Psychology in Spain, 2004, Vol. 8. No 1, 77-88

STATISTICAL ANALYSIS AND CONSIDERATIONS OF POWER IN PROGRAM EVALUATION USING TWO-STAGE SAMPLING DESIGNS

Guillermo Vallejo Seco, José Ramón Fernández Hermida and Roberto Secades Villa
University of Oviedo

La evaluación de programas de prevención acarrea errores de decisión derivados principalmente de la dificultad de asignar al azar las unidades individuales a las condiciones de investigación. La elección de la unidad de análisis apropiada a la hora de evaluar la efectividad del impacto está determinada por la naturaleza de la intervención y por el diseño de investigación seleccionado. Cuando las unidades de asignación y de observación difieren entre sí, esto es, cuando entidades colectivas más que individuales son asignadas al azar a los tratamientos, los análisis realizados en los niveles más bajos de la jerarquía proporcionan estimaciones ineficientes de los parámetros y a menudo conducen a que las pruebas de significación sean inadecuadas. La meta de este trabajo es doble. Por un lado, presentar un método analítico que permite utilizar los datos de cualquier nivel del diseño sin inflar las tasas de error. Y, por otro lado, determinar el número de grupos y el tamaño de éstos en función de la variabilidad existente y de los costos.

Statistical analysis and considerations of power in program evaluation using two-stage sampling designs. Evaluation of prevention programs involves decision errors due to the difficulty of randomly assigning individuals to research conditions. The nature of the intervention and design of the study determine the choice of the appropriate unit of analysis in impact assessments. When units of assignment and units of observation differ, that is, when clusters of people rather individuals are assigned at random to treatments, the analyses conducted at lower levels of the study hierarchy provide inefficient parameter estimates, and often result in inappropriate significance tests. Therefore, the purpose of this paper is (a) to present an analytical method that permits the use of data at all levels of design without increasing Type I error rates, and (b) to determine the number of clusters and the sample size per group according to variability and cost.

The data from many research designs based on program evaluation have a structure very similar to that observed in cluster sampling designs with two or more stages. In the simplest case, the researcher initially selects a random sample of clusters or primary sampling units, and then randomly selects the secondary sampling units within each of the clusters. The primary units can be schools, classes, clinics or any other type of entity, and the secondary units, teachers nested within schools, students nested within classes or patients nested within clinics.

In all the cases mentioned above, collective units of analysis, more than individual units of analysis, constitute the observational reference to which the treatment or social intervention program is directed. When researchers use collective units of analysis for reasons of logistics, political viability or ecological validity, or for any other reason, what they usually do is randomly assign some units configured prior to the intervention to the treatment condition and others to the control condition. A researcher who has proceeded in line with the above will rarely select at random individual units of the primary sampling units to then assign them at random to the program. In any case, even though it were possible to assign people within the groupings to the program, it does not appear to be a desirable option, due, among other reasons, to the probable diffusion of treatments. It is more frequent for the researcher to create the problem of the unit of analysis, administering the treatment collectively to units originally assigned to the groups on an individual basis. Researchers should therefore proceed with caution in these matters. It is clearly not the same to assign collective units at random to the prevention program as it is to do so individually. Such confusion not only restricts the researcher's ability to understand the research design employed, but may also invalidate the use of techniques based on the general or generalized linear model for evaluating the consequences of a program's application.

The original Spanish version of this paper has been previously published in *Psicothema*, 2003, Vol. 15. No 2, 300-308

.....

Correspondence concerning this article should be addressed to Guillermo Vallejo Seco, Facultad de Psicología, Universidad de Oviedo, 33003 Oviedo Spain. E-mail: gvallejo@correo.uniovi.es

The application of the techniques based on the models mentioned requires the satisfaction of certain assumptions, particularly that of independence between observations. When groups constitute the unit of analysis, it is reasonable to think that the specific characteristics of the groups are reflected in the data, since observations that are close in time, space or both dimensions at the same time tend to be more homogeneous than observations that are farther apart. It is highly probable that the data drawn from natural groupings such as towns, health communities or schools bear a certain similarity to each other, since they are exposed to common influences. For example, the pupils in a class talk to each other all the time, share the same types of experiences and are subject to the same educational factors. Thus, as Shadish, Cook and Campbell (2002) point out, the observations recorded from each unit will reflect both the effects of the individuality itself on the behaviour and the effects of the collective variables on the individuals. The former effects will vary within the collective units and across them, while the latter effects will vary only among the different collective units. An indicator of the portion of total variability attributable to the unit of assignment is obtained by means of the intra-class correlation coefficient.

Given that the members of a collective unit tend to be more similar than those that are not part of such a unit, a set of correlated observations provides less information than a similar number of independent observations. Thus, when statistical models that assume independence between the units are applied to correlated data, there is an underestimation of standard errors of measurement (Carvajal, Baumler, Harrist & Parcel, 2001). In practice, this means that both tests based on the classical linear model and those based on the generalized linear model substantially increase the probability of rejecting the null hypothesis when it is in fact true, and thus of leading us to conclude that a program is effective when it is noncommittal actually totally ineffective. In sum, it leads to researchers capitalizing on chance more frequently than they should due to inefficient estimations (Rinndskopf & Saxe, 1998).

In addition to non-fulfilment of the assumption of independence, when the data are organized hierarchically there is more than one source of random variance in them. Thus, neither techniques based on the general linear model nor those based on the generalized linear model are appropriate, since, in all cases, they only allow us to determine the variation of a single component. If the data follow a normal distribution, the natural solution is provided by the linear mixed models, also known as random-effects models, random components models, multilevel linear models, hierarchical models or mixed-effects regression models (Goldstein, 1995, Aitkin & Longford, 1986; Laird & Ware, 1982; Raudenbush & Bryk, 2002, Oliver, Rosel & Jara, 2000). If, on the other hand, the data follow any other member of the exponential family, the natural solution is provided by the generalized mixed models (Breslow & Clayton, 1993; Wolfinger & O'Connell, 1993). All of these models recognize the nested structure of the data and allow estimation of the variances occurring in the different strata produced by the grouping, in both transversal and longitudinal studies.

