Published in final edited form as:

Med Sci Sports Exerc. 2004 May ; 36(5): 876–882.

School-Level Intraclass Correlation for Physical Activity in Adolescent Girls

DAVID M. MURRAY¹, DIANE J. CATELLIER², PETER J. HANNAN³, MARGARITA S. TREUTH⁵, JUNE STEVENS³, KATHRYN H. SCHMITZ⁴, JANET C. RICE⁶, and TERRY L. CONWAY⁷

1Department of Psychology, College of Arts and Sciences, University of Memphis, Memphis, TN

2Department of Biostatistics and School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC

3Departments of Nutrition and Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC

4Division of Epidemiology, School of Public Health, University of Minnesota, Minneapolis, MN

5Center for Human Nutrition, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

6Department of Biostatistics, School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA;

7Graduate School of Public Health, San Diego State University, San Diego, CA

Abstract

Purpose—The Trial for Activity in Adolescent Girls (TAAG) is a multi-center group-randomized trial to reduce the usual decline in moderate to vigorous physical activity (MVPA) among middle-school girls. In group-randomized trials, the group-level intraclass correlation (ICC) has a strong inverse relationship to power and a good estimate of ICC is needed to determine sample size. As a result, we conducted a substudy to estimate the school-level ICC for intensity-weighted minutes of MVPA measured using an accelerometer.

Methods—To estimate the ICC, each of six sites recruited two schools and randomly selected 45 eighth grade girls from each school; 80.7% participated. Each girl wore an Actigraph accelerometer for 7 d. Readings above 1500 counts per half minute were counted as MVPA. These counts were converted into metabolic equivalents (MET) and summed over 6 a.m. to midnight to provide MET-minutes per 18-h day of MVPA. Minutes of MVPA per 18-h day also were calculated ignoring the MET value.

Results—The unadjusted school-level ICC for minutes of MVPA was 0.0205 (95% CI: -0.0079, 0.1727) and for MET-minutes of MVPA was 0.0045 (95% CI: -0.0147, 0.1145). Adjustment for age and BMI had no measurable effect, whereas adjustment for ethnicity reduced both ICC; adjusted values were 0.0175 (95% CI: -0.0092, 0.1622) for minutes of MVPA and 0.0000 (95% CI: -0.0166, 0.0968) for MET-minutes of MVPA. This information was used to calculate the number of schools and girls needed for TAAG to have 90% power to detect a 50% reduction in the decline of MET-minutes of MVPA between sixth and eighth grade.

Conclusions—The results called for 36 schools in TAAG, with 120 girls invited for measurements at each school, and a minimum participation rate of 80%.

Keywords

GROUP-RANDOMIZED TRIAL; ACCELEROMETER; INTERVENTION EFFECT; POWER ANALYSIS

Group-randomized trials (GRT) are characterized by the random assignment of identifiable social groups to study conditions, with observations on their members to evaluate an intervention (3,8). Studies with different units of assignment and observation exist in many disciplines (10) and face two challenges not found in randomized clinical trials (2).

The first challenge is that the variation in the condition-level statistic used to estimate the intervention effect must be assessed against the variation in the same statistic estimated at the group level to protect the Type I error rate (2). Unfortunately, the variation of the group-level statistic is usually inflated in a GRT due to between-group variation that exists in addition to the usual within-group variation. The magnitude of this extra variation is indexed by an intraclass correlation coefficient (ICC), which is the average correlation for the dependent variable among observations taken on members in the same group (7). As the magnitude of the ICC increases, the power of the trial decreases, if other factors are held constant (3,8).

The second challenge is that there are often only a limited number of groups assigned to each condition. This limits the degrees of freedom (df) available for the test of the intervention effect, as those df are based on the number of groups, not on the number of members (2). When the df are limited, power is limited, if other factors are held constant. This can also reduce power in a GRT, even when the ICC is quite small.

In spite of these challenges, GRT remain the best comparative design available to evaluate an intervention that modifies the physical or social environment, is implemented at a group level, or cannot be delivered to individuals. As such, the best approach is to plan a trial large enough to allow for the extra variation and limited df (3,8).

