Published in final edited form as:
Comput Stat Data Anal. 2009 January 15; 53(3): 603–608. doi:10.1016/j.csda.2008.06.010.

Sample Sizes Required to Detect Interactions between Two Binary Fixed-Effects in a Mixed-Effects Linear Regression Model

Andrew C. Leon, Ph.D[1,2] and **Moonseong Heo, Ph.D**[1]

[1] Department of Psychiatry, Weill Medical College of Cornell University

[2] Department of Public Health, Weill Medical College of Cornell University

Summary

Mixed-effects linear regression models have become more widely used for analysis of repeatedly measured outcomes in clinical trials over the past decade. There are formulae and tables for estimating sample sizes required to detect the main effects of treatment and the treatment by time interactions for those models. A formula is proposed to estimate the sample size required to detect an interaction between two binary variables in a factorial design with repeated measures of a continuous outcome. The formula is based, in part, on the fact that the variance of an interaction is fourfold that of the main effect. A simulation study examines the statistical power associated with the resulting sample sizes in a mixed-effects linear regression model with a random intercept. The simulation varies the magnitude (Δ) of the standardized main effects and interactions, the intraclass correlation coefficient (ρ), and the number (k) of repeated measures within-subject. The results of the simulation study verify that the sample size required to detect a 2×2 interaction in a mixed-effects linear regression model is fourfold that to detect a main effect of the same magnitude.

Keywords

interaction; mixed-effects linear regression; statistical power; sample size

1. Introduction

The mixed-effects linear regression model (Harville, 1977; Laird and Ware, 1982) is widely used in observational studies and randomized controlled clinical trials (RCT) in which there are repeated measures over time. In designing a study, the Ethical Guidelines of the American Statistical Association (ASA, 1999) advise statisticians to provide informed recommendations for sample size such that a research protocol will neither propose an inadequate nor an excessive number of subjects to detect a scientifically noteworthy result with acceptable statistical power. Several authors have examined the sample sizes required to detect the main effects and interaction of treatment and time in longitudinal studies with repeated measures (e.g., Hsieh, 1988; Rochon, 1991; Overall and Doyle, 1994; Hedeker, Gibbons, and Waterneaux, 1999; Raudenbush and Liu, 2001; Diggle Heagerty, Liang, and Zeger, 2002). Yet a study that is designed to detect the main effect of treatment will not have sufficient power to detect the

Corresponding Author: Andrew C. Leon, Ph.D., Weill Medical College of Cornell University, Department of Psychiatry, Box 140, 525 East 68th Street, New York, NY 10021, telephone: (212)746-3872, fax (212)746-8754, email: acleon@med.cornell.edu.

interaction between two binary fixed effects. In a 2 × 2 factorial fixed-effects ANOVA with equal cell sizes and an assumption of independence among observations, for instance, the sample size required to detect an interaction is four times that for a main effect of the same magnitude (Fleiss, 1986). However, we are not aware of formulae to estimate the sample size needed to detect an interaction between two binary fixed effects in a mixed-effects linear regression model for analysis of repeatedly measured correlated data.

The objective of this manuscript is to examine the sample size required to detect a 2 × 2 interaction of two binary fixed effects in mixed-effects linear regression analyses. The model, described in detail in Section 2, also incorporates a time-varying covariate, but that covariate does not interact with group membership. We sought to determine if, as with the fixed-effects factorial ANOVA, the sample size needed to detect an interaction in a repeated measures design is fourfold that of a main effect. A formula for the sample size required to detect an interaction is presented below. A simulation study then examines the statistical power of the resulting sample sizes to detect interactions of various magnitudes in a 2 × 2 factorial design with repeated measures of a continuous outcome.

