Winter 12-2004

Assessing Intervention Effects in a School-Based Nutrition Intervention Trial: Which Analytic Model Is Most Powerful?

Jessica B. Janega
University of Memphis

David M. Murray
University of Memphis

Sherri P. Varnell
University of Minnesota

Jonathan L. Blitstein
University of Memphis

Amanda Birnbaum
Montclair State University, birnbauma@montclair.edu

See next page for additional authors

MSU Digital Commons Citation

Authors

Jessica B. Janega, David M. Murray, Sherri P. Varnell, Jonathan L. Blitstein, Amanda Birnbaum, and Leslie A. Lytle

Assessing Intervention Effects in a
School-Based Nutrition Intervention Trial:
Which Analytic Model Is Most Powerful?

Jessica B. Janega, PhD
David M. Murray, PhD
Sherri P. Varnell, MS, PhD
Jonathan L. Blitstein, MS
Amanda S. Birnbaum, PhD, MPH
Leslie A. Lytle, PhD

This article compares four mixed-model analyses valid for group-randomized trials (GRTs) involving a nested cohort design with a single pretest and posttest. This study makes estimates of intraclass correlations (ICCs) available to investigators planning GRTs addressing dietary outcomes. It also provides formulae demonstrating the potential benefits to the standard error of the intervention effect (σ_Δ) from adjustments for both fixed and time-varying covariates and correlations over time. These estimates will allow other researchers using these variables to plan their studies by estimating a priori detectable differences and sample size requirements for any of the four analytic options. These methods are demonstrated using data from the Teens Eating for Energy and Nutrition at School study. Mixed-model analyses of covariance proved to be the most powerful analysis in that data set. The formulae may be applied to any dependent variable in any GRT given corresponding information for those variables on the parameters that define the formulae.

Keywords: *group-randomized trial; analytic methods; power*

Group-randomized trials (GRTs) are comparative studies in which the units of assignment are identifiable groups and the units of observation are members of those groups.[1] GRTs are one of the best comparative designs available when investigators wish to explore the effects of interventions delivered at the group level. They are widely used in public health, education, and sociology.[2,1] An example would be a school-based health promotion study in which schools are randomly assigned to conditions (e.g., intervention and control) and students within those schools are followed over time to assess the effect of the intervention.

Jessica B. Janega, David M. Murray, and Jonathan L. Blitstein, Department of Psychology, University of Memphis. Sherri P. Varnell, Northrop-Grumman Mission Systems. Amanda S. Birnbaum and Leslie A. Lytle, Division of Epidemiology, School of Public Health, University of Minnesota.

Address reprint requests to David M. Murray, Department of Psychology, University of Memphis, 202 Psychology Building, Memphis, TN 38152-3230; phone: (901) 678-5714; fax: (901) 678-4995; e-mail: d.murray@mail.psyc.memphis.edu.

Such trials employ different units of assignment (e.g., school) and observation (e.g., student); this poses a number of design and analytic problems absent when individuals are randomized to conditions.[1] One problem is that observations taken from members of the same identifiable group are likely to have something in common, due to commonality in selection, exposure, or mutual interaction.[3] This commonality is indexed by an intraclass correlation (ICC) and reflects a component of variance attributable to the groups, σ_g^2, in addition to the usual variation attributable to the members within those groups, σ_m^2. Another problem is that the degrees of freedom *(df)* for the test of the intervention effect are defined by the number of groups per condition rather than the number of members per group, so that *df* are usually limited.

When the ICC is ignored in the analysis of the intervention effect in a GRT, simulation studies have shown that the Type I error rate can be badly inflated.[4-6] However, when the ICC is properly reflected in the analysis, power may be limited, due in part to the extra variation attributable to the groups and due in part to the limited *df* available to estimate that extra variation. Given that several valid analyses are available for GRTs, it is important to know how they compare in terms of power and whether it is possible to determine, in advance, the circumstances under which one method would be more powerful than another. We explore those issues using data from the Teens Eating for Energy and Nutrition at School (TEENS) study.[7]

We will first give a general description of four analytic models commonly used in GRTs that employ a pretest-posttest design. We will then present the methods and discuss the results from TEENS in light of these models. Finally, we will show how to use the results from TEENS to estimate detectable differences and sample size requirements; the most powerful model will yield the smallest detectable difference and will require the smallest sample size, other factors constant. Although the formulae in this article are applied to nutrition-oriented variables, they can be adapted for use in planning any GRT as long as appropriate estimates are available; fortunately, such estimates are increasingly available in the published literature.[8] We have previously reported estimates of ICC based on the pretest data from TEENS,[9] and this report builds directly on that work.

The Four Models

The TEENS study employed a nested cohort design with a single pretest and a single posttest. For this most common GRT design, several valid analytic models exist.[1] This article will focus on four: mixed-model analysis of variance (ANOVA), mixed-model analysis of covariance (ANCOVA), repeated-measures ANOVA, and repeated-measures ANCOVA. Given a pretest-posttest nested cohort design with two study conditions, all four of these models have 1 *df* for the test of the intervention effect, allowing us to compare the models in terms of power using the standard error of the intervention effect, σ_Δ. The value of σ_Δ is affected by several factors, including the ICC, variance reduction due to regression adjustment for covariates, and variance reduction due to over-time correlation within individuals and groups. As such, σ_Δ may increase or decrease according to the analytic model.

The simplest of these analyses is the mixed-model ANOVA, which uses posttest data only. In this analysis, the intervention effect is the unadjusted simple difference between the intervention and control condition means. The main advantage of this option is that it is the simplest of the four models and so requires the fewest assumptions. At the same time, it does not take advantage of any improvements in precision that may result from

regression adjustment for covariates or over-time correlation and so may not be very efficient.

The mixed-model ANCOVA is a second option. Here, we analyze the posttest data with regression adjustment for baseline covariates, including the baseline value of the dependent variable if appropriate. The intervention effect for the ANCOVA is the simple difference between the condition means at posttest after adjustment for any difference attributable to baseline covariates. Compared to the mixed-model ANOVA, the mixed-model ANCOVA carries additional assumptions related to regression adjustment, but they are easily tested. The main advantage of this model is that it often has a smaller σ_Δ and therefore greater efficiency. In addition, the ANCOVA estimates the regression coefficient relating the pretest and posttest based on the data, rather than based on any presumed structural relationship.