Determining sample size for a GRT requires a good estimate of the ICC (3,8). The validity of the estimate will depend on how well the conditions under which it is estimated match the conditions anticipated for the study being planned; the best estimate is one that is derived from the target population and is obtained using protocols that mirror those planned for the study. Unfortunately, it is often difficult to locate a good ICC estimate.

This was the situation facing the investigators in the Trial for Activity in Adolescent Girls (TAAG) in the fall of 2000 as they began a 6-yr study funded by the National Heart Lung and Blood Institute (NHLBI) to test the effectiveness of a school- and community-based intervention to reduce the decline in physical activity observed in middle-school girls. This paper presents the TAAG experience as a case study of the issues facing investigators in GRT where good estimates of ICC are not yet available. We describe the methods employed to address those issues during the phase I planning period and show how we used the information obtained to estimate power for the main trial planned in phase II.

TAAG is a multi-center GRT involving six sites (at the Universities of Arizona, Maryland, Minnesota, and South Carolina; San Diego State University; and Tulane University); the Coordinating Center is at the University of North Carolina, Chapel Hill. Schools will be randomly allocated to either an intervention or a control condition. Baseline measures will be collected from a cross-sectional sample of sixth grade girls in each school and the intervention will be implemented soon thereafter. Follow-up measures will be collected 2 yr later from a cross-sectional sample of eighth grade girls in the same schools. The primary endpoint will be the mean difference in intensity-weighted minutes (i.e., MET-minutes) of MVPA between

intervention and control schools in the eighth grade, controlling for sixth grade levels. Activity will be measured using an Actigraph accelerometer (Actigraph, Manufacturing Technologies, Inc., <http://www.mtiactigraph.com/>).

No published reports are available on the magnitude of school-level ICC for physical activity measured using an accelerometer in middle-school girls. Unpublished reports from two separate samples from the Amherst Health and Activity Study (17) suggest ICC estimates of 0.220 and 0.060. However, these estimates were based on just a few schools, so they are not very precise; in addition, neither focused solely on eighth grade girls, as TAAG will do in its posttest survey. Other studies have shown that school-level ICC for most health behaviors tend to be less than 0.05 (9), suggesting that the unpublished estimates from Amherst may not reflect the level of ICC to be expected in TAAG. In addition, neither of the ICC estimates reflected any adjustment for covariates, which often helps reduce the magnitude of the ICC (9).

Consequently, the TAAG investigators conducted a substudy to estimate the school-level ICC for intensity-weighted minutes of MVPA measured using an accelerometer in middle-school girls. Secondary goals were to determine: 1) whether the ICC was related to the number of weeks over which accelerometer measurements would be collected, 2) whether the ICC might be reduced by adjustment for girl- or school-level covariates, and 3) the minimum number of measurement days necessary to estimate the habitual level of physical activity reliably. This paper presents the methods and primary results of that substudy, describes the primary analysis planned for TAAG, shows how the results from the substudy were used in the power calculations for that analysis, and discusses the implications for other studies.

METHODS

Design and selection of participants

The substudy had to involve enough schools and girls per school to ensure that the true ICC fell within a relatively narrow range. With measurements from “m” students in each of “g” schools, a Fisher’s z-transformation defined as $z = (1/2)\ln[1 + (m - 1)ICC]/(1 - ICC)$ is approximately normally distributed with variance $m/[2(m - 1)(g - 2)]$ (4). With $g = 12$ schools and $m = 30$ girls per school, if the true ICC is 0.01 (or 0.03, or 0.05), the 95% confidence interval has an upper bound no greater than 0.068 (or 0.11, or 0.15). Allowing for 15% missing data caused by nonadherence or dropout, the number of girls required for the substudy was 36 per school.

Each of the six sites recruited two middle schools selected to be as similar as possible to the schools to be recruited for TAAG and obtained complete rosters of the eighth grade girls enrolled in those schools. The Coordinating Center selected 45 girls at random from each school for invitation to the substudy, with the expectation that 80% would participate. Inclusion criteria were: 1) informed consent signed by a parent or guardian and, 2) informed assent from the girl. Exclusion criteria were: 1) medical condition for which physical activity is contraindicated and 2) inability to complete a brief questionnaire.

At each school, data were collected during three consecutive 1-wk periods, with 12 girls selected at random for measurement in each period. This selection scheme and measurement schedule permitted examination of variation in physical activity from week-to-week within a school (e.g., due to factors such as weather, scheduling of field trips, etc.) Evidence of such variation would suggest that the ICC might be reduced by extending the period of observation to avoid measuring all girls from a given school during the same week.