2. Mixed-Effects Linear Regression Model and Sample Size Determination

A mixed-effects linear regression model of repeated measures of a continuous dependent variable, y_{ij}, is specified as:

$$y_{ij}=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{1i}x_{2i}+\beta_4 t_j+\upsilon_i+\varepsilon_{ij} \quad (1)$$

for subject i ($i = 1, \ldots, N$), at time j ($j = 1, \ldots, k$), where β_0 is the intercept term, x_1, represents the treatment contrast ($x_1 = -1/2$ if placebo; $x_1 = 1/2$ if investigational treatment), x_2 represents the moderator contrast ($x_2 = -1/2$ if effect moderator is absent; $x_2 = 1/2$ if effect moderator is present), x_1x_2 represents the treatment by moderator interaction. As defined by Kraemer et al., (2002), "… moderators identify on whom and under what circumstances treatments have different effects". Randomization to treatment assignment is stratified by the moderator. Note that N is the *total sample size*. Therefore $N/2$ subjects are randomized to each treatment and the sample size per cell is $N/4$ for the balanced design with two binary factors, which we consider here. The coefficients, β_1 to β_3, represent the magnitude of the corresponding main effects and interaction, t_j represents the time point of the *j-th* assessment and its coefficient β_4 represents the slope over time. This model assumes parallel slopes across treatment groups and that the slopes do not vary as a function of the moderator. These assumptions could be relaxed if either a treatment by time interaction or a treatment by moderator by time interaction were included in the model. However, here we have chosen to focus on the treatment by moderator interaction. Therefore, model (1) is an extension of the factorial fixed-effects ANOVA model, and can be described as a 2 × 2 factorial random intercept ANCOVA model with t_j as a time-varying covariate.

The subject-specific random intercept υ_i is assumed to be distributed $N(0, \sigma_\upsilon^2)$, and the conditional distribution of error term ε_{ij} for a given υ_i is assumed to be independent and identical with $N(0, \sigma_\varepsilon^2)$ across time points j within the i-th subject. The marginal distributions of υ_i and ε_{ij} are assumed to be mutually independent, that is $\mathrm{Cov}(\varepsilon, \upsilon) = 0$. It follows from those conditional and mutual independence assumptions that $Var(Y_{ij}) \equiv \sigma^2=\sigma_\upsilon^2+\sigma_\varepsilon^2$ and $\mathrm{corr}(Y_{ij}, Y_{ij'}) \equiv \rho=\sigma_\upsilon^2/\sigma^2=\sigma_\upsilon^2/(\sigma_\upsilon^2+\sigma_\varepsilon^2)$, the intraclass correlation coefficient (ICC), for $j \neq j'$. The standardized effects of β_1 to β_3 can be quantified as $\Delta_m = \beta_m/\sigma$, $m = 1,2,3$.

The variance of the estimated interaction is four times that of estimated main effect in the factorial fixed-effects ANOVA (section 4.2 in Fleiss, 1986). That relation also holds for the 2

× 2 factorial random intercept ANCOVA model (1) that we are considering here, since neither $Var(Y_{ij}) = \sigma^2$ nor the correlation, ρ, depends on subject i or time point j. Specifically, the following holds:

$$Var(\widehat{\beta}_1)=Var(\widehat{\beta}_2)=Var(\widehat{\beta}_3)/4$$

and therefore

$$Var(\widehat{\Delta}_1)=Var(\widehat{\Delta}_2)=Var(\widehat{\Delta}_3)/4, \quad (2)$$

where $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$, are corresponding maximum likelihood estimates of β_1, β_2, and β_3. It follows that the sample size needed to detect an interaction effect will be four times that for detecting a main effect of the identical magnitude because the sample size is a linear function of the variance of an effect estimate.

The total number of subjects, say $N(\Delta_1)$, required to detect a main effect with power $1-\beta$ (where β is the level of type II error) was presented elsewhere (Donner et al., 1981; Donner and Klar, 2000; Diggle et al., 2002):

$$N(\Delta_1)=\frac{4(z_{\alpha/2}+z_\beta)^2(1+(k-1)\rho)\sigma^2}{k\beta_1^2}=\frac{4(z_{\alpha/2}+z_\beta)^2(1+(k-1)\rho)}{k\Delta_1^2} \quad (3)$$

It follows that $N(\Delta_1) = N(\Delta_2)$ for $\Delta_1 = \Delta_2$. However, for effects of the same magnitude, $\Delta_1 = \Delta_3$, the total number of subjects, say $N(\Delta_3)$, required to detect an interaction effect with power $1-\beta$ can then be expressed as fourfold that of the main effect. Finally, combining the sample size determination (3) for the main effect with the fourfold increase in the variance of the mle of the interaction effect of interest (2), we propose the following for sample size determination for detecting the interaction:

$$N(\Delta_3)=\frac{16(z_{\alpha/2}+z_\beta)^2(1+(k-1)\rho)\sigma^2}{k\beta_3^2}=\frac{16(z_{\alpha/2}+z_\beta)^2(1+(k-1)\rho)}{k\Delta_3^2}=4N(\Delta_1). \quad (4)$$