In the sections below we offer a brief introduction to the general linear mixed model, we describe the estimation techniques and we specify the inference procedures for checking the hypotheses corresponding to the fixed effects and random effects of the model and the variance components. Finally, we use a randomized-groups hierarchical design with pre-test and post-test to illustrate how to determine optimum sample sizes for determining, in turn, the effects of the design. As Raudenbush (1997) points out, even though these models are particularly attractive, they tend arouse a degree of mistrust among researchers in view of their relative analytical complexity and possible lack of statistical accuracy caused by incorrect selection of the size of sampling units.

The general linear mixed model

The standard linear model for explaining *n* observations taken for each one of the *p* covariants (predictors) and/or factors (independent variables) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad (1)$$

where **y** is a vector of dimension *n x 1* containing the values of the response variable y_{ij} for unit *i* in group *j* (or for subject *i* in time *j*), **X** is a design matrix of dimension *n x p* that specifies the fixed-effects values corresponding to each parameter for each one of the observations (vectors of zeroes and ones denote the absence and presence of categorical effects for variables without metric structure, while numerical vectors denote the effects of the variables measured on a quantitative scale), **ß** is a vector of non-random parameters estimated from the data that may include variables of various types, and **e** is a vector of unknown errors of dimension *n x 1* distributed normally and independently with a mean of zero and constant variance. The coefficients of the vector **ß** are fixed-effects

parameters that describe the average behaviour of the population. However, it may occur that not all the terms of the model take constant values in the successive repetitions of the study, and that, rather, some are seen as the result of extracting samples at random from a normal distribution (random-effects models). Moreover, if the design matrix includes covariants, it is also possible that the parameters of vector **ß** do not represent correctly the relationship between **X** and **y** for some subjects or groups. Consequently, it is necessary to employ an approach that permits the researcher to establish a global relationship between the variables for all the subjects and to model separate relationships that vary randomly among the subjects (random-coefficients models).

The linear mixed model provides the appropriate solution for dealing with the problem in question, since it does not require the assumption that selection of the levels of the variables must be made arbitrarily, or that all the coefficients of the model are fixed constants. Moreover, the mixed model approach also extends the general linear model on permitting a more flexible specification of the covariance matrix of **e**. Specifically, it relaxes the assumptions of homogeneity of the variances and independence of the errors. Using matricial notation, the mixed model is represented as follows:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z}\,\mathbf{u} + \mathbf{e} \qquad (2)$$

where the fixed-effects component, **Xß**, is defined as in the previous equation and the random-effects component, **Zu**, allows the definition of different relationships between units, or between subjects in the longitudinal case. $\mathbf{Z}_j$ is a design matrix of dimension n_j *x k* for a given second-level unit, or for an individual subject in the longitudinal case (n_j being the number of first-level units nested within each second-level unit, or the number of times a subject is observed, and *k* the number of predictors included). For many models, the *k* predictors are a subset of the *p* predictors included in the matrix **X**, and the subset of predictors is the same or similar for each unit or for each subject. **Z** is a second block-diagonal design matrix of the order n_j *x Jk* for the random component (*J* being the number of level 2 units, or the number of subjects). $\mathbf{u}_j$ is a random-effects parameters vector of dimension *k x 1* associated with $\mathbf{Z}_j$. The vectors $\mathbf{u}_1$, $\mathbf{u}_2$,…, $\mathbf{u}_j$ come together within **u,** a vector of dimension *Jk x 1* that contains the specific random-effects parameters for all the units, or for all the subjects in the longitudinal case. Finally, **e** is a vector of unknown parameters of dimension *n x 1*, whose elements, in contrast to the case of the classical model, do not need to be independent or homogeneous.

The distributional assumptions of the model imply that the residual coefficients **u** are distributed normally and independently, with mean **0** and covariance matrix **G**, where **G** is a block-diagonal matrix of dimension *Jk x Jk*, with each block $\mathbf{G}_j$ of dimension *k x k* containing the variances and covariances of the random effects for each one of the second-level units, or for each one of the subjects where repeated measures are used. It is also assumed that the error vector **e**, as well as being independent of the vector **u**, is distributed normally with mean **0** and covariance matrix **R**, where **R** is a block-diagonal matrix of the order *n x n*, with each block $\mathbf{R}_i$ containing the variances and covariances of the errors within the subjects. If it is fulfilled that $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$ and $cov(\mathbf{r}, \mathbf{u}) = \mathbf{0}$, then

$$\mathbf{y} \sim N[(\mathbf{X\beta}, \mathbf{V}(\theta)] \qquad (3)$$

where $\mathbf{V}(\theta) = \mathbf{Z}\,\mathbf{G}\,\mathbf{Z}' + \mathbf{R}$ and θ refers to the variance components of the matrix **V**. The classical model approach is a particular case of the mixed model approach. When $\mathbf{R} = \sigma^2\,\mathbf{I}$ and $\mathbf{Z} = \mathbf{0}$ the two approaches coincide perfectly.

Estimation of the parameters **ß**, ***u*** *and* **V(θ)** *of the mixed model*

The standard procedure for obtaining estimations of **ß** and **u**, assuming that the matrices **G** and **R** are known, consists in solving the equations of Henderson's well-known mixed model

$$\begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Z} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z} + \hat{\mathbf{G}}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{y} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{y} \end{bmatrix} \qquad (4)$$

Applying the identities

$$\hat{\mathbf{R}}^{-1} - \hat{\mathbf{R}}^{-1}\,\mathbf{Z}(\mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{R}}^{-1} = (\mathbf{Z}\hat{\mathbf{G}}^{-1}\mathbf{Z}' + \hat{\mathbf{R}})^{-}$$
$$(\hat{\mathbf{G}}^{-1} + \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{R}}^{-1} = \hat{\mathbf{G}}\mathbf{Z}'\mathbf{V}(\boldsymbol{\theta})^{-1} \qquad (5)$$

the estimators of **ß** and **u** that solve the equations of Henderson's mixed model are