Human subjects approval

The substudy was approved by the Institutional Review Board of each participating site and collaborating institution. The substudy was also approved by the appropriate school and school district officials.

Variables of interest and their measures

The primary outcome of the substudy was the primary outcome for TAAG, intensity-weighted minutes (i.e., MET-minutes) of MVPA, measured using the equipment and protocol planned for TAAG. Other measures included demographic and anthropometric variables (ethnicity, age, weight, height, and body mass index (BMI)). All measurement staff completed and passed a 6-h measurement training and certification program.

The substudy required seven measurement days and two school visits by staff for each girl. During the first school visit, staff confirmed eligibility and consent, obtained baseline information including date of birth, height, weight, and ethnicity, and gave instructions. Using a standardized protocol, staff measured each girl's weight twice to the nearest 0.1 kg on an electronic scale (Seca, Model 770, Hamburg, Germany). Staff measured height twice to the nearest 0.1 cm using a portable stadiometer (Shorr Height Measuring Board, Olney, MD). Each girl was asked to wear the accelerometer for seven complete days, except at night while sleeping or during any activity that might get the accelerometer wet (e.g., bathing, swimming). Girls wore the accelerometer on their right hip, attached to a belt provided by the study and standardized across sites; girls returned the accelerometer 7 d later at the second school visit. Fridays and Mondays were avoided as starting days for a weekly cycle, because absenteeism is usually higher on those days. Girls were called at home at least once to encourage adherence to the protocol. The second school visit was used to retrieve the accelerometer.

Accelerometer readings were processed using methods similar to those reported by Puyau et al. (11). Readings above 1500 counts per half minute were treated as MVPA, whereas readings below that threshold were ignored; an earlier substudy had shown that threshold to have the optimal sensitivity and specificity for discriminating brisk walking from less vigorous activities in eighth grade girls (15). Half-minute counts were used instead of full-minute counts based on the expectation that they would be more sensitive to fluctuations in activity levels. Occasional missing accelerometry data within a girl's 6-d record were replaced via imputation based on the expectation maximization (EM) algorithm; details on the imputation methods are provided elsewhere (1). Counts above 1500 per half minute were converted into METs using a regression equation developed in the earlier substudy (13); METs were calculated based on each girl's own resting metabolic rate. The METs were summed over the 6 a.m. to midnight day to provide MET-minutes per day of MVPA, where 1 MET·min represents one metabolic equivalent of energy expended for 1 min. Minutes of MVPA also were calculated separately ignoring the MET value.

Data processing and analysis

The goal of the analysis was to estimate the variability due to sites, schools (within sites), calendar weeks (within school and site), and girls (within week, school and site) to allow calculation of the ICC. To do so, a general linear mixed model (6) was applied to predict the MVPA score, with separate analyses for minutes of MVPA and MET-minutes of MVPA. The general linear mixed model is appropriate for correlated data with normally distributed errors. The model was fit using SAS PROC MIXED, Version 8.1 (12). Site, school, week, and girl were modeled as random effects. Because previous studies have shown that ICC are often reduced with regression adjustment for member- and group-level covariates (9), we evaluated age, ethnicity and BMI as fixed effects. Confidence intervals for ICC were calculated based on the *F* distribution with 6 *df* for schools (6 sites $\times$ (2 schools/site - 1)) and 400 *df* for girls (6

sites $\times$ 2 schools/site $\times$ 3 wk/school $\times$ (12.11 girls/school - 1)) (14). The Spearman-Brown prophecy formula (5) was used to estimate the effect on reliability of increasing the number of days of monitoring. In particular, the reliability of the estimated average daily MVPA based on “k” days of monitoring was estimated as

$$R = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2 / k} \quad [1]$$

where σ_b^2 and σ_w^2 are the between- and within-subject variance components. Defined in this way, reliability expresses the precision of measuring average daily MVPA in terms of how faithfully the observed average reflects the unknown true average.