3. Simulation Study

The primary focus of this simulation study was to examine whether the statistical power to detect an interaction of two fixed effects in a 2 × 2 factorial design with repeated measures of a continuous outcome in model (1) is consistent with the sample sizes derived from (4). The statistical power to detect a main effect with the sample sizes derived from (3) was also examined. A Wald test with a two-tailed alpha-level of .05 was used to test each of two hypotheses:

$$H_{01}:\beta_1=0$$
$$H_{02}:\beta_3=0.$$

The simulations were specified such that the magnitude of either one main effect (Δ_1) or the interaction (Δ_3) ranged from 0.20 to 0.50 and the remaining two effects were null. Thus the results of the interaction (Δ_3) and only one main effect (Δ_1) will be discussed hereafter.

3.1. Simulations Specifications

The simulation was designed by varying following specifications:

1. Main effect, β_1, specified as standardized effects (Δ_1): .20, .25, .30, .35, .40, .45, .50
2. Interaction, β_3, specified as standardized effects (Δ_3): .20, .25, .30, .35, .40, .45, .50
3. Intraclass correlation coefficient (ICC) ρ : .20, .40, .60
4. Repeated measures, within subject, over time (k): 4, 6, 8
5. Total number of subjects, $N(\Delta_1)$, based on equation (3), to detect the respective main effects (Δ_1) with 80%, 90%, and 95% power
6. The total number of subjects, $N(\Delta_3)$, to detect the respective interactions (Δ_3) with 80% 90%, and 95% power, based on equation (4).

3.2. Data Generation

The simulated outcome variable for the four treatment by moderator cells was generated as a time-varying continuous variable (Y_{ij}) based on normal distributions. Specifically, we first generated from $N(0, \sigma_\varepsilon^2)$ and then for given υ_i we independently generated ε_{ij} from $N(0, \sigma_\upsilon^2)$. Those simulated random values were then added to the respective fixed main effect and interaction. As specified above, the magnitude of either the main effect (Δ_1) or the interaction of the two binary fixed effects (Δ_3) ranged from 0.20 to 0.50. For each of 63 combinations of simulation specifications for the interaction ($7\Delta_3 \times 3\rho \times 3k$) for each level of power, 6000 data sets were generated. Similarly, 6000 data sets were generated for each of 63 combinations of simulation specifications for the main effect ($7\Delta_1 \times 3\rho \times 3k$) for each level of power. We chose to generate 6000 data sets per combination of specifications based on the precision of the resulting power estimates. Specifically, based on 6000 simulations, the 95% confidence interval for 80% power ranges from 0.789 to 0.810, for 90% power it ranges from 0.892 to 0.908, and for 95% power it ranges from .945 to .956.

3.3. Evaluation of Statistical Power

For each data set, model *(1)* was fit to the simulated outcome data using the S-plus routine "lme" with maximum likelihood (ML) method and *p*-values for the effects were retained for estimation of empirical power. Specifically, the empirical statistical power was defined as the proportion of the 6000 analyses per simulation specification in which the null hypothesis was rejected at a two-tailed alpha-level of .05. S-plus 7.0 was used for all computations.

