Published in final edited form as:
J Am Stat Assoc. 1992 December 1; 87(420): 1209–1226. doi:10.1080/01621459.1992.10476281.

Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications

Keith E. Muller [Associate Professor],
Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599

Lisa M. LaVange [Principal Scientist],
Center for Medical, Environmental, and Energy Statistics, Research Triangle Institute, Research Triangle Park, NC 27709, and Adjunct Assistant Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599. This work was conducted while she was Head, Design and Statistics Unit, Frank Porter Graham Child Development Center, University of North Carolina, Chapel Hill, NC

Sharon Landesman Ramey [Director], and
Civitan International Research Center, University of Alabama at Birmingham, UAB Station, Birmingham, AL 35294

Craig T. Ramey [Director]
Civitan International Research Center, University of Alabama at Birmingham, UAB Station, Birmingham, AL 35294

Abstract

Recently developed methods for power analysis expand the options available for study design. We demonstrate how easily the methods can be applied by (1) reviewing their formulation and (2) describing their application in the preparation of a particular grant proposal. The focus is a complex but ubiquitous setting: repeated measures in a longitudinal study. Describing the development of the research proposal allows demonstrating the steps needed to conduct an effective power analysis. Discussion of the example also highlights issues that typically must be considered in designing a study. First, we discuss the motivation for using detailed power calculations, focusing on multivariate methods in particular. Second, we survey available methods for the general linear multivariate model (GLMM) with Gaussian errors and recommend those based on *F* approximations. The treatment includes coverage of the multivariate and univariate approaches to repeated measures, MANOVA, ANOVA, multivariate regression, and univariate regression. Third, we describe the design of the power analysis for the example, a longitudinal study of a child's intellectual performance as a function of mother's estimated verbal intelligence. Fourth, we present the results of the power calculations. Fifth, we evaluate the tradeoffs in using reduced designs and tests to simplify power calculations. Finally, we discuss the benefits and costs of power analysis in the practice of statistics. We make three recommendations:

Correspondence to: Craig T. Ramey.

1. Align the design and hypothesis of the power analysis with the planned data analysis, as best as practical.
2. Embed any power analysis in a defensible sensitivity analysis.
3. Have the extent of the power analysis reflect the ethical, scientific, and monetary costs.

We conclude that power analysis catalyzes the interaction of statisticians and subject matter specialists. Using the recent advances for power analysis in linear models can further invigorate the interaction.

Keywords

Analysis of variance; Multivariate linear models; Noncentral distribution; Repeated measures; Sample size determination

1. MOTIVATION

1.1 What Is the Best Power Analysis?

Helping design and plan research constitutes an important activity for many statisticians. If the planned analysis includes hypothesis testing, then power analysis may be used to help choose the design and testing strategy (see, for example, Cohen 1977; Kraemer and Thiemann 1987; Lipsey 1990; Muller, Barton, and Benignus 1984). Although multivariate models are widely used for data analysis, corresponding power methods are not. Recent work (Muller and Barton 1989, 1991; Muller and Peterson 1984) has made power calculations for the general linear multivariate model (GLMM) convenient and readily available. Consequently, the data analyst now has a broader range of tools with which to create the best power analysis to use in designing the best study.

The power of a repeated measures or other multivariate design can be approximated by the power of a reduced design and/or a reduced test. For example, consider a study involving repeated measurements across time on members of four treatment groups. The power analysis might be based on a test of group differences at the last time point, even though the research hypothesis involves changes across time. In considering the use of such an approximation, the question arises as to whether the power analysis and study design are adequately aligned.

Misalignment of the power analysis and data analysis can lead to either of the two possible mistakes in choosing a sample size. Selecting an insufficient sample size yields a study with inadequate sensitivity, whereas selecting an excessive sample size wastes resources. Using approximations of the design and test can lead to overestimating or underestimating power computed from methods for the appropriate design and test. The potential for misestimating power is illustrated in the example (see Section 5.2). Misalignment of power and data analysis may provide an example of what Kimball (1957) termed a type III error, which consists of providing the right answer to the wrong question. For example, with an ANOVA planned, if computing the power of a *t* test in lieu of the power of the ANOVA test provides a substantially incorrect sample size, then a type III error has been committed.

Practical considerations constrain power analysis, as they do all aspects of study design. For example, informed guesses may be the best available estimates of parameters needed for the power analysis. Conducting a sensitivity analysis by evaluating power for a plausible range of estimates contributes substantially to providing plausible cower calculations. A second example of a practical constraint is a design feature, such as the number of groups, which the subject matter specialist requires to be fixed at a particular level. For nonrandom determinants of power, such as the number of subjects or the number of groups, power analysis can at least allow the statistician to warn coinvestigators about inadequate or excessive power. A third example of a practical constraint is that the cost of a power analysis should be proportional to the projected cost and importance of the study.

In general the choice of the appropriate power analysis depends on the choice of the research question, measurement procedures, design, and analysis plan. Practical considerations may affect any of these, as well as the conduct of the power analysis itself. In light of these issues, we suggest that the best power analysis has many of the features of a sound data analysis. First, it answers the right question. Second, it yields a credible answer. Third, it is cost-effective in reflecting the monetary and ethical trade-offs inherent in any study. All aspects must be aligned in power analysis, as in data analysis, to avoid a type III error.

1.2 An Example: Planning a Longitudinal Study

As part of a program project grant submission, investigators at the Frank Porter Graham Child Development Center at the University of North Carolina proposed a prospective study examining maternally linked intergenerational retardation. A probability sample of births at high risk for mental retardation comprised the proposed study sample. The study design called for a three-year follow-up of the children and their mothers, with assessments of intellectual development, health status, social ecology, and home environment during the first three yeas of life. The two primary goals of the study were (1) to measure the incidence of mental retardation in this high-risk subgroup and (2) to compare intellectual developmental trajectories of children born to mothers of varying levels of competence.

The choice of sample size for the project clearly should depend on two key properties of the study. First, the incidence of retardation in the target population must be estimated with adequate precision. Second, the power of detecting differences in developmental change during the first three years of life among children born to mothers of different competence levels must be high. In this article we consider only the latter, although both properties were taken into account in the proposal. It happens in this case that a design that allows meeting the second goal also allows meeting the first goal.

The problem, as first posed by the child developmentalists, was to determine the sample size required to detect a difference in child intelligence quotient (IQ) at age 36 months in children born to mothers categorized into one of four groups: retarded (IQ < 70), borderline (IQ 70–85), low average (IQ 85–100), and above-average (IQ > 100) competence. Child IQ, measured with standardized tests, was to be used to evaluate a child’s intellectual performance. Mother’s IQ would be estimated by a standardized achievement test. Data from previous research would allow determining an estimate of the relative sizes of the four groups in the target population, the sizes of the hypothesized differences, and an estimate of common variance. Given this information, power calculations for testing a main effect in an ANOVA model are widely available from approximate formulas (Kupper and Hafner 1989), approximate tables (Cohen 1977), and commercial software packages (Goldstein 1989).

During further discussion we discovered that the question about group differences was merely a surrogate for the real hypothesis of interest. Experience with other data led the child developmentalists to propose that the amount of a child's retardation varies continuously with the mother's IQ. Furthermore, the regression function relating child IQ to mother IQ was thought to vary during development. Such variation would reflect the expected cumulative effect of advantageous or disadvantageous experiences. Hence a more accurate depiction of the hypothesized relationship would be a series of nonparallel growth curves modeling child's competence as a function of mother's competence and time. Given the preference for measuring all children at the same ages, we concluded that a repeated measures model (a special case of the GLMM) provides the best framework for such a study. In turn, the study hypothesis could be formally cast as a test of a time × mother's IQ interaction in predicting child IQ, in the repeated measures model.

Given this statement of the study hypothesis, a more appropriate power analysis could be conducted in the repeated measures model setting. This approach is more appealing for several reasons. First, the power analysis mirrors the data analysis to be performed for the study, thereby providing a closer match between the statistics and the science. With continuous IQ measurements on both mothers and children, it is unlikely that investigators would choose to fit ANOVA models to groups of mothers defined according to competence. Such an approach would clearly be associated with a loss of information. Rather than hypothesize group boundaries at which the relationship between mother and child competence changes, fitting a repeated measures GLMM allows the analyst to pinpoint the range of mother's competence in which the decline in child competence becomes severe. Second, interest lies in the developmental trajectory during the first three years of life as well as in the 36-month endpoint. A cross-sectional analysis would be useful only in differentiating groups with respect to the endpoint and would not allow for a test of the change over time in child competence. Testing for effects in a repeated measures model allows testing hypotheses concerning the trajectory. Third, such an analysis should provide considerably more accurate power for the study hypotheses than would the corresponding cross-sectional analysis.

2. STATISTICAL METHODS FOR POWER CALCULATIONS

2.1 Statement of the Model and Hypothesis

Power analysis requires paying attention to distinctions that often do not matter in data analysis. One such issue is the distinction between regression and correlation. For the univariate case see Neter, Wasserman, and Kutner (1989, Sec. 3.7 and Chap. 14) for a discussion of the distinction in data analysis. Gatsonis and Sampson (1989) and Sampson (1974) provided discussions of the distinction in univariate power analysis. For the more general multivariate case the distinctions can be made clear by grouping the traditional methods for the analysis of linear relationships into those for correlation, regression, or ANOVA/MANOVA. In the correlation setting the relationships between two sets of

Gaussian variables are of interest. In the regression setting one set of variables, the responses, are assumed to follow a Gaussian distribution, and the other set of variables, the predictors, are assumed to be fixed and known constants. From this perspective, fixed-effect ANOVA is a special case of regression. Any random-effect model can be recast as a fixed-effect model with a complex response covariance pattern. For the purposes of this discussion, some simple random-effect ANOVA models can be ignored, because when recast they can be treated with the methods discussed in this section. Unfortunately, many interesting models still have limited distributional results available. See Section 2.7 for further discussion.

For the purposes of estimation, testing, and prediction (data analysis), the traditional linear model results coincide for correlation and regression under two conditions:

1. The distribution of the responses, conditional on the predictors, follows an appropriate Gaussian distribution.
2. The predictors are independent random variables whose distribution functions do not involve the parameters of the distribution of the responses.

For power analysis the results for regression and correlation do not coincide, except asymptotically. Gatsonis and Sampson (1989) and Sampson (1974) provided computational methods and tables for exact (small sample) power calculations for a test of the correlation of one variable with many variables. No exact results are available for the multivariate correlation case.

The following discussion of power emanates from the regression setting and also applies to fixed-effect ANOVA/MANOVA. The equivalence of the null hypotheses of zero slopes and zero correlation is exploited in the discussion of both the null and nonnull settings. The power analysis methods apply directly to any regression model. The methods apply to the correlation setting only if all conclusions are conditioned on the particular realization of predictor values studied, or if the sample size is sufficiently large. In the correlation setting the results of Gatsonis and Sampson should be used for the univariate case. The lack of power methods for multivariate correlation settings lead us to suggest using the regression methods described in this section, while attempting to accurately reflect the predictor distribution in the design matrix and being especially suspect of results in very small samples. Some reassurance may be gained by comparing univariate special cases against the results of Gatsonis and Sampson. We follow this strategy in the example (see Section 5.2). Our results are consistent with the conclusion of Gatsonis and Sampson that only a small bias is introduced in this fashion, even for modest sample sizes.

In this section we review and recommend methods for power calculations for the GLMM with Gaussian errors. Methods for repeated measures analysis, required for the child development example, are a special case. Davidson (1972) and O’Brien and Kaiser (1985) provided brief overviews of repeated measures data analysis in a GLMM setting. More extensive treatments can be found in Arnold (1981), Kshirsagar (1972), and Timm (1975). Hocking (1985) and Searle (1971) provided thorough treatment of estimation, hypothesis testing, and power calculation for the univariate general linear model. Muller and Barton (1989, 1991) and Muller and Peterson (1984) provided detailed presentations of the univariate approach to repeated measures and of power calculations for the GLMM. O’Brien and Muller (1992) provided a tutorial on power, aimed at behavioral scientists, for linear models settings ranging from *t* tests through multivariate tests.