RESULTS

Across the six sites, 436 girls participated in the substudy, representing 80.7% of those invited. The number of girls per site ranged from 69 to 75; 40.8% were white, 32.6% were African-American, 9.4% were Hispanic, 9.6% were multiethnic, and 6.9% were Asian. Three of these girls were missing too much accelerometry data to be eligible for the imputation procedures and were dropped from the analyses of the activity data.

Table 1 provides descriptive statistics for age, BMI, weight, minutes of MVPA, and MET-minutes of MVPA. Table 1 also presents descriptive statistics for MVPA and MET-minutes of MVPA by week; there was no clear pattern by week for either variable. The reliability of MET-minutes of MVPA was low for a single day ($R = 0.42$); the predicted reliability with 4, 5, 6, and 7 d of measurements increased to 0.75, 0.79, 0.82, and 0.84, respectively.

Relative to the residual error, the other components of variance were quite small. The school-level ICC was calculated as the proportion of the total random variation (sum of girl, week, and school) accounted for by the school,

$$\frac{\sigma_{\text{school}}^2}{\sigma_{\text{school}}^2 + \sigma_{\text{week}}^2 + \sigma_{\text{girl}}^2} \quad [2]$$

The unadjusted school-level ICC for minutes of MVPA was 0.0205 (95%CI: -0.0079, 0.1727) and for MET-minutes of MVPA was 0.0045 (95% CI: -0.0147, 0.1145). Adjustment for age and BMI had no measurable effect, whereas adjustment for ethnicity (five categories) reduced both ICC; adjusted values were 0.0175 (95% CI: -0.0092, 0.1622) for minutes of MVPA and 0.0000 (95% CI: -0.0166, 0.0968) for MET-minutes of MVPA [Variance components that are very close to zero are often estimated as negative. When that happens, SAS PROC MIXED fixes the estimate at zero. Estimates reported in this paper as 0.000 were in fact fixed at zero by the software during the estimation process.]

We also estimated the ICC from the data available from the first week alone and from the first 2 wk taken together. The unadjusted ICC for MET-minutes of MVPA was 0.0291 when estimated from week 1 alone and 0.0281 from week 1 and 2 alone. The ethnicity-adjusted ICC for MET-minutes of MVPA was estimated as 0.0000 when estimated from week 1 alone and as 0.0321 when estimated from week 1 and 2 alone.

Application: power analysis for the main trial

We will demonstrate the application of the results of the substudy through an explication of the power analysis for the main trial. The critical step is the selection of a value, or values, for the ICC. Values of school-level ICC are typically in the range of 0.005-0.05, though there are exceptions at both ends of that range (9). Underestimation of the true ICC can produce quite

misleading results in a power analysis, so it is important to be conservative in the selection of an ICC estimate. The fiducial distribution of the true ICC given an estimated ICC of 0.000 suggests a 50% probability that the true value is below 0.003, so 0.003 is a more realistic estimate than 0.000. Prior experience with ICC for other variables in school-based studies suggests a value of 0.010 as a conservative estimate (9), and the fiducial distribution of the true ICC given an observed ICC of 0.000 suggests a 62% probability that the true value is below 0.01. With these considerations in mind, the power analyses for TAAG used these three ICC estimates (0.010, 0.003, 0.000)

Primary analysis plan for the main trial

Explication of the power analysis for the main trial requires explication of the primary analysis plan for the main trial. In TAAG, that analysis will be conducted in two stages as if there is no overlap among girls measured in the sixth and eighth grade cross-sectional samples. This two-stage approach serves two functions. First, it avoids many of the complexities inherent in a single-stage mixed-model regression analysis of data from a multi-center GRT, including simultaneous estimation of multiple random effects and distributional assumptions for those effects (8). Second, it mimics an ANCOVA performed on eighth grade data, with regression adjustment for sixth grade values on the primary end point.

In the first stage, the girl's MET-minutes of MVPA will be regressed on school, time (baseline, follow-up), their interaction, and ethnicity; study condition will not be included in that model:

$$\text{MVPA} = \text{intercept} + \text{school} + \text{time} + \text{time} \times \text{school} + \text{ethnicity} + \text{error} \quad [3]$$

Here, all terms will be modeled as fixed effects, except for residual error. Of interest in this first stage is the estimation of ethnicity-adjusted school means for MET-minutes of MVPA, estimated for each school at sixth grade and at eighth grade. By performing a pooled analysis of sixth and eighth grade data in the first stage, we will standardize the results for the two surveys against the same reference distribution for ethnicity, here the average ethnicity distribution over time. The result of the first stage will be two adjusted mean MVPA values (baseline, follow-up) for each of the 36 schools.

