Journal of Modern Applied Statistical Methods

Volume 5 | Issue 1 Article 19

5-1-2006

The Use Of Hierarchical ANCOVA In Curriculum Studies

Show-Mann Liou
National Taiwan Normal University, t11033@ntnu.edu.tw

Chao-Ying Joanne Peng
Indiana University-Bloomington

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 230-247

1538 – 9472/06/$95.00

The Use Of Hierarchical ANCOVA In Curriculum Studies

Show-Mann Liou
National Taiwan Normal University

Chao-Ying Joanne Peng
Indiana University-Bloomington

Many educational studies are carried out in intact settings, such as classrooms or groups in which individual data were collected before and after a treatment. Researchers advocate either the use of individual scores as the unit of analysis or class means. Both approaches suffer from conceptual and methodological limitations. In this article, the use of hierarchical ANCOVA for analyzing quasi-experimental data including baseline measures is designed and promoted. It is illustrated with a real-world data set collected from a curriculum study. Results showed that the hierarchical ANCOVA is a conceptually and methodologically sound approach, and is better than ANCOVA based on individual scores or ANCOVA based on class means. The potential of using hierarchical ANCOVA designs for curriculum studies is discussed in terms of statistical power and congruence with study plans.

Key words: Educational research methodology, hierarchical ANCOVA, Project Citizen, civic education, civic skills, civic dispositions, adolescent students

Introduction

Among educational research methods, true experiments are designed to investigate causes and consequences in behavior (Fraenkel & Wallen, 2000; McMillan & Schumacher, 2001). However, most circumstances in education prevent the possibility of random selection and random assignment of subjects into experimental and control conditions. Consequently, the use of true experiments is limited in educational research. Instead, quasi-experiments are much more prevalent.

Show-Mann Liou, Associate Professor of Civic Education & Leadership. Research interests include citizenship education and experimental design. C.-Y. Joanne Peng, Professor of Inquiry Research Methodology. Research interests include logistic regression and missing data. Send correspondence to Show-Mann Liou, Department of Civic Education and Leadership, National Taiwan Normal University, Hoping E. Rd. Sec. 1 No. 162, Taipei, Taiwan 106, phone: 886-2-2369-8673 ex.31, fax: 886-2-2363-8821, email: t11033@ ntnu.edu.tw.

Even with quasi-experiments, educational researchers are faced with another difficulty that weakens the internal validity of a study. Namely, students in the same classroom are often administered the same treatment by the same instructor making their performances not statistically independent. Consider a study in which a researcher is interested in studying the effectiveness of two instructional strategies on students' achievement in biology. To carry out this study, a researcher may randomly select intact classes and train teachers of these classes to implement the instructional strategies. Consequently, students in a classroom cannot be randomly assigned to learn from a particular strategy, nor can teachers teach students independently or in isolation. To account for the difference in students' achievement that already existed in the beginning of the study and to compensate for the lack of independence among students' performances, a researcher can administer a pretest to determine a baseline measure of the outcome (i.e., biology achievement in this case). A one-way analysis of covariance (ANCOVA) can be subsequently applied to posttest measures to test differences due to the two strategies while statistically controlling for pretest differences. The ANCOVA approach has been a method of choice since Lindquist (1940) brought to light the issues with non-independence in subjects' responses in intact groups.

It is generally agreed that ANCOVA is an appropriate statistical technique for analyzing quasi-experimental data with baseline measures as long as its assumptions—linearity and independence between the covariate and the independent variable—are met (Buser, 1995; Henson, 1998; Hines & Foil, 2000; Loftin & Madison, 1991). There is, however, one issue remaining: what is the proper unit of analysis in quasi-experimental studies, class means or individual scores? (Barcikowski, 1981; Blair & Higgins, 1986; Hopkins, 1982; Morran, Robison, & Hulse-Killacky, 1990; Peckham, Glass, & Hopkins, 1969).

The issue has generated and received considerable attention in the literature ever since Lindquist (1940) presented an argument and rationale for using group means as the unit of analysis for data collected from intact groups. At the heart of the disagreement is: what is the most appropriate unit for data analysis and interpretation? With the use of individual scores, it is assumed that students in the same classroom are unrelated, as far as treatments are concerned, and therefore statistically independent. This assumption and its computational approach could lead to an overestimation of treatment effects with sufficiently large samples. Conversely, using group means as the unit of analysis ensures that the independence assumption is met, at the individual level, and the interpretation of the data has internal validity (Peckham, Glass, & Hopkins, 1969). However, this approach results in a great loss in sample size; hence, a decrease in statistical power (Barcikowski, 1981). Furthermore, the use of group means limits the generalizability of the findings only to classes, and results may not be informative to educators in general. It is evident from the brief summary that each approach has its own conceptual and methodological limitations.

This article addresses the limitations raised above regarding the use of these two traditional ANCOVAs, one based on individual's scores and the other on group means, and proposes a third approach. This approach applies the hierarchical ANCOVA to data collected from intact settings such as classrooms. It will be shown that the hierarchical ANCOVA is a conceptually and methodologically sound analytical approach that is well suited to educational research. Specifically, this approach isolates the nuisance variable of classes and incorporates the inherent hierarchical nature of the data structure into the analysis. Consequently, this approach not only takes into account the independence assumption required of individuals' scores but also makes valid and meaningful inferences at the individual's level.