To provide a more concise report, we first state formally the GLMM and associated general linear hypothesis. For *N* sampling units, *p* responses, and *q* predictors, the usual GLMM is

$$\underset{(N\times p)}{\boldsymbol{Y}} = \underset{(N\times q\times p)}{\mathbf{XB}} + \underset{(N\times p)}{\mathbf{E}}, \quad (2.1.1)$$

with each row iid,

$$\text{row}_i(\mathbf{E}) \overset{d}{=} N_p(\mathbf{0},\boldsymbol{\Sigma}). \quad (2.1.2)$$

For a repeated measures model with one within-subject factor, *p* is the number of repeated measures, whereas with *w* within-subject factors $p=\prod_{f=1}^{w}p_f$. (Here P_f is the number of levels for within-subject factor *f*.) The usual null hypothesis in the multivariate model involves the secondary parameter $\boldsymbol{\Theta}$ = **CBU**:

$$H_0\colon \underset{a\times b}{\boldsymbol{\Theta}}=\boldsymbol{\Theta}_{\mathbf{0}}. \quad (2.1.3)$$

Recognizing certain properties simplifies discussion of the general linear hypothesis and associated power calculations. Each row of **C** defines a row of $\boldsymbol{\Theta}$ and corresponds to a contrast among predictors, often referred to as a between-subject contrast. Each column of **U** defines a column of $\boldsymbol{\Theta}$ and corresponds to a transformation of the responses, often referred to as a within-subject contrast. The transformed responses may be written as $\mathbf{Y}_* = \mathbf{YU}$, with $\mathcal{E}(\mathbf{Y}_*) = \mathbf{M}$, $\text{row}_i\,(\mathbf{Y}_*)' \;\underline{\underline{d}}\; N_b(\text{row}_i\,(\mathbf{M})', \boldsymbol{\Sigma}_*)$, with row *i* independent of row i', $i \neq i'$. The covariance matrix among response contrasts (which are within-subject) is

$$\boldsymbol{\Sigma}_*=\mathbf{U}'\boldsymbol{\Sigma}\mathbf{U}. \quad (2.1.4)$$

2.2 Test Statistics Under the Null Hypothesis

Tests of the general linear hypothesis (2.1.3) are based on

$$\ddot{\mathbf{B}}=(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\boldsymbol{Y}, \quad (2.2.1)$$

$$\hat{\boldsymbol{\Theta}}=\mathbf{C}\ddot{\mathbf{B}}\mathbf{U}, \quad (2.2.2)$$

$$\mathbf{H}=(\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0)'\left[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}'\right]^{-1}(\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0), \quad (2.2.3)$$

$$\mathbf{E}=\mathbf{U}'\hat{\boldsymbol{\Sigma}}\mathbf{U}\cdot(N-r)=\hat{\boldsymbol{\Sigma}}_*\cdot(N-r), \quad (2.2.4)$$

and

$$\mathbf{T}=\mathbf{H}+\mathbf{E}, \quad (2.2.5)$$

with r = rank($\mathbf{X}$). **E, H**, and **T** follow Wishart distributions, based on $\mathbf{\Sigma}_*$. **E** follows a central Wishart with $(N - r)$ df. Under the null hypothesis, **H** follows a central Wishart (if $b \le a$) or pseudo-Wishart (if $b > a$), with a df.

Let $s = \min(a, b)$ for $\mathbf{\Theta}$ of dimension $(a \times b)$. The usual multivariate test statistics can be expressed as functions of the eigenvalues of $\mathbf{HE}^{-1}$, of which at most s are nonzero. Furthermore, the set of eigenvalues are minimal sufficient statistics for the hypothesis. For purposes of explanation, it is preferable to express the statistics in terms of the eigenvalues of $\mathbf{HT}^{-1}$, which are the generalized, squared canonical correlations, $\{\hat{\rho}_k^2\}$, and one-to-one functions of the eigenvalues of $\mathbf{HE}^{-1}$. By expressing the multivariate test statistics as functions of these correlations, we can explain multivariate power analysis in terms of the corresponding univariate results.

For the univariate case, $b = 1$ and an optimal test of overall relationship in the model can be computed from the generalized squared multiple correlation, $\hat{\rho}^2 = \mathbf{HT}^{-1}$, which is the ratio of the sum of squares due to the hypothesis to the total sum of squares. The generalized correlation reduces to an ordinary correlation if the model spans an intercept and the hypothesis does not. In the univariate model only one test statistic is usually considered, namely,

$$F_{\text{obs}}=\frac{\hat{\rho}^2/a}{(1-\hat{\rho}^2)/(N-r)}. \quad (2.2.6)$$

This F statistic provides a test that is optimal in the sense that it is (1) the likelihood-ratio statistic, (2) the union-intersection principle statistic, and (3) the uniformly most powerful test of size α(UMP-α).

In multivariate models several candidate tests exist, and four are in common use: (1) Roy’s largest root (RLR), (2) Wilks’ likelihood ratio statistic (W), (3) Pillai–Bartlett trace (PB), and (4) Hotelling–Lawley trace (HLT). Table 1 gives the formulas for computing each of the four statistics, the criterion optimized by each, and a multivariate measure of association consonant with the statistic. All four provide a size α test (under the GLMM assumptions), but are equivalent (one-to-one functions) only if $s = \min(a, b) = 1$.

Although the statistics are simple to compute, the associated p values are not. No general, exact formulas are available for the distribution functions under the null hypothesis. But approximations are available that are both convenient and sufficiently accurate. Muller and Peterson (1984) briefly reviewed the approximations and recommended those based on a single F distribution for all but RLR.

The single F approximations for W, PB, and HLT can be expressed as functions of the corresponding measures of multivariate association (defined in Table 1). Recall that $s = \min(a, b)$ for $\mathbf{\Theta}$ of dimension $(a \times b)$. All three use (ab) as the numerator df, and denominator df of $df_2(\text{W}) = g[(N - r) - (b - a + 1)/2] - (ab - 2)/2$, $df_2(\text{PB}) = s[(N - r) - b +$

s], and df_2(HLT) = $s[(N - r) - b - 1] + 2$. Here $g = [(a^2b^2 - 4)/(a^2 + b^2 - 5)]^{1/2}$. Using $\boldsymbol{\eta}_m$ to indicate the measure of multivariate association for $m \in$ {W, PB, HLT}, the transformation to an approximate F is

$$F_{\text{obs}}(m)=\frac{\hat{\eta}_m/(ab)}{(1-\hat{\eta}_m)/df_2(m)}. \quad (2.2.7)$$

The univariate approach to repeated measures analysis is a simple by-product of the multivariate analysis (see Muller and Barton 1989, 1991; Wang 1983), as long as no repeated covariates are included. One additional assumption leads to a single test statistic being the optimal choice: The covariance matrix, $\boldsymbol{\Sigma}_*$, must be proportional to an identity matrix. The assumption is usually referred to as “sphericity,” because the transformed scores are required to be spherical normal, with covariance $\boldsymbol{\Sigma}_*=\sigma_*^2\mathbf{I}_b$. It is sufficient, but not necessary, to assume $\boldsymbol{\Sigma} = [\mathbf{11}'\rho + (1 - \rho)\mathbf{I}]\sigma^2$, which is an assumption of compound symmetry (of $\boldsymbol{\Sigma}$, not $\boldsymbol{\Sigma}_*$). Having assumed compound symmetry, the correct test statistic may be computed by requiring that $\mathbf{U}$ be an orthonormal matrix (or proportional to one), $b < p$, and if $b > 1$ then $\mathbf{U}'\mathbf{1} = \mathbf{0}$. If sphericity holds, then a size α test of H_0: $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0$ is provided by

$$F_{\text{obs}}(\text{REP})=\frac{\text{tr}(\mathbf{H})/(ab)}{\text{tr}(\mathbf{E})/[b(N-r)]}. \quad (2.2.8)$$

Here REP is the name of the statistic tr($\mathbf{H}$)/tr($\mathbf{E}$) (see Table 1). With all assumptions met, $F_{\text{obs}}(\text{REP}) \underline{d}\ F[ab, b(N - r)]$. With all assumptions met *except* sphericity, Box (1954a, b) suggested

$$F_{\text{obs}}(\text{REP}) \overset{d}{\approx} F[ab\boldsymbol{\varepsilon}, b(N-r)\boldsymbol{\varepsilon}], \quad (2.2.9)$$

with

$$\boldsymbol{\varepsilon}=\frac{\text{tr}^2(\boldsymbol{\Sigma}_*)}{b\text{tr}(\boldsymbol{\Sigma}_*^2)}=\left(\sum_{k=1}^{b}\lambda_k\right)^2 \Big/ \left(b\cdot\sum_{k=1}^{b}\lambda_k^2\right) \quad (2.2.10)$$

and λ_k, $k \in \{1, 2, \ldots, b\}$ being the ordered eigenvalues of $\boldsymbol{\Sigma}_*$. It is easy to prove that $l/b \le \boldsymbol{\varepsilon} \le 1$, with sphericity corresponding to the upper bound. Estimating $\boldsymbol{\varepsilon}$ in (2.2.9) by its lower bound leads to a conservative test, sometimes known as the Box test. Using the maximum likelihood estimate of $\boldsymbol{\varepsilon}$, (which may be computed by replacing $\boldsymbol{\Sigma}_*$ with $\hat{\boldsymbol{\Sigma}}_*$, or λ_k with $\hat{\lambda}_k$ in (2.2.10), leads to the Geisser–Greenhouse test (Geisser and Greenhouse 1958; Greenhouse and Geisser 1959).

2.3 Distribution Theory for the Alternative Hypothesis

For $b = 1$, which includes univariate models as a special case, the test statistic follows a noncentral F distribution, $F_{\text{obs}} \underline{\underline{d}}\ F(a, N - r, \omega)$. The noncentrality parameter for $b = 1$, ω, may be expressed as

$$\omega=(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}']^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)/\sigma^2 \quad (2.3.1)$$

$$=a \cdot F_{\mathrm{A}}. \quad (2.3.2)$$

Here F_A denotes the F value obtained if one were to observe exactly $\boldsymbol{\theta}$ and $\sigma_*^2=\mathbf{u}'\boldsymbol{\Sigma}\mathbf{u}$, the population values under H_A. F_A differs from F_{obs} in being a constant (a parameter) rather than a random variable. O'Brien (1982, who cited Graybill 1976) and O'Brien and Lohr (1984) recommended adopting the formulation in (2.3.2) to help demystify the noncentrality parameter.

Closed-form expressions are not available for distributions of the multivariate test statistics under H_A in the most general case ($s > 1$). Despite this problem, all candidate statistics can be expressed in terms of **H, E**, and the non-centrality matrix. Under the alternative hypothesis, **H** follows a non-central Wishart (if $b \le a$) or a noncentral pseudo-Wishart (if $b > a$), with df parameter a and ($b \times b$) noncentrality matrix

$$\boldsymbol{\Omega}=(\boldsymbol{\Theta}-\boldsymbol{\Theta}_0)'[\mathbf{C}(\mathbf{X}'\mathbf{X})^-\mathbf{C}']^{-1}(\boldsymbol{\Theta}-\boldsymbol{\Theta}_0)\boldsymbol{\Sigma}_*^{-1}. \quad (2.3.3)$$

Note that if sample values of means and covariances are equal to population values, then $\boldsymbol{\Omega} = \mathbf{HE}^{-1} \cdot (N - r)$. Consequently, the noncentrality matrix is a function of (1) the true difference, $\boldsymbol{\Theta}$; (2) the variance structure, $\boldsymbol{\Sigma}_*$; and (3) the sample size, $(N - r)$. The eigenvalues of $\boldsymbol{\Omega}$, $\{\omega_k, k = 1, 2, \ldots, s\}$, are the additional statistics needed to provide a minimal sufficient set. As with the eigenvalues of $\mathbf{HE}^{-1}$, at most $s = \min(a, b)$ are nonzero. Necessarily, rank($\boldsymbol{\Omega}$) ≤ rank($\mathbf{HE}^{-1}$).

The properties of the noncentral distributions of the test statistics for the GLMM depend on the parameters of the model. In turn, anything that affects the noncentral distributions also affects the power approximation discussed in this section. As the rank of $\boldsymbol{\Omega}$ decreases, the difference between the noncentral and central distributions of the test statistic decreases, and the accuracy of the power approximations increase. In practice, the data analyst usually chooses small rank ($\boldsymbol{\Omega}$), which helps improve the accuracy of power calculations. Properties of the noncentral distributions also vary substantially as a function of the dimensions of $\boldsymbol{\Theta}$, the ($a \times b$) matrix of secondary parameters, which embodies the general linear hypothesis being tested. The properties do not vary as a function of the dimensions of **B**, the ($p \times q$) matrix of primary parameters, which embodies the design.

Some special cases deserve mention. When $b = 1$, all four multivariate statistics, plus the univariate repeated measures statistic, can be expressed as the same exact noncentral F random variable and provide a UMP-α test. For $b > 1$ and $a = 1$, all four multivariate statistics (1) can be transformed exactly to a noncentral F with parameters different than those for $b = 1$, (2) are one-to-one functions of each other, and (3) provide the UMP-α test for unstructured $\boldsymbol{\Sigma}_*$. For $b > 1$, $a = 1$, and spherical $\boldsymbol{\Sigma}_*$, the REP statistic provides the UMP-α test. For $s = \min(a, b) > 1$ and unstructured $\boldsymbol{\Sigma}_*$, none is UMP-α.

2.4 Survey of Power Approximations for the GLMM

Muller and Peterson (1984) reviewed approximations previously available for noncentral distribution functions of multivariate test statistics. The approximations all involve computing pages of coefficients based on tr(Ω), tr(Ω^2), and so forth. For example, Sugiura and Fujikoshi (1969) provided an asymptotically correct χ^2 mixture approximation for HLT, and Lee (1971) provided the same for W, PB, and HLT. The approximations gave three digits of accuracy for power values in a small set of examples.

Muller and Peterson (1984) also presented new approximations for W, PB, and HLT based on single noncentral *F* random variables, without any analytic results pertaining to numerical accuracy. For the examples considered by the earlier authors, the methods appeared to provide nearly two digits of accuracy of power values. Barton and Cramer (1989) evaluated the methods via Monte Carlo studies, incidentally to developing missing data methods. Their extensive simulations provided strong support for the accuracy claims of Muller and Peterson.

Kulp and Nagarsenker (1984) reported an approximation for W, based on a mixture of noncentral *F* random variables. As for the other mixture approximations, the method requires substantial computation for each power evaluation. The Muller and Peterson method for W is not a special case.

For the univariate approach to repeated measures (and all assumptions met), F_{obs}(REP) follows a noncentral *F* distribution exactly. For situations without sphericity, Muller and Barton (1989, 1991) presented methods for approximating power. Their methods, based on single noncentral *F* random variables and an asymptotic approximation to the expected value of the **ε** estimator, appear to provide approximately two digits of accuracy in power values.

No accurate approximation based on a single *F* is available for RLR. Furthermore, as discussed in Section 2.7, simulation results indicate that it is relatively less robust (Olson 1974, 1976, 1979). Hence in the remainder of the article, RLR will not be considered.

