + \epsilon_{i:atm}$]			
Methods (*M*)	1	$\sigma^2_{p:atm} + 18\sigma^2_{t:m} + 216\sigma^2_m$	1887	.64
Teachers (*T*:*M*)	10	$\sigma^2_{p:atm} + 18\sigma^2_{t:m}$	2933	14.89**
AM/PM (*A*)	1	$\sigma^2_{p:atm} + 18\sigma^2_{at:m} + 216\sigma^2_a$	1734	10.02*
MA	1	$\sigma^2_{p:atm} + 18\sigma^2_{at:m} + 108\sigma^2_{ma}$	248	1.43
AT:*M*	10	$\sigma^2_{p:atm} + 18\sigma^2_{at:m}$	173	.88
Pupils (*P*:*ATM*)	408	$\sigma^2_{p:atm}$	197	

* $p < .05$
** $p < .001$

If teachers were viewed as a fixed factor, a nongeneralizable but "highly significant" method difference would have been obtained $F = 9.58, p < .001$. The same result would have been obtained if a model which ignored the teacher factor had been employed: $F = 1887/260 = 7.25, p < .001$.

The pooling option for testing the method effect in Model G is clearly contraindicated ($F = 14.89$ for the $T{:}M$ effect). The pooling option for testing the AM/PM effect, however, is legitimized because the teacher-by-AM/PM interaction ($AT{:}M$) yielded an F-ratio of less than 1. However, pooling is unnecessary for the AM/PM effect, since in the orthodox analysis the power was sufficient to reveal a significant AM/PM effect.

Observe in Table IV that no teacher-by-AM/PM interaction was evidenced, but that there were significant differences among teachers within methods. Both of these findings are typical, illustrating the fact that the pooling option is much more likely to be possible with effects that cross a random factor (e.g., AM/PM) than for those under which the levels of the random factor are nested (e.g., methods).

A Hierarchical Example

The author was involved in a study in which two schools receiving experimental treatment were compared with two other control schools. All students who had remained in the same school for three years took a standardized achievement test at the end of their third year of formal schooling. Results of the study are presented in Table V. Three obvious choices for the observational unit are school means, teacher (class) means, and student scores. If schools were used as the observational unit (Model H) the study would yield only four observations—the four school means. Model I defines both schools and teachers as random variables; Model J adds pupils as a random variable consistent with the desire to generalize the findings to other schools, teachers, and students "like these." Table V gives the three analyses, where the observations unit is the schools (Model H), teachers (Model I), and students (Model J).

The three F-ratios for the methods effect (and all other common effects) for the three models are identical. Notice, however, that the F-ratio for the schools-within-methods ($S{:}M$) effect is not significant (Models I and J); indeed, $F < 1$. Thus, $H_o{:}\sigma^2_{s:m} = 0$ is tenable, and $\sigma^2_{s:m}$ can be deleted in the expected mean squares (if the model simplification strategy is employed). Therefore, in Models I and J, $S{:}M$ and $T{:}SM$ sources of variation can both be viewed as estimating the same parameter, and thus can be pooled to provide a more powerful test of the principal hypothesis, $H_o{:}\mu_E = \mu_C$. The pooling procedure substantially reduces the critical F value (at $\alpha = .05$) for treatments, from 18.5 to 4.96. But the null hypothesis is sustained even when the more powerful test with the pooled error term is employed. Can the model be further simplified to provide a more sensitive test? If $H_o{:}\sigma^2_{t:sm} = 0$

TABLE V

Expected Mean Squares for a Balanced ANOVA Design Involving n Pupils Nested within the T Teachers, Who are Nested within the S Schools, which are Nested within the M Methods Using Models A, B, and C