The hierarchical ANCOVA is introduced and demonstrated using a real world data set (Liou, 2002). The Liou study was primarily interested in the effects of *We the People…Project Citizen* on civic skills and four dimensions of the civic dispositions of adolescent students. The study exemplified most educational research in which classrooms are randomly selected or even assigned to treatment conditions but students are not. Furthermore, students' levels of civic skills and civic dispositions were assessed both before and after the implementation of *Project Citizen*. Data were analyzed by three methods: ANCOVA based on individual scores, ANCOVA based on class means, and hierarchical ANCOVA based on individual scores. Results from the three methods were shown to be different; they were interpreted in terms of substantive implications and methodological considerations (i.e., statistical power, practical as well as statistical significance). Recommendations and implications for educational researchers are offered in light of the relative superiority of hierarchical ANCOVA over the other two methods.

Design Structures: Crossed and Nested (Hierarchical) Designs

To ensure the internal and external validities of statistical analysis of quasi-experiments, one should carefully plan two aspects of a study: the structure of the design and the unit of analysis. Specifically, two major structures are possible for a quasi-experimental design: crossed and nested (or hierarchical) (Peng, 2004). Likewise, two types of units of analysis need to be distinguished conceptually and computationally: the unit of research design and the unit of statistical analysis.

A crossed design employs all combinations of levels of two or more independent variables in a study. It is typically used to test differences in a dependent variable due to main effects of independent variables and their interactions. A nested design is a research design in which levels of one independent variable (say *B*) are hierarchically subsumed under (or nested within) levels of another independent variable (say *A*). As a result, assessing the complete combination of *A* and *B* levels is not possible in a nested design.

Nested design is alternatively called hierarchical design; it is used most often in quasi-experimental studies in which researchers have little or no control over random assignment of observations into treatment conditions. The design is popular, and sometimes necessary, among curriculum studies, clinical, sociological, and ethological research in which participants belong to intact groups (such as classes, therapeutic groups, etc.); these intact groups cannot be dismantled to allow for a random assignment of participants into different treatment conditions.

Many studies in education can be carried out only in nested designs. Consider the example mentioned earlier in which instructional strategies are administered in classroom settings. Even though students individually learn and are tested on their achievement in biology, their learning effects are to an extent dependent on the learning environment and dynamics of interactions among peers. Thus, students are nested within classrooms which in turn are nested within instructional strategies. In this case, a crossed design neglects the hierarchical nature of the data and produces incorrect interpretations of the results. According to Roberts (2000), neglecting a nested design leads to the following consequences:

> Neglecting a nested design when one actually exists will make the research: (1) wrongly attribute a main effect to an interaction effect when, in fact, no interaction exists; (2) divide by the wrong degrees of freedom when determining the mean square and *F*-value (and the statistical significance of the *F*-value); and (3) assume that a main effect has a smaller effect size (eta square) because the sum of squares for that effect is being partly attributed to the interaction effect. (Roberts, 2000, p. 6)

Unit of Research Design and Unit of Statistical Analysis

Another issue that should be taken into consideration when analyzing quasi-experiments is the unit of analysis. Valid statistical inferences from data depend on the compatibility between the unit of a research design and that of statistical analysis (Peckham, Glass & Hopkins, 1969; Glass & Stanley, 1970; Morran, Robinson, & Hulse-Killacky, 1990). Units of a research design refer to entities that are allocated to a condition of the independent variable, independently from other entities. Units of statistical analysis refer to entities whose measures or scores form the basis of statistical inferences. Clearly, a research design unit can be either individuals or classes. Even if classes are the research design units, students' scores can still be treated as units of statistical analyses.

When analyzing data in an ANOVA framework, educational researchers may, and frequently do, make an a priori decision to treat individuals as the unit of statistical analysis (Morran, Robinson, & Hulse-Killacky, 1990). Several reasons contribute to this decision. One is to ensure that the statistic, whether it is *F*, *q*, or *t*, is tested with the maximal *df* based on the sample. Another reason for regarding individuals as the unit of analysis is to retain the variability at the individual level, thus, maximizing information a research can glean from the data. This approach further affords researchers the opportunity to study the effects of certain organismic or demographic characteristics and their interactions with independent variables on the dependent variable (Hopkins, 1982; Morran, Robinson, & Hulse-Killacky, 1990; Peckham, Glass, & Hopkins, 1969). It is impossible to study these effects if group means are analyzed. Thus, the group means approach ignores the hierarchical nature of the data collected in typical educational settings and consequently impoverishes inferences that may be drawn at the individual level.

Yet, a few researchers advocate the use of group means on statistical grounds. They argue that participants studied in intact settings are not the appropriate unit of analysis since they fail to meet the independence assumption. The result of such a violation is deflated within-group variability, hence, inflated treatment effects. In a typical educational setting, the classroom provides a shared educational experience; thus, students are not entirely independent insofar as sampling errors are concerned. According to Peckham, Glass, and Hopkins (1969), "violating the assumption of independence of errors may substantially affect the validity of probability statements" (p.338). They concluded that the use of group means promotes "the greatest insurance that the independence assumption has been met" (p.344); and therefore statistical inferences from the result are valid. Some proponents went further in arguing that when the independent assumption is not tenable, treating individuals as the unit of statistical analysis leads to non-replicable findings.

As Hopkins (1982) showed that the recommendation of using class means proves to be restrictive, unnecessary, and less powerful than alternatives that are derived directly from individual data and proper statistical models. A better treatment of the inter-dependence among units of observation is to employ an efficient statistical modeling technique, such as the hierarchical ANCOVA, that adequately represents the condition under which data were collected and provides the greatest statistical power and external validity.