In our opinion two digits of accuracy in power values is adequate for planning a study. Hence we use and recommend the power computation methods based on single noncentral *F* random variables (described in sec. 2.5) due to their advantages. First, they are easy to program. (See Appendix A for directions to obtain a free copy of the software used for this article.) Second, the approximations are short and non-iterative, which allows them to be evaluated many times for plotting or computing an inverse solution (e.g., to solve for an unknown sample size, given a power). Third, when applied to special cases they provide exact results without additional effort. Fourth, they are easy to understand and explain to nonstatisticians.

2.5 Implementing Single *F* Approximations to Noncentral Distributions

Computing approximate power for the various tests of the multivariate general linear hypothesis can be most easily understood as a generalization of computing power for the univariate general linear hypothesis. The method for a univariate hypothesis can be reduced to the following four steps:

1. Specify α, σ^2, $\mathbf{X}$, $\boldsymbol{\beta}$, $\mathbf{C}$, and θ_0.

2. Find the critical value from an inverse (central) *F* distribution function, say

$$F_{\text{crit}}{=}\text{FINV}(1-\alpha,a,N-r).$$

3. Compute the noncentrality parameter, ω, as denned by (2.3.1).

4. Compute power with a noncentral *F* distribution function as

$$\text{Power}{=}1-\text{FPROB}(F_{\text{crit}},a,N-r,\omega){=}1-\text{FPROB}(F_{\text{crit}},a,N-r,a\cdot F_{\text{A}}).$$

   FINV(*p*, df_1, df_2) represents the value of an *F* statistic based on df_1 numerator and df_2 denominator degrees of freedom such that Pr{$F \leq F_{\text{crit}}$} = *p*. Furthermore, FPROB(*f*, df_1, df_2, *nc*) represents the noncentral *F* distribution function, namely Pr{$F \leq f$}, for a noncentral *F* statistic based on df_1 numerator, df_2 denominator degrees of freedom, and non-centrality parameter *nc*.

Some additional results are needed to generalize to the multivariate case. Here *m* stands in for W, HLT, or PB. Define

$$F_{\text{A}}(m){=}\frac{\boldsymbol{\eta}_m/(ab)}{(1-\boldsymbol{\eta}_m)/df_2(m)},\quad (2.5.1)$$

which is the *F* value that would arise if H_{A} were observed; that is, if $\hat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$.

Muller and Peterson (1984) suggested that under H_{A}, $F_{\text{obs}}(m)\overset{d}{\approx}F[ab,df_2(m),\omega_m]$, with

$$\omega_m{=}(ab)\cdot F_{\text{A}}(m){=}\frac{\boldsymbol{\eta}_m}{(1-\boldsymbol{\eta}_m)/df_2(m)}.\quad (2.5.2)$$

This approach generalizes the one recommended by O’Brien (1982) for the univariate case.

The four steps for computing power approximations for the various tests of the multivariate general linear hypothesis are the same as for computing power for the test of a univariate hypothesis:

1. Specify α, **Σ**, **X, B, C, U**, and $\boldsymbol{\Theta}_0$.

2. Find the approximate critical value from an inverse (central) *F* distribution function, say

$$\begin{aligned}&F_{\text{crit}}(m)\approx\text{FINV}[1-\alpha,ab,df_2(m)],\text{with}\\&\quad df_2(\text{W}){=}g[(N-r)-(b-a{+}1)/2]-(ab-2)/2\text{for}\\&\quad g{=}[(a^2b^2-4)/(a^2{+}b^2-5)]^{1/2},\\&\quad df_2(\text{PB}){=}s[(N-r)-b{+}s],\text{or}\\&\quad df_2(\text{HLT}){=}s[(N-r)-b-1]{+}2,\end{aligned}$$

   as defined in Section 2.2 for the null hypothesis.

3. Compute the noncentrality in terms of $F_{\text{A}}(m)$, from (2.5.2), as

$$\omega_{\mathrm{W}}=(ab)\cdot F_{\mathrm{A}}(\mathrm{W})=\frac{\boldsymbol{\eta}_{\mathrm{w}}}{(1-\boldsymbol{\eta}_{\mathrm{w}})/df_2(\mathrm{W})}=\frac{1-\mathrm{W}_{\mathrm{A}}^{1/g}}{\mathrm{W}_{\mathrm{A}}^{1/g}/df_2(\mathrm{W})}, \quad (2.5.3)$$

$$\omega_{\mathrm{PB}}=(ab)\cdot F_{\mathrm{A}}(\mathrm{PB})=\frac{\boldsymbol{\eta}_{\mathrm{PB}}}{(1-\boldsymbol{\eta}_{\mathrm{PB}})/df_2(\mathrm{PB})}=\frac{(\mathrm{PB}_{\mathrm{A}}/s)}{(1-\mathrm{PB}_{\mathrm{A}}/s)(df_2(\mathrm{PB})}, \quad (2.5.4)$$

or

$$\omega_{\mathrm{HLT}}=(ab)\cdot F_{\mathrm{A}}(\mathrm{HLT})=\frac{\boldsymbol{\eta}_{\mathrm{HLT}}}{(1-\boldsymbol{\eta}_{\mathrm{HLT}})/df_2(\mathrm{HLT})}=\frac{\mathrm{HLT}_{\mathrm{A}}/s}{1/df_2(\mathrm{HLT})}. \quad (2.5.5)$$

W_{A} is the value of W that would be observed if $\hat{\mathbf{\Theta}} = \mathbf{\Theta}$ and $\hat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$, and PB_{A} and $\mathrm{HLT}_{\mathrm{A}}$ have parallel definitions for the other two multivariate statistics.

4. Compute approximate power with a noncentral *F* distribution function as

$$\mathrm{Power}(\mathrm{W}) \approx 1-\mathrm{FPROB}[F_{\mathrm{crit}}(\mathrm{W}), ab, df_2(\mathrm{W}), ab\cdot F_{\mathrm{A}}(\mathrm{W})],$$
$$\mathrm{Power}(\mathrm{PB}) \approx 1-\mathrm{FPROB}[F_{\mathrm{crit}}(\mathrm{PB}), ab, df_2(\mathrm{PB}), ab\cdot F_{\mathrm{A}}(\mathrm{PB})],$$

or

$$\mathrm{Power}(\mathrm{HLT}) \approx 1-\mathrm{FPROB}[F_{\mathrm{crit}}(\mathrm{HLT}), ab, df_2(\mathrm{HLT}), ab\cdot F_{\mathrm{A}}(\mathrm{HLT})].$$

The method produces the appropriate exact results for $s = 1$, such as fixed-effect ANOVA, fixed-effect regression, and two-group discriminant analysis, among others. A single general approach provides the same convenience in power analysis as the GLMM does in data analysis.

Power approximations for the univariate approach to repeated measures are also straightforward. If sphericity holds, then the exact result under H_{A} is that

$$F_{\mathrm{obs}}(\mathrm{REP}) \overset{d}{=} F[ab, b(N-r), ab\cdot F_{\mathrm{A}}(\mathrm{REP})]. \quad (2.5.6)$$

The noncentrality for the uncorrected test is

$$\omega_{\mathrm{REP}}=(ab)\cdot F_{\mathrm{A}}(\mathrm{REP}) \quad (2.5.7)$$

$$=\frac{\boldsymbol{\eta}_{\mathrm{REP}}}{(1-\boldsymbol{\eta}_{\mathrm{REP}})/df_2(\mathrm{REP})}=\frac{\mathrm{REP}_{\mathrm{A}}}{1/df_2(\mathrm{REP})}. \quad (2.5.8)$$

If sphericity does not hold (see Muller and Barton 1989, 1991), then an approximate result under H_{A} is

$$F_{\mathrm{obs}}(\mathrm{REP}) \overset{d}{\approx} F[ab\boldsymbol{\varepsilon}, b(N-r)\boldsymbol{\varepsilon}, ab\cdot F_{\mathrm{A}}(\mathrm{REP})\boldsymbol{\varepsilon}]. \quad (2.5.9)$$

Without sphericity, the noncentrality parameter is multiplied by $\boldsymbol{\varepsilon}$. Because sphericity corresponds to $\boldsymbol{\varepsilon} = 1$, (2.5.7) is unambiguously a special case of

$$\omega_{\mathrm{REP}} = (ab) \cdot F_{\mathrm{A}}(\mathrm{REP})\boldsymbol{\varepsilon}. \quad (2.5.10)$$

The steps for computing power approximations for the various tests arising from the univariate approach to repeated measures are the same as those for the general linear hypothesis and for the univariate hypothesis:

1. Specify α, $\boldsymbol{\Sigma}$, $\mathbf{X}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{U}$, and $\boldsymbol{\Theta}_0$.

2. Find the approximate critical value from an inverse (central) F distribution function, namely

$$\begin{aligned}
&F_{\mathrm{crit}}(\mathrm{REP}) \approx \mathrm{FINV}[1-\alpha, ab, b(N-r)] \text{for the uncorrected test,}\\
&\quad F_{\mathrm{crit}}(\mathrm{HF}, \mathscr{E}\tilde{\varepsilon}) \approx \mathrm{FINV}[1-\alpha, (ab)\delta\tilde{\varepsilon}, b(N-r)\delta\tilde{\varepsilon}] \text{for the Huynh–Feldt test,}\\
&\quad F_{\mathrm{crit}}(\mathrm{GG}, \mathscr{E}\hat{\varepsilon}) \approx \mathrm{FINV}[1-\alpha, (ab)\mathscr{E}\hat{\varepsilon}, b(N-r)\mathscr{E}\hat{\varepsilon}] \text{for the Geisser–Greenhouse test,}
\end{aligned}$$

or

$$F_{\mathrm{crit}}(\mathrm{Box}) \approx \mathrm{FINV}[1-\alpha, a, (N-r)] \text{for the conservative test.}$$

Muller and Barton (1989, 1991) provided practical approximations for $\mathscr{E}(\tilde{\boldsymbol{\varepsilon}})$ and $\mathscr{E}(\hat{\boldsymbol{\varepsilon}})$.

3. Compute the same noncentrality in terms of F_{A} (REP), from (2.5.10), for all four tests.

4. Compute approximate power with a noncentral F distribution function as

$$\begin{aligned}
&\mathrm{Power}(\mathrm{REP}) \approx 1-\mathrm{FPROB}[F_{\mathrm{crit}}(\mathrm{REP}), ab\boldsymbol{\varepsilon}, b(N-r)\boldsymbol{\varepsilon}, ab \cdot F_{\mathrm{A}}(\mathrm{REP})\boldsymbol{\varepsilon}],\\
&\quad \mathrm{Power}(\mathrm{HF}) \approx 1-\mathrm{FPROB}[F_{\mathrm{crit}}(\mathrm{HF}, \mathscr{E}\tilde{\varepsilon}), ab\boldsymbol{\varepsilon}, b(N-r)\boldsymbol{\varepsilon}, ab \cdot F_{\mathrm{A}}(\mathrm{REP})\boldsymbol{\varepsilon}],\\
&\quad \mathrm{Power}(\mathrm{GG}) \approx 1-\mathrm{FPROB}[F_{\mathrm{crit}}(\mathrm{GG}, \mathscr{E}\hat{\varepsilon}), ab\boldsymbol{\varepsilon}, b(N-r)\boldsymbol{\varepsilon}, ab \cdot F_{\mathrm{A}}(\mathrm{REP})\boldsymbol{\varepsilon}],
\end{aligned}$$

or

$$\mathrm{Power}(\mathrm{Box}) \approx 1-\mathrm{FPROB}[F_{\mathrm{crit}}(\mathrm{Box}), ab\boldsymbol{\varepsilon}, b(N-r)\boldsymbol{\varepsilon}, ab \cdot F_{\mathrm{A}}(\mathrm{REP})\boldsymbol{\varepsilon}].$$

Some properties of the results for the univariate approach to repeated measures deserve mention. Notice that the four critical values are ordered from smallest to largest, whereas the four powers (and type I error rates) are ordered from largest to smallest. Also notice that special care must be taken to define the appropriate critical value for a particular application. For the uncorrected test one uses the same critical value, $F_{\mathrm{crit}}(\mathrm{REP}) = \mathrm{FINV}[1 - \alpha, ab, b(N - r)]$, for both data analysis and power analysis. A similar result holds for the conservative test. In contrast, for the Geisser–Greenhouse test one uses a critical value of $F_{\mathrm{crit}}(\mathrm{GG}, \hat{\boldsymbol{\varepsilon}}) = \mathrm{FINV}[1 - \alpha,$

$(ab)\hat{\boldsymbol{\varepsilon}}$, $b(N - r)\hat{\boldsymbol{\varepsilon}}]$ for data analysis and $F_{\text{crit}}(\text{GG}, \mathcal{E}\hat{\boldsymbol{\varepsilon}}) = \text{FINV}[1 - \alpha, (ab)\mathcal{E}\hat{\boldsymbol{\varepsilon}}, b(N - r)\mathcal{E}\hat{\boldsymbol{\varepsilon}}]$ for power analysis. A parallel result holds for the Huynh–Feldt test.

2.6 Some Possible Design Factors in a Power Study

Applying the methods of power analysis to a proposed study requires choosing which factors to vary in the design of the power study (in contrast to the factors in the research study). Some design characteristics that might be included as factors in a power study are the (1) responses, (2) predictors, (3) definition of the target population, (4) response time sampling scheme, and (5) predictor sampling scheme. Ideally, the choice of level for each factor would closely mimic the characteristic expected in the proposed study. Because the information available is necessarily limited, the consequences of various values of most factors will typically be studied.