The mixed-model repeated-measures ANOVA is a third option. This analysis models time explicitly, and the intervention effect is the net difference between intervention and control condition pretest and posttest means. Compared to the mixed-model ANOVA, this model carries an additional assumption of constant correlation over time, but that assumption is safe with only two time intervals. The mixed-model ANOVA has the advantage of estimating the intervention effect as a net difference, which is often intuitively attractive to investigators. Another advantage is that it will have increased precision when the over-time correlation is large. One disadvantage is that σ_Δ is larger by a factor of $\sqrt{2}$, so that the over-time correlation must be greater than .5 to realize any advantage in precision, other factors constant. Another disadvantage is that the regression coefficient relating the pretest and posttest is presumed to be 1.0, and that may not be correct given the data.

The fourth option is a mixed-model repeated-measures ANCOVA. The intervention effect is the net difference among the adjusted condition pretest and posttest means. This analysis combines the advantages and disadvantages of the mixed-model repeated-measures ANOVA and the mixed-model ANCOVA.

Formulae for Detectable Difference and Sample Size

The ICC is the proportion of the total random variance that is due to the group. For a simple mixed-model ANOVA, the total random variance for the dependent variable *y* is

$$\sigma_y^2 = \sigma_m^2 + \sigma_g^2 \tag{1}$$

Therefore, the ICC from that analysis is calculated as

$$\mathrm{ICC_{m:g:c}} = \frac{\sigma_g^2}{\sigma_m^2 + \sigma_g^2} \tag{2}$$

where *m:g:c* reflects the nesting of members within groups and groups within conditions.[1]

Both the mixed-model ANOVA and ANCOVA use the same formula for σ_Δ. For the ANOVA and ANCOVA, a general formula for σ_Δ is provided by Murray[1] as

$$\sigma_\Delta = \sqrt{2\left\{\frac{\sigma_y^2(1-\mathrm{ICC}_{m:g:c})\theta_m + m(\sigma_y^2(\mathrm{ICC}_{m:g:c}))\theta_g}{mg}\right\}} \qquad (3)$$

As noted above, the total variance in the dependent variable y is σ_y^2. The $\mathrm{ICC}_{m:g:c}$ is calculated as shown in Equation 2. The member component of variance can be written as $\sigma_m^2 = \sigma_y^2(1 - \mathrm{ICC}_{m:g:c})$ and the group component as $\sigma_g^2 = \sigma_y^2(\mathrm{ICC}_{m:g:c})$. The θ_m and θ_g reflect any change in σ_m^2 or σ_g^2 due to regression adjustment for covariates, if any. Adjustments for covariates are defined as the ratio of the unadjusted and adjusted components of variance, either for members or for groups. If there is no regression adjustment for covariates, then θ_m and θ_g are equal to 1. For θ_m and θ_g to reduce the σ_Δ in the ANCOVA, they must have values less than 1. Because the simple ANOVA does not employ regression adjustment for covariates, θ_m and θ_g would automatically be set equal to 1 in using Equation 3.

For the mixed-model repeated-measures ANOVA and ANCOVA, a general formula for σ_Δ is provided by Murray[1] as

$$\sigma_\Delta = \sqrt{2\times 2\left\{\frac{\sigma_y^2(1-\mathrm{ICC}_{m:g:c})(1-r_{yy(m)})\theta_m + m(\sigma_y^2(\mathrm{ICC}_{m:g:c}))(1-r_{yy(g)})\theta_g}{mg}\right\}} \qquad (4)$$

Here, most of the terms are defined as in Equation 3, but they will have different values because they are estimated from the pretest and posttest data rather than just from the posttest data. New in Equation 4 are the $r_{yy(g)}$ and $r_{yy(m)}$, which are correlations over time for groups and members within groups, respectively; those correlations cannot be estimated from the posttest data alone. There is also an additional 2 in the numerator of Equation 4 because the intervention effect is defined as a net difference among four means (intervention and control conditions at pretest and posttest). This additional 2 is what makes it necessary for correlations over time to exceed .5 before the repeated-measures analyses will be more efficient than the nonrepeated-measures analyses.

Power in a GRT is reflected in the ability of the analysis to find a small detectable difference, defined as

$$\Delta = \sqrt{\sigma_\Delta^2(t_{\mathrm{critical}:\alpha/2} + t_{\mathrm{critical}:\beta})^2} \qquad (5)$$

Here, the critical t values are chosen based on acceptable Type I and Type II error rates and the available df, and the variance of the intervention effect is the squared value of σ_Δ taken either from Equation 3 or 4. Expanding Equation 5 using Equation 3 gives

$$\Delta = \sqrt{\frac{2(\sigma_y^2(1-\mathrm{ICC}_{m:g:c})\theta_m + m(\sigma_y^2(\mathrm{ICC}_{m:g:c}))\theta_g)(t_{\mathrm{critical}:\alpha/2} + t_{\mathrm{critical}:\beta})^2}{mg}} \qquad (6)$$

Expanding Equation 5 using Equation 4 gives

$$\Delta = \sqrt{2\times 2\frac{(\sigma_y^2(1-\mathrm{ICC}_{m:g:c})(1-r_{yy(m)})\theta_m + m(\sigma_y^2(\mathrm{ICC}_{m:g:c}))(1-r_{yy(g)})\theta_g)(t_{\mathrm{critical}:\alpha/2} + t_{\mathrm{critical}:\beta})^2}{mg}} \qquad (7)$$

Df are calculated as $2(g - 1)$ for all four models, as long as there are just two time intervals (pretest and posttest) and two conditions (intervention and control).

Power in a GRT is also reflected in having a small required sample size of groups per condition. To calculate the number of groups required to achieve the desired power given the mixed-model ANOVA or ANCOVA, we can rewrite Equation 6 to solve for *g:*

$$g = \frac{2(\sigma_y^2 (1 - \text{ICC}_{m:g:c})\theta_m + m(\sigma_y^2(\text{ICC}_{m:g:c}))\theta_g)(t_{\text{critical}:\alpha/2} + t_{\text{critical}:\beta})^2}{m\Delta^2} \qquad (8)$$

Here, θ_m and θ_g are set equal to 1 in the ANOVA.

