

Using Multilevel Statistical Models in Social Work Intervention Research

James K. Nash
Lawrence L. Kupper
Mark W. Fraser

ABSTRACT. Statistical analyses of data from a classroom-based study illustrate the need to account for intra-class clustering in studies involving schools, classrooms, and other higher order units of analysis. Students were clustered in homerooms that were assigned to intervention and comparison conditions. Standard multiple linear regression analysis yielded a significant group effect but incorrectly ignored intra-cluster response correlations. A multilevel model appropriately accounting for the dependency among responses in the same cluster yielded a nonsignificant group effect. Implications for the analysis of intervention research data are discussed. *[Article copies available for a fee from The Haworth Document Delivery Service: 1-800-HAWORTH. E-mail address: <docdelivery@haworthpress.com> Website: <<http://www.HaworthPress.com>> © 2004 by The Haworth Press, Inc. All rights reserved.]*

James K. Nash, PhD, is Assistant Professor, Graduate School of Social Work, Portland State University. Lawrence L. Kupper, PhD, is Alumni Distinguished Professor, Department of Biostatistics, School of Public Health, University of North Carolina at Chapel Hill. Mark W. Fraser, PhD, is Tate Distinguished Professor for Children in Need, School of Social Work, University of North Carolina at Chapel Hill.

Address correspondence to: James K. Nash, Graduate School of Social Work, Portland State University, P.O. Box 751, Portland, OR 97207-0751 (E-mail: nashj@pdx.edu).

This project is supported by grants from the North Carolina State Division of Mental Health, Developmental Disabilities, and Substance Abuse Services, the Z. Smith Reynolds Foundation, and the UNC-CH Center for Injury Prevention. The authors thank the teachers, staff, and students who participated in this study.

Journal of Social Service Research, Vol. 30(3) 2004
<http://www.haworthpress.com/web/JSSR>
© 2004 by The Haworth Press, Inc. All rights reserved.
Digital Object Identifier: 10.1300/J079v30n03_03

KEYWORDS. Multilevel statistical models, clustered data, intervention research

Well-designed and carefully-executed research studies are key to building an evidence base for social work practice (Fraser, in press; Gambrill, 1999). Longitudinal cohort and other survey-based studies yield important knowledge of human behavior and the social environment, and this knowledge guides the design and development of social work interventions. As interventions are developed, outcome studies provide information about their efficacy and effectiveness. In both survey and intervention research, accurate analysis of data is essential for producing knowledge to guide practice. Of particular importance is the need for analysis methods to fit the structure of the data that result from the design and procedures of a study.

Survey and intervention research often employ sample selection methods that result in grouped or *clustered* data. Clustered data may also be present in an outcome study of a group-based social work intervention. For example, factors such as group composition or the training of the group leader are likely to make the intervention experience similar for participants within a particular group. This may cause them to receive similar scores, relative to participants in other groups, on an outcome measure used to evaluate the effects of the intervention. Thus, observations on important variables within a cluster are likely to be positively correlated (i.e., dependent). This is problematic, because independence of observations is a key assumption of many common statistical analysis methods. Biased or inaccurate results may occur if researchers analyze clustered data using methods that assume statistical independence of observations. Failure to account for dependence among clustered observations may result in underestimation of standard errors used to calculate test statistics, such as a *t* ratio. Depending on the degree of dependence that is present, this may lead to incorrect statistical inference, for example, rejecting a null hypothesis that is true (i.e., making a Type I error). In survey research, social work researchers are increasingly conducting analyses that account for clustered data. However, use of methods that account for clustered data is not as common in research on social work interventions.

The purpose of this paper is to describe and illustrate the importance of accounting for clustering when analyzing data from studies of social work interventions. Data from a pilot study of a classroom-based skills-

training program targeting children's behavior are used to illustrate key concepts.

CLUSTERED DATA IN SOCIAL WORK RESEARCH

Survey Research

Clustered data are common in survey research, where respondents are selected using complex multistage sampling methods (Kish, 1995). Multistage sampling is a strategy for selecting a probability sample from a population when direct selection of the study's unit of analysis (e.g., individuals) is not feasible, for example, because no list of all individuals exists. Instead, researchers begin by selecting sampling units that comprise multiple units of analysis. Thus, a researcher might select a probability sample of U. S. high school students by, first, selecting a sample of high schools, then selecting a sample of students within each school.

Social work researchers are, increasingly, using specialized methods to account for a study's sampling design in survey research. For example, Pottick and colleagues analyzed data from a nationally representative survey of patients (unit of analysis) in specialty mental health clinics (cluster), using methods that accounted for the multistage sampling design. The goal was to identify predictors of inpatient length of stay in adolescents and children (Pottick, Hansell, Miller, & Davis, 1999). In another study, Farmer (2000) used multilevel covariance analysis (MCA) of data from the National Educational Longitudinal Survey of 1988, which employed a two-stage sampling design. MCA was used to fit a structural equation model, using clustered data, of relationships among individual-level characteristics, school-level characteristics, and students' perceptions of school safety.