2.7 Limitations of the Methods

Many kinds of linear models with Gaussian errors are not covered by this treatment. Random effects models (Jennrich and Schluchter 1986) are not covered, except for very special cases such as the univariate approach to repeated measures without repeated covariates. Similarly, designs in which elements of $\mathbf{X}$ are not fixed but are sampled from a Gaussian (or other) distribution are not covered. (See the discussion of Gatsonis and Sampson 1989 and Sampson 1974 in Sec. 2.1). Additional examples of models not covered are those for multiple-design matrix problems (Srivistava 1967), seemingly unrelated regression (Zellner 1962), and time series models. In general, for models more flexible than the GLMM, estimation methods have been demonstrated to be accurate in small samples, whereas standard hypothesis testing methods have been shown to be badly biased in small samples (Freedman and Peters 1984; Rocke 1989). Acceptable nonnull case methods (for models more general than the GLMM) must wait on acceptable null case methods.

As with all parametric methods, robustness of the power calculations is of concern. In practice the statistics are all computed as functions of $\mathbf{H}$, defined in (2.2.3), and $\mathbf{E}$, defined in (2.2.4). The accuracy of the power approximations depends on $\mathbf{H}$ being noncentral Wishart, independent of the central Wishart $\mathbf{E}$, both based on the same covariance matrix, $\boldsymbol{\Sigma}_* = \mathbf{U}'\boldsymbol{\Sigma}\mathbf{U}$. In the current setting it is sufficient for the multivariate tests to consider violation of (1) independence of sampling units, (2) correct model specification (linearity), (3) homogeneity of covariance between sampling units, and (4) Gaussian distribution. For the uncorrected test for the univariate approach to repeated measures, one must also consider violation of sphericity of $\boldsymbol{\Sigma}_*$. Entry to the extensive literature on violation of sphericity can be had by consulting Muller and Barton (1989) and their reference list. Note that multiplying the degrees of freedom by an estimator of $\boldsymbol{\varepsilon}$ constitutes a correction for violation of sphericity.

The remaining four assumptions fall into two groups. The logical properties of the sampling scheme usually allows determining the validity of the independence assumption by inspection. Violations of linearity, homogeneity, and Gaussian distribution usually can be treated together both in diagnosis and cure. The interested reader should consult the work of Olson (1974, 1976, 1979) and comments about Olson's work by Stevens (1979, 1980).

Recall that no single test can be uniformly most powerful (size α), so that the optimal choice depends on the alternative hypothesis. Olson drew the following conclusions from a series of extensive simulations. First, the largest root (RLR) test provides the least protection against violation of homogeneity and Gaussian distribution. Second, the Pillai–Bartlett trace statistic (PB) was the most robust against the same two violations. Third, the relative power of the tests depends on whether the noncentrality is concentrated (one nonzero eigenvalue for **Ω**) or diffused (more than one nonzero eigenvalue of **Ω**, all of equal size). Fourth, the multivariate tests fall into two equivalence classes: (1) the largest root test, powerful for concentrated structures, and (2) the remaining tests, all of which average across the structure and thereby provide good power for diffuse cases. Fifth, balancing robustness and power supports using the PB trace. Stevens (1979) interpreted Olson's results to support choosing the likelihood ratio statistic (W). Olson (1979) reaffirmed his earlier conclusions.

The choice of test statistic depends on (1) the alternative hypothesis, (2) covariance structure, (3) robustness, and (4) personal preference. The power methods discussed here constitute a significant aid in comparing the alternate statistics for various alternative hypotheses and covariance structures. Such a comparison may also include consideration of a series of univariate analyses with a Bonferroni correction (conduct each analysis with nominal type I error rate set to α/p, for one analysis per response). Naturally, such an approach introduces additional uncertainty about the power of the set of tests. Combining knowledge of power differences for various analysis strategies with knowledge of robustness will help the analyst create a good balance between type I and type II error rates.

3. POWER ANALYSIS DESIGN FOR CHILD DEVELOPMENT EXAMPLE

3.1 Basic Research Study Design

As noted previously, the power analysis described here centers on the second goal of the study, which was to compare intellectual developmental trajectories of children born to mothers of varying levels of competence. This was translated into a wish to evaluate the time × mother's IQ interaction in predicting child IQ. This formulation was derived from a first draft of the study design. Practical considerations usually dictate some research study characteristics, which in turn affect the choice of power study factors and factor levels. This was true here.

For the proposed study two disparate catchment areas were chosen to ensure substantial diversity in culture and educational resources. The target population consisted of all live births in the catchment areas during a 12-month recruitment period among mothers whose children would be considered at risk of low competence. Because maternal IQ scores would not be available, a history of receiving special education or less than high school education provides a useful surrogate for low IQ. A systematic sample would then be selected independently in each hospital, with screening on educational history. It was initially thought that mothers would be screened for risk by measuring their IQ prior to sampling for inclusion in the study; however, this was found to be infeasible. Nevertheless, we considered several subject sampling schemes for the power analysis to assess the effect of such a screening process.

Home environment evaluations and behavioral assessments of the children and mothers would be obtained during each of three follow-up visits at specified ages during the first three years of life. Because the primary study hypothesis concerned retardation, child intelligence measurements provided the outcomes of dominant interest. The overwhelming important of intelligence led to restricting the power study to always involve child IQ as the response variable.

Various response time sampling schemes were considered in designing the study. Originally, follow-up visits were to be scheduled at the commonly studied ages of 12, 24, and 36 months. Interest in other time periods, arising from current developmental theory, concern about the validity of the scales used prior to age 18 months, and resource constraints, induced the child developmentalists to consider alternate measurement times. Therefore, the response sampling scheme was treated as a multilevel factor in the power analyses.

3.2 Limitations of the Power Methods for the Example

Two limitations of the power methods should be recognized for the child development example. The first limitation stems from the inability to incorporate any allowance for missing data in the analysis. Two kinds of missing data will be encountered: data missing due to study attrition and single observations missing due to missed visits or measurement problems. The possibility of nonrandom attrition must always be examined during design, data collection, and data analysis of any longitudinal study. Extensive experience in similar studies led the researchers to expect 20% attrition during a three-year study and to expect that the attrition would likely be nearly uniform. Hence the investigators intended to begin data collection on 20% more subjects than would be needed to provide the desired power. More accurate power calculations would be based on methods for data missing at random. See Little and Rubin (1987) for a general discussion of missing data, and Jennrich and Schluchter (1986) for linear models applications. As noted, acceptable approximations are currently not available. Missingness that is not at random presents a source of possible nonrobustness. The second limitation stems from the Gaussian nature of some of the predictors. Gatsonis' and Sampson's results apply only to the univariate case. Hence in Section 4.2 we examine the effect on power of assuming fixed rather than Gaussian predictors for univariate subhypotheses.

3.3 Sources of Data for Parameter Choices

Once the research study is defined, the next step in planning a power analysis is to identify appropriate choices of the model parameters required for power computations. Most analysts use existing data to compute estimates of the parameters. For a GLMM analysis $\mathbf{B}$ and $\boldsymbol{\Sigma}$, the primary parameters of the GLMM, and the choice of $\mathbf{X}$ values must be specified.

Data used to provide choices of parameters should match the study proposal as closely as possible with respect to the power analysis factors previously discussed. Possible sources for our application included previously published results available in the literature, extant data bases available in public use files, and other earlier studies conducted at Frank Porter Graham Child Development Center. Few published sources included population estimates of the frequency of mental retardation, and even fewer treated intergenerational transmission. Two collections of data contained linked measures of mother and child competence.

The National Longitudinal Survey of Labor Force Behavior–Youth Cohort (NLSY) consists of yearly interviews with 12,686 people. A child-mother longitudinal data file exists for public use. Unfortunately, neither the measures nor the design are strictly comparable with those of the proposed study. Despite the problems, we considered these data as the basis for estimating the relative frequencies of mother's IQ values.

Data from untreated children in the Infant Health and Development Program (IHDP) (1990; also see Ramey et al. 1992) proved to be more useful. The program consisted of a trial of an intervention program for low-birth-weight babies, all born at Level III hospitals. The entire study pool of 985 subjects was evaluated for three years on cognitive, health, and behavior performance. Data include mother's IQ and child IQ at age 12, 24, and 36 months. The control group (N = 474) provided a source for estimating the parameters needed for the power calculations. The major drawback is the nonrepresentativeness of the sample. Low-birth-weight babies are considered to be at significantly higher risk for mental retardation than are other babies (McCormick 1985). Overall, however, the IHDP study appeared to be the best candidate for input to the power analysis, in that its characteristics most closely approximated the corresponding primary dimensions of the proposed study. The data analyst often must tolerate substantial mismatch, because if there were a close match there would be no need for a similar study. See Section 3.5 for further discussion of appropriateness of using the IHDP data in the manner described.

3.4 Research Model Formulation

The next step in the planning phase consisted of a formal statement of the model and hypothesis for which power was desired. Discussions among the investigators, described in Section 1, led to proposing a repeated measures model framework for the major study hypothesis. The model is an application of (2.2.1) in which $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2 \cdots \mathbf{Y}_p]$ denotes the matrix of child IQ measurements with rows corresponding to subjects and columns to measurement times, $\mathbf{X}$ denotes the design matrix consisting of polynomials in mother's IQ, and $\mathbf{B}$ denotes the matrix of unknown polynomial regression coefficients. The rows of $\mathbf{B}$ correspond to between-subject effects, here polynomials in mother's IQ, and the columns correspond to time points. For example, with responses at 12, 24, and 36 months (p = 3), and with intercept, linear, quadratic, and cubic trends in mother's IQ as predictors (q = 4), this yields

$$\mathbf{B}=\begin{bmatrix} \beta_{0,12} & \beta_{0,24} & \beta_{0,36} \\ \beta_{\mathrm{L},12} & \beta_{\mathrm{L},24} & \beta_{\mathrm{L},36} \\ \beta_{\mathrm{Q},12} & \beta_{\mathrm{Q},24} & \beta_{\mathrm{Q},36} \\ \beta_{\mathrm{C},12} & \beta_{\mathrm{C},24} & \beta_{\mathrm{C},36} \end{bmatrix}. \quad (3.4.1)$$

For $t \in \{12, 24, 36\}$, $\beta_{0,t}$ is the intercept for time t, whereas $\beta_{\mathrm{L},t}$, $\beta_{\mathrm{Q},t}$, and $\beta_{\mathrm{C},t}$ are the corresponding coefficients of linear, quadratic, and cubic values of the predictor, mother's IQ.

General linear hypotheses of the form **CBU** can be tested via either multivariate or adjusted univariate *F* statistics, as described in Section 2. Given the focus on one hypothesis in one analysis, a type I error rate of .05 was deemed acceptable. The degree of the polynomial trend in mother's IQ that provides the best fit can be determined by specifying an appropriate choice for **C**. With the use of orthogonal polynomial contrasts in **U**, hypotheses about the degree of the trend in child IQ over time can be tested. In turn, the interaction between trends in time and mother's IQ can be tested by combining **C** and **U**. The hypothesized relationship between mother and child competence of interest here corresponds to a test of the time × mother's IQ interaction. The parameters defined by the between-subjects contrast matrix, for the model in (3.4.1), are

$$\mathbf{CB}=\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} \beta_{0,12} & \beta_{0,24} & \beta_{0,36} \\ \beta_{\mathrm{L},12} & \beta_{\mathrm{L},24} & \beta_{\mathrm{L},36} \\ \beta_{\mathrm{Q},12} & \beta_{\mathrm{Q},24} & \beta_{\mathrm{Q},36} \\ \beta_{\mathrm{C},12} & \beta_{\mathrm{C},24} & \beta_{\mathrm{C},36} \end{bmatrix} = \begin{bmatrix} \beta_{\mathrm{L},12} & \beta_{\mathrm{L},24} & \beta_{\mathrm{L},36} \\ \beta_{\mathrm{Q},12} & \beta_{\mathrm{Q},24} & \beta_{\mathrm{Q},36} \\ \beta_{\mathrm{C},12} & \beta_{\mathrm{C},24} & \beta_{\mathrm{C},36} \end{bmatrix}. \quad (3.4.2)$$

In this case the between-subjects contrast matrix simply extracts the predictor (mother's IQ) trend coefficients. The **U** matrix consists of two columns that generate the linear and quadratic trends across time. Multiplying the result in (3.4.2) by **U** yields the matrix of secondary parameters, **Θ**:

$$\boldsymbol{\Theta}=\mathbf{CBU}=\begin{bmatrix} \beta_{\mathrm{L},12} & \beta_{\mathrm{L},24} & \beta_{\mathrm{L},36} \\ \beta_{\mathrm{Q},12} & \beta_{\mathrm{Q},24} & \beta_{\mathrm{Q},36} \\ \beta_{\mathrm{C},12} & \beta_{\mathrm{C},24} & \beta_{\mathrm{C},36} \end{bmatrix} * \begin{bmatrix} -1 & 1 \\ 0 & -2 \\ 1 & 1 \end{bmatrix} * \begin{bmatrix} 2 & 0 \\ 0 & 6 \end{bmatrix}^{-1/2}. \quad (3.4.3)$$

With *C* having *a* = 3 rows and **U** having *b* = 2 columns, **Θ** is 3 × 2 and contains the six mother's IQ × time interaction terms. For example, $\boldsymbol{\theta}_{\mathrm{L,Q}}$ is the parameter for the linear (in mother's IQ) × quadratic (in time) interaction, and so forth:

$$\boldsymbol{\Theta}=\begin{bmatrix} \boldsymbol{\theta}_{\mathrm{L,L}} & \boldsymbol{\theta}_{\mathrm{L,Q}} \\ \boldsymbol{\theta}_{\mathrm{Q,L}} & \boldsymbol{\theta}_{\mathrm{Q,Q}} \\ \boldsymbol{\theta}_{\mathrm{C,L}} & \boldsymbol{\theta}_{\mathrm{C,Q}} \end{bmatrix}. \quad (3.4.4)$$

Any change to the design or hypothesis of interest will lead to corresponding changes in one or more of **C, B, U**, and **Θ**.