Hierarchical ANCOVA

In light of the issues raised in the preceding two sections, it is not without understanding that the two ordinary ANCOVA's – one based on class means and the other on individual scores – are unlikely to yield satisfactory interpretation of data collected from hierarchical settings that include pretests or baseline measures. In their places, researchers have proposed that nested or hierarchical ANCOVA be used in order to account for variances due to treatments, classes, and individual students nested within classrooms (Hopkins, 1982; Lindman, 1992; Morran, Robison & Hulse-Killacky, 1990; Robert, 2000). Hierarchical ANCOVA combines features from a hierarchical research design with those of analysis of covariance.

Assume that a researcher wishes to study the effect of Internet search strategies (Factor *A*) on college students' information seeking efficiency (the dependent variable). Six classes of freshmen English at a state college are randomly selected; three classes are assigned to the linear search condition and the other three to the nonlinear search condition. At the onset of the study, all freshmen are assessed in terms of their information seeking efficiency. These measures will be treated as covariates in analysis of covariance. Figure 1 illustrates the research design.

Because freshmen enrolled in these classes form intact groups, they cannot be randomly assigned to the two treatment conditions on an individual basis. Furthermore, their learning processes and behaviors are likely to be mutually dependent; differences in students' information seeking behavior among classes are embedded within each treatment condition. This restriction makes this design a nested design rather than a fully crossed design. In addition, the pretest measures taken from all participants can serve as a covariate in the hierarchical ANCOVA model presented below:

$$Y_{ijk} = \mu_y + beta(X_{ijk} - \mu_x) + \alpha_j + \beta_{k(j)} + e_{i(jk)}, \qquad (1)$$

Where

i = 1, …,n (number of freshman in a class, say, 20);

j = 1, …,p (number of treatment condition=2 in this example);

k= 1,…,q (number of classes=3 in this example);

		Factor *A*: Internet Search Strategy			
		Treatment 1 Linear		Treatment 2 Nonlinear	
		Pretest	Posttest	Pretest	Posttest
Factor *B*: Freshman English Class	Class 1	$\bar{X}_{1(1)}$	$\bar{Y}_{1(1)}$		
	Class 2	$\bar{X}_{2(1)}$	$\bar{Y}_{2(1)}$		
	Class 3	$\bar{X}_{3(1)}$	$\bar{Y}_{3(1)}$		
	Class 4			$\bar{X}_{4(2)}$	$\bar{Y}_{4(2)}$
	Class 5			$\bar{X}_{5(2)}$	$\bar{Y}_{5(2)}$
	Class 6			$\bar{X}_{6(2)}$	$\bar{Y}_{6(2)}$

Figure 1

Y_{ijk} is the dependent score of the *i*th participant in the *j*th level of Factor *A* and *k*th level of Factor *B*;

μ_y is the population mean of the dependent scores;

beta is the pooled within-group regression coefficient derived from regressing the covariate score, X_{ijk} on the dependent score Y_{ijk};

X_{ijk} is the covariate measure (such as the pretest score) of the *i*th participant in the *j*th level of Factor *A* and *k*th level of Factor *B*;

μ_x is the population mean of the covariate measures;

α_j is the effect of the *j*th treatment condition of Factor *A*; algebraically, it equals the deviation of the *j*th population mean (μ_{y_j}) from the grand mean (μ_y). It is a constant for all participants' dependent scores in the *j*th condition, subject to the restriction that all α_j sum to zero across all conditions.

$\beta_{k(j)}$ is the effect of the *k*th condition under Factor *B*, nested within the *j*th level of Factor *A*; algebraically, it equals the deviation of the population mean ($\mu_{y_{jk}}$) in the *k*th and *j*th combined level from the grand mean (μ_y). It is a constant for all observations' dependent scores in the *k*th condition, nested within Factor *A*'s *j*th condition. The effect is assumed to be normally distributed in its underlying population.

$e_{i(jk)}$ is the random sampling error associated with the *i*th participant in the *j*th condition of Factor *A* and *k*th condition of Factor *B*. It is a random variable that is normally distributed in the underlying population and is independent of $\beta_{k(j)}$. In comparison, the ordinary ANCOVA model based on individual scores does not examine nor acknowledge the nested effect, $\beta_{k(j)}$ in its model as follows:

$$Y_{ijk} = \mu_y + beta(X_{ijk} - \mu_x) + \alpha_j + e_{ijk}, \quad (2)$$

where all terms are defined as previously, except that there is no $\beta_{k(j)}$ effect and no nested effect of classes within treatment conditions.

A third approach, i.e., the ordinary ANCOVA based on class means, follows the same model as model (2) except that data are aggregated over the entire class before they are analyzed by the ANCOVA model as stated below:

$$Y_{jk} = \mu_y + beta(X_{jk} - \mu_x) + \alpha_j + \varepsilon_{jk}, \quad (3)$$

where

j = 1, …,p (number of treatment condition=2 in this example);

k= 1,…,q (number of classes=3 in this example);

Y_{jk} is the average dependent score of the kth class in the jth level of Factor A;

μ_y is the population mean of average class dependent scores;

beta is the within-group regression coefficient derived from regressing the covariate score, X_{jk} on the dependent score Y_{jk};

X_{jk} is the average covariate measure (such as the pretest score) of the kth class in the jth level of Factor A;

μ_x is the population mean of average class covariate measures;

α_j is the effect of the jth treatment condition of Factor A; algebraically, it equals the deviation of the jth population mean (μ_{y_j}) from the grand mean (μ_y). It is a constant for all class average dependent scores in the jth condition, subject to the restriction that all α_j sum to zero across all conditions;

e_{jk} is the random sampling error associated with the kth class in the jth condition of Factor A. It is a random variable that is normally distributed in the underlying population.