For the repeated-measures ANOVA and ANCOVA, we can rewrite Equation 7 to solve for *g:*

$$g = 2 \times 2 \frac{(\sigma_y^2(1 - \text{ICC}_{m:g:c})(1 - r_{yy(m)})\theta_m + m(\sigma_y^2(\text{ICC}_{m:g:c}))(1 - r_{yy(g)})\theta_g)(t_{\text{critical}:\alpha/2} + t_{\text{critical}:\beta})^2}{m\Delta^2} \qquad (9)$$

Here too, θ_m and θ_g are set equal to 1 in the repeated-measures ANOVA.

With the appropriate estimates, we can identify the analysis that provides the smallest σ_Δ. That analysis will give us the smallest detectable difference and require the fewest groups per condition, other factors constant. As such, it will be the most powerful analysis among the four under consideration.

MATERIALS AND METHODS

The TEENS Study Design and Survey Procedures

TEENS is a GRT involving 16 Minneapolis–St. Paul, Minnesota, middle schools.[7] TEENS implemented school-, classroom-, and family-level interventions to reduce cancer-related dietary risk behaviors by increasing fruit and vegetable servings and reducing total fat consumption in seventh and eighth graders. TEENS focused on a lower income population and included only districts in which at least 20% of the students were eligible for free or reduced-price lunches. Participating schools were also required to have both seventh and eighth graders in the same building and to enroll at least 30 students per grade. In 14 districts, 33 schools were eligible, and of those, 20 schools in 9 districts agreed to participate. Schools that refused to participate did so due to time constraints, personnel changes, and lack of interest in the school food-environment component of the intervention. Of the 20 schools that agreed to participate, 1 was used in a pilot study and 3 others were excluded due to scheduling conflicts. The remaining 16 schools were randomly assigned from matched pairs to control and intervention conditions; schools were pair-matched based on the proportion of seventh graders expected to receive the school-based components of the intervention and on the proportion of seventh graders receiving free or reduced-price meals. Evaluation of the TEENS project included classroom surveys at baseline and follow-up and 24-hour dietary recall interviews from a random subsample of those students at baseline and follow-up.

Baseline data were gathered in the fall semester of 1998 (fall seventh grade). Trained staff members administered baseline surveys to students in a required seventh-grade class. Staff members noted absences and made one follow-up visit per school to reach as

many absent students as possible. Of the 4,050 eligible seventh graders, 95 (2.3%) were absent from school during two survey attempts, 77 (1.9%) were excluded due to parental or student refusal, and 3,878 (95.9%) completed the in-class survey.

Trained interviewers conducted 24-hour recalls using laptop computers and the Minnesota Nutrition Data System (NDS Version 2.6/8a/23, Nutrition Coordinating Center, University of Minnesota). Eight hundred twenty students were randomly selected and invited to participate. Of these students, 174 (21.3%) were missed due to absence or scheduling problems, 6 (0.7%) were excluded due to parental or student refusal, and 640 (78.0%) completed the recall. An average of 30 and 50 students were measured in each of the control and intervention schools, respectively, in 15- to 30-minute interviews. Students participated in dietary recalls Monday through Thursday, with absentees seen later in the week whenever possible. Trained staff members conducted dietary recalls individually in a quiet location. Dietary recalls consisted of students reporting the type and amount of all foods and beverages consumed in the previous day. Experimenters provided food models and other portion-size prompts for students to increase accuracy in reporting.

The research team collected follow-up survey data and 24-hour recall data in the spring semester of 2000 (spring eighth grade) using the same procedures as those used for seventh graders at baseline. Of the 3,878 seventh graders who provided baseline survey data, 3,010 (77.6%) provided follow-up survey data as eighth graders. Thus, the loss to follow-up rate for the survey was 22.4%. An analysis of the baseline data showed that students who were lost to follow-up were more likely to report minority status, single-parent households, and eligibility for the free- or reduced-price lunch program; they were less likely to report two parents working full time and parents with higher levels of educational attainment.

At follow-up, 113 (17.7%) of the 640 students who completed baseline 24-hour recalls were no longer enrolled. Of the remaining 527, follow-up recalls were obtained for 509 (96.6%). Thus, the loss to follow-up rate was 20.5% for the recalls. Students who provided a follow-up 24-hour dietary recall were no different than those students who did not provide follow-up recall data on any of the 24-hour dietary recall variables measured at baseline. Even so, those who were lost to follow-up for the 24-hour recall were more likely to report Native American or mixed ethnicity and single-parent households and less likely to report two parents working full time and parents with higher levels of education. Those who were lost to follow-up also had higher vegetable scores and lower scores for the "choices" measure from the survey.

Variables of Interest and Their Measures

The 24-hour dietary recall data were coded to compute total energy intake and total cholesterol consumed as well as percentage of energy from total fat and saturated fat. The 24-hour recall also provided counts of fruit servings and vegetable servings as well as a total fruit and vegetable count. The NDS program initially included french fries in the vegetable counts; we computed modified vegetable and total fruit and vegetable count variables by excluding french fries from the vegetable category.

Other variables were taken from the school survey.[10] Demographic variables included gender (male or female), race (White, Black, Hispanic, Asian, Native American, mixed, or other), household structure (two parents at home or another arrangement), parents' education attainment (student reports that both parents have at least some high school education, one parent has vocational school or some college training, one parent has com-

pleted college, both parents have at least completed college, data are missing or student does not know parental education status, or other), parents' employment status (both parents work full time, one parent works full time, or other), and eligibility for the free or reduced-price lunch program (yes or no).

Fruit and vegetable intake was measured using a modified and validated version of the Behavioral Risk Factor Student Survey.[11] This version measured the frequency of consuming fruit juices, fruit, green salad, potatoes, carrots, and vegetables during the past year. The six items were weighted and summed to estimate the average number of daily servings of fruits, vegetables, and fruits and vegetables combined.[9] Baseline measurements of typical fruit and vegetable intake had good internal consistency (Cronbach's alpha = .75).[10] Separate subscores for fruit and vegetable intake were also estimated.