3.5 Choosing Hypothesized B and Σ

For reasons cited earlier, the IHDP data were selected to provide estimates of **B** and **Σ** for the power analysis. The calculations reported here were obtained for the specific, purpose of power calculations and in no sense should be considered a definitive analysis. The absence of additional covariates and the failure to embed the analysis in an overall plan for the study are obvious limitations. Subsequent to the research proposal preparation, we conducted a more appropriate analysis of the IHDP data to examine intergenerational transmission of competence (Ramey et al. 1991). This analysis incorporated adjustments for covariates, including the stratification variables, site and birthweight, model cross-validation, and an assessment of intervention effects on the child development trajectories. It should be noted that power analysis results based on the parameter estimates resulting from this more complex analysis did not differ appreciably from those reported in Section 4.

A GLMM was fitted to data from 474 children in the IHDP low-educated subsample of the control group (chosen to be most compatible to the proposed target population), according to the model formulation given in Section 3.4. A third-degree polynomial in mother's IQ was found to provide adequate fit, based on tests of polynomials up through degree six. Linear and quadratic (in mother's IQ) models generated surfaces judged unsatisfactory due to inadequate approximation to monotonicity. Table 2 contains the model parameter estimates, $\hat{\mathbf{B}}$ and $\hat{\Sigma}$. Note that mother's IQ was expressed in terms of population mean zero and unit variance (rather than mean 100 and standard deviation 15) for numerical accuracy with polynomials. Table 3 contains corresponding hypothesis tests for between-subject effects, within-subject effects, and their interactions. Power would be computed for the test generated by the **C** matrix in (3.4.2), which concatenates the three **C** matrices used to generate the three interaction tests in the last three (within-subject) rows of Table 3.

A similar approach was used to choose the **U** matrix in (3.4.3). Both the linear and quadratic trends in time were significant as main effects and also interacted significantly with the cubic term in mother's IQ. Our choice of models was partially based on the principle that some overfitting has only a minimal cost in efficiency, whereas underfitting can bias the results. The combination of (1) the exploratory analysis results, (2) the wish to avoid underfitting, and (3) the commonness of growth phenomena involving two plateaus led us to select a cubic polynomial in mother's IQ and quadratic polynomial in time.

To illustrate the hypothesized relationship between child and mother competence, predicted values were computed under Model (2.1.1) for the first three years of life. (See Appendix B for a sketch of the algorithm.) Figure 1 contains the resulting response surface. The reader should recognize that the surface displayed in Figure 1 constitutes an hypothesized relationship, rather than a description of conclusions about data. The graph concretely embodies the alternative hypothesis, namely that the relationship of child IQ to mother's IQ changes substantially over time. The model assumes that the negative impact of low maternal IQ on child IQ increases dramatically over time. The slight decline for children with high-IQ mothers can probably be attributed to differences between tests. (The Bayley Scale of Infant Development was administered at ages 12 and 24 months, whereas the Stanford–Binet Intelligence Scale was administered at age 36 months.) No single test covers the entire age range of interest. Both tests provide scores on essentially the same scale and constitute the most widely used tests for their applicable age groups. We found that the graph clearly depicted the hypothesized relationship by embodying the complexity not captured in the original four-group formulation.

3.6 Power Study Factor Levels

Having chosen the GLMM parameters, we could conveniently compute power for a variety of research study designs. We limited our selections to a representative sampling of scenarios. Table 4 gives the (incomplete) factorial design used in the power study and indicates the combinations actually evaluated. Some additional variations are considered later in this article to compare the results to those from simpler approaches (see Table 11 and the associated discussion). The alternate research study designs were compared on the basis of power calculations as part of preparing the grant. As mentioned earlier, type I error rate was fixed at α = .05 due to the narrow focus of the planned analysis.

The three factors varying in Table 4 are response time sampling scheme, predictor values assumed, and covariance structure. Four sets of measurement times were of interest: {12, 24, 36}, {12, 18, 36}, {6, 18, 36}, and {18, 36}. The first of these corresponds to the response sampling times present in the IHDP data; hence the parameter estimates of Table 2 apply. The other three choices required new estimates of the GLMM parameters. Estimates for the columns of **B** that would be associated with 6- and 18-month measurements were computed using the formulas (given in Appendix B) for predicted values under a repeated measurements model. We used the covariance structure (given in Table 2) for {12, 24, 36}, {12, 18, 36}, and {6, 18, 36}. The lower right (2 × 2) submatrix of $\hat{\Sigma}$ was used for {18, 36}.

The second factor, predictor values assumed, involved determining the values of mother's IQ expected in the proposed study sample. The first choice was to assume the relative frequency of values observed in the IHDP control group, restricted to mothers with less than a high school education, to approximate the sampling plan for the proposed study. The second choice was to use the relative frequency of values observed for the NLSY sample of mothers. The third choice corresponded to a hypothetical situation in which mother's IQ could be measured in the hospital, which would allow a sample with four equal-sized groups of mothers with low, borderline, normal, or high IQ's. Within each group we assumed frequencies of values proportional to those observed for that particular subgroup of the IHDP low-educated control group subsample. Note that although sampling was to be conducted in terms of four equal-sized groups, the analysis would be based on continuous values of mother's IQ in a polynomial model. Although this was not practical, we were interested in assessing the expected gain in power associated with oversampling extremes of mother's IQ values.

We considered two choices for the third factor, covariance structure: (1) unstructured and (2) compound-symmetric. There are numerous other possibilities, including autoregressive and moving average structures; however, we thought these two provided the most realistic choices in the proposed setting. We used $\hat{\Sigma}$ from Table 2 for the unstructured case. For compound symmetry the common variance and correlation were assumed to be 238.3 and . 440, which constituted the maximum likelihood estimates (Morrison 1976, p. 250) based on $\hat{\Sigma}$ from Table 2. The correlation estimate is the ratio of the average covariance to the average variance, and the variance estimate is the average variance. Additional power calculations were made with the common correlation set to .220 or .880.

4. RESULTS FOR THE EXAMPLE

4.1 Power Calculations

The results of the primary power computations are given in Tables 5–10. The hypothesis for which power was computed corresponds to the hypothesis of no interaction between time and mother's IQ in Model (2.1.1). A test of this hypothesis evaluates $\mathbf{\Theta}$, in (3.4.4), for which rows correspond to linear, quadratic, and cubic trends in mother's IQ and columns correspond to linear and quadratic tends in time. Because the hypothesized model parameters only approximate what would be collected under the proposed study design, we considered cases in which the hypothesized regression relationships and covariance parameters differed from their estimated sizes by factors of .5 and 2.0. We thought that these multipliers adequately reflected our uncertainty about parameter values and, when coupled with the other factors in the power design, provided an adequate sensitivity analysis. Multiplying **Σ** by a constant multiplies all variances by the constant, whereas the error correlations remain unchanged because the covariances are also multiplied by the same constant. Note that changes in correlation are considered later (see Table 10 and the associated discussion). Approximate power was computed for overall sample sizes of 100, 200, and 400. In some cases the actual sample size considered varied up to 2% of the target value due to discreteness introduced in creating the **X** matrix.

Table 5 presents approximate power for child IQ measured at ages 12, 24, and 36 months, assuming the IHDP low-educated subsample frequencies of mother's IQ, and the IHDP control group estimates of **B** and **Σ** from Table 2. A total sample of 400 would yield an expected Geisser–Greenhouse power of .99 for detecting differences at least as large as those suggested by the IHDP data. In this case the power appears to be much more sensitive to the strength of the regression relationship than to the variance of child IQ scores. For example, if the error variances were twice as large and regression coefficients were unchanged, the power would be .82 ($N = 400$, Geisser–Greenhouse test). In turn, if the variances were unchanged but the regression coefficients were twice as large, then power would be at least .99 ($N \geq 100$).

Table 6 contains the approximate power expected for group-based sampling and a polynomial model. It was assumed that through screening the sample would be evenly distributed across the four groups of mother's IQ; namely retarded, borderline, normal, and high. It was also assumed that the spread of predictor scores within each group would follow the IHDP pattern. Note that the actual IQ scores would be treated as continuous values in the analysis. As would be expected, an increase in power would be associated with oversampling of extreme values of mother's IQ. The sampling scheme would generate better estimates of the relationship between mother and child IQ, particularly at the low end of the range of mother's IQ, where the slopes change most over time.

The investigators considered replacing the 24-month assessment with one at age 18 months, which led to Table 7. At first reading, the results surprised us. An increase in power would not initially be expected for a response sampling scheme with such unequal spacing. But further inspection of the values in Table 2 and the response surface in Figure 1 stimulated the following thoughts. We recognized that for the assumed model the relationship between mother and child competence changes more during the first 18 months of life than during the second 18 months. The power associated with a test for interaction would increase under a sampling scheme that yielded more data during a period of greater change.

The investigators also considered dropping the 12-month assessment altogether, which led to Table 8. The proposal was based on the plausible speculation that intellectual development may not reflect any environmental influence at age 12 months. For this case the $\mathbf{C}$ in (3.4.2) was used, $\mathbf{U} = [0\ 1\ -1]'$, and Θ; was 3 × 1, with elements corresponding to a linear trend in time interacting with the linear, quadratic, and cubic trends in mother's IQ. As would be expected, reducing to two measurement times decreases the power in detecting the hypothesized interaction. In addition, three measurements allow estimation of a quadratic term in a situation apparently involving substantial nonlinearity. However, the lower reliability, lower predictive validity, and different content of the 12-month instrument all provide support for dropping the assessment.

Table 9 provides results for a best-case scenario with respect to Σ, namely compound symmetry. In the calculation of power, the Geisser–Greenhouse test reduces to the un-corrected test, and hence to the UMP-α test. Only modest differences can be seen between Table 9 (compound symmetric Σ) and Table 5 (unstructured Σ), with the largest values in the far right column of Table 9. Hence for the child development example, little power would be lost by using one of the more general and robust tests. The similarities arise from a combination of forces. First, the smallness of Σ_* constrains $\boldsymbol{\varepsilon}$ from below. Second, sample sizes are substantial. Third, and perhaps most importantly, the unstructured Σ chosen gave $\boldsymbol{\varepsilon} = .90$ ($\boldsymbol{\varepsilon}$ is a function of eigenvalues of $\Sigma_* = \mathbf{U}'\Sigma\mathbf{U}$), a value not far from the upper bound of 1.0, which corresponds to sphericity.

Table 10 allows investigating the effects of varying the common correlation under compound symmetry. Table 9 was based on $\rho = .440$ and $\sigma^2 \in \{119.15, 238.3, 476.6\}$, whereas results in Table 10 are based on $\sigma^2 = 238.3$ and $\rho \in \{.220, .440, .880\}$. As ρ increases, the power increases. Doubling the correlation would dramatically reduce the sample size required, whereas halving the correlation does not substantially increase the sample size required. The child psychologists involved in the project thought that the correlation was extremely unlikely to be as high as .880.

4.2 Bias in Assuming X Fixed

To investigate the amount of bias introduced by assuming $\mathbf{X}$ fixed rather than Gaussian, powers were calculated for the univariate test of predictor trend at age 36 months. The $b = 1$ special case of methods in Section 2.5 gave sample sizes assuming fixed $\mathbf{X}$ that agreed within 3% of those found from tables in Gatsonis and Sampson (1989). The discrepancies seemed completely acceptable, given the many unknowns in study planning. Some of the error likely arose from interpolating in the tables of Gatsonis and Sampson. The fact that the smallest sample size studied was 100 undoubtedly increased accuracy.

4.3 Final Choice of Sample Size

The sample size proposed in the grant was based on the results presented in Table 5. We felt that the standard errors of the IHDP parameter estimates indicated good precision and that the IHDP low-educated control subsample was a reasonable approximation to the research target population. Consequently, an initial enrollment of 500, which, due to nonresponse and attrition, could be expected to yield an analysis sample size of 400, was proposed as sufficient for detecting the hypothesized interaction.

5. ALTERNATE POWER APPROXIMATIONS

5.1 Seeking a Simpler Method

The methods presented earlier, especially when coupled with the software described in Appendix A, allow straightforward calculation of power for multivariate and repeated measures models. As demonstrated by the example, however, substantial work may be required to conduct a power analysis for multivariate hypotheses. The question naturally arises as to whether simpler methods might be adequate. In the context of the GLMM, the amount of work may be reduced by simplifying the proposed design matrix, **X**, or proposed hypothesis matrices, **C** and **U**. In this section we contrast the results of such simplifications with those of the previous section. We also emphasize the critical need for a sensitivity analysis in power computations.

5.2 Effect of Simplifying the Hypothesis

Reducing the design matrix may lead to a simpler hypothesis and, in turn, simpler power calculations. Differences in power values due to changing design may be inferred by comparing rows of Table 11. The four sets of three rows correspond to a 2 × 2 factorial, with one factor being distribution of predictor values within interval and the second factor being relative frequency between intervals.