Note in model (3), the i subscript is no longer present due to the fact that individuals are not the unit of analysis. Instead, class means are used; they are denoted by the k subscript.

Statistical Assumptions and Tests

The null hypothesis (H_0) for all the three models is identical, namely, the parameter α_j equals zero in the population for all conditions (or linear search and nonlinear search according to the present example). The alternative hypothesis (H_1) states that some of the α_j's do not equal zero. To test the null hypothesis according to models (1), (2), or (3), data are organized to form a ratio of mean squares treatment (MS_t) over mean squares error (MS_e). The ratio is distributed as a central F distribution under the null hypothesis but non-central F distribution under the alternative, provided that statistical assumptions are met. For all three models, it is assumed that random sampling errors [$e_{i(jk)}$, e_{ijk}, or e_{jk}] are normally distributed, homogeneous in variances, and independent from each other in the population. Furthermore, the covariate (pretest in the example) is assumed by three models to be linearly related with the dependent variable, independent of the independent variable, homogeneous in regression slopes and variances, and measured without errors. Finally, for Model (1) alone, it is assumed that the $\beta_{k(j)}$ effect is normally distributed in its underlying population, as stated earlier.

It might be asked why researchers need three models when any of the three can be used to test the null hypothesis. The answer lies in selecting a model that renders the greatest statistical power and the least bias. In terms of statistical power, the hierarchical ANCOVA model in (1) enables a researcher to separate the nuisance variable of classrooms that may affect the participant's performance on the dependent variable, from the sampling error. The inclusion of the nested effect $\beta_{k(j)}$ in Model (1) effectively removes a portion of the sum of squares due to this effect from the error sum of squares (or SS_e). Consequently, the magnitude of SS_e in Model (1) is smaller than that in Model (2). The reduction in SS_e is accompanied by a reduction in degrees of freedom for the error term as well. As it will be shown with real world data in the next section, if the reduction in SS_e is sizeable, it can offset the loss in degrees of freedom. Hence, the MS_e ($=SS_e/df_e$) is made smaller in Model (1) than in Model (2). A smaller MS_e in the denominator of an F-ratio inevitably leads to a greater F statistic and potentially more powerful F test. Compared with Models (1) and (2), Model (3) has the lowest statistical power because it aggregates data over all participants in a classroom. This approach reduces the sample size (in terms of number of classes, rather than number of individuals) and therefore the statistical power.

All three models employ a covariate to statistically adjust differences due to covariates in nonrandomized studies, or to provide a more precise estimation of the treatment effect (i.e., α_j) in randomized studies. Thus, three models are comparable in these regards. In the next section, the application of hierarchical ANCOVA is illustrated in a curriculum study. Results of this application will be contrasted with those obtained from two ordinary ANCOVA's based on individual scores and class means, respectively. The empirical evidence based on real data will support the recommendation for the hierarchical ANCOVA as a conceptually sound and analytically powerful method for interpreting data gathered from intact groups that also include a pretest or baseline measure.

An Illustration

To help illustrate the superiority of hierarchical ANCOVA modeling over two ordinary ANCOVA's, a real world data set with all three methods was analyzed. Results will be shown to be different. They are discussed in terms of interpretability, generalizability, and statistical power.

Data Set and Its Related Study

Data came from a curriculum study by Liou (2002), which was carried out in Taiwan. There were dramatic political changes in Taiwan in recent years. These political changes created a society that is becoming politically more open and democratic than ever before. In order to prepare citizens for future developments of a truly democratic society and the rule of law, the civic curricula in the Taiwanese educational system aim at cultivating in students the knowledge, skills, and dispositions indispensable for such developments and fostering a participatory perspective. However, civic education faces formidable barriers, most notably a gap between pedagogical theory and classroom practice, and a conventional emphasis on the acquisition of factual knowledge regarding the political system instead of actual civic participation. Consequently, the goal of adequately preparing democratic citizens through education is not being fulfilled.

Project Citizen is a civic education program for middle school students. The program actively engages students in learning how to monitor and influence public policy through an interactive and cooperative process. It is typically implemented as a class project. For the project, students work together to identify and study a public policy issue, eventually developing an action plan for implementing their policy solution. According to its developers, the goal of *Project Citizen* is to motivate and empower adolescents to exercise their rights and to accept the responsibilities of democratic citizenship through the intensive study of a local community problem. Specifically, *Project Citizen* is designed to help adolescents:

- learn how to monitor and influence public policy in their communities;
- learn the public policy-making process;
- develop concrete skills and the foundation needed to become responsible participating citizens;
- develop effective and creative communication skills; and
- develop more positive self-concepts and confidence in exercising the

> rights and responsibilities of citizenship. (Center for Civic Education, 2000)

In light of the goals of *Project Citizen* and problems facing Taiwan's civic education, it seems that *Project Citizen* can be used as a curriculum supplement to remedy some of the weaknesses of Taiwan's civic education and to help Taiwan prepare participatory citizens. Consequently, Liou conducted the study to evaluate the effects of *Project Citizen* on the civic skills and dispositions of adolescent students in Taiwan.

Research Design

For administrative reasons, it was deemed impractical to randomly assign students into different pedagogical conditions. Therefore, the study employed a pretest-posttest quasi-experimental design with one treatment and one comparison conditions. Twelve Taiwanese high school teachers, each teaching one experimental and one comparison class, participated in this research. Classes taught by the same teacher were randomly assigned to either the treatment or the comparison condition. In the fall of 2001, students in the experimental classes received instruction in *Project Citizen* as an adjunct to the traditional instruction of *Civics* or *Three Principles of the People*. The comparison students received traditional, discipline-based instruction that focused on the hierarchical model of knowledge acquisition. Liou collected data from 942 students on the pre- and post-treatment assessment of their civic skills and civic dispositions along with their demographic, experiences, teacher-related, and school-related information.