A 9-item scale assessed typical food choices. Students were asked to indicate which item in nine different pairs of foods they would choose to eat *most of the time*. Examples include "cookies OR popcorn without butter" and "plain baked potato OR french fries." Responses were coded and summed so that higher scores reflect a greater tendency to choose lower fat foods. This measure was adapted from a previously published 17-item scale used with adolescents.[12] Both test-retest reliability and internal consistency at baseline were moderate (Spearman r = .65 and Cronbach's alpha = .67).[10]

A similar scale, adapted from a 17-item scale used with adolescents,[12] was used to assess nutrition knowledge. This scale presented students with 10 food-item pairs and asked them to choose which food they thought was "better for your health." Test-retest reliability for this scale was low (Spearman r = .43), whereas internal consistency at baseline was high (Cronbach's alpha = .94).[10] An additional single item asking "How many servings of fruits and vegetables should teens have each day?" offered three possible responses ranging from "1-2 servings" to "5 or more servings." Test-retest reliability for this item was low (Spearman r = .41).[10]

Two items assessed physical activity.[13] First, students were asked to indicate how often they were physically active for at least 20 minutes at a time, with a five-category Likert-type response scale ranging from *most of the time* to *never*. The second item asked the students how hard they breathed during physical activity, with a four-category Likert-type response scale ranging from *breathing much harder than usual* to *breathing the same as usual*. The physical activity measure combined these items on a 0 to 9 scale where higher values indicated greater physical activity (Spearman r = .66).[10]

Four items assessed common sedentary activities. Two items measured the number of hours per day of TV watching on weekdays and weekends using a five-category Likert-type scale ranging from *I don't watch TV during the weekdays (weekend)* to *more than 4 hours per day*. The other two items measured hours per weekday and weekend day spent playing video games, with similar response categories. The four items combined formed a scale ranging from 4 to 20 where higher values indicated more sedentary activity. Test-retest reliability for sedentary behavior was good (Spearman r = .86); internal consistency at baseline was also good (Cronbach's alpha = .73).[10]

Analysis Methods

To be able to compare the four analytic models, they had to be run on the same individuals. As a result, individuals who did not provide complete information at baseline and at follow-up for all variables of interest were excluded from the analysis. Separate exclusions were made for analyses involving the 24-hour recall variables because those variables were only measured in a subset of the students eligible for the school survey.

After those exclusions, we applied the four mixed-model analyses to each of the dependent variables using either PROC MIXED or the GLIMMIX macro using SAS version 8.2.[14] MIXED implements the general linear mixed-model and is appropriate for models that specify multiple sources of random variation, all distributed Gaussian.[15] The GLIMMIX macro implements the generalized linear mixed-model and is appropriate for models that specify multiple random effects, all distributed Gaussian, except for a non-Gaussian residual error distribution.[15] The GLIMMIX macro iteratively calls the MIXED procedure until it converges on a solution for the fixed- and random-effect parameter estimates. The GLIMMIX macro allows the specification of a variety of link and variance functions. To conduct Poisson regression for count variables, we specified a log link and a Poisson error function. PROC MIXED was used for analyses of normally distributed dependent variables: total energy, total cholesterol, percentage of energy from total fat, percentage of energy from saturated fat, food choices, nutrition knowledge, physical activity, and sedentary behavior. We specified a log link and Poisson error function in the GLIMMIX macro for count variables (fruit and vegetable servings).

The mixed-model ANOVA was performed on the posttest data only. From that analysis, we obtained estimates of the group and member components of variance for each dependent variable. As noted in Equation 1, the sum of these two variance components is the total variance for the dependent variable (σ_y^2). We estimated the $ICC_{m:g:c}$ for each dependent variable using Equation 2, with a preliminary step for the count variables to rescale the parameter estimates from a log scale to a linear scale, adapting procedures described in Murray[1] (pp. 239-240). We estimated σ_Δ using Equation 3.

In the mixed-model ANCOVA, we adjusted for the baseline value of the dependent variable, gender, ethnicity, age at baseline, whether the child had one or two parents in the home at baseline, whether the child was eligible to receive free or reduced-price lunches at baseline, parental employment status at baseline, and parental education attainment at baseline. From that analysis, we estimated for each variable σ_m^2, σ_g^2, and σ_y^2, adjusted for the covariates. We calculated θ_m and θ_g as the ratio of the adjusted and unadjusted σ_m^2 and σ_g^2, per Murray,[1] drawing on the results of the mixed-model ANOVA for the unadjusted estimates. We estimated σ_Δ using Equation 3.

The mixed-model repeated-measures ANOVA was performed on the pretest and posttest data. This analysis models time explicitly and provides estimates of σ_m^2, σ_g^2, and σ_y^2, as well as estimates of the over-time correlation at the group, $r_{yy(g)}$, and member, $r_{yy(m)}$, levels. The estimates of σ_m^2, σ_g^2, and σ_y^2 usually differ from those obtained in the mixed-model ANOVA, because they are based on the pretest and posttest data, rather than on the posttest data alone. We estimated σ_Δ using Equation 4.

In the mixed-model repeated measures ANCOVA, we adjusted for several time-varying covariates: age, whether the child reported one or two parents in the home, whether the child was eligible for free or reduced-price lunches, parental employment status, and parental education level. The mixed-model repeated measures ANOVA was also performed on the pretest and posttest data. We calculated θ_m and θ_g as the ratio of the adjusted and unadjusted σ_m^2 and σ_g^2, per Murray,[1] drawing on the mixed-model repeated-measures ANOVA for the unadjusted estimates. We estimated σ_Δ using Equation 4.

RESULTS

All 16 participating schools completed the study. Of the 640 students who provided baseline 24-hour recall, 21 were judged unreliable by the study nutritionist and 21 did not

have baseline survey data, leaving 598 reliable recalls with survey data. Of that number, 314 were included in the analyses reported here and 214 (35.8%) were excluded, either because they were missing at follow-up or missing a value for at least one variable of interest. Students who were excluded had significantly higher values on all dependent variables from the 24-hour recall than those who were included. Students excluded also were less likely to report living with two parents.

Of the 3,878 students who provided baseline survey data, 3,009 were included in the analyses reported here and 869 (22.4%) were excluded; of the exclusions, 868 did not complete the follow-up survey and 1 student was missing a value for at least one variable of interest. Students excluded were significantly different from those included on all dependent variables from the survey except sedentary behavior: those included were reported significantly higher nutrition knowledge and physical activity and significantly lower values for the other dependent variables. Students excluded also were less likely to report living with two parents, more likely to report eligibility for the free and reduced-price lunch program, less likely to report both parents employed full time, less likely to report higher levels of parents' educational attainment, and more likely to self-identify as a member of a minority ethnic group.