Recently, Rodgers-Farmer and Davis (2001) illustrated the importance of specialized analysis methods using data from the 1994 AIDS Knowledge and Attitudes Supplement to the National Health Interview Survey, which employed multistage cluster sampling. The authors fit several regression models to identify predictors of respondent's knowledge of AIDS. They contrasted results from a model that accounted for the sampling design with results from models that failed to account for the sampling design. Taken together, the analyses revealed that failure to account for the multistage sampling design resulted in a Type I error—models that did not account for clustered data incorrectly identi-

fied perceived risk of HIV infection as a significant predictor of AIDS knowledge.

Outcome Studies

Effects of the sample selection method. Multistage probability sampling is rarely used in studies on the effects of social work interventions. Instead, researchers identify a target population and then recruit participants into the study based on factors such as location, willingness to collaborate, and convenience. Similar to survey research, individuals often serve as the unit of analysis in social work outcome studies. The effects of interventions are typically assessed by observing, and contrasting across treatment and comparison groups, individual-level outcomes such as behavior, employment status, or symptomatology. However, such research often occurs in a setting where a collection of individuals—rather than the individuals themselves—serves as the unit of assignment to treatment and comparison groups.

For example, evaluating the effects of a welfare reform program in a particular state might involve selecting a sample of counties in which to implement the program. All eligible consumers within the selected counties would comprise an intervention group, and consumers in the remaining counties would make up a comparison group. Individual consumers would be the units of analysis in this study, but assignment to groups would occur at the county level. Researchers might use similar procedures when evaluating a skills-training program for students in schools. A sample of classrooms in a school would be selected to participate in the program and students in these classrooms would make up an intervention group. Students in the remaining classrooms would comprise a comparison group.

In both examples, data resulting from observations on individuals reflect the grouped (or clustered) nature of the research procedures. Counties serve as clusters in the first example, classrooms in the second. Unless assignment to clusters (e.g., placing students in classrooms) is random, research participants within a cluster are likely to share background characteristics and to respond similarly to an intervention, relative to participants in other clusters.

Thus, similar to survey research, clustered data in intervention research may result from the method used to select sample members. For example, counties served as the unit of assignment to pilot and comparison conditions in an examination of the effects of family-centered out-of-home care (Lewandowski & Pierce, 2002). Multiple outcomes

related to reunification were measured at the child (i.e., individual) level. As a result of the design, however, children were essentially clustered within counties.

Whenever a particular individual is assigned to a treatment condition by virtue of membership in a larger “sampling unit” (e.g., classroom, county), the effect is similar to that produced by multistage sampling in survey research. Correct statistical analyses of the resulting clustered data require specialized procedures. A fundamental statistical issue involves how to account appropriately for the correlation among outcomes (or responses) from research participants who are members of the same cluster. Ignoring within-cluster correlation can bias study conclusions. In particular, failure to account for within-cluster correlation can result in drawing incorrect conclusions about the statistical significance of parameter estimates, and thus about the effects of an intervention under evaluation.

Group-based interventions. Delivering an intervention via a group modality in a research study introduces the potential for clustered data, even if assignment to condition (i.e., experimental or control) is random. This is because members of a particular treatment group within each condition may become more similar on characteristics that are of interest in a research study. For example, students within a particular classroom in a school-based study may display similar scores on posttest outcome measures by virtue of common experience, such as exposure to an intervention as implemented by the same teacher.

Outcome studies of group-based interventions are common in social work. Moote, Smyth, and Wodarski (1999) reviewed and synthesized research on social skills interventions with children. Of 25 reviewed studies, 16 reported on interventions that utilized a group modality, and two reported on classroom-based interventions. Numerous outcome studies of interventions delivered to groups of participants were published in the past several years (see, e.g., Auslander, Haire-Joshu, Houston, Williams, & Krebill, 2000; McKay, Gonzales, Quintana, Kim, & Abdul-Adil, 1999; Mitchell, 1999; Pomeroy, Kiam, & Abel, 1999; Rice, 2001; Rotherman-Borus, Murphy, Fernandez, & Srinivasan, 1998). Children who are research participants are often clustered into schools or classrooms (Abbott et al., 1998; Spoth et al., 1998).

Other sources of clustering. Features of service delivery can introduce clustering even when an intervention itself is not group-based. For example, Lie and Moroney (1992) used an experimental design to examine the effects of intensive case management services for young women receiving Aid to Families with Dependent Children (AFDC).

Although the intervention was delivered on an individual basis, participants in the treatment group who received services from different case managers essentially formed separate clusters. That is, their common experience—sharing the same case manager—produced tacit clusters. Similarly, consumers in the control group assigned to different “regular” AFDC caseworkers represented distinct clusters. In another study, Gibbs and Sinclair (1999) examined treatment outcomes of 141 youths living in 48 children’s homes. Predictors of child-level outcomes included child characteristics, but also home-level characteristics, such as emphasis on building a child-family relationship. The latter would be expected to vary across homes, but would be similar for youths within a home.

Accounting for clustered data. Grouping research participants into clusters, whether for clinical, administrative, or research purposes, makes it quite likely that participants within a cluster will be more alike with respect to outcomes and/or to (possibly unmeasured) characteristics that affect outcomes, relative to participants in other clusters. The resulting intra-cluster response correlations are often inappropriately ignored in standard analyses that assume mutual independence among responses. As exceptions, Abbott and colleagues (1998), as well as Spoth and colleagues (1998), utilized analysis strategies that explicitly accounted for classroom and school clusters. Although ignoring clustering effects does not always mean that global conclusions are flawed, analyses that account for clustering generally yield more accurate and precise statistical inferences about treatment effects.