Measurements

To help illustrate the hierarchical ANCOVA approach, students' pre-test and post-test of the civic skills and four dimensions of civic dispositions as a function of their group (treatment versus comparison) information were analyzed; all extracted from Liou's study (2002). Civic skills are those intellectual and participatory capacities that enable active involvement in civic life (Vontz, et al., 2000). Civic dispositions are those traits of public and private character that contribute to both the political efficacy of the individual and the common good of society (Vontz, et al., 2000). Civic dispositions in the Liou study were operationalized by summing the mean scores derived from four subscales of Adolescent Student Civic dispositions Scale (ASCDS): Politic Interest, Propensity to Participate in Future Political Life, Commitment to Rights and Responsibilities of Citizenship, and Sense of Political Efficacy.

Means on the civic skills and dispositions ranged from 1 to 6; the higher the score, the better was the performance. Descriptive information about the pre-test and the post-test of civic skills and civic dispositions is presented in Table 1. The post-test means were adjusted for the pre-test scores using the ANCOVA approach based on individual scores. The group information was coded dichotomously, 1 for the experimental group (participated in *Project Citizen*) and 2 for the comparison group (did not participate in *Project Citizen*). There were equal numbers of students in each group.

Research Hypothesis and Data Analyses

The research hypothesis posted to data was: there was significant difference between experimental and comparison students in their civic skills and four dimensions of civic disposition, namely, political interest, propensity to participate, commitment of rights and responsibilities of citizenship, and sense of political efficacy due to the implementation of *Project Citizen*. To test this research hypothesis, three statistical procedures were applied to the data: ANCOVA based on individual scores, ANCOVA based on class means, and hierarchical ANCOVA based on individual scores. The statistical model underlying ANCOVA based on individual scores was Model (2); Model (3) underlay ANCOVA based on class means, and Model (1) for hierarchical ANCOVA based on individual scores. All three ANCOVA's treated the post-test scores of the five outcome variables as the dependent

Table 1. Descriptive Information about the Sample Data.

Outcome variables	Group	Pretest		Adjusted Posttest
		Mean	SD	Mean
Civic skills	Experiment	3.45	.85	3.62
	Comparison	3.60	.80	3.45
Political interest	Experiment	3.40	.87	3.47
	Comparison	3.55	.86	3.38
Propensity to participate	Experiment	3.61	.78	3.64
	Comparison	3.67	.72	3.56
Commitment of rights and responsibilities of citizenship	Experiment	5.22	.51	5.11
	Comparison	5.19	.53	4.97
Sense of political efficacy	Experiment	4.47	.84	4.49
	Comparison	4.41	.81	4.42

Note. Full sample: N=942. Females: n_f = 475 (50.4%). Males: n_m = 467 (49.6%). Experimental group: n_e = 471 (50%). Comparison group: n_c = 471 (50%).

variables and the pre-test scores as the covariate. The independent variable was the implementation (or lack of) of *Project Citizen* in civic education curriculum. Prior to analyses, statistical assumptions such as normality, equal variance, independence of errors, linearity between pretest (the covariate) and posttest scores, and common slope for all treatment conditions were examined. All assumptions associated with the three procedures were satisfactorily met. Appendix A lists SAS® programming codes for examining these assumptions.

Based on the rationale and previous research, it was hypothesized that *Project Citizen* would have a positive impact on adolescent's civic skills and civic dispositions. Hence, statistical tests pertaining to the research hypothesis were conducted as one-tailed at an alpha level of .025. It was also decided that univariate tests were preferred over multivariate tests of all five dependant variables because the objective of this article was to compare models, instead of accounting for underlying relationships among these dependant variables. The data were analyzed using SAS® version 8.2 (SAS Institute Inc., 1999) and SPSS® version 10 (SPSS Inc., 1999) in the Windows 2000 environment.

ANCOVA Results Based on Individual Scores

Data of the 942 observations were submitted to the GLM procedure in SPSS® version 10 to determine the effect of *Project Citizen* on the civic skills and dispositions of Taiwanese adolescents. Univariate ANCOVA results based on individual scores are shown in Table 2. The five *F*-tests were carried out using MS_{error} as the denominator. An examination of the results indicated that students participating in *Project Citizen* significantly outperformed students in the comparison group on civic skills and three dimensions of civic dispositions including political interest, propensity to participate, and commitment to rights and responsibilities of citizenship. The two groups were comparable on the fourth dimension of civic disposition, namely, sense of political efficacy.