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\hat{\mathbf{V}}(\boldsymbol{\theta})^{-1}\mathbf{X}]^{-}\,\mathbf{X}'\hat{\mathbf{V}}(\boldsymbol{\theta})^{-1}\mathbf{y}$$
$$\hat{\mathbf{u}} = \mathbf{G}\,\mathbf{Z}'\hat{\mathbf{V}}(\boldsymbol{\theta})^{-1}\,[\mathbf{y} - \mathbf{X}'\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})] \qquad (6)$$

and **û** have the properties of being the best unbiased linear estimator and the best unbiased linear predictor (also referred to as Bayes' empirical estimator or the shrunken estimator), respectively, of **ß** and **u** (McCulloch and Searle, 2001). In turn, the variances of and **û** are obtained as

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\hat{\mathbf{V}}(\boldsymbol{\theta})^{-1}\mathbf{X})^{-}$$
$$\text{var}(\hat{\mathbf{u}}) = \mathbf{G} - \mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G} \quad (7)$$

where the projection matrix

$$\mathbf{P} = \hat{\mathbf{V}}(\boldsymbol{\theta})^{-1} - \hat{\mathbf{V}}(\boldsymbol{\theta})^{-1}\mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}(\boldsymbol{\theta})^{-1}\mathbf{X})^{-}\mathbf{X}'\hat{\mathbf{V}}(\boldsymbol{\theta})^{-1}.$$

If the sample size is small and the design not perfectly balanced, the standard errors corresponding to the fixed and random effects may be negatively biased. In order to adjust the bias, several solutions have been suggested. For example, Kenward and Roger (1997) proposed calculating the specific inflation factor and adjusting the degrees of freedom. Alternatively, Liang and Zeger (1986) suggest computing the standard errors for the fixed-effects parameters using an asymptotically consistent estimator, known as the empirical variance estimator or sandwich estimator. Essentially, the sandwich estimator involves using the observed covariance pattern of the data, instead of a pattern of covariance selected statistically. Both procedures are available in the PROC MIXED module of the SAS program (2001, SAS Institute, version 8.2).

The symbol $(\cdot)^{-}$ used in the expression $[\mathbf{X'V}^{\wedge}(\theta)=\mathbf{X}]$ of the equation referring to the fixed effects indicates that a generalized inverse is required if **X** is not of full rank. The vector θ contains the unique elements of **G** and the parameters in **R**. It can clearly be seen that estimation of a multilevel model is equivalent to estimation of a mixed or combined model, since, although separate models can be formulated for each level, these models are linked statistically. On observing the equations referring to the estimators **ß** and **u**, it can be appreciated that the estimation of the fixed-effects vector depends on the matrix of variance components, while the estimation of the random-effects vector depends on both the matrix **V(θ)**, and the estimator of generalized squared minima **ß**.

With very few exceptions, the matrices **G** and **R** are unknown, which obliges us to determine the variance components of **V(θ)** from the data by means of one of the different estimation procedures available. If the research design is balanced, we could use algebraic procedures based on the method of moments (Searle, Casella and McCulloch, 1992). However, the traditional method of moments consisting in solving systems of simultaneous equations, relating the expected values with the observed ones, is difficult to accommodate in program evaluation, since, in applied contexts, the sampling units are usually nested in non-balanced groups with arbitrarily parameterized dispersion matrices. When the research design is unbalanced, the components of the matrix **V(θ)** are estimated iteratively using numerical procedures. As a general rule, these procedures are based on maximum-likelihood (ML) estimation techniques, or those of restricted maximum likelihood (RML), so as to avoid obtaining biased estimations. Another procedure available for estimating the elements of the matrix **V(θ)** is based on the Bayesian approach. Nevertheless, van der Leeden (1998), points out that the computational effort required by this procedure may be considerable when the models are complex and the sample sizes of the levels are large. In addition to the procedures mentioned, PROC MIXED incorporates other methods.

The ML estimators of θ are obtained by maximizing the natural logarithm of the likelihood function corresponding to the density of the vector **y** for **ß** and θ, where

$$l_c(\boldsymbol{\beta},\boldsymbol{\theta}) = -\tfrac{1}{2}[(n\log(2\pi) + \log|\mathbf{V}(\boldsymbol{\theta})| + \mathbf{e}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{e}] \quad (8)$$

with **e= y - Xß**. If *n* is small, what may be of interest, more then estimating the variance components from the global likelihood, is maximizing the part of the likelihood that is invariant of the model's fixed effects by means of the RML method. Specifically, in accordance with Harville's (1977) derivations, maximizing the likelihood function logarithm

$$l_r(\boldsymbol{\beta},\boldsymbol{\theta}) = -\tfrac{1}{2}[(n-h)\log(2\pi) + \log|\mathbf{V}(\boldsymbol{\theta})| + \log|\mathbf{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}| + \mathbf{e}'\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{e}] \quad (9)$$

Under the normal model, the ML or RML estimators of **ß** and θ are usually determined by means of the Newton-Raphson (NR) algorithm or the Expectation-Maximization (EM) algorithm described by Dempster, Laird and Rubin (1977). Nevertheless, there are some reasons for preferring the NR algorithm to the EM. According to Lindstrom and Bates (1988), the NR algorithm requires a smaller number of iterations for converging than the EM algorithm. Another advantage of the NR algorithm over the EM resides in the fact that the

former computes the standard errors of the elements of θ from the empirical information matrix (inverse of the Hessian matrix with a change of sign). When the EM algorithm is used this matrix is not calculated, and nor, therefore, are the standard errors for the elements of θ (Jennrich & Schluchter, 1986). Details of the matricial derivatives of the NR algorithm implemented in the PROC MIXED module are available in the work of Wolfinger, Tobias and Sall (1994).