Methods for conducting analyses that take into account the clustered nature of data are increasingly available and accessible to social work researchers. Below we describe and present the results of two separate analyses of outcome data from a pilot test of a social work intervention. The first analysis represents a common approach to analyzing such data, but one that fails to account for the presence of clustered data. The second analysis takes into account the clustered nature of the data.

OVERVIEW OF STUDY PROCEDURES

All sixth grade students ($N = 181$) enrolled in a public middle school participated in the pilot study (see Nash, 1999 for details of the intervention, sample characteristics, and study procedures). Prior to the study and for administrative and educational reasons, each student was assigned to

one of 11 homerooms. Students in five homerooms ($n_1 = 76$) comprised the intervention group. In each intervention homeroom, students received 12 weeks of instruction in a prototype group-based problem-solving skills training program designed to promote positive behavior in the classroom, with peers, and in other settings (Fraser, Nash, Galinsky, & Darwin, 2000). Students in the remaining six homerooms ($n_0 = 105$) received the regular homeroom program and made up the comparison group. Incomplete data resulted in dropping 17 students from the analysis, yielding a final sample size of 164 students: $n_1 = 70$ in the intervention group and $n_0 = 94$ in the comparison group.

Teachers completed pretest and posttest behavior checklists on all participants in their homerooms. Based on these checklists, each student received pretest and posttest scores on subscales measuring multiple behavioral domains (Nash, 1999). A student's posttest score on the Cognitive Concentration subscale of the Social Health Profile, an adaptation of the Teacher Observation of Classroom Adaptation-Revised, serves as the outcome variable for the analyses reported here (Fast Track Project, 1997; Karns & Conduct Disorders Prevention Research Group, 1997; Werthamer-Larsson, Kellam, & Wheeler, 1991).

The Cognitive Concentration subscale has demonstrated reliability and validity as a measure of behaviors and attributes thought to promote academic success in the classroom (Fast Track Project, 1997). Subscale items include "stays on task," "works hard," and "easily distracted." Item scores range from 0 ("almost never") to 5 ("almost always").

For these analyses, all items were coded such that higher scores reflected more competent behavior. Item scores were summed to yield pretest and posttest Cognitive Concentration scores for each student. Thus, higher scores on Cognitive Concentration at posttest relative to pretest suggested improved behavior. In addition to behavior scores, demographic and background data on all students were gathered from the school's central database.

ANALYSIS 1: MULTIPLE LINEAR REGRESSION MODEL

Rationale

A common strategy for analyzing data from outcome studies involves fitting multiple linear regression models to the data. There are numerous applications of multiple linear regression (e.g., comparing

models with different numbers of predictors, assessing the relative effects of individual predictors in the presence of other predictors, and assessing the relative effects of groups of predictors). A range of strategies are available for building and evaluating multiple linear regression models (Kleinbaum, Kupper, Nizam, & Muller, 1998). In this article, the focus is on a particular application of multiple linear regression: to assess the effect of a predictor variable (intervention group versus comparison group) on a targeted outcome, adjusting for important covariates.

Analysis Strategy

The first analysis of data from the pilot study involves fitting a multiple linear regression model to predict student's posttest scores on Cognitive Concentration. Predictor variables in this model include pretest score, sex, race/ethnicity, and group status. This model can be written as

$$Y_i = \beta_0 + \beta_1(X_{i1}) + \beta_2(X_{i2}) + \beta_3(X_{i3}) + \beta_4(X_{i4}) + E_i, \text{ where}$$

- Y_i is the posttest Cognitive Concentration score for the i^{th} student ($i = 1, 2, \dots, N$);
- X_{i1} is the pretest Cognitive Concentration score for the i^{th} student;
- X_{i2} is a dummy variable representing the sex of the i^{th} student ($X_{i2} = 0$ if female, 1 if male);
- X_{i3} is a dummy variable representing the race/ethnicity of the i^{th} student ($X_{i3} = 0$ if European-American student, 1 if student of color);
- X_{i4} is the group status of the i^{th} student ($X_{i4} = 0$ if comparison group, 1 if intervention group); and
- E_i is the error term for the i^{th} student.

Note that the intercept term, β_0 , is the adjusted (for pretest score) mean posttest Cognitive Concentration score for a European-American female student in the comparison group (i.e., $X_{i2} = X_{i3} = X_{i4} = 0$).

Several assumptions are needed for valid analysis conclusions using this multiple linear regression model. First, the model itself is assumed to be correctly specified (e.g., no important covariates have been omitted). Second, the error terms, E_1, E_2, \dots, E_N , are assumed to be mutually independent, and E_i is assumed to be a normally distributed random variable with mean zero and variance σ^2 . The error term, E_i , represents

the difference between the i^{th} student's observed response Y_i (here, posttest Cognitive Concentration score) and the expected value dictated by the assumed regression model. Third, all predictor variables, X_{i1} to X_{i4} , are assumed to be measured without error. Under these assumptions, Y_i is modeled as the sum of a fixed component, $\beta_0 + \beta_1(X_{i1}) + \beta_2(X_{i2}) + \beta_3(X_{i3}) + \beta_4(X_{i4})$, and a random component E_i . The latter component of the model varies randomly from student to student.