The bottom half of the table was computed using a cubic polynomial regression model, with the IHDP spread of predictor scores within each of four IQ intervals. The cubic polynomial design involves treating mother's IQ as a continuous predictor (fixed effect, interval scale). The last three rows (10, 11, and 12) are based on the same design as for Tables 5, 7, 8, 9, and 10, as described in Section 3.6. Recall that $\hat{\mathbf{B}}$ and $\hat{\mathbf{\Sigma}}$ from Table 2 were taken to be **B** and **Σ**. The relative frequencies of mother's IQ values were those observed in the IHDP control group, restricted to mothers with less than a high school education. Rows 7, 8, and 9 of Table 11 differ from the last three rows only in being based on a design that would result if mothers were screened on IQ and selected to yield four equal-sized groups (low, borderline, normal, or high IQ). Here **C** is the same as for Tables 5, 7, 8, 9, and 10; see equation (3.4.2.).

The top half of Table 11 is based on a fixed-effect ANOVA/MANOVA model, with response variable means assumed equal to those observed within the four mother's IQ intervals (groups) for the IHDP control data. The first three rows are based on assuming equal numbers in each group, whereas rows 4, 5, and 6 are based on cell sizes proportional to the mother's IQ frequencies in the IHDP control data. Here **C** is chosen to test the usual ANOVA hypothesis, H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$.

Differences in power values due to changing the **U** matrix (within-subject contrast matrix) may be inferred by comparing columns of Table 11. The first column ("Last time") results from using **U** = [0 0 1]′, and evaluates the effect of mother's IQ on child IQ at the end of the study (age 36 months). This generates a univariate analysis that may be characterized as cross-sectional. The second column ("Linear only") results from using only the linear trend across time. For the example, this corresponds to using $\mathbf{U}=[\begin{array}{ccc}-1 & 0 & 1\end{array}]'/\sqrt{2}$, see equation

(3.4.3). The third column ("Linear and quadratic") corresponds to using both trends, as in Equation (3.4.3). Note that the numbers in the lower right corner of the table duplicate results from the middle of Table 5, and that the numbers immediately above come from the middle of Table 6.

Comparisons of rows and columns of Table 11 reveal that (1) simplifying the design and/or hypothesis may substantially change calculated power, (2) substantial bias in power values may result if power analysis and subsequent data analysis are mismatched, and (3) some simplifications have little effect. Three particular conclusions may be drawn. First, for this application all simplifications increase power. Moving to the left or up in the table corresponds to simplification and to more power. Second, for this example basing power analysis on an ANOVA (or MANOVA) model when, in fact, data analysis will involve a polynomial model would lead to a study with very poor sensitivity. This difference arises because the polynomial model incorporates the natural heterogeneity of predictor values. An additional concern is the random nature of the predictors. Recall that this question was addressed in Section 4.2. Within the limits of the resolution of the tables provided by Gatsonis and Sampson, using the methods described in Section 2 for power approximation would change the sample size required by only approximately 3%. This result applies to the left two columns of Table 11. Simulations or analytic results would be required to evaluate the impact on the rightmost column (which involves a multivariate hypothesis). Third, the power of the linear-trend-only test closely approximates that of the all-trends tests, because for the example the higher-order terms were relatively small. Although common, this is obviously not always true.

The relationship between simplification and power is not universal. In the child development example, the last time yielded power substantially larger than the planned test of trends. In contrast, consider the results of a power analysis (using the same methods as here) conducted by two of the authors (Muller and LaVange 1991). For a clinical trial of alternate kidney therapies, the corresponding test of effect at the last time yielded power substantially lower than the planned test of trends. We suspect that situation-specific power analysis provides the best insurance against such surprises. Drawing any general conclusions about power orderings among alternate power approximation methods will require substantial additional work.

6. DISCUSSION

6.1 Recommendations for Choosing Power Analysis Designs

1. Align the design and hypothesis of the power analysis with the planned data analysis, as best as practical. Kimball (1957) described a type III error as "the error committed by giving the right answer to the wrong problem [emphasis in the original]." As the power analysis and data analysis become more disparate, the risk of a type III error increases.

2. Embed any power analysis in a defensible sensitivity analysis. The dimensions along which power should be studied depend on the particular analysis method, design, hypothesis, and research goals.

3. Have the extent of the power analysis reflect the ethical, scientific, and monetary costs of type I, II, and III errors. Having software available such as that reviewed by Goldstein (1989) or described in Appendix A can greatly expand the range of situations in which well-tailored power analysis may be applied.

6.2 Benefits of Power Analysis in Study Design

By reviewing and demonstrating power calculation methods for the GLMM, we hope to encourage other statisticians to align their power analysis with the planned data analysis and to conduct appropriate sensitivity analysis. The value of a power analysis depends on these two features. Such a power analysis stimulates the iterative refinement of the scientific hypothesis, design, and analysis plan, thereby catalyzing the interactions of statisticians and subject matter specialists.

Power calculation increases client satisfaction not only by increasing the chances of research success, but also by providing a natural and effective vehicle for deepening the client’s understanding of the analysis methods to be used. Although not convenient for the child development example, plots of power as a function of effect size or sample size often provide great insight to subject matter specialists. In all cases one should conduct some form of sensitivity analysis. The nature of the analysis depends on which parameters have the most effect on power and/or are known least certainly. As for the example, with linear models one usually examines the effects of varying (1) sample size, (2) true difference, and (3) variance—the three components of the noncentrality parameter.

Power calculation increases statistician satisfaction by increasing quality of advice given and by helping formalize the front-loading of design and analysis planning. Power analysis coerces the specification of an analysis in advance of data collection. We considered only one particular data analysis in the example. More globally, power analysis allows evaluating the tradeoffs among type I error rate, type II error rate, choice of variables, choice of analysis, and choice of tests. See Muller et al. (1984) for a tutorial designed to explain these issues to nonstatisticians.

6.3 Costs of Power Analysis in Study Design

We think that the effort allocated to power analysis should be proportional to all costs of the study, including (1) money to be spent, (2) personnel time of statisticians and subject matter specialists, (3) time to complete the study (opportunity costs), and (4) ethical costs in the research. Note that the statistician shares ethical responsibility with the subject matter specialist. For example, an excessive sample size may expose too many subjects to risk, whereas an inadequate sample size may delay the discovery of an efficacious intervention. The example research was planned to last five years, with one year for recruitment, three for data collection, and one for analysis. The budget was approximately 1.3 million dollars. The power analysis for the example consumed roughly two work weeks of a doctoral-level statistician, one work week of a masters-level statistician, and one work week of data manager time. We believe that these times translate to a necessary expense in any billing environment. For us as statisticians, it is an ethical imperative to conduct an appropriate and affordable power analysis.

6.4 Obstacles to Power Analysis

Some data analysts have been stopped from conducting multivariate power calculations by not having immediately apparent choices for **X, B**, or **Σ**. Thoughtful discussion with subject matter specialists can usually elicit plausible ranges for **X** and **B**. For example, without data available we would have created a **B** with only a linear × linear interaction term. The test power could then have been plotted and studied as a function of the scalar noncentrality that would result. Finding a choice for Σ usually provides the biggest obstacle. Because Σ describes errors, residuals from any study using the same measures and similar target population should provide a usable choice. Unfortunately, few authors publish residual covariance or correlation matrices. One may be forced to resort to more speculative choices. Assuming either an autoregressive or compound symmetric pattern reduces the task to choosing only two parameters, not $p(p + l)/2$. Systematically varying the two parameters allows evaluation of the sensitivity of the conclusions to a particular choice.

6.5 Other Uses of Power Analysis

The example comprises a “prospective power analysis” in that the study had yet to be done. In some cases investigators wish to declare a difference zero after computing a test with a large *p* value. In such a case “retrospective power analysis“ allows demonstrating high power against any hypothesis of scientific interest, if the power is present. Such an approach greatly enhances the credibility of claims of no difference. It may be argued that just as we require a small α (type I error rate) to declare nonzero differences, so should we require a small β (type II error rate) to declare zero differences. See Benignus, Kafer, Muller, and Case (1987), Harbin, Benignus, Muller, and Barton (1988) or Benignus, Muller, Smith, Pieper, and Prah (1990) for examples of this use of power analysis. Muller and Benignus (1992) discussed the issue in the context of toxicology.

By assuring clients that excellent power will be available even with greatly reduced nominal type I error rates, we have nearly always succeeded in convincing clients to balance type I and type II error rates. Given the focus on a single hypothesis in a single analysis, a type I error rate of .05 was deemed acceptable at the planning stage of the research study. The presence of many other outcomes and many other predictors in the final study proposal should have provided strong pressure for a more conservative choice. Looking back, we failed to recognize the problem while planning the power analysis.

The power approximations reviewed here may also be used to study which test statistic to choose for the multivariate hypothesis. Incidentally to the example power analysis, Wilks’ test and the Geisser–Greenhouse test were compared. The importance of the observation lies not in the superiority of one statistic in a particular example, but rather that convenient approximations allow choosing the best for any application.

Acknowledgments

This research was supported in part by NICHD Mental Retardation Research Center Grant 2-P30-HD03110-21. Muller’s work was supported in part by NCI Program Project Grant P01 CA47 982-04 and NIH Clinical Research Center Grant RR00046. The authors gratefully acknowledge the helpful comments of Professor D. Quade and the reviewers of earlier versions of this article. The authors also thank V. A. Benignus, Research Psychologist, USEPA,

and Research Professor, Department of Psychology, University of North Carolina, Chapel Hill, NC 27599, for stimulating our consideration of the issues in Section 6.5.

APPENDIX A

SOFTWARE FOR POWER CALCULATIONS

Readers may obtain a free copy of a general purpose program that implements the methods described and used in this paper. The program is written entirely in the IML matrix language of SAS©. The program requires the user to specify the design matrix, model parameters, and contrasts as matrices. Moderate flexibility and good error checking are included. The program has been run successfully on PC-DOS, Sun Unix, and IBM MVS machines; however, no guarantees or support of any kind can be supplied.

Beginning with Version 6.09 of SAS, the program has been included with the installation package. Consult your local SAS site representative for documentation of the example library for IML. Documentation is included with the program. If you do not yet have version 6.09 or later of SAS, you may mail a Bitnet address to either POWER.FPG@MHS.UNC.EDU or Statistical Power Software, Design and Statistics Unit, Frank Porter Graham Child Development Center, CB#8180, University of North Carolina, Chapel Hill, NC 27599. This distribution path will be supported only long enough to ensure that all sites have access to the program.

The current (third) version was written by L. L. Keyes in the Design and Statistics Unit, Frank Porter-Graham Child Development Center, and K. E. Muller. B. L. Peterson and K. E. Muller wrote the first version, which included only the multivariate test statistics, in PROC MATRIX. C. N. Barton and K. E. Muller added the treatment of the univariate approach to repeated measures to create the second version in PROC MATRIX.

APPENDIX B

PREDICTED VALUES IN REPEATED MEASURES MODELS

B. 1 Predicted Values at Observed Times

For the models of interest the columns of **Y** and the columns of **B** are indexed and ordered by time, $t \in \{t_1, t_2, \ldots, t_p\}$. For a one-group design, $\mathbf{X} = \mathbf{1}_N$ and $\mathbf{B} = \boldsymbol{\mu}'$ is $(1 \times p)$, with $\boldsymbol{\mu}_j$ the mean response at t_j. Any other design may be treated by applying the results to $\mathbf{CB} = \{\boldsymbol{\mu}_{gj}\}$, with **C** being the contrast matrix that generates the set of group, $g \in \{1\ 2 \cdots G\}$, or marginal means of interest. This corresponds to projecting into a cell mean coding parameterization. Note that for estimation (of expected value parameters), the same formulas are used for both multivariate and univariate repeated measures analysis.

A “full” model involves a situation in which the model across time is of order $k = (p - 1)$ and can be expressed as a function of unknown regression coefficients, $\{\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{p-1}\}$:

$$\mathscr{E}Y_t = \sum_{j=0}^{p-1} t^j \cdot \boldsymbol{\gamma}_j. \quad \text{(B.1)}$$

The predicted values at the observed times are simply $\hat{\mathbf{Y}} = \mathbf{X}\ddot{\mathbf{B}}$.