The second stage analysis will be conducted on the adjusted means from the first stage. We will first look for evidence of a differential effect of the intervention among the six sites by testing for an interaction between site and condition. If there is a significant interaction between site and condition, the data cannot be pooled across sites, and instead the results will be reported separately for each site. However, we do not anticipate any evidence of heterogeneity, so that we can remove the interaction and proceed with a main-effects model.

In the main-effects model, we will regress the follow-up school mean MET-minutes of MVPA on condition, adjusting for the baseline school mean and stratifying on site:

$$\text{MVPA}_{\text{followup}} = \text{intercept} + \text{condition} + \text{MVPA}_{\text{baseline}} + \text{site} + \text{school} \quad [4]$$

Here, intercept (1 *df*), condition (1 *df*), and baseline MVPA (1 *df*) are fixed effects, whereas site (6 - 1 = 5 *df*) and school (36 - 8 = 28 *df*) are random effects; in this model, there is no residual error beyond school, and we use the term school instead of residual error to avoid confusion between Eq. 1 and Eq. 2. Given a proper randomization and a well-executed study, this model provides an unbiased test of the intervention effect. It also provides the statistical basis for inferences to sites and schools like those included in TAAG. The test of the intervention effect is given by the *F*-test for condition with 1 and 28 *df*

Assumptions for the power analysis

The assumptions underlying the power analysis for MET-minutes of MVPA are summarized in Table 2. The mean and SD for MET-minutes of MVPA are taken from the substudy reported here and as reported in Table 1. The test of the TAAG intervention effect will be two-tailed, with a Type I error rate of 5%. To be conservative, the power analysis ignores any correlation over time at the level of the girl or school. Eighteen schools will be randomized to each arm of the study.

The power calculations assume an attrition rate of 36% over 2 yr. The power calculations also assume that occasional missing accelerometry data within a girl's 6-d records are replaced via imputation based on the EM algorithm (1); details on the imputation methods are available from the second author upon request.

We assumed that 20% of girls would refuse at the baseline and follow-up surveys. Even so, the power calculations ignore refusals. Refusals that are unrelated to study condition or physical activity level will not bias the estimate of the intervention effect. We anticipate that no more than 25% of the refusals will be nonignorable. Separate analyses suggest that power will decline by about 3% if we impute values for 25% of the refusals, assuming no intervention effect for those cases. At the same time, the intervention effect is overestimated by only about 5% if those refusals are ignored. Given this minor adverse effect in terms of bias, we do not believe it is necessary to impute for refusals in the primary analysis, though we will check that assumption.

TAAG investigators anticipate that girls in the intervention schools during seventh and eighth grade will display an intervention effect of 15.6 more MET-minutes of MVPA than girls in the control schools; this effect is equivalent to 10% of the mean and to a 50% reduction in the decline anticipated between the sixth and eighth grades. TAAG investigators also anticipate that girls enrolled in the intervention schools only in the eighth grade will display an intervention effect of 9.4 more MET-minutes of MVPA; this effect is equivalent to 6% of the mean and to a 30% reduction in the decline anticipated from the sixth to the eighth grades. Combining these expectations with those made for attrition, we anticipate that the average intervention effect observed in the girls measured in the eighth grade will be 9.2% of the mean or 14.4 MET·min.

Methods for the power analysis

The test for the intervention effect in the primary analysis is given by the F -test for condition from the second stage of that analysis. Because that test involves a 1 df contrast, it is convenient to use the equivalent t -test for the power analysis, written generically as $\hat{t} = \hat{\Delta} / \hat{\sigma}_{\Delta}$ where $\hat{\Delta}$ is the estimate of the intervention effect and $\hat{\sigma}_{\Delta}$ is the estimate of the SE for the intervention effect. We estimate power using Eq. 9.7 from Murray (8):

$$\text{power} = \text{prob}(t \leq t_{\beta}), \text{ where } t_{\beta} = \frac{\hat{\Delta}}{\hat{\sigma}_{\Delta}} - t_{\text{critical}; \alpha/2} \quad [5]$$