Each of the model parameters β_1 to β_4 can be interpreted as the effect of a particular variable on a student's posttest Cognitive Concentration score, controlling for the effects of other variables in the model. The primary focus of the analysis is to make valid statistical inferences about the parameter β_4 , which measures the effect of the variable X_{i4} delineating group status. A statistically significant positive parameter estimate provides evidence that students in the intervention group have, on average, significantly higher posttest scores on Cognitive Concentration than do students in the comparison group, after controlling for the effects of pretest scores, sex, and race/ethnicity.

Results

Results based on fitting this particular multiple linear regression model appear in Table 1. Group status is a significant positive predictor of student's posttest Cognitive Concentration scores ($p < .05$), controlling for pretest scores, sex, and race/ethnicity. These results indicate that membership in the intervention group is associated with significantly higher—i.e., better—posttest scores on this measure of behaviors and attributes thought to promote classroom academic success. Student's pretest scores are also highly predictive of posttest scores in this model ($p < .01$), and this is not surprising given the brief duration of intervention. Sex is a marginally significant predictor of posttest scores ($p < .10$), with boys having marginally lower posttest scores, relative to girls, after controlling for other effects in the model. Race/ethnicity is a nonsignificant predictor in this model.

Given the quasi-experimental design of the study that produced the data, these results provide reasonably strong evidence that the intervention had a significant positive effect on student's Cognitive Concentration. This positive effect is evident after controlling for the other variables in the model (e.g., pretest scores) that may influence students' posttest scores. However, this model inappropriately ignores the proba-

TABLE 1. Results of Multiple Regression Analysis Predicting Posttest Cognitive Concentration Scores

Variable	<i>B</i>	<i>SE B</i>	β	<i>t</i>
Intercept	10.55	2.61		
Pretest Cognitive Concentration Score	0.77	0.05	0.78	16.38***
Sex ^a	-2.35	1.23	-.09	-1.91*
Race/Ethnicity ^b	-1.83	1.30	-.07	-1.41
Group Status ^c	2.61	1.19	0.10	2.18**
Model	<i>F</i> = 79.44; <i>df</i> = (4, 159); <i>p</i> < .01			<i>R</i> ² = .67

Note: ^aCoded 1 if male, 0 if female. ^bCoded 0 if European-American student, 1 if a student of color. ^cCoded 1 for intervention students, 0 for comparison students. **p* < .10, ***p* < .05, ****p* < .01.

ble positive correlation (i.e., the dependency) among the Cognitive Concentration scores of students in the same homeroom.

Implications of Clustered Data

Using standard multiple linear regression analysis methods to assess the separate effects of individual predictor variables (e.g., group status) on the outcome variable often involves testing a separate null hypothesis for each predictor (e.g., $H_0: \beta_4 = 0$ for the variable, group status). Standard multiple linear regression statistical software packages typically employ ordinary least squares (OLS) methods to produce model parameter estimates and associated standard errors, leading to a *t* statistic and a *p* value for each predictor variable in the fitted multiple linear regression model. Each *t* statistic is simply the ratio of the OLS parameter estimate divided by its estimated standard error. For example, in the model predicting posttest Cognitive Concentration scores reported earlier (see Table 1), the parameter estimate of group status was 2.61, with estimated standard error 1.19, yielding a *t* statistic of $2.61/1.19 = 2.18$ (*df* = 1, *p* < .05).

As noted earlier, an assumption of multiple linear regression is that the error terms, E_1, E_2, \dots, E_N are mutually independent. If subjects are clustered in units (such as homerooms), and if the intra-cluster correlation between any pair of responses is positive, this assumption is violated. To put it another way, error terms within a cluster tend to be more similar to each other than do error terms in other clusters. As described

in detail below, one can envision the error term for each individual as being partitioned into two components: an individual-level error component that varies randomly from subject to subject, and a common error component that is shared by all members of the same cluster. It is this shared error term that introduces a positive correlation between pairs of responses in the same cluster. Potential reasons for positive intra-cluster correlation in our example include the fact that assignment to intervention and comparison groups occurred at the homeroom level, with homerooms monitored by specific homeroom teachers, and that students in the intervention group participated in the intervention within particular homerooms. Moreover, students in both groups received pretest and posttest subscale scores based on assessments done by these specific homeroom teachers.

Standard multiple linear regression procedures ignore intra-cluster correlation. When error terms are not independent, the estimated standard errors of parameter estimates (the denominators of t statistics) resulting from OLS estimation will be incorrect, sometimes being too high and sometimes too low. In our example, the OLS estimated standard error of the estimate of β_4 would tend to be too low when there is a positive correlation between pairs of responses in the same cluster. Thus, the t statistic will be inflated, resulting in an increased chance of incorrectly rejecting the null hypothesis, $H_0: \beta_4 = 0$, when it is true (a Type I error).