ANCOVA Results Based on Class Means

The second ANCOVA procedure used class means instead of individual scores as the unit of statistical analysis. In order to perform ANCOVA based on class means, data were first aggregated by classes resulting in 24 classroom means (12 treatment class means with 471 students and 12 comparison class means with 471 students). ANCOVA was subsequently

Table 2. ANCOVA Results Of Civic Skills And Four Civic Dispositions Subscales Using Individual Scores As The Unit Of Analysis

Source	SS	*df*	MS	*F*	*p*
Civic skills					
Group	7.93	1	7.93	19.89	< .001**
Error	374.352	939	.399		
Political interest					
Group	1.62	1	1.62	4.15	.011*
Error	365.45	939	.389		
Propensity to participate					
Group	1.17	1	1.17	4.29	.010*
Error	255.78	939	.272		
Commitment to rights and responsibilities of citizenship					
Group	4.98	1	4.98	17.12	< .001**
Error	273.26	939	.291		
Sense of political efficacy					
Group	1.22	1	1.22	2.44	NS [a]
Error	468.86	939	.499		

* $p < .025$ (one-tailed), ** $p < .01$ (one-tailed).
[a] Not significant at $\alpha = .025$.

applied to these 24 class means using the GLM procedure in SPSS® version 10. Results are shown in Table 3. According to Table 3, students participating in *Project Citizen* significantly outperformed students in the comparison group on civic skills. Furthermore, two dimensions of civic dispositions, namely, propensity to participate and commitment to rights and responsibilities of citizenship were also found to be significant with experimental students outperforming comparison students.

Table 3. ANCOVA Results of Civic Skills and Four Civic Dispositions Subscales with Class Means as The Unit Of Analysis.

Source	SS	*df*	MS	*F*	*p*
Civic skills					
Group	.19	1	.19	10.77	.001**
Error	.37	21	.018		
Political interest					
Group	.037	1	.037	2.66	NS [a]
Error	.288	21	.014		
Propensity to participate					
Group	.039	1	.039	3.21	.022*
Error	.254	21	.012		
Commitment to rights and responsibilities of citizenship					
Group	.111	1	.111	5.40	.008*
Error	.431	21	.021		
Sense of political efficacy					
Group	.020	1	.020	1.07	NS [a]
Error	.393	21	.019		

* $p < .025$ (one-tailed), ** $p < .01$ (one-tailed).
[a] Not significant at $\alpha = .025$.

Hierarchical ANCOVA Results

The results of the hierarchical ANCOVA are presented in Table 4 that treated intact classes as nested in the two experimental conditions and students nested in classes. As shown in Table 4, students participating in *Project Citizen* significantly outperformed students in the comparison group in civic skills and also in three dimensions of civic dispositions, namely, political interest, propensity to participate, and commitment to rights and responsibilities of citizenship.

SAS® programming codes for performing the hierarchical ANCOVA is provided in Appendix A for each of the dependent variables. Note that for each dependent variable (such as civic skills); two statistical procedures in SAS® were applied to data: PROC REG and PROC GLM, twice. The purpose of each statistical analysis is explained in the TITLE statement immediately preceding the RUN; statement. For example, the purpose of REG procedure was to test the linearity assumption regarding the linear relationship between the covariate and the dependent variable. The linear relationship was assumed within each condition as well as for the entire data set. The first GLM procedure was to apply the ANCOVA model to the data according to equation (1) presented earlier. The second GLM procedure was to test the equal slope assumption assumed by the ANCOVA model. This assumption was tested via the interaction between the covariate (i.e., pretest) and the independent variable (participating in *Project Citizen* or not). Non-significant *F* test results were obtained for all five dependent variables indicating that the equal slope assumption was met.

Table 4. Hierarchical ANCOVA Results for Civic Skills And Four Civic Dispositions Subscales Using Individual Scores as The Unit of Analysis

Source	SS	*df*	MS	*F*	*p*
Civic skills					
Group	7.37	1	7.37	10.89	< .001**
Class (Group)	14.90	22	.677	1.73	.0201
Error	359.46	417	.391		
Political interest					
Group	1.803	1	1.803	3.53	.019*
Class (Group)	11.233	22	.511	1.32	.1466
Error	354.219	917	.386		
Propensity to participate					
Group	1.280	1	1.280	2.81	.024*
Class (Group)	10.031	22	.156	1.70	.0232
Error					
Commitment to rights and responsibilities of citizenship					
Group	4.8855	1	4.885	6.03	.006*
Class (Group)	17.815	22	.810	2.91	< .001**
Error	255.441	917	.279		
Sense of political efficacy					
Group	1.062	1	1.062	1.43	NS [a]
Class (Group)	16.315	22	.742	1.50	.0643
Error	452.549	917	.494		

* $p < .025$ (one-tailed), **$p < .01$ (one-tailed).
[a] Not significant at $\alpha = .025$.

Comparison of Three Results

Results obtained from three statistical approaches regarding the research question are contrasted in Table 5. For civic skills, propensity to participate, commitment to rights and responsibilities of citizenship, and sense of political efficacy, there was agreement among the three approaches. For the political interest of Taiwanese adolescent students, ANCOVA based on class means yielded a non-significant result; this contrasted with a significant finding ($p < .025$) obtained from the hierarchical ANCOVA and ANCOVA based on individual scores. As stated earlier, ANCOVA based on class means aggregated scores into class means leading to great loss in units of analysis and therefore, statistical power, compared to the other two approaches. Further, findings from the means approach limit the interpretation and generalizability to class averages only—a result not useful or relevant to most educators or parents.

The hierarchical ANCOVA approach yielded results comparable to those obtained from ANCOVA based on individual scores. Yet, the hierarchical approach uncovered additional class differences that could not be found by ANCOVA based on individual scores due to its model configuration. As shown in Table 4 in gray areas, the 12 classes nested in each treatment condition exhibited statistically significant differences ($p < .05$, two tailed) on civic skills, propensity to participate, and commitment to rights and responsibilities of citizenship. On sense of political efficacy, class differences were significant at the $p < .10$ (two-tailed) level but not at .05.