4. Simulation Results

Empirical power estimates for each specification of the main effect models (Table 1 for 80% power; Table 2 for 90% power; Table 3 for 95% power) are consistent with the sample size N (Δ_1) calculation based on equation (3). Furthermore, the required sample sizes N(Δ_3) for an interaction are indeed fourfold that of a main effect of the same magnitude. For example, for 80% power, with ρ = 0.20 and k=4 observations per subject, N(Δ_3)=808 subjects in total (or 202/cell) are needed for power of 80% to detect an interaction effect (Δ_3) of .25; N(Δ_3)=560 subjects are needed for Δ_3=0.30, 320 subjects for Δ_3=0.40 and N(Δ_3)=208 subjects for Δ_3=0.50. Similar patterns hold for ρ = 0.40, 0.60 and k = 6, 8, as shown in Table 1, yet the required sample sizes increase with greater ρ . The required N(Δ_3)'s are fourfold N(Δ_1) for the main effects for all values of k, Δ and ρ, For example, the corresponding sample size for a main effect with ρ = 0.20 and k=4 are $N(\Delta_1)$=202 (Δ_1=0.25), $N(\Delta_1)$=140 (Δ_1=0.30), $N(\Delta_1)$=80 (Δ_1=0.40) and $N(\Delta_1)$=52 (Δ_1=0.50). The same relation holds true for power of .90 (Table 2) and .95 (Table 3). Thus, a multiplicative factor of four can be used to estimate the required

sample size for an interaction effect, given the $N(\Delta_1)$ for a main effect of the same magnitude based on the equation (3).

5. Application

There is a recent NIH initiative (NIH: RFA-MH-09-010) to identify personalized treatments by designing clinical trials that test not only the effect of treatment, but moderators of the treatment effect. The goal of such a trial would be to test whether an hypothesized subject characteristic (i.e., the moderator) is associated with enhanced or inhibited treatment response. In either case, a treatment by moderator could test an important clinical question, in that it would help the clinician provide a targeted intervention to patients in need.

Consider, for example, an RCT of an antidepressant that is hypothesized to be more effective in the subgroup of subjects who carry the short allele of the serotonin transporter gene polymorphism (5-HTTLPR). Subjects meeting criteria for major depressive disorder will be randomized to either fluoxetine or placebo and evaluated weekly with the Quick Inventory of Depressive Symptomatology-Self-Rated (QIDS-SR; Rush et al., 2003) over a 6 week trial (k=6). The sample will be equally divided by recruiting half of the subjects having the short allele and the other half without the short allele. Randomization will then stratified by allelic variation. The study will be designed to detect an interaction effect as small as Δ_3=0.35. For example, that would represent a difference in response between the two allele groups, within a treatment cell, of about one-third of a standard deviation on the QIDS-SR, which will represent about 6 points, or a clinically meaningful effect. The total sample size required for power of 80% will vary with the intraclass correlation coefficient: $N(\Delta_3)$ =344 (ρ =0.20), $N(\Delta_3)$=520 (ρ =0.40), and $N(\Delta_3)$=688 (ρ =0.60). In contrast, the total sample size for power of 90% is $N(\Delta_3)$ =464 (ρ =0.20), $N(\Delta_3)$=688 (ρ =0.40), and $N(\Delta_3)$=920 (ρ =0.60) and, for power of .95%, $N(\Delta_3)$ =568 (ρ =0.20), $N(\Delta_3)$=856 (ρ =0.40), and $N(\Delta_3)$=1136 (ρ =0.60).

6. Discussion

This simulation study examined required sample sizes for the main effects and interaction of two binary fixed effects in a mixed-effects linear regression model with a random intercept. The results indicate that, for a given set of design specifications, four times as many subjects are required to detect an interaction as for a main effect, as specified in our formula (4). The formula was verified by simulation for 80%, 90%, and 95% statistical power. This relationship did not depend on the standardized effect size Δ_m, the number of observations per subject k, or the intraclass correlation coefficient ρ.

The simulation results indicate that required sample sizes for the main effect were in accord with estimates based on equation (3). It is worth noting that linear interpolation of $N(\Delta_3)$ appears to be accurate across ICCs, for a given k and Δ_3. However, interpolation is not warranted across Δ3's or k's.

The simulation study examined statistical power of the interaction of two binary fixed effects in a mixed-effects linear regression model with a random intercept. Equation (4) does not necessarily apply to a model with a random slope. Furthermore we did not examine the required sample size in the presence of a treatment by time interaction or a treatment by moderator by time interaction. Similarly, the results presented here do not apply to sample sizes needed to detect interactions among categorical covariates with more than two levels. An investigation into that issue would involve a likelihood ratio test, not the normal approximation that was used here.