Including dummy variables representing classrooms as covariates in a multiple linear regression model might seem to be an appropriate strategy for estimating the effects associated with classrooms. Not shown, we fit such a model to predict students' posttest Cognitive Concentration scores. Results were consistent with those shown in Table 1 (i.e., group status and pretest score were significant predictors of posttest scores, and sex was a marginally significant predictor of posttest scores) and revealed significant effects of a single intervention classroom and a single comparison classroom. However, this approach rests on the assumptions of multiple linear regression described earlier and fails to account for intra-cluster correlation. Thus, it carries increased risk of a Type I error to the extent that positive intra-cluster correlation of outcomes is present.¹

In contrast to other assumptions for using multiple linear regression, these procedures are not robust to violations of the assumption of independence among observations (Bryk & Raudenbush, 1992; Goldstein, 1995). To the extent that the intra-cluster correlation is positive, esti-

mated standard errors of the parameter estimates will tend to be too low. To be sure, this will not always lead to incorrect statistical inference. For example, a very small intra-cluster correlation will have little effect on estimated standard errors. However, the only way to know whether positive intra-cluster correlation will lead to incorrect inference is to compare results of multiple linear regression to the results of procedures that appropriately account for clustered data correlation effects.

ANALYSIS 2: MULTILEVEL MODEL

Overview

Valid estimates of standard errors of model parameter estimates can be obtained by using more general analysis methods. These appropriately account for intra-cluster correlation. One approach involves the use of multilevel modeling methods (Bryk & Raudenbush, 1992; Goldstein, 1995). The term *multilevel* reflects the idea that the data arise from a research setting involving one or more levels of clustering (e.g., students within a classroom, classrooms within a school, schools within a school district, etc.). Research participants and participant-level characteristics typically represent level 1 of a multilevel model.

For example, consider the model

$$Y_{ijk} = \beta_{0ij} + \beta_1(X_{ijk1}) + \beta_2(X_{ijk2}) + \beta_3(X_{ijk3}) + E_{ijk}, \text{ where}$$

- Y_{ijk} is the posttest Cognitive Concentration score for the k^{th} student in the j^{th} homeroom in the i^{th} group;
- X_{ijk1} is the pretest Cognitive Concentration score of the k^{th} student in the j^{th} homeroom in the i^{th} group;
- X_{ijk2} is a dummy variable representing the sex of the k^{th} student in the j^{th} homeroom in the i^{th} group ($X_{ijk2} = 0$ if female, 1 if male);
- X_{ijk3} is a dummy variable representing the race/ethnicity of the k^{th} student in the j^{th} homeroom in the i^{th} group ($X_{ijk3} = 0$ if European-American student, 1 if student of color);
- E_{ijk} is the random error term associated with the score of the k^{th} student in the j^{th} homeroom in the i^{th} group. It is assumed that E_{ijk} is a normally distributed random variable with mean 0 and variance σ_e^2 , and that the E_{ijk} s are mutually independent.

At level 2 of this multilevel model, one takes the intercept term, β_{0ij} , and treats it as a new random variable. One fits a second-level model

describing β_{0ij} for each i-j (intervention-homeroom) combination as $\beta_{0ij} = \beta_0 + \beta_4(I) + U_{ij}$, where

- I is an indicator variable that takes the value 1 for students in the intervention group and 0 for students in the comparison group, and
- U_{ij} represent an additional error term that can be interpreted as the *random* effect due to being in the j^{th} homeroom in the i^{th} group [note that U_{ij} is shared by all members in cluster (i-j)]. It is assumed that U_{ij} is normally distributed with mean 0 and variance σ_u^2 , and that the U_{ij} s are mutually independent and are independent of the E_{ijk} s.

Note that β_0 can be thought of as the adjusted (for pretest score) mean (overall comparison group homerooms) posttest Cognitive Concentration score for a European-American female student (i.e., $X_{ijk2} = X_{ijk3} = I = 0$). The corresponding adjusted mean posttest Cognitive Concentration score specific to the j^{th} comparison homeroom differs from this overall comparison group mean by the random effect U_{0j} . Similarly, the term $(\beta_0 + \beta_4)$ is the adjusted (for pretest score) mean (overall intervention group homerooms) posttest Cognitive Concentration score for a European-American female student (i.e., $X_{ijk2} = X_{ijk3} = 0$ and $I = 1$). The corresponding adjusted mean posttest Cognitive Concentration score for the j^{th} intervention homeroom differs from this overall intervention group mean by the random effect U_{1j} .

Combining these two models results in the following multilevel model: $Y_{ijk} = \beta_0 + \beta_1(X_{ijk1}) + \beta_2(X_{ijk2}) + \beta_3(X_{ijk3}) + \beta_4(I) + U_{ij} + E_{ijk}$. In this model, the coefficients associated with the variables pretest score, sex, and race/ethnicity are fixed student-level effects. However, in contrast to the multiple linear regression model considered earlier, this multilevel model contains two independent random effects, U_{ij} and E_{ijk} . As a result, student responses in the same intervention-homeroom combination are modeled to be positively correlated because each involves the same random effect U_{ij} . In particular, for $k \neq k'$, correlation $(Y_{ijk}, Y_{ijk'}) = [\text{Variance}(U_{ij})] / [\text{Variance}(U_{ij}) + \text{Variance}(E_{ijk})] = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2)$.