SV	*E(MS)*	*v*	*MS*	*F*
a)	[Model H: $X_{sm} = \mu + \alpha_m + \beta_{s:m}$]			
Methods (*M*)	$\sigma^2_{s:m} + 2\sigma^2_m$	1	.04	.06
Schools (*S*:*M*)	$\sigma^2_{s:m}$	2	.71	
b)	[Model I: $X_{tsm} = \mu + \alpha_m + \beta_{s:m} + \gamma_{t:sm}$]			
Methods (*M*)	$\sigma^2_{t:sm} + 3\sigma^2_{s:m} + 6\sigma^2_m$	1	.12	.06
Schools (*S*:*M*)	$\sigma^2_{t:sm} + 3\sigma^2_{s:m}$	2	2.13	.46
	$\sigma^2_{t:sm}$	8	4.64	
c)	[Model J: $X_{igtsm} = \mu + \alpha_m + \beta_{s:m} + \gamma_{t:sm} + \delta_g + \alpha\delta_{gm} + \beta\delta_{gs:m} + \gamma\delta_{gt:sm} + \epsilon_{i:gtsm}$]			
M	$\sigma^2_{p:gtsm} + 24\sigma^2_{t:sm} + 72\sigma^2_{s:m} + 144\sigma^2_m$	1	1.4	.06
S:*M*	$\sigma^2_{p:gtsm} + 24\sigma^2_{t:sm} + 72\sigma^2_{s:m}$	2	25.5	.46
T:*SM*	$\sigma^2_{p:gtsm} + 24\sigma^2_{t:sm}$	8	55.7	8.19*
Sex (*G*)	$\sigma^2_{p:gtsm} + 12\sigma^2_{gt:sm} + 18\sigma^2_{gs:m} + 144\sigma^2_g$	1	72.2	4.29
MG	$\sigma^2_{p:gtsm} + 12\sigma^2_{gt:sm} + 18\sigma^2_{gs:m} + 72\sigma^2_{mg}$	1	5.1	.30
GS:*M*	$\sigma^2_{p:gtsm} + 12\sigma^2_{gt:sm} + 18\sigma^2_{gs:m}$	2	16.8	.62
GT:*SM*	$\sigma^2_{p:gtsm} + 12\sigma^2_{gt:sm}$	8	27.1	3.99*
Pupils: *GTSM*	$\sigma^2_{p:gtsm}$	120	6.8	

Note. The author wishes to express gratitude to Carol Vojir for performing these analyses.
* $p < .05$.

is tenable, then this term could also be dropped from the model and the three sources, *S*:*M*, *T*:*SM*, and *p*:*GTSM* could be pooled to form an error term which would then have 129 degrees of freedom, and the critical value of *F* would drop to 3.92. But in Model J, the one model in which this effect can be tested, teacher differences within schools are highly significant ($F = 8.19$, $p < .001$); thus, the second stage pooling is contraindicated.

What then has been gained in this study by using individual scores rather than group means as the observational unit? First, in comparing Models J with H and I one can be more confident that a Type-II error has not been made regarding the method effect since the null hypothesis continued to be tenable even with a much more powerful test following pooling in Models I and J. Indeed, even when the study is viewed with restricted generalizability to see if the treatment effects are significant, even for *these* schools and *these* teachers, i.e., when schools and teachers are viewed as fixed factors, the treatment conclusion is not altered. In this inferentially impoverished fixed-effects model context, where the variance within ($\hat{\sigma}^2_{p:gtsm}$) is the denominator for all *F* tests, the treatment *F* ratio is only .21. Thus, our confidence that a Type-II error of consequence has not been made is strengthened.

In addition, the student-level analysis (Model J) yields information on several additional sources of variation.

SUMMARY

This paper has shown that the common recommendation to use group means where there may be nonindependence among observational units is unnecessary, unduly restrictive, impoverishes the analysis, limits the questions that can be addressed in a study, and does not insure that the relevant independence assumption has been met. When random factors are properly identified and included in the analysis, the results for all common effects (*F*s and critical *F*s) are identical in balanced ANOVA designs, regardless of the observational unit employed. The use of individual observations, however, also allows other interesting questions pertaining to interaction and generalizability to be explored. In addition, if students are used as the observational unit, the pooling option can be explored. The question of the proper observational unit (or unit of analysis) is answered directly, correctly, and implicitly when the proper statistical model is employed.