An RCT that is specifically designed to test a treatment by moderator interaction could yield valuable information to guide clinical decision making regarding appropriate interventions for

subgroups of those with the diagnosis of interest. However, given the sheer number of subjects that is needed to detect that interaction, a researcher might consider an alternative design. For instance, if the objective of a study is to demonstrate efficacy in a particular subgroup, one that has been identified in preliminary research, the RCT inclusion criteria might be designated to enroll only that subgroup. Thus the focus would no longer be on a moderating effect, but instead on treatment of a group of particular interest.

The results of this simulation study provide sample size estimates for statistical power of 80%, 90%, and 95% to detect various standardized main effects and interactions between two binary fixed effects in a mixed-effects linear regression model with a random intercept. The range of the magnitude of those effects, the number of repeated observations, and the ρ 's should be useful for broad application. However, because the sample size required to detect an interaction is four times that of a main effect, equations (3) and (4) can be used to estimate sample size for research designs with specifications that were not examined here.

Acknowledgments

This research was supported, in part, by grants from the National Institute Health (MH060447 and MH068638).

References

American Statistical Association. Ethical guidelines for statistical practice: Executive summary. Amstat News. 1999 April, 12–15;

Diggle, PJ.; Heagerty, P.; Liang, K-Y.; Zeger, SL. Analysis of Longitudinal Data. Vol. 2. Oxford: Oxford University Press; 2002.

Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. American Journal of Epidemiology 1981;114:906–914. [PubMed: 7315838]

Donner, A.; Klar, N. Design and Analysis of Cluster Randomization Trials in Health Research. London: Arnold; 2000.

Fleiss, JL. The Design and Analysis of Clinical Experiments. NY: Wiley and Sons; 1986.

Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 1977;72:320–340.

Hedeker D, Gibbons RD, Waternaux C. Sample size estimation for longitudinal designs with attrition: comparing time-related contrasts between two groups. Journal of Educational and Behavioral Statistics 1999;24:70–93.

Hsieh FY. Sample size formulae for intervention studies with the cluster as unit of randomization. Stat Med 1988;7:1195–201. [PubMed: 3201045]

Kraemer HC, Wilson T, Fairburn CG, Agras WS. CG et al: Mediators and moderators of treatment effects in randomized clinical trials. Arch Gen Psychiatry 2002;59:877–883. [PubMed: 12365874]

Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963–974. [PubMed: 7168798]

Overall JE, Doyle SR. Estimating sample sizes for repeated measurement designs. Control Clin Trials 1994;15:100–23. [PubMed: 8205802]

Raudenbush SW, Liu X. Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. Psychological Methods 2001;6:387–401. [PubMed: 11778679]

Rochon J. Sample size calculations for two-group repeated-measures experiments. Biometrics 1991;47:1383–1398.

Rush AJ, Trivedi MH, Ibrahim HM, et al. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 2003;54:573–83. [PubMed: 12946886]

Table 1

Sample Size Required for Theoretical Statistical Power of 80% to Detect the Main Effect and the Interaction of Two Binary Fixed Effects in a Mixed-Effects Linear Regression Model with a Random Intercept