Advantages of a Multilevel Model

Compared with the multiple linear regression model presented earlier, the multilevel model better reflects the fact that students are clus-

tered within homerooms and that assignment to groups (intervention versus comparison) occurs at the homeroom level and not at the student level. In the multilevel model, each student's total error is the sum of an individual-level component E_{ijk} that varies randomly from student to student, and an error component U_{ij} that is shared with all other students within the same homeroom. The variable U_{ij} is the homeroom-level random error that reflects random variation from cluster to cluster (i.e., homeroom to homeroom).

Under the assumptions of multiple linear regression, the error terms are independent and there is no shared error component. OLS procedures would ignore such a shared error component when calculating the estimated standard errors of parameter estimates, leading to incorrect estimated standard errors of parameter estimates and hence incorrect t statistics. Incorporating homeroom-level error terms (U_{ij} s) via multilevel models can result in more reliable estimated standard errors of parameter estimates. It should be noted that the parameter estimates themselves often will not differ too greatly when using either standard multiple linear regression or multilevel modeling methods.

Analysis Strategy

To account for possible within-homeroom correlation of student responses, a multilevel model was fit to predict posttest Cognitive Concentration scores as a function of possibly important covariates. This analysis utilized the MIXED procedure in SAS (Littell, Milliken, Stroup, & Wolfinger, 1996). The MIXED procedure allows one to specify that students were clustered within homerooms and that assignment to groups (intervention versus comparison) occurred at the homeroom level.

Results

Table 2 displays results of fitting the multilevel model predicting student's posttest Cognitive Concentration scores, with fixed effects for pretest score, sex, race/ethnicity, and group status, and with homeroom treated as a random effect. As can be seen in Tables 1 and 2, parameter estimates are similar for the two modeling procedures. For example, the parameter estimate of β_4 , which measures the effect of the variable delineating group status, is 2.89 in the multilevel model, as compared to the value 2.61 for the multiple linear regression model. However, in contrast to the results shown

TABLE 2. Results of Multilevel Analysis Predicting Posttest Cognitive Concentration Scores

Fixed Effects	<i>B</i>	<i>SE B</i>	<i>t</i>
Intercept	8.58	2.89	
Pretest Cognitive Concentration Score	0.82	0.05	17.29**
Sex ^a	-2.40	1.16	-2.07*
Race/ethnicity ^b	-2.37	1.21	-1.95*
Group Status ^c	2.89	2.31	1.25 ^e
Random Effects ^d	Estimate	<i>SE</i>	<i>t</i>
Homeroom 0-1	-2.24	1.96	-1.15
Homeroom 0-2	3.85	1.93	2.00*
Homeroom 0-3	-1.82	1.99	-0.91
Homeroom 0-4	1.29	1.98	0.65
Homeroom 0-5	-0.18	1.97	-0.09
Homeroom 0-6	-0.90	1.98	-0.46
Homeroom 1-1	2.49	2.10	1.19
Homeroom 1-2	0.34	2.08	0.16
Homeroom 1-3	2.90	2.09	1.38
Homeroom 1-4	-6.14	2.12	-2.90**
Homeroom 1-5	0.40	2.08	0.19

Note. ^aCoded 1 if male, 0 if female. ^bCoded 0 if European American student, 1 if student of color. ^cCoded 1 if intervention, 0 if comparison. ^dHomeroom 0-1 refers to comparison group homeroom number 1, Homeroom 1-1 refers to intervention group homeroom number 1, etc. ^e $p = .24$. * $p < .05$, ** $p < .01$.

in Table 1, results of the multilevel model analysis provide no evidence of a statistically significant effect of group status ($p = .24$).

Inspection of Tables 1 and 2 reveals that the ratio of the parameter estimate for group status divided by its estimated standard error (the t statistic) is 2.18 for the multiple regression model; however, this ratio is 1.25 for the multilevel model. The lower value of the t statistic in the multilevel model resulted in a higher p value, namely $p = .24$, which exceeds the usual .05 cutoff for statistical significance. The lower value of the t statistic in the multilevel model results from a generally more accurate (i.e., higher) estimate of the standard error of the parameter estimate in question. The MIXED procedure utilizes restricted maximum likelihood estimation with a Newton Raphson algorithm (Littell et al., 1996). It accounts for the positive intra-homeroom response correlation resulting from clustered data, calculating more appropriate estimated standard errors of parameter estimates.

Parameter estimates of other fixed effects in the multilevel model were reasonably close in value to those in the multiple linear regression model. Pretest scores were highly significant predictors of posttest scores in both analyses. In contrast to the multiple linear regression model, the parameter estimates of sex and race/ethnicity were both statistically significant in the multilevel model ($p < .05$ in each case). This illustrates an important general phenomenon. Assuming that the multilevel model under consideration is valid, the effects of variables that vary within clusters (e.g., pretest scores, sex, and race/ethnicity) tend to be understated (i.e., estimated standard errors are often too high and so t statistic values are often too low) using estimation methods that inappropriately ignore positive intra-cluster correlation. In contrast, the effects of variables that vary across clusters but not within clusters (e.g., group membership status) tend to be overstated (i.e., estimated standard errors are often too low and so t statistic values are often too high) using estimation methods that incorrectly assume that responses within a cluster are statistically independent.