Table 5. Comparison Of Three ANCOVA Results For Civic Skills And Four Civic Dispositions Subscales

Source	Hierarchical ANCOVA	ANCOVA (Individual Scores)	ANCOVA (Class Means)
	p	*p*	*p*
Civic skills	< .001**	< .001**	<.001**
Political interest	.019*	.011*	NS [a]
Propensity to participate	.024*	.010*	.022*
Commitment to rights and responsibilities of citizenship	.006*	< .001**	.008*
Sense of political efficacy	NS [a]	NS [a]	NS [a]

* $p < .025$ (one-tailed), **$p < .01$ (one-tailed).

[a] Not significant at $\alpha = .025$.

These differences merited further investigation as to why and how these differences existed, as well as to what extent these differences were due to teacher-related, school-related, or student-related characteristics. Research into these class differences can be a worthy endeavor; findings may suggest curricula or cultural changes to schools or classes in order to bring about equality.

Implications for Educational Researchers

In this article, the application of hierarchical ANCOVA for analyzing quasi-experimental data including baseline measures is demonstrated. This procedure is illustrated with a real-world data set to investigate the effect of *Project Citizen* on Taiwan adolescent students' civic skills and four dimensions of civic dispositions, namely, political interest, propensity to participate, commitment of rights and responsibilities of citizenship, and sense of political efficacy. Results obtained from the hierarchical ANCOVA and ANCOVA based on individual scores were comparable. Both statistical approaches were shown to be more powerful than ANCOVA based on class means. Additional statistically significant differences among classes assigned to either the treatment or the comparison condition were uncovered by the hierarchical ANCOVA, but not by ANCOVA based on individual scores. On the basis of statistical power, interpretability, and generalizability, it was concluded that the hierarchical ANCOVA was superior to ANCOVA based on individual scores or class means. The latter two approaches suffered from conceptual and methodological limitations.

In accounting for effects associated with *Project Citizen*, the hierarchical ANCOVA approach incorporated the hierarchical (or nested) nature of Liou's (2002) quasi-experimental design into the analysis of covariance model. Consequently, data analysis was congruent with the way the study was actually carried out. It retained the maximum number of degrees of freedom for testing pertinent population parameters. It employed the pretest score as a covariate in order to control for pre-existing differences in students that were unrelated to *Project Citizen.* The hierarchical ANCOVA was shown in this article to be well suited to educational research in which data are collected from intact settings (such as classrooms) in quasi-experimental designs that also include one or more baseline measures.

To ensure credibility and to minimize, if not eliminate, potential bias in the findings reported in quasi-experimental research, it is necessary that educational researchers keep the following recommendations in mind.

First and the foremost, efforts should be exerted to randomly assign subjects to treatments. By so doing, educational researchers exclude the confounding issue of unit of analysis from their research and therefore, reduce bias and distortion in estimating population parameters or testing pertinent hypotheses. Researchers are advised to achieve random assignment whenever possible.

Second, data collected in intact groups deserve a rigorous examination. In educational research, it is possible to randomly assign subjects to treatment conditions and to establish circumstances in which the outcome measures are isolated from systematic carryover effects or threats to the independence assumption. Yet, it is often impossible or even undesirable to administer treatments individually in isolation. To account for the hierarchical nature of research designs and to maintain the interpretation of results at the individual level, an appropriate statistical model such as hierarchical ANCOVA should be employed.

Lastly, it should be noted that, even though the hierarchical ANCOVA has been proven to be a conceptually and methodologically sound procedure, this approach should be regarded as a viable approach that exercises only statistical control of biases. Moreover, the hierarchical ANCOVA is computationally more complex than an ordinary ANCOVA; it requires a set of restrictive statistical assumptions (Kirk, 1995). These assumptions must be met before valid inferences can be drawn from data analysis.

References

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 6* (3), 267-285.

Blair, R. C., & Higgins, J. J. (1986). Comment on "Statistical Power with Group Mean as the Unit of Analysis". *Journal of Educational Statistics, 11*(2), 161-169.

Buser, K. (1995, April). *Dangers in using ANCOVA to evaluate special education program effects*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 384 654).

Center for Civic Education. (2000). *We the People…Project Citizen: A professional development manual*. Calabasas, CA: Center for Civic Education.

Fraenkel, J. R., & Wallen, N. E. (2000). *How to design and evaluate research in education* (4th ed.). New York: McGraw Hill.

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

Henson, R. K. (1998, November). *ANCOVA with intact groups: Don't do it*! Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 426 086).

Hines, J. L., & Foil, C. R. (2000, January). *Covariance corrections: What they are and what they are not.* Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX. (ERIC Document Reproduction Service No. ED 445 076)

Hopkins, K. D. (1982). The units of analysis: Group means versus individual observations. *American Educational Research Journal, 19* (1), 5-18.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Lindman, H. R. (1992). *Analysis of variance in experimental design*. New York: Springer-Verlag.

Lindquist, E. F. (1940). *Statistical analysis in educational research*. New York: Houghton Mifflin.

Liou, S.-M. (2002). *The effect of We the People…Project Citizen on the civic skills and dispositions of Taiwanese adolescent students*. Unpublished doctoral dissertation, Indiana University-Bloomington.

Loftin, L., & Madison, S. (1991). The extreme dangers of covariance corrections. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological development* (Vol. 1, pp. 133-147.). Greenwich, CT: JAI Press Inc.

McMillan, J. H., & Schumacher, S. (2001). Research in education: A conceptual introduction (5th ed.). New York: Longman.

Morran, D. K., Robinson, F. F., & Hulse-Killacky, D. (1990). Group research and the unit of analysis problem: The use of ANOVA designs with nested factors. *The Journal for Specialists in Group Work, 15*(1), 10-14.