Table 1 presents results for the four mixed-models for the 24-hour recall dependent variables. To illustrate, consider the variable energy. The mean energy intake during the previous 24 hours in the sample was 2,217.5 kcal. The $ICC_{m:g:c}$ was estimated as 0.0364 in the mixed-model ANOVA, which is fairly large for a school-based GRT. Regression adjustment for covariates was helpful, as reflected in the fractional values for the θ_g and θ_m in the mixed-model ANCOVA. The $ICC_{m:g:c}$ was estimated as 0.0297 in the mixed-model repeated-measures ANOVA. As noted earlier, the estimates from the mixed-model ANOVA and from the mixed-model repeated-measures ANOVA often differ somewhat, because they are taken from overlapping but different data sets (repeated measures analyses include both pretest and posttest measures). Regression adjustment for covariates was not helpful in the mixed-model repeated-measures ANCOVA, as reflected in the values for the θ_g and θ_m. Energy intake was correlated over time, though much more so at the level of the school than at the level of the student. The mixed-model ANCOVA produced the lowest value for σ_Δ, suggesting that it would be the most powerful of the four analyses for this variable.

This pattern held quite generally in Table 1, where for eight of the nine dependent variables, the mixed-model ANCOVA was the most powerful analysis, producing the lowest σ_Δ. In Table 1, the σ_Δ from the mixed-model ANCOVA was 5.5% smaller on average than the next smallest σ_Δ. For the ninth variable (servings of vegetables), the mixed-model ANCOVA produced the second-lowest σ_Δ, but the difference in σ_Δ between the mixed-model ANCOVA and the mixed-model ANOVA was negligible (0.3%). Several of the models estimated negative variance components for several of the variables; we will return to this point in the discussion.

Table 2 provides the corresponding estimates for the school survey variables. For five of the seven dependent variables, the mixed-model ANCOVA was the most powerful analysis, producing the lowest σ_Δ. In Table 2, the σ_Δ from the mixed-model ANCOVA was 15.6% smaller on average than the next smallest σ_Δ. Of the remaining two dependent variables, the mixed-model repeated-measures ANCOVA was the most powerful model for physical activity, whereas the mixed-model repeated-measures ANOVA was the most powerful model for nutrition knowledge. For these variables, the σ_Δ from the mixed-model ANCOVA was 5.0% larger on average than that of the model with the smallest σ_Δ.

Table 1.   Intraclass Correlations (ICCs), Variance Components, Over-Time Correlations, Thetas, and the Standard Error of the Intervention Effect for Dependent Variables From the 24-Hour Recall Analyses

Variable	Mean Units	Model[a]	$ICC_{m;g;c}$	σ^2_g	σ^2_m	$r_{yy(g)}$	$r_{yy(m)}$	θ_g	θ_m	σ_Δ
Energy (kcal)	2,217.5	ANOVA	0.0364	0.0254	0.6710					0.1155
		ANCOVA						0.2615	0.8562	0.0874
		RM ANOVA	0.0297	0.0178	0.5792	0.8884		1.0552	0.9934	0.0958
		RM ANCOVA					0.3223			0.0960
Cholesterol (mg/dl)	227.0	ANOVA	0.0351	0.0012	0.0322					0.0250
		ANCOVA						0.7151	0.9313	0.0228
		RM ANOVA	0.0226	0.0006	0.0266	0.7850		1.0586	0.9982	0.0234
		RM ANCOVA					0.1337			0.0231
Percentage of energy as fat	30.3	ANOVA	−0.0117	−0.6755	58.4640					0.6634
		ANCOVA						1.6173	0.9378	0.5459
		RM ANOVA	0.0217	1.1956	53.9422	1.3501		0.8138	0.9871	0.8289
		RM ANCOVA					0.2024			0.8523
Percentage of energy as saturated fat	11.2	ANOVA	−0.0005	−0.0064	13.2555					0.3694
		ANCOVA						−4.7828	0.9224	0.3674
		RM ANOVA	0.0134	0.1727	12.7224	0.4204		1.3038	0.9886	0.5051
		RM ANCOVA					0.2264			0.5030
Servings of fruit and vegetables	3.60	ANOVA	−0.0013	−0.0086	6.6966					0.2600
		ANCOVA						1.1364	0.9662	0.2549
		RM ANOVA	0.0007	0.0060	8.4924	2.5184		−10.0000	1.0175	0.3089
		RM ANCOVA					0.4352			0.2882
Servings of fruit	1.45	ANOVA	−0.0065	−0.0192	2.9801					0.1620
		ANCOVA						0.9647	0.9754	0.1602
		RM ANOVA	0.0069	0.0189	2.7014	0.2278		0.1070	0.9989	0.2212
		RM ANCOVA					0.2604			0.2169

(continued)

765

Table 1. (continued)

Variable	Mean Units	Model[a]	$ICC_{m;g;c}$	σ^2_g	σ^2_m	$r_{.yy(g)}$	$r_{.yy(m)}$	θ_g	θ_m	σ_Δ
Servings of vegetables	2.15	ANOVA	0.0098	0.0357	3.6158					0.2158
		ANCOVA						1.0899	0.9858	0.2165
		RM ANOVA	0.0124	0.0441	3.4970	1.3365		0.7335	1.0182	0.2372
		RM ANCOVA					0.1257			0.2334
Servings of fruit and vegetables without french fries	2.55	ANOVA	−0.0097	−0.0455	4.7308					0.1947
		ANCOVA						1.3782	0.9615	0.1781
		RM ANOVA	0.0004	0.0024	5.4546	15.9311		−8.7383	1.0090	0.2183
		RM ANCOVA					0.4251			0.2151
Servings of vegetables without french fries	1.10	ANOVA	−0.0183	−0.0276	1.5333					0.0952
		ANCOVA						1.2304	1.0525	0.0912
		RM ANOVA	−0.0051	−0.0073	1.4355	−1.9917		1.1242	0.9977	0.1314
		RM ANCOVA					0.0573			0.1358

a. Mixed models coded as follows: ANOVA = analysis of variance; ANCOVA = analysis of covariance; RM ANOVA = repeated-measures ANOVA; RM ANCOVA = repeated-measures ANCOVA.