		***k* = 4**				***k* = 6**				***k* = 8**			
		Main Effect		**Interaction**		**Main Effect**		**Interaction**		**Main Effect**		**Interaction**	
ICC (ρ)	**Standardized Effect (Δ_m)**	**N(Δ_1)**	**Empirical Power**	**N(Δ_3)**	**Empirical Power**	**N(Δ_1)**	**Empirical Power**	**N(Δ_3)**	**Empirical Power**	**N(Δ_1)**	**Empirical Power**	**N(Δ_3)**	**Empirical Power**
0.20	**.20**	314	0.808	1256	0.796	262	0.808	1048	0.803	236	0.803	944	0.798
	.25	202	0.796	808	0.804	168	0.806	672	0.801	152	0.809	608	0.799
	.30	140	0.806	560	0.801	118	0.802	472	0.814	106	0.813	424	0.815
	.35	104	0.813	416	0.815	86	0.811	344	0.795	78	0.810	312	0.803
	.40	80	0.796	320	0.800	66	0.791	264	0.800	60	0.801	240	0.811
	.45	64	0.811	256	0.810	52	0.799	208	0.811	48	0.815	192	0.814
	.50	52	0.817	208	0.809	42	0.798	168	0.804	38	0.798	152	0.799
0.40	**.20**	432	0.798	1728	0.795	394	0.788	1576	0.795	374	0.807	1496	0.799
	.25	278	0.805	1112	0.812	252	0.805	1008	0.803	240	0.803	960	0.808
	.30	192	0.797	768	0.801	176	0.805	704	0.803	166	0.806	664	0.798
	.35	142	0.796	568	0.803	130	0.804	520	0.811	122	0.801	488	0.811
	.40	108	0.808	432	0.798	100	0.808	400	0.816	94	0.808	376	0.799
	.45	86	0.808	344	0.808	78	0.807	312	0.799	74	0.805	296	0.797
	.50	70	0.804	280	0.808	64	0.810	256	0.806	60	0.794	240	0.805
0.60	**.20**	550	0.796	2200	0.797	524	0.817	2096	0.796	512	0.810	2048	0.798
	.25	352	0.798	1408	0.793	336	0.797	1344	0.802	328	0.800	1312	0.802
	.30	246	0.799	984	0.808	234	0.804	936	0.808	228	0.801	912	0.803
	.35	180	0.799	720	0.803	172	0.800	688	0.800	168	0.803	672	0.806
	.40	138	0.798	552	0.807	132	0.801	528	0.801	128	0.794	512	0.800
	.45	110	0.811	440	0.806	104	0.800	416	0.812	102	0.801	408	0.803
	.50	88	0.809	352	0.797	84	0.801	336	0.801	82	0.809	328	0.808

Notes:

[1] *k* represents the number of observations per subject.

[2] The sample sizes required to detect a main effect N(Δ_1) or an interaction N(Δ_3) represent the total sample size, based on equations (3) and (4), respectively and assume power of 80% and a two-tailed alpha-level of .05.

[3] Empirical power is based on analyses of 6000 simulated data sets for each combination of parameter specifications.

Table 2

Sample Size Required for Theoretical Statistical Power of 90% to Detect the Main Effect and the Interaction of Two Binary Fixed Effects in a Mixed-Effects Linear Regression Model with a Random Intercept

		k=4				*k*=6				*k*=8			
		Main Effect		Interaction		Main Effect		Interaction		Main Effect		Interaction	
ICC (ρ)	**Standardized Effect (Δ_m)**	**N(Δ_1)**	**Empirical Power**	**N(Δ_3)**	**Empirical Power**	**N(Δ_1)**	**Empirical Power**	**N(Δ_3)**	**Empirical Power**	**N(Δ_1)**	**Empirical Power**	**N(Δ_3)**	**Empirical Power**
0.20	**.20**	422	0.897	1688	0.890	352	0.900	1408	0.903	316	0.898	1264	0.903
	.25	270	0.903	1080	0.902	226	0.897	904	0.896	202	0.895	808	0.898
	.30	188	0.903	752	0.902	156	0.901	624	0.905	142	0.905	568	0.909
	.35	138	0.897	552	0.902	116	0.901	464	0.905	104	0.904	416	0.902
	.40	106	0.897	424	0.905	88	0.900	352	0.903	80	0.901	320	0.900
	.45	84	0.902	336	0.900	70	0.897	280	0.902	64	0.911	256	0.913
	.50	68	0.907	272	0.902	58	0.912	232	0.915	52	0.910	208	0.919
0.40	**.20**	578	0.897	2312	0.909	526	0.899	2104	0.902	500	0.902	2000	0.902
	.25	370	0.894	1480	0.907	338	0.900	1352	0.899	320	0.905	1280	0.902
	.30	258	0.896	1032	0.907	234	0.901	936	0.902	222	0.902	888	0.902
	.35	190	0.907	760	0.897	172	0.900	688	0.899	164	0.894	656	0.900
	.40	146	0.905	584	0.899	132	0.903	528	0.903	126	0.905	504	0.898
	.45	116	0.904	464	0.907	104	0.902	416	0.904	100	0.906	400	0.900
	.50	94	0.904	376	0.901	86	0.909	344	0.906	80	0.898	320	0.900
0.60	**.20**	736	0.901	2944	0.893	702	0.907	2808	0.907	684	0.899	2736	0.897
	.25	472	0.903	1888	0.898	450	0.897	1800	0.914	438	0.901	1752	0.903
	.30	328	0.895	1312	0.903	312	0.900	1248	0.900	304	0.895	1216	0.889
	.35	242	0.905	968	0.902	230	0.901	920	0.904	224	0.901	896	0.902
	.40	184	0.899	736	0.907	176	0.904	704	0.898	172	0.900	688	0.904
	.45	146	0.902	584	0.899	140	0.908	560	0.906	136	0.905	544	0.907
	.50	118	0.901	472	0.894	114	0.906	456	0.905	110	0.908	440	0.903