Peckham, P. D., Glass, G. V., & Hopkins, K. D. (1969). The experimental unit in statistical analysis. *The Journal of Special Education, 3* (4), 337-349.

Peng, C.-Y. J. (2004). Nested Design. In *The SAGE Encyclopedia of Social Science Research Methods, 2*, 717-719. Thousand Oaks, CA: Sage Publications.

Roberts, J. K. (2000). *Nested ANOVA vs. crossed ANCOVA: When and how to use which*. (ERIC Document Reproduction Service No. ED 440 112).

SAS Institute Inc. (1999). *SAS/STAT® user's guide, version 8, volume 2*. Cary, NC: SAS Institute Inc.

SPSS Inc. (1999). *SPSS base 10.0 user's guide*. Chicago: SPSS Inc.

Appendix A SAS® Programming Codes

```
*--------------------------------------------------Test of Civic Skills--------------------------------------------------------;
PROC   REG;
       MODEL q2_ski=q1_ski;
       PLOT q2_ski*q1_ski;
       BY q1_group;
TITLE  'TEST OF LINEARITY ASSUMPTION: Civic Skills';
RUN;

PROC   GLM;
       CLASS q1_group class;
       MODEL q2_ski=q1_ski q1_group class(q1_group)/SOLUTION;
       TEST H=q1_group E=class(q1_group);
       Means q1_group;
       LSMEANS q1_group/E=class(q1_group) ADJUST=BON E STDERR PDIFF;
TITLE  'Hierarchical ANCOVA for Civic Skills';
RUN;

PROC   GLM;
       CLASS q1_group;
       MODEL q2_ski=q1_ski q1_group q1_ski*q1_group;
TITLE  'TEST OF EQUAL SLOPE ASSUMPTION: Civic Skills';
RUN;

*--------------------------------------------------Test of Political Interest----------------------------------------------------;
PROC   REG;
       MODEL q2_int=q1_int;
       PLOT q2_int*q1_int;
       BY q1_group;
TITLE  'TEST OF LINEARITY ASSUMPTION: Political Interest';
RUN;

PROC   GLM;
       CLASS q1_group class;
       MODEL q2_int=q1_int q1_group class(q1_group)/SOLUTION;
       TEST H=q1_group E=class(q1_group);
       Means q1_group;
       LSMEANS q1_group/E=class(q1_group) ADJUST=BON E STDERR PDIFF;
TITLE  'Hierarchical ANCOVA for Political Interest';
RUN;
PROC   GLM;
       CLASS q1_group;
       MODEL q2_int=q1_int q1_group q1_int*q1_group;
TITLE  'TEST OF EQUAL SLOPE ASSUMPTION: Political Interest';
RUN;
```

```
*--------------------------------------------------Test of Propensity to Participate-----------------------------------------;
PROC  REG;
      MODEL q2_par=q1_par;
      PLOT q2_par*q1_par;
      BY q1_group;
TITLE 'TEST OF LINEARITY ASSUMPTION: Propensity to Participate';
RUN;

PROC  GLM;
      CLASS q1_group class;
      MODEL q2_par=q1_par q1_group class(q1_group)/SOLUTION;
      TEST H=q1_group E=class(q1_group);
      Means q1_group;
      LSMEANS q1_group/E=class(q1_group) ADJUST=BON E STDERR PDIFF;
TITLE 'Hierarchical ANCOVA for Propensity to Participate';
RUN;

PROC  GLM;
      CLASS q1_group;
      MODEL q2_par=q1_par q1_group q1_par*q1_group;
TITLE 'TEST OF EQUAL SLOPE ASSUMPTION: Propensity to Participate';
RUN;

*---------------------------------Test of Commitment to Rights and Responsibilities--------------------------------------;
PROC  REG;
      MODEL q2_right=q1_right;
      PLOT q2_right*q1_right;
      BY q1_group;
TITLE 'TEST OF LINEARITY ASSUMPTION: Commitment to Rights and Responsibilities' ;
RUN;

PROC  GLM;
      CLASS q1_group class;
      MODEL q2_right=q1_right q1_group class(q1_group)/SOLUTION;
      TEST H=q1_group E=class(q1_group);
      Means q1_group;
      LSMEANS q1_group/E=class(q1_group) ADJUST=BON E STDERR PDIFF;
TITLE 'Hierarchical ANCOVA for Commitment to Rights and Responsibilities';
RUN;

PROC  GLM;
      CLASS q1_group;
      MODEL q2_right=q1_right q1_group q1_right*q1_group;
TITLE 'TEST OF EQUAL SLOPE ASSUMPTION: Commitment to Rights and Responsibilities';
RUN;
```

```
*--------------------------------------------------Test of Political Efficacy--------------------------------------------------
-;
PROC   REG;
       MODEL q2_effic=q1_effic;
       PLOT q2_effic*q1_effic;
       BY q1_group;
TITLE  'TEST OF LINEARITY ASSUMPTION: Political Efficacy';
RUN;

PROC   GLM;
       CLASS q1_group class;
       MODEL q2_effic=q1_effic q1_group class(q1_group)/SOLUTION;
       TEST H=q1_group E=class(q1_group);
       Means q1_group;
       LSMEANS q1_group/E=class(q1_group) ADJUST=BON E STDERR PDIFF;
TITLE  'Hierarchical ANCOVA for Political Efficacy';
RUN;

PROC   GLM;
       CLASS q1_group;
       MODEL q2_effic=q1_effic q1_group q1_effic*q1_group;
TITLE  'TEST OF EQUAL SLOPE ASSUMPTION: Political Efficacy';
RUN;
```