The multilevel modeling software also provides predicted values for the (random) effects of comparison group and intervention group homerooms (see Table 2). There was a significant negative effect for intervention homeroom 4 (estimate = -6.14 , $t = -2.90$, $p < .01$), and a significant positive effect for comparison homeroom 2 (estimate = 3.85 , $t = 2.00$, $p < .05$). As described earlier, the homeroom-level random effect for a particular homeroom in the intervention group can be interpreted as the difference between that particular homeroom's mean adjusted posttest Cognitive Concentration score and the mean (overall intervention group homerooms) adjusted posttest score. Each random effect estimate must be interpreted relative to an overall group mean, and the random effect estimates within a group are not independent of each other. Here, the mean adjusted posttest Cognitive Concentration score for intervention homeroom 4 is estimated to be 6.14 points lower than the mean adjusted posttest Cognitive Concentration score averaged over all intervention homerooms. Similarly, the mean adjusted posttest Cognitive Concentration score for comparison homeroom 2 is estimated to be 3.85 points higher than the mean adjusted posttest Cognitive Concentration score over all comparison homerooms.

DISCUSSION

Using data from a quasi-experimental pilot test of a skills-training intervention for children, we have illustrated that ignoring the presence of

positive correlation among cluster-specific responses can lead to meaningfully different conclusions about the statistical significance of an intervention effect. In particular, the use of standard multiple linear regression methods, which ignore such intra-cluster positive correlation and treat cluster-specific responses as being statistically independent, can lead to incorrect standard error estimates and hence to invalid t statistic values.

A key goal of intervention research, especially during the pilot-study phase, is to demonstrate that a prototype intervention, delivered as intended, produces expected outcomes in participants who receive the intervention. To accomplish this, researchers should complete an analysis whose assumptions fit the structure of the data arising from the research design. We have illustrated the use of a simple multilevel model as one approach to account for positive intra-cluster correlation. Compared with standard multiple linear regression procedures, multilevel modeling leads to a different conclusion regarding the statistical significance of the intervention effect in our example.

Accurate estimation of random effects permits intervention researchers to identify contextual factors that affect outcomes. Here, we found significant homeroom-level estimated random effects for two homerooms. These effects may have resulted from a range of homeroom-level characteristics, such as aggregated background characteristics of students within homerooms (e.g., the within-classroom average pretest score on Cognitive Concentration), or teacher characteristics (e.g., teaching style, year's experience). Multilevel modeling also permits estimation of homeroom-level fixed effects through the inclusion of additional covariates at the second level of the model (Bryk & Raudenbush, 1992; Goldstein, 1995).

We have considered a simple multilevel model which assumes that the response variable is approximately normally distributed and that any pair of responses within any cluster has exactly the same correlation. Although such a simple model is reasonably appropriate for our dataset, there are clearly many other realistic situations where such a simple model would be inadequate. For example, we might want to allow the intra-cluster correlation for all intervention group clusters to be different from that for comparison group clusters, or to allow the correlation structure to be even more complicated. Also, it is possible to estimate effects for more than two levels of clustering (e.g., children within classrooms, classrooms within schools, schools within districts, districts within states, etc.) and to measure responses that are not normal

and that are possibly even discrete (e.g., a dichotomous or ordinal response). It is beyond the scope of this paper to discuss these more complicated methods of analysis. Instead, we refer readers to the books by Bryk and Raudenbush (1992) and Goldstein (1995) and to the recent paper by Sashegyi, Brown, and Farrell (2000).

The design and development of interventions is central to social work research. Arguably, intervention research should be the basis of knowledge in the profession. The gold standard of intervention studies is the randomization of research participants—typically ranging from individuals to organizations—to experimental and control groups. Ethical and feasibility concerns often force social work researchers to use a quasi-experimental design, where assignment to treatment condition is non-random. In both situations, assignment to treatment condition may occur at the individual level or at an aggregate level. In the latter case, the units of assignment may be classrooms, caseloads of specially trained workers, units in a residential treatment center, wards of a hospital, or other aggregations. We have shown that traditional methods for analyzing data resulting from such studies fail to account for possible intra-cluster positive correlation on the response variable. Although the degree of correlation may be trivial, our findings demonstrate that it may be non-trivial and may produce biased findings. As a general guideline, we believe multilevel modeling is useful whenever the intra-cluster correlation of the response variable is expected to be positive, which is almost always the case.

To be useful, intervention research must reflect the nature of human services practice (e.g., using group-based interventions). To be feasible, intervention research often utilizes design features such as assigning participants to treatment and comparison conditions based on membership in a larger unit. Such features increase the likelihood that outcome data will be clustered. The effects of clustered data on the existing body of knowledge that informs social work practice—and practice in other applied fields—are largely unknown. In the sense that many treatment effects may have been overestimated, it is potentially large. Due in part to the relatively recent development of multilevel modeling procedures, accounting for clustered data in intervention research is the exception and not the rule. However, these advances in data analysis now permit the estimation of models that better conform to common practices in intervention research (Bryk & Raudenbush, 1992; Bryk, Raudenbush, & Congdon, 1996; Goldstein, 1995; Littell et al., 1996). Multilevel regression procedures, sometimes called hierarchical linear modeling, result

in models that more precisely account for the intra-cluster correlation of responses from research participants who are assigned to treatment and control (or comparison) conditions in intervention studies. Accounting for within-cluster correlation in such studies has been a major challenge in research. Multilevel modeling holds promise for increasing the precision of research and, ultimately, the effectiveness of services.