We calculate $\hat{\sigma}_{\Delta}$ using Eq. 9.25 from Murray (8):

$$\hat{\sigma}_{\Delta} = \sqrt{\frac{2(\hat{\sigma}_m^2 + m\hat{\sigma}_g^2)}{mg}} \quad [6]$$

The value 2 in the numerator reflects the two means that define the intervention effect, $\hat{\sigma}_m^2$ is the estimate of the unadjusted girl component of variance, $\hat{\sigma}_g^2$ is the estimate of the unadjusted school component of variance, m is the number of girls per school included in the

analysis, and g is the number of schools per condition included in the analysis. The two components of variance sum to the total variation in the dependent variable,

$$\hat{\sigma}_y^2 = \hat{\sigma}_m^2 + \hat{\sigma}_g^2 \text{ and are related to the ICC as shown in Eq. 9.74 and 9.75 in Murray (8):}$$

$$\hat{\sigma}_m^2 = \hat{\sigma}_y^2(1 - \hat{\text{ICC}}) \text{ and } \hat{\sigma}_g^2 = \hat{\sigma}_y^2(\hat{\text{ICC}}) \quad [7]$$

Table 2 provides the estimate of the standard deviation for MET-minutes of MVPA from the data gathered for this substudy; squared, we have a value for $\hat{\sigma}_y^2$ of 9000.32. To be conservative, we assumed that there would be no overtime correlation in MET-minutes of MVPA at the girl or school level, so those correlations are not reflected in Eq. 6 or 7. Girl-level correlation cannot affect these calculations, given the serial cross-sectional surveys; the best estimate of the school-level correlation is about 0.2, so that any loss in efficiency from assuming that correlation is zero is a few percentage points at best. We do expect to benefit from regression adjustment for ethnicity, but that is reflected in the estimate of the ICC; as noted above, we used values of 0.000, 0.003, and 0.010 for the power analysis. The df available for the test of the intervention effect are 28, so that $t_{\text{critical};\alpha/2}$ is 2.05.

To illustrate these methods, consider power for an ICC of 0.01. Substituting the estimates for the parameters in Eq. 7 yields:

$$\hat{\sigma}_m^2 = 9000.32(1 - 0.01) = 8910.31 \text{ and } \hat{\sigma}_g^2 = 9000.32(0.01) = 90.00 \quad [8]$$

If we invite 120 girls to participate in the eighth grade survey, and 80% complete the measures, we will have data on 96 girls in each of 18 schools from the eighth grade survey. Substituting these estimates into Eq. 6 yields:

$$\hat{\sigma}_\Delta = \sqrt{\frac{2(8910.31 + 96 \times 90.00)}{96 \times 18}} = 4.507 \quad [9]$$

Substituting that estimate into Eq. 5 yields:

$$\begin{aligned} \text{power} &= \text{prob}(t \leq t_{\beta}), \text{ where } t_{\beta} = \frac{14.4}{4.507} - 2.05 \\ &= \text{prob}(t \leq 1.145) \\ &= 0.869 \end{aligned} \quad [10]$$

So given the assumptions outlined above, TAAG will have 86.9% power to detect a full dose intervention effect of 10%.

Results of the power analysis

Table 3 summarizes power for TAAG based on these assumptions and methods, presented for the three levels of ICC as a function of the number of girls per school invited to participate at followup. These results indicate that even if the ICC is as high as 0.01, if we invite 120 girls from each school to participate in the follow-up survey, and no more than 20% refuse to provide data, power for the intervention effect is quite good, given the assumptions noted in Table 2.

DISCUSSION

The unadjusted school-level ICC for MVPA was 0.0205 and for MET-minutes of MVPA was 0.0045. These values are appreciably smaller than the unpublished values from the Amherst team but consistent with the values observed for other school-level ICC (9). The differences between the values observed in this substudy, and in the studies by the Amherst team (17) are likely due to the very limited number of schools included in the previous studies (two and four schools), and to the fact that the earlier estimates were based on fewer measurement days in each school, which would increase the expected correlation among the observations. In

contrast, the current estimates were derived from many more schools with 3 wk of measurements in each school.