REFERENCES

ADDELMAN, S. Variability of treatment and experiment units in the design and analysis of experiments. *Journal of the American Statistical Association*, 1970, *65,* 1,095–1,108.

BARCIKOWSKI, R. S. *Statistical power with group mean as the unit of analysis.* Final Report of contract NIE-G-78-0072 to the National Institute of Education, 1980.

BURSTEIN, L. *The role of levels of analysis in the specification of educational effects.* Education Finance and Productivity Center, undated. Department of Education, University of Chicago.

CAMPBELL, D. T., & STANLEY, J. C. Experimental and quasi-experimental designs in research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand-McNally, 1963.

CRONBACH, L. J. Research on classrooms and schools: *Formulation of questions, design, and analysis.* Stanford, Calif.: Stanford University, 1976.

DEROSIA, P. *A comparative study of pupil achievement and attitudes and involvement of parents of children enrolled in extended-day and half-day kindergarten programs.* Doctoral thesis, University of Colorado, 1980.

FLETCHER, H. J. Possible interpretive problems in analyses using group means as the experimental unit. *Psychological Bulletin*, 1968, *69,* 157–60.

GEHLER, T. M. G. An analysis of kindergarten achievement including the effects of day and sex. Doctoral thesis, University of Colorado, 1979.

GLASS, G. V, PECKHAM, P. D., & SANDERS, J. R. Consequences of failure to meet assumptions underlying the fixed effects 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.

GREEN, B. F. & TUKEY, J. Complex analysis of variance: General problems. *Psychometrika*, 1960, *25,* 127–152.

KEMPTHORNE, O. *The design and analysis of experiments.* New York: Wiley, 1952.

KEMPTHORNE, O. The design and analysis of experiments with some reference to educational research. In *Research design and analysis, second annual Phi Delta Kappa Symposium on educational research*, 1961.

KIRK, R. E. *Experimental design: Procedures for the behavioral sciences.* Belmont, Calif.: Brooks/Cole, 1968.

KOOP, J. B. *A description of the effects of the use of calculators in the community college arithmetic class.* Doctoral thesis, University of Colorado, 1978.

LINDQUIST, E. F. *Statistical analysis in educational research.* New York: Houghton-Mifflin, 1940.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton-Mifflin, 1953.

LUMSDAINE, A. A. Instruments and media of instruction. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally, 1963.

MCNEMAR, Q. *Psychological Bulletin*, 1940, *37,* 747.

MYERS, J. L. *Fundamentals of experimental design* (3rd ed.). Boston: Allyn & Bacon, 1979.

PAGE, E. B. *Recapturing the richness within the classroom.* Paper presented at the annual meeting of AERA, Chicago, February, 1965.

PAGE, E. B. Statistically recapturing the richness within the classroom. *Psychology in the Schools*, 1975, *12,* 339–344.

PECKHAM, P. D., GLASS, G. V, & HOPKINS, K. D. The experimental unit statistical analysis. *Journal of Special Education*, 1969, *3,* 337–349.

RATHS, J. The appropriate experimental unit. *Educational Leadership*, 1967, *12,* 263–266.

SCHEFFÉ, H. *The analysis of variance.* New York: Wiley, 1959.

STECK, J. C. *The independence of observations obtained in classroom research.* Master's thesis, University of Maryland, 1966.

WINER, B. J. *Statistical Principles in Experimental Design* (2nd ed.). New York: McGraw-Hill, 1971.

WRIGHT, D. J. *Groups and experimental units in educational research.* Paper presented at the annual meeting of AERA, Los Angeles, February, 1969.

AUTHOR

KENNETH D. HOPKINS, Director, Laboratory of Educational Research, University of Colorado, Boulder, Colorado 80309. *Specializations:* Research methodology, statistics, measurement.