NOTE

1. Details and results of this model are available from the first author.

REFERENCES

- Abbott, R. D., O'Donnell, J., Hawkins, J. D., Hill, K. G., Kosterman, R., & Catalano, R. F. (1998). Changing teaching practices to promote achievement and bonding to school. *American Journal of Orthopsychiatry*, *68*, 542-552.
- Auslander, W., Haire-Joshu, D., Houston, C., Williams, J. H., & Krebill, H. (2000). The short-term impact of a health promotion program for low-income African American women. *Research on Social Work Practice*, *10*, 78-97.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage Publications.
- Bryk, A. S., Raudenbush, S. W., & Congdon, R. (1996). *Hierarchical linear and non-linear modeling with the HLM/2L and HLM/3L programs*. Chicago: Scientific Software International.
- Farmer, G. L. (2000). Use of multilevel covariance structure analysis to evaluate the multilevel nature of theoretical constructs. *Social Work Research*, *24*, 180-191.
- Fast Track Project. (1997) *Psychometric properties of the Social Health Profile (SHP)*. Durham, NC: Duke University, Department of Psychology.
- Fraser, M. W. (2003). Intervention research in social work: A basis for evidence-based practice and practice guidelines. In A. Rosen & E. K. Proctor (Eds.), *Developing practice guidelines for social work intervention: Issues, methods, and research agenda* (pp. 17-36). New York: Columbia University Press.
- Fraser, M. W., Nash, J. K., Galinsky, M. J., & Darwin, K. M. (2000). *Making choices: Social problem-solving skills for children*. Washington, DC: NASW Press.
- Gambrill, E. (1999). Evidence-based practice: An alternative to authority-based practice. *Families in Society: The Journal of Contemporary Human Services*, *80*, 341-350.
- Gibbs, I., & Sinclair, I. (1999). Treatment and treatment outcomes in children's homes. *Child and Family Social Work*, *4*, 1-8.
- Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). New York: Halsted Press.
- Karns, A., & Conduct Disorders Prevention Research Group (1997). *Social Health Profile and TOCA-R: Technical Report*. Unpublished Manuscript, Pennsylvania State University at University Park.

- Kish, L. (1995). *Survey sampling*. New York: John Wiley & Sons, Inc.
- Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied regression analysis and other multivariable methods* (3rd ed.). Pacific Grove, CA: Duxbury Press.
- Lewandowski, C. A., & Pierce, L. (2002). Assessing the effect of family-centered out-of-home care on reunification outcomes. *Research on Social Work Practice, 12*, 205-221.
- Lie, G., & Moroney, R. M. (1992). A controlled evaluation of comprehensive social services provided to teenage mothers receiving AFDC. *Research on Social Work Practice, 2*, 429-447.
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute, Inc.
- McKay, M. M., Gonzales, J., Quintana, E., Kim, L., & Abdul-Adil, J. (1999). Multiple family groups: An alternative for reducing disruptive behavioral difficulties of urban children. *Research on Social Work Practice, 9*, 593-607.
- Mitchell, C. G. (1999). Treating anxiety in a managed care setting: A controlled comparison of medication alone versus medication plus cognitive-behavioral group therapy. *Research on Social Work Practice, 9*, 188-200.
- Moote, G. T., Smyth, N. J., & Wodarski, J. S. (1999). Social skills training with youth in school settings: A review. *Research on Social Work Practice, 9*, 427-465.
- Nash, J. K. (1999). *Understanding and preventing youth violence: A pilot study of the Making Choices skills-training program*. Unpublished doctoral dissertation, University of North Carolina, Chapel Hill.
- Pomeroy, E. C., Kiam, R., & Abel, E. M. (1999). The effectiveness of a psychoeducational group for HIV-infected/affected incarcerated women. *Research on Social Work Practice, 9*, 171-187.
- Pottick, K. J., Hansell, S., Miller, J. E., & Davis, D. M. (1999). Factors associated with inpatient length of stay for children and adolescents with serious mental illness. *Social Work Research, 23*, 213-224.
- Rice, A. H. (2001). Evaluating brief structured group treatment of depression. *Research on Social Work Practice, 11*, 53-78.
- Rodgers-Farmer, A. Y., & Davis, D. (2001). Analyzing complex survey data. *Social Work Research, 25*, 185-192.
- Rotherman-Borus, M., Murphy, D. A., Fernandez, M. I., & Srinivasan, S. (1998). A brief HIV intervention for adolescents and young adults. *American Journal of Orthopsychiatry, 68*, 553-564.
- Sashegyi, A. I., Brown, K. S., & Farrell, P. J. (2000). Application of a generalized random effects regression model for cluster-correlated longitudinal data to a school-based smoking prevention trial. *American Journal of Epidemiology, 152*, 1192-1200.
- Spoth, R., Redmond, C., Shin, C., Lepper, H., Haggerty, K., & Wall, M. (1998). Risk moderation of parent and child outcomes in a preventive intervention: A test and replication. *American Journal of Orthopsychiatry, 68*, 565-579.

RECEIVED: 08/02

REVISED: 04/03

ACCEPTED: 04/03