Checking of hypotheses in the mixed model

The procedures presented in this section for testing the fixed and random effects and the variance components are general. For checking the fixed and random effects, SAS uses statistical tests based on the distributions *F* or *t*, while for testing the variance components, SAS uses Wald's *Z* statistic. Once the dispersion matrix has been identified and its parameters duly estimated, **ß** and **u** are estimated using the procedures defined above and it is checked whether the *h* functions to be estimated have the value specified in the following null hypotheses: . For example, the different coefficients corresponding to the fixed and random effects of the model are checked by calculating the ratio between the ML (or RML) estimators and their respective standard errors, as follows:

$$t(\hat{\boldsymbol{\beta}}) = \frac{\mathbf{L}'\hat{\boldsymbol{\beta}}}{\sqrt{\mathbf{L}'[\mathbf{X}'\mathbf{V}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{X}]^{-}\mathbf{L}}} \quad \text{y} \quad t(\hat{\mathbf{u}}) = \frac{\mathbf{L}'\hat{\mathbf{u}}}{\sqrt{\mathbf{L}'(\mathbf{G} - \mathbf{GZ}'\mathbf{PZG})\mathbf{L}}} \qquad (10)$$

Each one of the hypotheses referring to the fixed and random effects of the model are rejected at level α if *t* >, where is the $100_{(1-\alpha/2)}{}^{th}$ percentile of the distribution *t* with n degrees of freedom. When the data are not balanced, as is usually the case, and the number of level 2 units is small, the above test may offer liberal results. With a view to solving this problem, SAS provides the possibility of using the options DDFM= SATTERTH or DDFM= KENWARDROGER for adjusting n.

In turn, the null hypotheses corresponding to the variance components (**R** and **G**) are of the form H_0: θ= **0**. In order to test this type of hypothesis, SAS provides Wald's *Z* statistic. This statistic is obtained by dividing each one of the parameters estimated via ML (or RML) by its corresponding standard error:

$$Z = \frac{\hat{\boldsymbol{\theta}}}{[\mathbf{I}(\boldsymbol{\theta})]^{1}}, \qquad (11)$$

where $\mathbf{[I(\theta)]^{-1}}$ is the asymptotic covariance matrix (inverse of the information matrix) corresponding to the ML (or RML) solution. The Wald statistic, computed by default by SAS, is only exact asymptotically (Wolfinger, 1996), that is, for relatively large samples. Consequently, when the number of second-level units is small, caution should be exercised on interpreting the results.

It should be underlined, finally, that when analyzing longitudinal data recorded regularly and consistently for a small number of subjects, it is usual to fix the matrix **Z**= **0** and model the pattern of covariance that follows the matrix **R**. Obviously, this will also depend on whether researchers are interested in the global interceptors and tendencies (fixed effects) or the individual ones (random effects). If the researcher is interested in the fixed effects, more than testing a simple covariance parameter, the appropriate procedure is to verify whether a given pattern of covariance causes a significant improvement with respect to another pattern. When PROC MIXED is used for discriminating between nested models without changes in the fixed effects, the researcher should use the residual likelihood ratio test; on the other hand, if the aim is to compare nested models with different fixed effects, the full likelihood ratio test should be used (see Singer, 2002, for a detailed explanation of this topic). For non-nested models, selection criteria, such as the Akaike Information Criterion (AIC) or Schwarz's Bayesian Information Criterion (BIC), have usually been adopted (Wallace & Green, 2002). However, it is also possible to continue using the likelihood ratio test if we compare each one of the models of interest with one that is simpler, but nested simultaneously in the two of them, and select the model that offers the greatest improvement (Brown and Prescott, 1999).

POWER ANALYSIS FOR DETECTING THE EFFECTS OF A MULTI-LEVEL DESIGN

More and more researchers consider it useful to know the probability that an impact of a given size will be statistically significant. Thus, methods aimed at detecting optimum sample sizes, differences between treatments and test power (henceforth referred to as power analysis) are essential to the process of planning research. However, the relatively widespread belief that the information required for carrying out a power analysis is difficult to obtain, and that merely conjectural work is irresponsible, has led to power analysis being largely neglected.

For this reason, it is not unusual to come across studies with inadequate sample sizes, most frequently in the form of insufficient numbers of participants. As is well known, the use of small sample sizes involves the risk of undue acceptance of null hypotheses, when they are in fact false in that population. In these cases we say that the design lacks testing power, as there is a high probability of accepting the model of chance as the most plausible explanation of the differences found. Thus, a key issue for research design is the correct choice of sample size.

Power analysis is useful both *a priori* and *a posteriori*. However, in our view, when it is crucial is before carrying out the research. And while we agree that inadequate work on power may lead to disappointment, we also firmly believe that detailed analysis of the factors that determine power gives researchers the capacity to design better studies. We coincide with D'Amico, Neilands, and Zambarano (2001) in their assertion that rigorous *a priori* analysis of power makes it possible to verify whether the effort, time and cost required by the research design are fully justified.

As illustrated through the examples mentioned earlier, there are a variety of research areas in which treatments are administered to groups of people, rather than individuals. Moreover, in the majority of cases group members have not been assigned to their groups in accordance with the rules of chance. In the best of cases intact groups are assigned at random to the treatment conditions. Consequently, in these circumstances it is necessary to determine precisely the size of the units of observation and assignment in order to detect differences between treatments and interactions, or the moderating effects of the characteristics of subjects and/or groups on variability of impact, since the units of assignment used in such cases are not subjects, but groups. Furthermore, as the resources involves in sampling the two types of units differ substantially, sample sizes should be determined not only according to the effect size of interest and the variations within and across groups, but also in line with the costs involved in detecting such effects.

In the remainder of this work we shall illustrate how to determine optimum sample sizes in order to show the main effect of the intervention and of its interaction with time in a hierarchical random-groups design with pretest and post-test. It can clearly be seen by means of the procedure we shall use that this design is uniformly superior for detecting differences between treatments to the standard hierarchical random-groups design. In order to achieve the objectives set we shall follow a procedure similar to that presented by Cohen (1988) and Raudenbush and Liu (2000). Specifically, we shall use a standardized model combining small (0.3), medium (0.5) and large (0.70) measures of effect size, variances within the groups equal to unit and variances across the groups equal to the square of the standardized effect sizes we have just specified.