Similar results may be provided for a reduced model. Assume that one has chosen

$$\mathbf{U}_{\boldsymbol{A}}=\mathbf{1}/\sqrt{p}=\mathbf{u}_0 \quad \text{(B.2)}$$

and

$$\text{áźť}_{\boldsymbol{T}}=[\text{áźţ}_1\cdots\text{áźţ}_{p-1}] \quad \text{(B.3)}$$

to be the *orthonormal* polynomial trends matrices and has decided that the order of model across time is $k \le (p - 1)$. This implies that, for $\mathbf{U}_p = [\mathbf{U}_A\ \mathbf{U}_T]$, the model of interest involves

$$_*\gamma'=\mathbf{B}\mathbf{U}_p=[_*\gamma_0{}_*\gamma_1\cdots{}_*\gamma_k 0\cdots 0]. \quad \text{(B.4)}$$

Here $_*\boldsymbol{\gamma}_{\boldsymbol{j}}$ is the orthogonal polynomial coefficient of order j. Also, define

$$_*\boldsymbol{\gamma}_k'=\mathbf{B}[\mathbf{u}_0\mathbf{u}_1\cdots\mathbf{u}_k]=\mathbf{B}\mathbf{U}_k. \quad \text{(B.5)}$$

Then

$$\hat{\mathbf{Y}}_k=\mathbf{X}\ddot{\mathbf{B}}[\mathbf{U}_k\mathbf{0}]\mathbf{U}_p^{-1} \quad \text{(B.6)}$$

is the predicted value set for the model of order k. Obviously the full-model results are a special case of these results

B.2 Prediction Model Coefficients

For times $\mathbf{t} = [t_1, t_2 \cdots t_p]'$, define the full-rank matrix

$$\underset{p\times p}{\mathbf{T}}=[\ \mathbf{1}\quad \mathbf{t}\quad \mathbf{t}^2\cdots\mathbf{t}^{p-1}\ ]. \quad \text{(B.7)}$$

Regression coefficients are sought for model (B.1). Again consider a one group design, with $\mathbf{B} = \boldsymbol{\mu}'$. Then

$$[\gamma_0\gamma_1\cdots\gamma_{p-1}]\mathbf{T}'=[\boldsymbol{\mu}_1\cdots\boldsymbol{\mu}_p], \quad \text{(B.8)}$$

and

$$\gamma'=\boldsymbol{\mu}'\mathbf{T}^{-t}. \quad \text{(B.9)}$$

Note that $_*\boldsymbol{\gamma}' = \boldsymbol{\mu}'\mathbf{U}_p$ are the orthogonal polynomial regression coefficients, which are the secondary parameters used for choosing the order of the model. Furthermore, the natural polynomial coefficients are

$$\gamma'={}_*\gamma'\mathbf{U}_p^{-1}\mathbf{T}^{-t}={}_*\gamma'[\mathbf{T}'\mathbf{U}_p]^{-1}. \quad \text{(B.10)}$$

For the reduced model, following earlier results,

$$\hat{\gamma}_k' = \ddot{\mathbf{B}}[\mathbf{U}_k\mathbf{0}]\mathbf{U}_p^{-1}\mathbf{T}^{-t} = \ddot{\mathbf{B}}[\mathbf{U}_k\mathbf{0}][\mathbf{T}'\mathbf{U}_p]^{-1}. \quad \text{(B.11)}$$

Under the reduced model, γ_k' will have $[p - (k + 1)]$ trailing zero elements. Also, in general the nonzero elements of $\hat{\gamma}_k'$ are not equal to the first $k + 1$ elements of $\hat{\boldsymbol{\gamma}}$, for particular sample estimates. In turn, for a single time, say t_0,

$$\hat{y}_{t_0} = \ddot{\mathbf{B}}[\mathbf{U}_k\mathbf{0}](\mathbf{T}'\mathbf{U}_p)^{-1}\begin{bmatrix} 1 \\ t_0 \\ t_0^2 \\ \vdots \\ t_0^{p-1} \end{bmatrix}. \quad \text{(B.12)}$$

One may prefer to use $[\begin{array}{ccc} 1 & t_0 \cdots t_0^k & 0 \cdots 0 \end{array}]'$. As always, polynomials may cause numerical problems. Centering time values may help (see Kleinbaum, Kupper, and Muller 1988, chap. 13).

REFERENCES

Arnold, SF. The Theory of Linear Models and Multivariate Analysis. New York: John Wiley; 1981.

Barton CN, Cramer EC. Hypothesis Testing in Multivariate Linear Models With Randomly Missing Data. Communications in Statistics-Simulation. 1989; 18:875–895.

Benignus VA, Kafer ER, Muller KE, Case MW. Absence of Symptoms With Carboxyhemoglobin Levels of 16–23%. Neurotoxicology and Teratology. 1987; 9:345–348. [PubMed: 3696105]

Benignus VA, Muller KE, Smith MV, Pieper KS, Prah JD. Compensatory Tracking in Humans With Elevated Carboxyhemoglobin. Neurotoxicology and Teratology. 1990; 12:105–110. [PubMed: 2333061]

Box GEP. Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems: I. Effects of Inequality of Variance in the One-Way Classification. The Annals of Mathematical Statistics. 1954a; 25:290–302.

Box GEP. Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems: II. Effects of Inequality of Variance and of Correlation Between Errors in the Two-Way Classification. The Annals of Mathematical Statistics. 1954b; 25:484–498.

Cohen, J. Statistical Power Analysis for the Behavioral Sciences. rev. ed.. New York: Academic Press; 1977.

Davidson ML. Univariate Versus Multivariate Tests in Repeated Measures Experiments. Psychological Bulletin. 1972; 77:446–452.

Freedman DA, Peters SC. Bootstrapping a Regression Equation: Some Empirical Results. Journal of the American Statistical Association. 1984; 79:97–106.

Gatsonis C, Sampson AR. Multiple Correlation: Exact Power and Sample Size Calculations. Psychological Bulletin. 1989; 106:516–524. [PubMed: 2813654]

Geisser S, Greenhouse SW. An Extension of Box's Results on the Use of the *F* Distribution in Multivariate Analysis. The Annals of Mathematical Statistics. 1958; 29:885–891.

Goldstein R. Power and Sample Size Via MS/PC-DOS Computers. The American Statistician. 1989; 43:253–260.

Graybill, FA. Theory and Applications of the Linear Model. Scituate, MA: Duxbury Press; 1976.

Greenhouse SW, Geisser S. On Methods in the Analysis of Profile Data. Psychometrika. 1959; 24:95–112.

Harbin TJ, Benignus VA, Muller KE, Barton CN. The Effects of Low-Level Carbon Monoxide Exposure Upon Evoked Cortical Potentials in Young and Elderly Men. Neurotoxicology and Teratology. 1988; 10:93–100. [PubMed: 3398828]

Hocking, RR. The Analysis of Linear Models. Belmont, CA: Wadsworth; 1985.

Infant Health and Development Program. Enhancing the Outcomes of Low-Birth-Weight, Premature Infants. Journal of the American Medical Association. 1990; 263:3035–3042. [PubMed: 2188023]

Jennrich RI, Schluchter MD. Unbalanced Repeated-Measures Models With Structured Covariance Matrices. Biometrics. 1986; 42:805–820. [PubMed: 3814725]

Kimball AW. Errors of the Third Kind in Statistical Consulting. Journal of the American Statistical Association. 1957; 52:133–142.

Kleinbaum, DG.; Kupper, LL.; Muller, KE. Applied Regression Analysis and Other Multivariate Methods. 2nd ed.. Boston: PWS-Kent; 1988.

Kraemer, HC.; Thiemann, S. How Many Subjects? Statistical Power Analysis in Research. Newbury Park, CA: Sage; 1987.

Kshirsagar, AM. Multivariate Analysis. New York: Marcel Dekker; 1972.

Kulp RW, Nagarsenker BN. An Asymptotic Expansion of the Nonnull Distribution of Wilks' Criterion for Testing the Multivariate Linear Hypothesis. The Annals of Statistics. 1984; 12:1576–1583.

Kupper LL, Hafner KB. How Appropriate are Popular Sample Size Formulas? The American Statistician. 1989; 43:101–105.

Lee YS. Distribution of the Canonical Correlations and Asymptotic Expansions for Distributions of Certain Independence Test Statistics. The Annals of Mathematical Statistics. 1971; 42:526–537.

Lipsey, MW. Design Sensitivity: Statistical Power for Experimental Research. Newbury Park, CA: Sage; 1990.

Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. New York: John Wiley; 1987.

McCormick MC. The Contribution of Low Birth Weight to Infant Mortality and Childhood Morbidity. New England Journal of Medicine. 1985; 312:82–90. [PubMed: 3880598]

Morrison, DF. Multivariate Statistical Methods. 2nd ed.. New York: McGraw-Hill; 1976.

Muller KE, Barton CN. Approximate Power for Repeated Measures ANOVA Lacking Sphericity. Journal of the American Statistical Association. 1989; 84:549–555.

Muller KE, Barton CN. Correction to "Approximate Power for Repeated Measures ANOVA Lacking Sphericity". Journal of the American Statistical Association. 1991; 86:255–256.

Muller KE, Barton CN, Benignus VA. Recommendations for Appropriate Statistical Practice in Toxicology. Neurotoxicology. 1984; 5:113–126. [PubMed: 6542184]

Muller KE, Benignus VA. Increasing Scientific Power With Statistical Power. Neurotoxicology and Teratology. 1992; 14:211–219. [PubMed: 1386138]

Muller, KE.; LaVange, LM. Sample Size Determination for Multivariate Models. invited presentation at ENAR Spring Meetings, 1991; 1991.

Muller KE, Peterson BL. Practical Methods for Computing Power in Testing the Multivariate General Linear Hypothesis. Computational Statistics and Data Analysis. 1984; 2:143–158.

Neter, J.; Wasserman, W.; Kutner, MH. Applied Linear Regression Model. 2nd ed.. Homewood, IL: Irwin; 1989.

O'Brien RG. Performing Power Sensitivity Analyses on General Linear Model Hypotheses. Proceedings of the Statistical Computing Section, American Statistical Association. 1982:114–118.

O'Brien RG, Kaiser MK. MANOVA Method for Analyzing Repeated Measures Designs: An Extensive Primer. Psychological Bulletin. 1985; 97:316–333. [PubMed: 3983301]

O'Brien, RG.; Lohr, VI. Power Analysis for Linear Models: The Time Has Come. Proceedings of the Ninth Annual SAS Users Group International Conference; SAS Institute; Cary, NC. 1984.

O'Brien, RG.; Muller, KE. A Unified Approach to Statistical Power for *t* Tests to Multivariate Models. In: Edwards, LK., editor. Applied Analysis of Variance in Behavioral Sciences. New York: Marcel Dekker; 1992.

Olson CL. Comparative Robustness of Six Tests in Multivariate Analysis of Variance. Journal of the American Statistical Association. 1974; 69:894–908.

Olson CL. On Choosing a Test Statistic in Multivariate Analysis. Psychological Bulletin. 1976; 83:579–586.

Olson CL. Practical Considerations in Choosing a MANOVA Test Statistic: A Rejoinder to Stevens. Psychological Bulletin. 1979; 86:1350–1352.

Ramey, CT.; Ramey, SL.; Sparling, JJ.; LaVange, LM.; Bryant, DM.; Wasik, BH. Altering Intergenerational Transmission of Competence. paper presented at Society for Research in Child Development Biennial Meeting; April; Seattle, Washington. 1991.

Ramey CT, Bryant DM, Wasik BH, Sparling JJ, Fendt KH, LaVange LM. The Infant Health and Development Program for Low Birthweight, Premature Infants: Program Elements, Family Participation. Pediatrics. 1992; 89:454–465. [PubMed: 1371341]

Rocke DM. Bootstrap Bartlett Adjustment in Seemingly Unrelated Regression. Journal of the American Statistical Association. 1989; 84:598–601.

Sampson AR. A Tale of Two Regressions. Journal of the American Statistical Association. 1974; 69:682–689.

Searle, SR. Linear Models. New York: John Wiley; 1971.

Srivistava JN. On the Extension of Gauss–Markov Theorem to Complex Multivariate Linear Models. Annals of the Institute of Statistical Mathematics. 1967; 19:417–437.

Stevens, JP. Comment on Choosing a Test Statistic in Multivariate Analysis. In: Olson, CL., editor. Psychological Bulletin. Vol. 86. 1979. p. 355-360.

Stevens JP. Power of the Multivariate Analysis of Variance Tests. Psychological Bulletin. 1980; 88:728–737.

Sugiura N, Fujikoshi Y. Asymptotic Expansions of the Non-null Distributions of the Likelihood Ratio Criteria for Multivariate Linear Hypothesis and Independence. The Annals of Mathematical Statistics. 1969; 40:942–952.

Timm, NH. Multivariate Analysis. Monterey, CA: Brooks/Cole; 1975.

Wang CM. On the Analysis of Multivariate Repeated Measures Designs. Communications in Statistics Part A—Theory and Methods. 1983; 12:1647–1659.

Zellner A. An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. Journal of the American Statistical Association. 1962; 57:348–368.

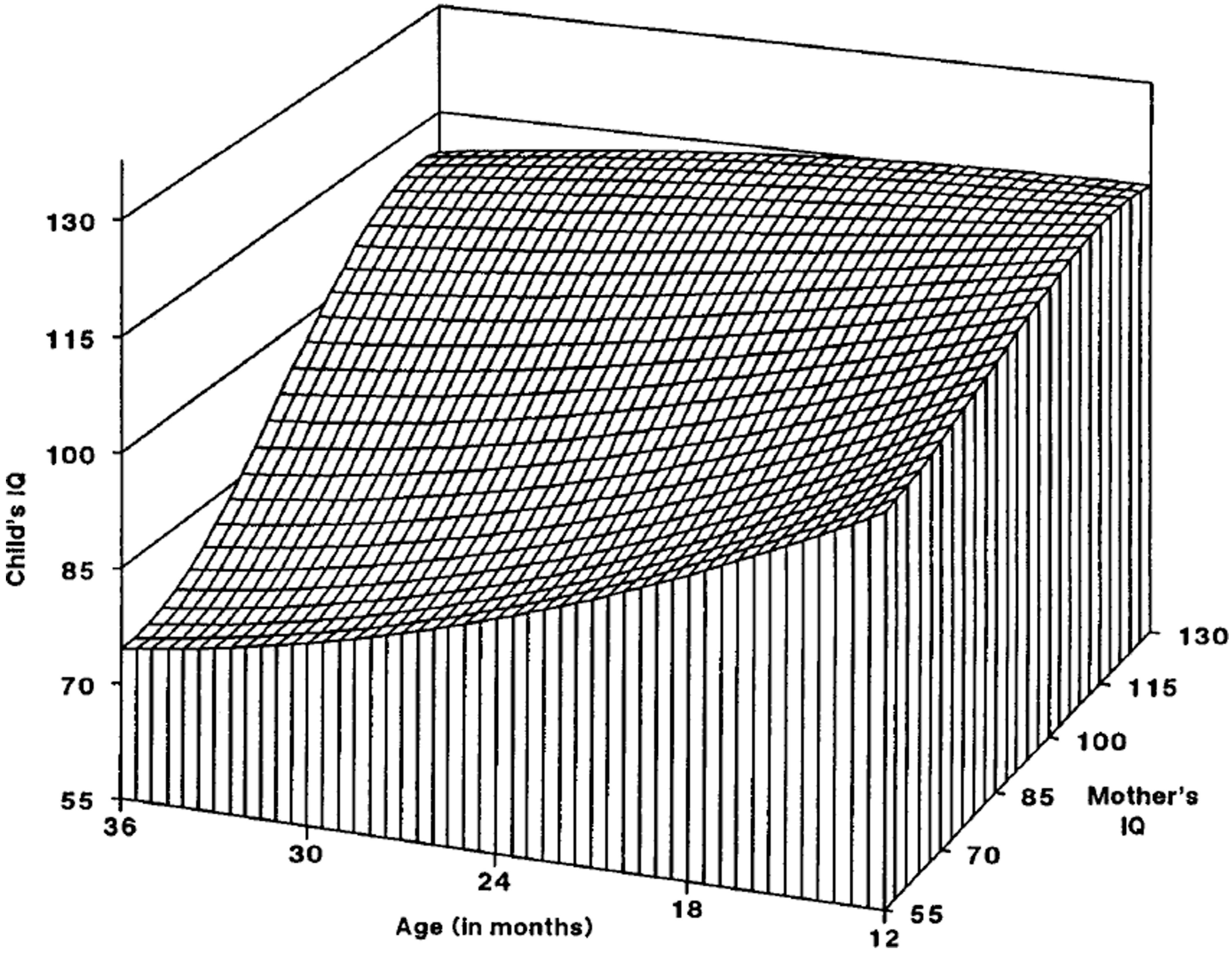

Figure 1.