766

Table 2. Intraclass Correlations (ICCs), Variance Components, Over-Time Correlations, Thetas, and the Standard Error of the Intervention Effect for Dependent Variables From the School Survey Analyses

Variable	Mean Units	Model[a]	$\text{ICC}_{m;g;c}$	σ_g^2	σ_m^2	$r_{yy(g)}$	$r_{yy(m)}$	θ_g	θ_m	σ_Δ
Average daily servings of fruit and vegetables	4.29	ANOVA	0.0073	0.0986	13.4123					0.2075
		ANCOVA				0.8072		0.6479	0.8183	0.1761
		RM ANOVA	0.0058	0.1820	31.0619			0.8900	0.9826	0.1976
		RM ANCOVA					0.7476			0.1918
Average daily servings of fruit	2.45	ANOVA	0.0041	0.0177	4.2750					0.1015
		ANCOVA				0.8827		0.3353	0.8602	0.0808
		RM ANOVA	0.0063	0.0426	6.7029			1.0301	0.9840	0.0983
		RM ANCOVA					0.6103			0.0951
Average daily servings of vegetables	1.84	ANOVA	0.0068	0.0294	4.2963					0.1151
		ANCOVA				0.7455		0.7249	0.8250	0.1010
		RM ANOVA	0.0087	0.0482	5.5055			0.7643	0.9893	0.1170
		RM ANCOVA					0.4996			0.1149
Healthy choices	6.11	ANOVA	0.0171	0.0790	4.5443					0.1606
		ANCOVA				0.7819		0.4394	0.6912	0.1134
		RM ANOVA	0.0211	0.0955	4.4397			0.6654	0.9938	0.1269
		RM ANCOVA					0.5182			0.1243
Nutrition knowledge	17.0	ANOVA	0.0325	0.4403	13.0940					0.3570
		ANCOVA				1.0052		0.0616	0.8713	0.1481
		RM ANOVA	0.0403	0.5068	12.0604			0.5671	0.9804	0.1436
		RM ANCOVA					0.3163			0.1437

(continued)

Table 2. (continued)

Variable	Mean Units	Model[a]	$ICC_{m;g;c}$	σ_g^2	σ_m^2	$r_{.yy(g)}$	$r_{.yy(m)}$	θ_g	θ_m	σ_Δ
Physical activity	5.63	ANOVA	0.0380	0.3013	7.6202					0.2928
		ANCOVA						0.1786	0.9026	0.1512
		RM ANOVA	0.0390	0.3157	7.7879	0.9685				0.1419
		RM ANCOVA					0.2891	0.6284	0.9796	0.1415
Sedentary behavior	11.0	ANOVA	0.0181	0.2080	11.2746					0.2604
		ANCOVA						0.1581	0.6551	0.1363
		RM ANOVA	0.0257	0.2824	10.6832	0.8174				0.1992
		RM ANCOVA					0.5357	0.6433	0.9888	0.1973

a. Mixed models coded as follows: ANOVA = analysis of variance; ANCOVA = analysis of covariance; RM ANOVA = repeated-measures ANOVA; RM ANCOVA = repeated-measures ANCOVA.

None of the survey variables were estimated to have negative variance components under any of the four models.

DISCUSSION

Comparison of the Four Models

The mixed-model ANCOVA resulted in the lowest σ_Δ for 13 of the 16 variables examined for this article. The advantage of the mixed-model ANCOVA over the simple unadjusted model and the repeated-measures models was due to the use of fixed covariates, particularly the baseline measure of the dependent variable, and the simple comparison between two means used in calculating σ_Δ. As seen in Equations 3 and 4, the repeated-measures models have an additional 2 in the numerator of σ_Δ because that effect involves four means; unless the over-time correlations are strong enough (greater than .5 on average), the additional 2 will result in a larger σ_Δ than in the mixed-model ANOVA or ANCOVA performed on posttest data alone. Three dependent variables did not fit this pattern: servings of vegetables (recall data), nutrition knowledge, and physical activity (survey data). In the case of servings of vegetables, the mixed-model ANOVA was the most powerful model, suggesting that neither adjustment for covariates nor over-time correlations significantly reduced σ_Δ. The mixed-model repeated-measures ANOVA yielded the smallest σ_Δ for nutrition knowledge, which benefited from an explicit adjustment for time, but not from adjustments due to covariates. Finally, the mixed-model repeated measures ANCOVA yielded the smallest σ_Δ for physical activity, which benefited from both explicit adjustment for time and adjustment for time-varying covariates.

Negative Variance Estimates

When variance component estimates are negative, correlations over time can fall outside the range of –1 to 1. Although negative variance components are a nuisance in this way, it is necessary to allow these components to be estimated as negative to preserve the nominal Type I error rate in GRTs.[1] At the same time, for purposes of sample size or power calculations, it would be prudent to fix negative estimates at zero to avoid undersizing the study.

Previous TEENS ICC Estimates

We previously published estimates of $ICC_{m:g:c}$ developed from the pretest data from the TEENS study.[9] In that article, we reported estimates for $ICC_{m:g:c}$ from a mixed-model ANOVA (labeled there as Model 1), which was quite similar to the mixed-model ANOVA applied in the analyses reported here. A comparison of the unadjusted $ICC_{m:g:c}$ from the previous and current articles will identify a number of variables for which the estimates of $ICC_{m:g:c}$ are quite different in the two reports, even when the variables were measured using the same procedures; examples include energy, measured via the 24-hour recall; and average daily servings of fruits and vegetables, measured via school survey. That comparison will identify other variables for which the estimates of $ICC_{m:g:c}$ are quite similar in the two reports; an example is physical activity, measured via school survey. Moreover, there is no clear pattern as to which estimates are smaller. It is very natural then to

ask both why some of the estimates are so different between the two articles, whereas others are so similar, and which set of estimates is preferred.

The answer to the first question is quite simple: The estimates are often different because they were based on different data and slightly different models, so that it is quite plausible that they would not agree. The ICCs reported in the earlier paper were based on pretest data alone (fall seventh grade), and they were based on the entire sample providing useable data at pretest; the ICCs reported here for the analogous mixed-model ANOVA were based on posttest data alone (spring eighth grade), and they were based on the sample providing useable data at both pretest and at posttest. In addition, study condition was included in the analysis reported here, where we expected some condition effect, whereas it was omitted from the analysis reported earlier because there was no expectation for any condition effect in the pretest data. That the differences in the ICCs were observed for some variables and not for others simply means that they were not systematically related to these methodological differences.