Illustration of how to obtain sample sizes

For the design referred to in the previous section, the model can be formulated in scalar terms as follows:

$$y_{ijkl} = \mu + \alpha_j + \beta_{k(j)} + \gamma_l + (\alpha\gamma)_{jl} + (\beta\gamma)_{k(j)l} + \varepsilon_{i(jkl)} \quad (12)$$

where the observed value of the i^{th} subject nested within the j^{th} condition and of the k^{th} group in the l^{th} time (y_{ijkl}) is expressed as a function of the general measure (μ), of the effect of the j^{th} treatment condition (α_j), of the random effect of the k^{th} group nested within condition j ($\beta_{k(j)}$), of the effect of the l^{th} time (γ_l), of the combined effect of the j^{th} condition and the l^{th} time ($(\alpha\gamma)_{jl}$), of the random combination of the k^{th} group and the l^{th} time and of the random variation among the group members ($\varepsilon_{i(jkl)}$).

Alternatively, Equation 12 can be rewritten in terms of a multilevel model. To do so, we begin by writing at the first level a model similar to that of classical regression, incorporating time as explanatory variable

$$y_{ij} = b_{0j} + b_{1j} T_{ij} + e_{ij}, \quad (13)$$

where y_{ij} denotes the score of the i^{th} subject in the j^{th} group, the interceptor, b_{0j}, is equal to the mean of the group *j*, the slope, b_{1j}, represents average change on the post-test associated with a unit of change in the pre-test and e_{ij} denotes the difference between the score of the ij^{th} subject and the mean of the j^{th} group. For the sake of simplicity, we assume that the error follows a normal distribution with mean zero and constant variance across the groups, that is, .

Next, we incorporate the hierarchical nature of the data into the model. To do so we shall consider the regression coefficients b_{0j} and b_{1j} as dependent variables that fluctuate across the groups as a function of one mean plus the treatment and the error. Specifically, the regression coefficients are related to the treatment as follows:

$$b_{0j} = \beta_{00} + \beta_{01} Trat_j + u_{0j}$$
$$b_{1j} = \beta_{10} + \beta_{11} Trat_j + u_{1j} \quad (14)$$

In the level 2 models it is assumed that in each group the parameters b_{0j} and b_{1j} are distributed normally with means and , respectively, and matrix of variances-covariances

$$G = \begin{bmatrix} var(\beta_{00}) & cov(\beta_{00},\beta_{10}) \\ cov(\beta_{00},\beta_{10}) & var(\beta_{10}) \end{bmatrix} = \begin{bmatrix} var(u_{0j}) & cov(u_{0j},u_{1j}) \\ cov(u_{1j},u_{0j}) & var(u_{1j}) \end{bmatrix} = \begin{bmatrix} \omega_{00} & \omega_{01} \\ \omega_{10} & \omega_{11} \end{bmatrix} \quad (15)$$

It is also assumed that the errors corresponding to levels 1 and 2 are independent of one another, that is,

Substituting the expressions corresponding to Equation 15 in Equation 14, we obtain the following mixed model:

$$y_{ij} = \beta_{00} + \beta_{01} Trat_j + u_{0j} + \beta_{10} T_{ij} + \beta_{11} T_{ij} Trat_j + u_{1j} T_{ij} + e_{ij} \quad (16)$$

where y_{ij} denotes the score of the i^{th} subject in the j^{th} group, represents the value resulting from averaging the means of the groups, represents the difference of means in the response of interest between the groups receiving treatment and those that do not, u_{0j} indicates whether there are differences between the means of the groups in the dependent variable controlling the effect of the treatment, represents the average difference between pre-test and post-test, β_{11} represents the mean difference in the pre-test/post-test relationship between the groups receiving treatment and the control groups, u_{1j} indicates whether the relationship between the pre-test and post-test within the group varies across the groups when the effect of the treatment is kept constant, and e_{ij} denotes the difference between the score of the ij^{th} subject and the mean of the j^{th} group. The coding system assumed is as follows: 1 for the constant, |0.5| for the treatment and time, and |0.25| for their interaction.

Determination of sample size without considering the costs of the research

Apart from specifying the form and magnitude of the effect of the design, three other aspects are of key interest in a power analysis (see also Murray, 1998):

a. Selecting a statistical test for evaluating the effects of the design.
b. Determining the distribution of the statistical test selected.
c. Developing the non-centrality parameters of the effects of interest, together with their corresponding variances.

In the case of the main effect of the treatment, a valid test of is provided by the statistical test

$$F_{\beta_{01}} = \frac{E(\hat{\beta}_{01})}{E(\hat{u}_{0j})} \quad (17)$$

When the hypothesis tested is true, the distribution of the F statistic is approximated by means of a central F distribution with and degrees of freedom for the numerator and denominator, respectively. In the case of interaction, the F test is constructed in a similar way, specifically

$$F_{\beta_{11}} = \frac{E(\hat{\beta}_{11})}{E(\hat{u}_{1j})} \quad (18)$$

with B_{01} and B_{11} defined as in (22) and (26).

If is true, the F statistic follows a central F distribution with and degrees of freedom for the numerator and denominator, respectively.