Adjustment for age and BMI had no measurable effect on the ICC, whereas adjustment for ethnicity reduced both ICC; the adjusted values were 0.0175 for minutes of MVPA and 0.0000 for MET-minutes of MVPA. These reductions are consistent with previous studies that have shown the benefit of regression adjustment for covariates at the member or group level (9).

The ICC for MET-minutes of MVPA were somewhat smaller than for minutes of MVPA, whether unadjusted or adjusted for ethnicity. Duration of activity may be correlated among girls in the same school, due to common schedules, whereas the intensity of activity may be more idiosyncratic, even among girls following the same activity schedule, so that it is less well correlated among girls in the same school.

There was no clear pattern in the MET-minutes of MVPA or minutes of MVPA over the 3 wk of measurement, either for the means or the SD (cf. Table 1). This suggests that the sampling scheme was effective in generating weekly groups that were comparable. The unadjusted ICC for MET-minutes of MVPA was appreciably higher when based on the first week, or on the first 2 wk, compared with the full 3 wk of data; that pattern supported the hypothesis that the ICC would be smaller if data collection were spread out over time. The adjusted ICC for MET-minutes of MVPA was appreciably higher when based on 2 wk of data but no different when based on just 1 wk compared with 3 wk; that pattern was not expected and is difficult to explain. Nonetheless, because the ICC estimates were at their lowest level when estimated based on three weeks of data, the protocol for TAAG will encourage a 3-wk schedule for data collection.

As expected, the reliability of the estimate of MET-minutes of MVPA was low for a single day of measurement. We determined that 5 or 6 d of monitoring would be sufficient to obtain acceptable reliability of about 0.8. As a result, the protocol for TAAG requires only six full days of accelerometry data. This will also facilitate the logistics of measurement, as it will provide 1 d each week to be used for removing the accelerometers, downloading the data, and reinitializing the accelerometers. By comparison, in African-American girls aged 8-9 and participating in the GEMS study wearing the same accelerometer, 7 d of monitoring were required to attain a reliability of 0.80 (16). In a separate study, between 4 and 5 d of accelerometer measurements were necessary to achieve a reliability of 0.80 in both male and female children and adolescents spanning grades 1-12 (17).

This substudy affirms the importance of developing an accurate estimate of the ICC expected to apply in the primary analysis as part of the planning for any new GRT. Two unpublished estimates for the primary outcome in TAAG were available, but both appeared high in comparison to school-based ICC for other health behaviors (9). The estimate obtained in the substudy was considerably smaller, particularly after adjustment for ethnicity. Had the investigators planned TAAG based on the estimates from the Amherst team, a much larger and more expensive trial would have been required.

This experience also affirms the need for studies that collect data in schools and other identifiable social groups to analyze their data and report ICC for variables that could serve as primary endpoints for in other studies. Had better estimates been available, it might have been possible to power TAAG based on those estimates, thereby avoiding the time and cost of the substudy.

Absent such information, other investigators are encouraged to follow the strategy adopted by the TAAG investigators: to plan a pilot study designed to provide a valid and precise estimate of the ICC likely to operate in their trial. That estimate is as critical in the planning of a GRT

as is a valid and precise estimate of the variance of the primary endpoint in a randomized clinical trial.

Acknowledgements

This research was funded by grants from the National Heart, Lung, and Blood Institute (U01HL66858, U01HL66857, U01HL66845, U01HL66856, U01HL66855, U01HL66853, and U01HL66852).

None of the authors has any personal or professional relationship with any company or manufacturer that might benefit from the results of this study.