The answer to the second question is also simple: Use the estimates reported in this article, because they are based on the circumstances most likely to be obtained at the end of another trial. Most investigators will focus on their follow-up data, will include study condition in their analysis, and will limit their sample to those with useable data at both pretest and posttest.

The fact that for so many variables, the estimates based on pretest data alone were so different from the estimates based on follow-up data raises questions about the utility of estimates based on pretest data alone. Unfortunately, there are very few articles that have published $ICC_{m:g:c}$ estimates based on follow-up data,[16-18] whereas there are many articles based on pretest data alone. We have long assumed that estimates based on pretest data were conservative and so encouraged investigators to use those estimates absent better information. The fact that several of the variables had much lower $ICC_{m:g:c}$ estimates in our earlier article compared to the analyses reported here suggests that assumption is not necessarily safe and underscores the need to plan future studies using estimates that reflect the design and analysis proposed for the study. They also underscore the need to publish estimates based on follow-up data, as we are doing here.

Detectable Difference and Sample Size Calculations

To demonstrate how to use these results, we will consider an example from the school survey data, though these formulae are applicable to both the school survey data and the 24-hour recall data. Let us consider a study designed to increase the average number of daily servings of fruits and vegetables among junior high school students. Assume a pretest-posttest control group design following a cohort of students in schools randomized to conditions as in the design employed in TEENS. The study could be planned based on a simple mixed-model ANOVA, mixed-model ANCOVA, mixed-model repeated-measures ANOVA or mixed-model repeated-measures ANCOVA. Using the results from the mixed-model ANCOVA analysis of the school survey variable "average daily servings of fruit and vegetables" available in Table 2, we can compute a detectable difference using Equation 6. Degrees of freedom for the analysis are equal to $c(g - 1)$ where c represents the number of conditions and g represents the number of groups per condition; if we plan on 10 schools per condition, df are 18 and the critical t values are 2.101 and 0.862 for a two-tailed Type I error rate of 5% and 80% power. Given 100 members per group, the detectable difference is computed using Equation 6 as

$$\Delta = \sqrt{\frac{2(\sigma_y^2(1-\text{ICC}_{m:g:c})\theta_m + m(\sigma_y^2(\text{ICC}_{m:g:c}))\theta_g)(t_{\text{critical}:\alpha/2} + t_{\text{critical}:\beta})^2}{mg}}$$

$$= \sqrt{\frac{2(13.5109(1-0.0073)\,0.8183 + 100(13.5109(0.0073))0.6479)(2.101+0.862)^2}{100(10)}}$$

$$= 0.5522$$

Thus, based on this example, there would be adequate power to detect a change in average daily servings of fruits and vegetables of 0.5522, or just over half of a serving per day. The same analysis using the results from the mixed-model ANOVA gives a detectable difference of 0.6393, which is a 15.8% increase in the size of the detectable difference from the mixed-model ANCOVA. Likewise, the mixed-model repeated-measures ANOVA gives a detectable difference of 0.6309, a 14.3% increase, whereas the mixed-model repeated-measures ANCOVA gives a detectable difference of 0.6162, which is 10.4% higher than the mixed-model ANCOVA.

If we were interested in detecting an even smaller change in daily servings of fruits and vegetables, for example, an increase of half a serving per day ($\Delta = .5$), we could use the sample size calculation formula in Equation 8 to determine the number of groups necessary for an ANCOVA. To do so, we would select the appropriate ICC and calculate σ_y^2 from Table 1. Assume that there are 100 members per group:

$$g = \frac{2(\sigma_y^2(1-\text{ICC}_{m:g:c})\theta_m + m(\sigma_y^2(\text{ICC}_{m:g:c}))\theta_g)(t_{\text{critical}:\alpha/2} + t_{\text{critical}:\beta})^2}{m\Delta^2}$$

$$= \frac{2(13.5109(1-0.0073)\,0.6479 + 100(13.5109)\,0.0073))0.8183)(2.101+0.862)^2}{100(0.5)^2}$$

$$= 12.196$$

This calculation is not final, as the t values were based on $g = 10$ and $c(g-1) = 18$ df. The calculated result is 12.196, which must be rounded up to 13. Df based on 13 schools per condition would be $2(13-1) = 24$, and the proper t values would be 2.06 and 0.857. Recalculating with those t values gives

$$g = \frac{2(\sigma_y^2(1-\text{ICC}_{m:g:c})\theta_m + m(\sigma_y^2(\text{ICC}_{m:g:c}))\theta_g)(t_{\text{critical}:\alpha/2} + t_{\text{critical}:\beta})^2}{m\Delta^2}$$

$$= \frac{2(13.5109(1-0.0073)\,0.6479 + 100(13.5109(0.0073)\,0.8183))(2.064+0.8569)^2}{100(0.5)^2}$$

$$= 11.851$$

Rounding up, we have a result of $g = 12$. Revising the t values based on $2(12-1) = 22$ df gives a calculated g of 11.943, which is consistent with the value used to select the t values, so we can stop; two or three iterations are often necessary in this kind of calculation, for exactly this reason. The result of $g = 12$ means that to detect an intervention effect of half a serving in a mixed-model ANCOVA using these estimates, we would randomize a minimum of 12 groups to each condition to have 80% power with a Type I error rate of 5%. By comparison, both the mixed-model ANOVA and the mixed-model repeated-

measures ANOVA would require 16 groups per condition for the same half-serving effect. The mixed-model repeated-measures ANCOVA would require 15 groups per condition. This pattern again illustrates the increased power associated with the mixed-model ANCOVA for this endpoint.

Recommendations

In the case of the dietary variables in the pretest-posttest TEENS data set, it is clear that the mixed-model ANCOVA was the most powerful analytic method for most of the variables examined. However, we must also keep in mind that these results are based on a single data set and so this pattern may not hold for other data sets. Other models may be preferred for other variables or under other circumstances. When proper estimates are available, informed decisions regarding the proper analysis are possible.