However, under an alternative hypothesis, $F_{\beta 01}$ and $F_{\beta 11}$ follow a non-central F distribution with the specified degrees of freedom and the following non-centrality parameters:

$$\lambda_{\beta_{01}} = \frac{n Q r \beta_{01}^2}{4(\sigma_{\varepsilon}^2 + n r \omega_{00})} \quad (19)$$

$$\lambda_{\beta_{11}} = \frac{n Q \beta_{11}^2}{8(\sigma_{\varepsilon}^2 + n \omega_{11})} \quad (20)$$

Having specified the non-centrality parameters, selected the statistical test and determined the distribution, we can obtain the power corresponding to the fixed effects of the design by calculating the probability that a non-central F with degrees of freedom and and non-centrality parameter for exceeds the corresponding critical value (Muller, La Vange Ramey & Ramey, 1992). Formally

$$Power = 1 - Prob[F(v_1, v_2; \lambda) < Finv(1 - \alpha, \omega_1, \omega_2)] \quad (21)$$

where represents the critical value obtained from the inverse central F distribution function. The power values can be revealed by using appropriate computational rou-

tines. For example, the following expressions from the SAS program can be used to obtain the power corresponding to the effects of the random-groups hierarchical design with pre-test and post-test:

$$Power_{\beta_{01}} = 1 - \text{Prob}\, f\,(\text{Finv}(1\text{-alpha}, \nu_1, \nu_2), \nu_1, \nu_2, \lambda_{\beta_{01}})$$
$$Power_{\beta_{11}} = 1 - \text{Prob}\, f\,(\text{Finv}(1\text{-alpha}, (\nu_1, \nu_2), \nu_1, \nu_2, \lambda_{\beta_{11}})$$

Table 1
Power for the main effect of the treatment without taking cost into account

Q	N	ω_g	δ	Power	Q	Power	Q	Power
20	20	0.15	0.2	0.299	30	0.432	40	0.549
20	20	0.15	0.3	0.574	30	0.765	40	0.878
20	20	0.15	0.4	0.816	30	0.947	40	0.986
20	20	0.10	0.2	0.395	30	0.562	40	0.693
20	20	0.10	0.3	0.718	30	0.887	40	0.959
20	20	0.10	0.4	0.922	30	0.988	40	0.999
20	20	0.05	0.2	0.589	30	0.779	40	0.889
20	20	0.05	0.3	0.906	30	0.984	40	0.998
20	20	0.05	0.4	0.992	30	1.000	40	1.000
20	30	0.15	0.2	0.311	30	0.449	40	0.570
20	30	0.15	0.3	0.594	30	0.784	40	0.893
20	30	0.15	0.4	0.834	30	0.956	40	0.990
20	30	0.10	0.2	0.418	30	0.591	40	0.723
20	30	0.10	0.3	0.748	30	0.907	40	0.969
20	30	0.10	0.4	0.938	30	0.992	40	0.999
20	30	0.05	0.2	0.640	30	0.825	40	0.922
20	30	0.05	0.3	0.935	30	0.991	40	0.999
20	30	0.05	0.4	0.996	30	1.000	40	1.000
20	40	0.15	0.2	0.318	30	0.458	40	0.580
20	40	0.15	0.3	0.605	30	0.795	40	0.900
20	40	0.15	0.4	0.843	30	0.960	40	0.991
20	40	0.10	0.2	0.431	30	0.606	40	0.739
20	40	0.10	0.3	0.763	30	0.917	40	0.974
20	40	0.10	0.4	0.946	30	0.994	40	0.999
20	40	0.05	0.2	0.668	30	0.849	40	0.936
20	40	0.05	0.3	0.948	30	0.994	40	0.999
20	40	0.05	0.4	0.998	30	1.000	40	1.000
20	50	0.15	0.2	0.322	30	0.464	40	0.587
20	50	0.15	0.3	0.612	30	0.801	40	0.905
20	50	0.15	0.4	0.848	30	0.962	40	0.992
20	50	0.10	0.2	0.439	30	0.616	40	0.748
20	50	0.10	0.3	0.772	30	0.922	40	0.976
20	50	0.10	0.4	0.950	30	0.995	40	1.000
20	50	0.05	0.2	0.685	30	0.863	40	0.945
20	50	0.05	0.3	0.955	30	0.996	40	1.000
20	50	0.05	0.4	0.998	30	1.000	40	1.000

Note: Q = number of groups; N= members within the group, α= .05; ω_g= Variance due to the groups + variance due to the interaction. ω_{11}= 0.10

By way of illustration, Table 1 shows the power obtained for different values of n, Q, effect sizes and variances across the groups.

The results in Table 1 suggest the appropriateness of planning the design with a larger number of groups than of members in each group, especially when the impact size postulated is small and the variance of treatments across groups is large. However, from an economic point of view, this conclusion may be unrealistic, since, as a general rule, sampling of groups is more costly than sampling of group members. Thus, it is important to carry out the power analysis considering also the costs involved in the sampling process.

Determination of sample size according to sampling costs

In order to carry out this analysis we need to know the variance of the effects of the design. Following a procedure similar to that described by Murray (1998) and Raudenbush and Liu (2000), the standard error of the main effect

$$\hat{\beta}_{01} = \overline{Y}_E - \overline{Y}_C \quad (22)$$

can be easily obtained if we express the variance of the group mean based on *n* dependent observations and *r* repeated measures as

$$\sigma^2_{\bar{y}} = \frac{\sigma^2_\varepsilon}{nr} + \omega_{00} \quad (23)$$

and the variance of the treatment condition j based on q groups of the same size

$$\sigma^2_{y_j} = \left(\frac{\sigma^2_\varepsilon}{nr} + \omega_{00} \right) / q \quad (24)$$

Thus, assuming that the variances are homogeneous across the groups, we have

$$Var(\hat{\beta}_{01}) = \frac{4(\sigma^2_\varepsilon + nr\omega_{00})}{nQr} \quad (25)$$

Operating in the same way, we find that the variance corresponding to the interaction effect

$$\hat{\beta}_{11} = (\overline{Y}_{ED} - \overline{Y}_{EA}) - (\overline{Y}_{CD} - \overline{Y}_{CA}) \quad (26)$$

gives

$$Var(\hat{\beta}_{11}) = \frac{8(\sigma_\varepsilon^2 + n\omega_{11})}{nQ} \quad (27)$$

From Equations 25 and 27 it can be appreciated that both the number of groups and the number of members in each group affect the accuracy of the estimations. However, lack of statistical accuracy will be greater when Q is reduced that when *n* is reduced. Thus, given that the units of assignment affect the sensitivity of the design more than the units of observation, researchers should negotiate carefully, according to their costs, the sizes of Q and *n* they include in the study for obtaining appropriate power.