REFERENCES

1. Catellier JD, Murray DM, Hannan PJ, et al. Imputation of missing data when measuring physical activity by accelerometry. *Obes. Res* 2003;11(Suppl):A32.
2. Cornfield J. Randomization by group: a formal analysis. *Am. J. Epidemiol* 1978;108:100–102. [PubMed: 707470]
3. Donner, A.; Klar, N. Design and Analysis of Cluster Randomization Trials in Health Research. Arnold; London: 2000. p. 178
4. Fisher, RA. Statistical Methods for Research Workers. Hafner Press; New York: 1970. p. 221
5. Fleiss, JL. The Design and Analysis of Clinical Experiments. John Wiley & Sons; New York: 1986. p. 370-385.
6. Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc* 1977;72:320–338.
7. Kish, L. Survey Sampling. John Wiley & Sons; New York: 1965. p. 643
8. Murray, DM. Design and Analysis of Group-Randomized Trials. Oxford University Press; New York: 1998. p. 467
9. Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. *Eval. Rev* 2003;27:79–103. [PubMed: 12568061]
10. Murray DM, McKinlay SM, Martin D, et al. Design and analysis issues in community trials. *Eval. Rev* 1994;18:493–514.
11. Puyau MR, Adolph AL, Vohra FA, Butte NF. Validation and calibration of physical activity monitors in children. *Obes. Res* 2002;10:150–157. [PubMed: 11886937]
12. SASInstitute. SAS/STAT Software: Changes and Enhancements. SAS Institute, Inc; Cary, NC: 2000. p. 154Release 8.1
13. Schmitz KH, Treuth MS, McMurray RG, et al. Predicting energy expenditure for accelerometry counts in adolescent girls. *Obes. Res* 2003;11(Suppl):A33.
14. Snedecor, GW.; Cochran, WM. Statistical Methods. 8th Ed.. The Iowa State University Press; Ames, IA: 1989. p. 244-245.
15. Treuth MS, Schmitz K, Catellier DJ, et al. Defining accelerometer thresholds for physical activity intensities in adolescent girls. *Obes. Res* 2003;11(Suppl):A33.
16. Treuth MS, Sherwood NE, Butte NF, et al. Validity and reliability of activity measures in African-American Girls for GEMS. *Med. Sci. Sports Exerc* 2003;35:532–539. [PubMed: 12618587]
17. Trost C, Pate RR, Freedson PS, Sallis JF, Taylor WC. Using objective physical activity measures with youth: how many days of monitoring are needed? *Med. Sci. Sports Exerc* 2000;32:426–431. [PubMed: 10694127]

TABLE 1
Descriptive statistics for BMI, weight, minutes of MVPA, and MET-minutes of MVPA.

	N	Mean	SD	Median	Site Mean	
					Minimum	Maximum
Age (yr)	436	14.1	0.45	14.1	14.0	14.3
BMI (kg·m ⁻²)	436	22.7	4.87	21.5	21.5	24.1
Weight (kg)	436	58.4	14.10	55.2	55.6	61.5
Minutes of MVPA	433	25.8	13.32	23.0	21.8	29.8
Week 1	135	26.3	13.54	23.4		
Week 2	145	25.0	13.11	22.5		
Week 3	153	26.1	13.38	23.2		
MET-minutes of MVPA	433	156.4	94.87	136.5	131.2	178.7
Week 1	135	157.6	85.54	138.5		
Week 2	145	153.2	93.86	133.5		
Week 3	153	158.4	103.76	137.8		

BMI, body mass index; MVPA, moderate to vigorous physical activity; MET, metabolic equivalent.
Site mean minima and maxima were identified from the six site means available for each measure.

TABLE 2

Assumptions underlying the power analysis for MET-minutes of MVPA.

Mean	156.4
SD	94.87
Two-tailed Type I error rate	5%
Over-time correlation among schools	None
Over-time correlation among girls	None
Schools per condition	18
Attrition rate over 2 yr	36%
Refusal rate at follow-up for accelerometry measurements	20%
Intervention effect	
For girls enrolled in intervention schools in 7th and 8th grade	15.6 MET·min (10% of mean)
For girls enrolled in intervention schools only in 8th grade	9.4 MET·min (6% of mean)
Accelerometry measurement schedule	
Days per girl	Six 18-h days
Weeks per school	3
Imputation of occasionally missing accelerometry data	EM algorithm
Imputation for nonignorable refusals	None
School-level Intraclass Correlation	
Observed estimate	0.000
Realistic estimate	0.003
Conservative estimate	0.010

MVPA, moderate to vigorous physical activity; MET, metabolic equivalent; EM, expectation maximization, implemented as described elsewhere (1).

TABLE 3
Power for the main effect of 10% in the TAAG main trial as a function of the number of girls per school invited to participate and the school-level ICC.

	Girls/School Invited to Participate at Follow-up			
	90	105	120	150
	Girls/School Measured at Follow-up (80% Participation)			
ICC	72	84	96	120
0.010	81%	85%	87%	90%
0.003	92%	95%	97%	98%
0.000	96%	98%	99%	100%