Hypothesized Values of Child's IQ From Ages 12 to 36 Months as a Function of Mother's IQ.

Table 1

GLMM Test Statistic Properties

Name	Statistic	Principle	Association ($\hat{\eta}$)
RLR	$\max \hat{\rho}_k^2 = \max \ \ \mathrm{eval}(\mathbf{HT}^{-1})$	Union-Intersection	RLR
W[*]	$\prod (1 - \hat{\rho}_k^2) = \lvert \mathbf{ET}^{-1} \rvert$	Likelihood Ratio	$1 - W^{1/g}$
PB	$\Sigma \rho_k^2 = tr(HT^{-1})$	Total Sqrd Correlation	PB/s
HLT	$\sum \hat{\rho}_k^2 / (1 - \hat{\rho}_k^2) = \mathrm{tr}(\mathbf{HE}^{-1})$	ANOVA Analog	(HLT/s)/(1 + HLT/s)
REP	tr(**H**)/tr(**E**)	Sphericity	(REP)/(1 + REP)

[*] Here $g = [(a^2b^2 - 4)/(a^2 + b^2 - 5)]^{1/2}$.

Table 2

B and $\hat{\Sigma}$ for IHDP Control Data (N = 474) Response Is Child's IQ

$\hat{\boldsymbol{B}}$	*12 Months*	*24 Months*	*36 Months*
Intercept	114.46	104.66	98.83
(Mother's IQ – 100)/15	2.88	8.77	10.67
$[(\text{Mother's IQ} - 100)/15]^2$	−0.71	−0.90	−1.30
$[(\text{Mother's IQ} - 100)/15]^3$	−0.21	−0.54	−0.72
$\hat{\boldsymbol{P}}/\hat{\boldsymbol{\Sigma}}^{\dagger}$	*12 Months*	*24 Months*	*36 Months*
12 Months	218.48	83.66	72.19
24 Months	0.36	251.92	158.60
36 Months	0.31	0.64	244.58

† Covariances above the diagonal and correlations below.

Table 3

IHDP Control Group Data (N = 474) ANOVA Tests

Between-subject effects				Within-subject effects					
Source	df	F_{obs}[*]	P value	Source	df	F_{obs}(REP)	G–G P value	F_{obs}(W)	W P value
Mother's IQ	1	106.39	<.0001	Time	2	71.40	<.0001	59.00	<.0001
Mother's IQ^2	1	6.97	.0086	Time × Mother's IQ	2	35.46	<.0001	27.49	<.0001
Mother's IQ^3	1	14.23	.0002	Time × Mother's IQ^2	2	0.75	.4729	0.78	.4585
				Time × Mother's IQ^3	2	4.52	.0112	3.66	.0265

[*] Between *F* tests have 470 error df, F_{obs}(REP) tests have 940 error df, $\hat{\boldsymbol{\varepsilon}}$ = .902, F_{obs}(G–G) tests have 847.9 error df, and F_{obs}(W) tests have 469 error df.

Table 4

Power Analysis Design Factors

	Predictor values assumed		
Covariance response times	**IHDP ratios of four intervals, IHDP spread within**	**NLSY**	**Balanced four intervals, IHDP spread within**
Unstructured Σ			
12, 24, 36	×	×	×
12, 18, 36	×		
6, 18, 36	×		
18, 36	×		
Compound-Symmetric Σ			
12, 24, 36	×	×	×
12, 18, 36			
6, 18, 36			
18, 36			

Table 5

Approximate Power (×100) for Interaction of Time × Mother's IQ: Unstructured Σ for Time $\in \{12, 24, 36\}$, IHDP Ratios of Four Intervals, IHDP Spread Within

Σ multiplier	N	B multiplier	Wilks's LR	Gesser–Greenhouse
.5	100	.5	25	30
.5	100	1.0	83	91
.5	100	2.0	>99	>99
.5	200	.5	41	48
.5	200	1.0	97	99
.5	200	2.0	>99	>99
.5	400	.5	74	82
.5	400	1.0	>99	>99
.5	400	2.0	>99	>99
1.0	100	.5	14	16
1.0	100	1.0	49	59
1.0	100	2.0	99	>99
1.0	200	.5	21	25
1.0	200	1.0	74	83
1.0	200	2.0	>99	>99
1.0	400	.5	41	48
1.0	400	1.0	97	99
1.0	400	2.0	>99	>99
2.0	100	.5	09	10
2.0	100	1.0	25	30
2.0	100	2.0	83	91
2.0	200	5	12	14
2.0	200	1.0	41	48
2.0	200	2.0	97	99
2.0	400	.5	21	24
2.0	400	1.0	74	82
2.0	400	2.0	>99	>99

Table 6

Approximate Power (×100) for Interaction of Time × Mother's IQ: Unstructured Σ for Time ∈ {12, 24, 36}, Balanced Four Intervals, IHDP Spread Within

Σ multiplier	N	B multiplier	Wilks's LR	Geisser–Greenhouse
.5	100	.5	51	61
.5	100	1.0	99	>99
.5	100	2.0	>99	>99
.5	200	.5	85	92
.5	200	1.0	>99	>99
.5	200	2.0	>99	>99
.5	400	.5	>99	>99
.5	400	1.0	>99	>99
.5	400	2.0	>99	>99
1.0	100	.5	26	31
1.0	100	1.0	84	92
1.0	100	2.0	>99	>99
1.0	200	.5	51	61
1.0	200	1.0	99	>99
1.0	200	2.0	>99	>99
1.0	400	.5	87	93
1.0	400	1.0	>99	>99
1.0	400	2.0	>99	>99
2.0	100	.5	14	17
2.0	100	1.0	51	61
2.0	100	2.0	99	>99
2.0	200	.5	26	31
2.0	200	1.0	85	92
2.0	200	2.0	>99	>99
2.0	400	.5	53	62
2.0	400	1.0	>99	>99
2.0	400	2.0	>99	>99

Table 7

Approximate Power (×100) for Interaction of Time × Mother's IQ: Unstructured Σ for Time ∈ {12, 18, 36}, IHDP Ratios of Four Intervals, IHDP Spread Within

Σ multiplier	N	B multiplier	Wilks's LR	Geisser–Greenhouse
.5	100	.5	30	29
.5	100	1.0	89	89
.5	100	2.0	>99	>99
.5	200	.5	48	47
.5	200	1.0	99	99
.5	200	2.0	>99	>99
.5	400	.5	82	80
.5	400	1.0	>99	>99
.5	400	2.0	>99	>99
1.0	100	.5	16	15
1.0	100	1.0	57	56
1.0	100	2.0	99	>99
1.0	200	.5	24	24
1.0	200	1.0	82	81
1.0	200	2.0	>99	>99
1.0	400	.5	48	46
1.0	400	1.0	99	99
1.0	400	2.0	>99	>99
2.0	100	.5	10	10
2.0	100	1.0	30	29
2.0	100	2.0	89	89
2.0	200	.5	14	13
2.0	200	1.0	48	47
2.0	200	2.0	99	99
2.0	400	.5	24	23
2.0	400	1.0	82	80
2.0	400	2.0	>99	>99

Table 8

Approximate Power (×100) for Interaction of Time × Mother's IQ: Unstructured Σ for Time ∈ {18, 36}, IHDP Ratios of Four Intervals, IHDP Spread Within

Σ multiplier	N	B multiplier	Wilks's LR	Geisser–Greenhouse
.5	100	.5	22	22
.5	100	1.0	74	74
.5	100	2.0	>99	>99
.5	200	.5	35	35
.5	200	1.0	93	93
.5	200	2.0	>99	>99
.5	400	.5	64	64
.5	400	1.0	>99	>99
.5	400	2.0	>99	>99
1.0	100	.5	13	13
1.0	100	1.0	42	42
1.0	100	2.0	97	97
1.0	200	.5	19	19
1.0	200	1.0	65	65
1.0	200	2.0	>99	>99
1.0	400	.5	35	35
1.0	400	1.0	93	93
1.0	400	2.0	>99	>99
2.0	100	.5	09	09
2.0	100	1.0	22	22
2.0	100	2.0	74	74
2.0	200	.5	11	11
2.0	200	1.0	35	35
2.0	200	2.0	93	93
2.0	400	.5	19	19
2.0	400	1.0	64	64
2.0	400	2.0	>99	>99

Table 9

Approximate Power (×100) for Interaction of Time × Mother's IQ: Compound Symmetric Σ for Time ∈ {12, 18, 36}, ρ = .440, σ^2 ∈ {119.15, 238.3, 476.6}, IHDP Ratios of Four Intervals, IHDP Spread Within

Σ multiplier	N	B multiplier	Wilks's LR	Geisser– Greenhouse
.5	100	.5	31	32
.5	100	1.0	91	93
.5	100	2.0	>99	>99
.5	200	.5	51	51
.5	200	1.0	99	>99
.5	200	2.0	>99	>99
.5	400	.5	85	85
.5	400	1.0	>99	>99
.5	400	2.0	>99	>99
1.0	100	.5	17	17
1.0	100	1.0	60	62
1.0	100	2.0	>99	>99
1.0	200	.5	26	26
1.0	200	1.0	85	85
1.0	200	2.0	>99	>99
1.0	400	.5	50	51
1.0	400	1.0	99	99
1.0	400	2.0	>99	>99
2.0	100	.5	10	10
2.0	100	1.0	31	32
2.0	100	2.0	91	93
2.0	200	.5	14	14
2.0	200	1.0	51	51
2.0	200	2.0	99	>99
2.0	400	.5	26	26
2.0	400	1.0	85	85
2.0	400	2.0	>99	>99

Table 10

Approximate Power (×100) for Interaction of Time × Mother's IQ: Compound Symmetric Σ for Time ∈ {12, 24, 36}, $\sigma^2 = 238.3$ $\rho \in \{.220, .440, .880\}$, IHDP Ratios of Four Intervals, IHDP Spread Within

ρ multiplier	N	B multiplier	Wilks's LR	Geisser– Greenhouse
.5	100	.5	13	13
.5	100	1.0	45	46
.5	100	2.0	98	99
.5	200	.5	19	19
.5	200	1.0	69	70
.5	200	2.0	>99	>99
.5	400	.5	37	37
.5	400	1.0	96	96
.5	400	2.0	>99	>99
1.0	100	.5	17	17
1.0	100	1.0	60	62
1.0	100	2.0	>99	>99
1.0	200	.5	26	26
1.0	200	1.0	85	85
1.0	200	2.0	>99	>99
1.0	400	.5	50	51
1.0	400	1.0	99	99
1.0	400	2.0	>99	>99
2.0	100	.5	68	70
2.0	100	1.0	>99	>99
2.0	100	2.0	>99	>99
2.0	200	.5	90	91
2.0	200	1.0	>99	>99
2.0	200	2.0	>99	>99
2.0	400	.5	>99	>99
2.0	400	1.0	>99	>99
2.0	400	2.0	>99	>99

Table 11

Approximate Power (×100) of W for Alternate Designs and Contrasts, N = 100, **B** and Σ based on IHDP Control Data

		Time contrast, U matrix		
Design (analysis, C matrix)	**B multiplier**	**Last time**	**Linear only**	**Linear and quadratic**
Balanced Four Intervals,	.5	83	43	36
Subjects at IHDP Means	1.0	>99	97	95
(Four Group ANOVA, Overall)	2.0	>99	>99	>99
IHDP Ratios of Four Intervals,	.5	82	40	33
Subjects at IHDP Means	1.0	>99	96	92
(Four Group ANOVA, Overall)	2.0	>99	>99	>99
Balanced Four Intervals,	.5	83	35	26
IHDP Spread Within	1.0	>99	93	84
(Cubic Polynomial, All Slopes)	2.0	>99	>99	>99
IHDP Ratios of Four Intervals,	.5	51	18	14
IHDP Spread Within	1.0	99	62	49
(Cubic Polynomial, All Slopes)	2.0	>99	>99	>99