In accordance with Cochran (1977), in many two-stage sampling designs the cost involved in collecting data can be approximated by an expression of the form

$$C = C_1 nQ + C_2 Q \quad (28)$$

where C refers to the total cost of the study, C1 to the cost involved in sampling the members within each one of the groups and C2 the cost associated with each one of the groups.

Having determined the total cost of the study, the researcher is in a position to select the value of *n* that minimizes the variance of the design's effects. For this, just two simple operations are necessary. On the one hand, to express the variances of the effects of the design taking into account the costs of the study

$$Var(\hat{\beta}_{01}) = \frac{4(\sigma_\varepsilon^2 + nr\omega_{00})(nC_1 + C_2)}{nrC} \quad (29)$$

and

$$Var(\hat{\beta}_{11}) = \frac{8(\sigma_\varepsilon^2 + n\omega_{11})(nC_1 + C_2)}{nC} \quad (30)$$

And, on the other, to discover the value of *n* that minimizes the variances of Equations 29 and 30. Deriving with respect to *n*, we find (see Appendix)

$$n_{\hat{\beta}_{01}}(optimum) = \left[\frac{\sigma_\varepsilon^2 C_2}{r\omega_{00} C_1}\right]^{1/2} \quad (31)$$

and

$$n_{\hat{\beta}_{11}}(optimum) = \left[\frac{\sigma_\varepsilon^2 C_2}{\omega_{11} C_1}\right]^{1/2} \quad (32)$$

We would have obtained identical values by maximizing the non-centrality parameters of Equations 19 and 20 with respect to *n*.

Assuming that the relative cost between C2/C1 is estimated at 2, 6, 8 and 10, Table 2 shows the values of *n*, Q and power, for diverse effect sizes, variances across groups and cost ratios.

With regard to the results in Table 2, four aspects are worthy of mention. First, keeping effect size and variance across groups constant, power increases as the ratio of costs decreases. Secondly, medium and large effects produce powers that approach the value considered ideal. However, it can also be seen that when treatment groups are separated by 0.2 standard units, variances lower than 0.10 will be required to provide powers that detect the treatment effect, at least in 50% of cases, since power increases as variance decreases. Third, the more costly it is to sample the groups in relation to the number of members making up the group, the greater the size of *n* and the smaller that of Q. Finally, it should be stressed that detection of the interaction effect requires larger sample sizes than detection of the main effect. Nevertheless, from the qualitative point of view it can be seen how the power functions of the main effect and interaction effect are identical.

CONCLUSIONS

The derivations presented in the present work show that the general linear model cannot be used to estimate the parameters of the mixed model in Equation 2, since the ordinary squared minima procedure assumes that the errors are independent, with a mean of zero and constant variance. However, in a model such as that of Equation 2 there are multiple sources of random variance, the

errors are not necessarily independent and the variances can differ among one another. In these cases neither the general nor the generalized linear model are appropriate. The natural solution to the problems indicated is provided by the general linear mixed model if the probability distribution of the response variable does not deviate from normality, or the generalized linear mixed model if the data follow any other member of the exponential family of distributions. For example, if we focus on the model in Equation 16, we can see identify three sources of random variation, which are duly estimated and interpreted by means of the mixed model approach. Moreover, if there is dependence between the first-level units nested within the second-level units, this is obtained independently of the error by estimating the variation in the second-level units induced by the grouping (Carvajal, Baumler, Harrist & Porcel, 2001).

Table 2
Power for effect of treatment and interaction as a function of costs

C_2/C_1	VAR	ES	N(T)	Q(T)	POWER(T)	N(I)	Q(I)	POWER(I)
2	.16	.1	3	100	0.232	4	83	0.169
2	.16	.3	3	100	0.865	4	83	0.700
2	.16	.4	3	100	0.998	4	83	0.978
2	.06	.1	4	83	0.312	6	63	0.211
2	.06	.3	4	83	0.957	6	63	0.822
2	.06	.4	4	83	1.000	6	63	0.996
2	.01	.1	10	42	0.447	14	31	0.266
2	.01	.3	10	42	0.995	14	31	0.915
2	.01	.4	10	42	1.000	14	31	1.000
5	.16	.1	4	56	0.164	6	45	0.128
5	.16	.3	4	56	0.682	6	45	0.527
5	.16	.4	4	56	0.973	6	45	0.901
5	.06	.1	6	45	0.229	9	36	0.168
5	.06	.3	6	45	0.859	9	36	0.696
5	.06	.4	6	45	0.998	9	36	0.977
5	.01	.1	16	24	0.371	22	19	0.235
5	.01	.3	16	24	0.983	22	19	0.869
5	.01	.4	16	24	1.000	22	19	0.998
10	.16	.1	6	31	0.121	8	28	0.104
10	.16	.3	6	31	0.488	8	28	0.393
10	.16	.4	6	31	0.870	8	28	0.770
10	.06	.1	9	26	0.172	13	22	0.135
10	.06	.3	9	26	0.711	13	22	0.561
10	.06	.4	9	26	0.981	13	22	0.922
10	.01	.1	22	16	0.308	32	12	0.194
10	.01	.3	22	16	0.952	32	12	0.775
10	.01	.4	22	16	1.000	32	12	0.991
20	.16	.1	8	18	0.092	11	16	0.082
20	.16	.3	8	18	0.321	11	16	0.260
20	.16	.4	8	18	0.666	11	16	0.557
20	.06	.1	13	15	0.126	18	13	0.104
20	.06	.3	13	15	0.514	18	13	0.395
20	.06	.4	13	15	0.890	18	13	0.771
20	.01	.1	32	10	0.234	45	8	0.157
20	.01	.3	32	10	0.863	45	8	0.644
20	.01	.4	32	10	0.998	45	8	0.957

Note: C_2/C_1= ratio of costs; VAR= variance of impact across groups; ES=standardized effect size; N(T)= value of n required for main effect; Q(T)= number of groups required for main effect; POWER (T)= test power corresponding to main effect; N(I)= value of n required for the interaction; Q(I)= number of groups required for the interaction; POWER (I)= test power corresponding to the interaction

The present study also shows, for the hierarchical design of groups at random with pre-test and post-test (one of the most commonly-used designs, according to Murray [1998], in the evaluation of prevention programs based on organizations) how to maximize power for highlighting the effects of treatments by selecting optimum sample sizes, in terms of both number of groups and their size. Although for achieving this aim we have used, like Raudenbush and Liu (2000), a standardized model, there will be plenty of cases in which researchers are able to anticipate the value of the variance components and the effect size using data from some previous study or a pilot study. In any case, the results presented in the tables show quite clearly some of the guidelines for researchers using this type of design in the planning of their work, with a view to obtaining sufficient statistical power. Maintaining design, type of analysis and number of replications constant, statistical accuracy would probably be improved by the inclusion of some auxiliary variable and using more rounds of observations.