This study makes ICCs and other important parameter estimates available to investigators planning GRTs with dietary measures as the outcomes of interest. It also provides formulae demonstrating the potential benefits to σ_Δ of both adjustments for fixed and time-varying covariates, as well as correlations over time. These formulae can be used with any endpoint; our example is limited to dietary outcomes, though good estimates from other data sets can be used to compute a priori power analyses for other outcomes, and those estimates are increasingly available.[8] The estimates presented here will allow other researchers using these variables to plan their studies by performing a priori power analyses for any of the different analysis options, and weighing the potential benefits, to choose the most appropriate analysis for their budget and available means.

Implications for Research

When researchers have access to estimates of over-time correlations and potential adjustments for both time-varying and fixed covariates, two main guidelines can be useful in planning future studies. First, correlations over time must be about .5 or larger to be beneficial, as they must compensate for the additional 2 in the numerator of the variance of the intervention effect. When estimates of over-time correlation do not exceed .5, there is generally no benefit to power to selecting a mixed-model repeated-measures analysis. Second, in terms of adjustments for covariates, θ as defined below should be less than 1 for a significant gain in power:

$$\theta_m = \frac{\text{adjusted } \sigma_m^2}{\text{unadjusted } \sigma_m^2} \quad \text{and} \quad \theta_g = \frac{\text{adjusted } \sigma_g^2}{\text{unadjusted } \sigma_g^2} \tag{10}$$

Researchers would do well to design studies with endpoints that are accurately measured. Researchers should also design studies with covariates that are appropriately related to the endpoints of the study. Measuring fixed covariates at baseline (including the baseline measurement of the dependent variable) allows the use of a mixed-model ANCOVA. The better the covariates, the more σ_Δ is potentially reduced, and the better the power in the study. Time-varying covariates should also be measured at pretest and posttest so that the option of using a mixed-model repeated-measures ANCOVA is open.

Publishing estimates of ICC, θ_m and θ_g, and $r_{yy(m)}$ and $r_{yy(g)}$ will allow other researchers to use these estimates in choosing the most appropriate analysis model for their data set. Performing a priori power and sample size calculations is vitally important for planning

research. Without such calculations, even the best designs may be undersized and yield disappointing results. By choosing the best analysis model via detectable difference or sample size calculations, researchers can determine the best way to plan a study.

Implications for Practice

Practitioners should be alert as they read and evaluate research based on the allocation of identifiable social groups to study conditions. Health education and behavior change studies are now commonly conducted in hospitals, clinics, schools, worksites, churches, and other identifiable social groups. Studies that allocate such groups to study conditions face extra design and analytic challenges, as described at the beginning of this article. Absent clear evidence that the authors were cognizant of those issues and planned for them both in their design and their analysis, practitioners should be cautious about taking any reported findings at face value. The wrong analyses can lead to p values that overstate the significance of the findings, and readers who skim too lightly over the methods run the risk of disseminating intervention programs that may not be as effective as they appear. Authors who include an open discussion of the issues presented here are more likely to have planned and executed their study properly and more likely to be presenting results that can be taken at face value.

References

1. Murray DM: *Design and Analysis of Group-Randomized Trials*. New York: Oxford University Press, 1998.
2. Donner A, Klar N: *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold, 2000.
3. Kish L: *Survey Sampling*. New York: John Wiley, 1965.
4. Murray DM, Hannan PJ, Baker WL: A Monte Carlo study of alternative responses to intraclass correlation in community trials: Is it ever possible to avoid Cornfield's penalties? *Evaluation Review* 20(3):313-337, 1996.
5. Murray DM, Wolfinger RD: Analysis issues in the evaluation of community trials: Progress toward solutions in SAS/STAT MIXED. *Journal of Community Psychology*, CSAP Special Issue:140-154, 1994.
6. Zucker DM: An analysis of variance pitfall: The fixed effects analysis in a nested design. *Educational and Psychological Measurement* 50:731-738, 1990.
7. Lytle LA, Perry CL: Applying research and theory in program planning: An example from a nutrition education intervention. *Health Promotion Practice* 2(1):68-80, 2001.
8. Murray DM, Blitstein JL: Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review* 27(1): 79-103, 2003.
9. Murray DM, Phillips GA, Birnbaum AS, Lytle LA: Intraclass correlation for measures from a school-based nutrition intervention study: Estimates, correlates and applications. *Health Educ Behav* 28(6):666-679, 2001.
10. Birnbaum A, Lytle LA, Murray DM, Story M, Perry CL, Boutelle KN: Survey development to assess correlates of young adolescents' eating. *American Journal of Health Behavior* 26(4): 284-295, 2002.
11. Field AE, Colditz GA, Fox MK, Byers T, Serdula M, Bosch RJ, Peterson KE: Comparison of 4 questionnaires for assessment of fruit and vegetable intake. *Am J Public Health* 88(8):1216-1218, 1998.
12. Edmundson E, Parcel GS, Feldman HA, Elder J, Perry CL, Johnson CC, Williston BJ, Stone EJ, Yang M, Lytle LA, Webber L: The effects of the child and adolescent trial for cardiovascular

health upon psychosocial determinants of diet and physical activity behavior. *Prev Med* 25(76):442-454, 1996.

13. Perry CL, Kelder SH, Murray DM, Klepp K-I: Community-wide smoking prevention: Long-term outcomes of the Minnesota heart health program and the class of 1989 study. *Am J Public Health* 82(9):1210-1216, 1992.

14. SAS Institute, Inc: *SAS/STAT User's Guide* (version 8, Vols. 1-3). Cary, NC: SAS Institute, Inc., 1999.

15. Littell RC, Milliken GA, Stroup WW, Wolfinger RD: *SAS System for MIXED Models*. Cary, NC: SAS Institute, Inc., 1996.

16. Feng Z, Diehr P, Yasui Y, Evans B, Beresford S, Koepsell TD: Explaining community-level variance in group randomized trials. *Stat Med* 18:539-556, 1999.

17. Murray DM, Alfano CM, Zbikowski SM, Padgett LS, Robinson LA, Klesges R: Intraclass correlation among measures related to cigarette use by adolescents: Estimates from an urban and largely African American cohort. *Addict Behav* 27(4):509-527, 2002.

18. Murray DM, Clark MH, Wagenaar AC: Intraclass correlations from a community-based alcohol prevention study: The effect of repeat observations on the same communities. *J Stud Alcohol* 61(6):881-890, 2000.