Finally, we should bear in mind some of the limitations of the present work. Specifically, all the derivations refer to the power analysis for a relatively simple design with two experimental conditions (treatment and comparison), in which it was assumed that the dependent variable was continuous, with regularly-recorded data, balanced groups and no lost observations. Although it lies outside the brief of this work to extend the derivations found to more real situations, such as non-balanced designs, it would not be much more complicated to do so.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to all those whose comments contributed to improving the present work. The study was supported by a research grant from the Spanish Ministry of Science and Technology (MCT) (Ref.: BOS-2000-0410).

APPENDIX

Differentiating the with respect to *n* and equalling to zero, we have:

$$Var(\hat{\beta}_{11}) = 4\left(\frac{\sigma_e^2}{nr} + \frac{\omega_{11}}{n} + \omega_{00}\right)(nC_1 + C_2)/C$$

$$\frac{\partial Var(\hat{\beta}_{11})}{\partial n} = -\left\{4\left[\left(\frac{r\sigma_u^2}{n^2r^2}\right)\left(\frac{\omega_{11}}{n^2}\right)\right](nC_1 + C_2) + 4\left(\frac{\sigma_e^2}{nr} + \frac{\omega_{11}}{n} + \omega_{00}\right)C_1\right\}/C$$

$$\frac{4\omega_{00}C_1}{C} - \frac{4\sigma_e^2C_2}{n^2rC} + \frac{4\omega_{11}C_2}{n^2C}$$

$$\frac{\omega_{00}C_1n^2rC}{C} = \sigma_e^2C_2 + \omega_{11}C_2r$$

$$n = \left|\frac{(\sigma_e^2 + \omega_{11}r)C_2}{\omega_{00}rC_1}\right|^{1/2}$$

Whilst differentiating the with respect to *n* and equalling to zero, we have:

$$Var(\hat{\beta}_{11}) = 8\left(\frac{\sigma_e^2}{n} + \omega_{11}\right)(nC_1 + C_2)/C$$

$$\frac{\partial Var(\hat{\beta}_{11})}{\partial n} = -\left\{8\left[\left(\frac{\sigma_u^2}{n^2r^2}\right)\right](nC_1 + C_2) + 4\left(\frac{\sigma_e^2}{n} + \omega_{11}\right)C_1\right\}/C$$

$$\frac{8\omega_{11}C_1}{C} = \frac{8\sigma_e^2C_2}{n^2C}$$

$$n = \left[\frac{\sigma_e^2C_2}{\omega_{11}C_1}\right]^{1/2}$$

REFERENCES

Aitkin, M. A., & Longford, N. T. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A, 149*, 1-43.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*, 9-25.

Brown, H., & Prescott, R. (1999). *Applied mixed models in medicine*. New York: Wiley.

Carvajal, S. C., Baumler, E., Harrist, R. B., & Parcel, G. S. (2001). Multinivel models and unbiased tests for group based interventions: Examples from the safer choices study. *Multivariate Behavioral Research, 36(2)*, 185-205.

Cochran, W. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

D'Amico, E. J., Neilands, T. B., & Zambarano, R. (2001). Power analysis for multivariate and repeated measures designs: A flexible approach using SPSS MANOVA procedure. *Behavior Research Methods, Instruments, and Computer, 33*, 479-484.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Ser. B, 39*, 1-38.

Goldstein, H. (1995). *Multilevel Statistical Models* (2nd ed.). London: Edward Arnold.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association, 72*, 320-338

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated–measures models with structured covariance matrices. *Biometrics, 42*, 805-820.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*, 983-997.

Lair, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*, 963-974.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13-22.

Lindstrom, M. J., & Bates, D. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association, 83*, 1014-1022.

McCulloch, C.E., & Searle, R. S. (2001). *Generalized, linear, and mixed models*. New York: Wiley.

Muller, K. E., La Vange, L. M., Ramey, S. L., & Ramey, C. T. (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association, 87*, 1209-1226.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.

Oliver, J. C., Rosel, J., & Jara, P. (2000). Modelos de regresión multinivel: Aplicación en psicología escolar. *Psicothema, 12*, 487-494.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173-185.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data*. Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials.

Psychological Methods, 5, 199-213.

Rinndskopf, D., & Saxe, L., (1998). Zero effects in substance abuse programs: Avoiding false positives and false negatives in the evaluation of community-based programs. *Evaluation Review, 22*, 78-94

SAS Institute (2001). *SAS/STAT software: Version 8.2 (TS M0)*. Cary, NC: SAS Institute.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.

Shadish, W. R., Cook, T. D., & Campbell, D. J. (2002). *Experimental and quasi experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Singer, D. J. (2002). Fitting individual growth models using SAS PROG MIXED. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data: Methods and applications* (pp.135-170). Mahwah, NJ: Erlbaum.

Vallace, D., & Green, B. S. (2002). Analysis of repeated measures designs with linear mixed models. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data: Methods and applications* (pp.135-170). Mahwah, NJ: Erlbaum.

Van der Leeden, R. (1998). Multilevel analysis of longitudinal data. In C. H. J. Bijleveld & L. J. Th. van der Kamp (Eds.), *Longitudinal data analysis: Designs, models and methods* (pp. 269-317). London: Sage.

Wolfinger, R. D. (1996). Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics, 1*, 205-230

Wolfinger, R., & O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation, 48,* 233-243.