Notes:

[1] *k* represents the number of observations per subject.

[2] The sample sizes required to detect a main effect N(Δ_1) or an interaction N(Δ_3) represent the total sample size, based on equations (3) and (4), respectively and assume power of 90% and a two-tailed alpha-level of .05.

[3] Empirical power is based on analyses of 6000 simulated data sets for each combination of parameter specifications.

Table 3

Sample Size Required for Theoretical Statistical Power of 95% to Detect the Main Effect and the Interaction of Two Binary Fixed Effects in a Mixed-Effects Linear Regression Model with a Random Intercept

		k=4				*k*=6				*k*=8			
		Main Effect		Interaction		Main Effect		Interaction		Main Effect		Interaction	
ICC (ρ)	**Standardized Effect (Δ_m)**	**N(Δ_1)**	**Empirical Power**	**N(Δ_3)**	**Empirical Power**	**N(Δ_1)**	**Empirical Power**	**N(Δ_3)**	**Empirical Power**	**N(Δ_1)**	**Empirical Power**	**N(Δ_3)**	**Empirical Power**
0.20	**.20**	520	0.953	2080	0.944	434	0.952	1736	0.948	390	0.947	1560	0.950
	.25	334	0.948	1336	0.953	278	0.953	1112	0.949	250	0.947	1000	0.953
	.30	232	0.951	928	0.954	194	0.955	776	0.949	174	0.951	696	0.953
	.35	170	0.954	680	0.951	142	0.954	568	0.949	128	0.949	512	0.949
	.40	130	0.954	520	0.950	110	0.956	440	0.948	98	0.950	392	0.953
	.45	104	0.952	416	0.956	86	0.954	344	0.951	78	0.955	312	0.954
	.50	84	0.951	336	0.947	70	0.945	280	0.955	64	0.957	256	0.957
0.40	**.20**	716	0.952	2864	0.950	650	0.956	2600	0.952	618	0.946	2472	0.950
	.25	458	0.952	1832	0.947	416	0.948	1664	0.952	396	0.948	1584	0.953
	.30	318	0.947	1272	0.951	290	0.954	1160	0.949	276	0.946	1104	0.949
	.35	234	0.952	936	0.952	214	0.954	856	0.950	202	0.952	808	0.952
	.40	180	0.948	720	0.948	164	0.951	656	0.952	156	0.955	624	0.954
	.45	142	0.949	568	0.952	130	0.950	520	0.950	122	0.950	488	0.952
	.50	116	0.952	464	0.952	104	0.955	416	0.956	100	0.956	400	0.952
0.60	**.20**	910	0.942	3640	0.953	868	0.950	3472	0.953	846	0.953	3384	0.952
	.25	584	0.950	2336	0.952	556	0.949	2224	0.946	542	0.952	2168	0.952
	.30	406	0.956	1624	0.951	386	0.950	1544	0.946	376	0.948	1504	0.950
	.35	298	0.953	1192	0.942	284	0.957	1136	0.960	276	0.946	1104	0.951
	.40	228	0.950	912	0.952	218	0.951	872	0.952	212	0.953	848	0.948
	.45	180	0.951	720	0.946	172	0.948	688	0.949	168	0.949	672	0.955
	.50	146	0.948	584	0.952	140	0.952	560	0.953	136	0.952	544	0.955

Notes:

[1] *k* represents the number of observations per subject.

[2] The sample sizes required to detect a main effect N(Δ_1) or an interaction N(Δ_3) represent the total sample size, based on equations (3) and (4), respectively and assume power of 95% and a two-tailed alpha-level of .05.

[3] Empirical power is based on analyses of 6000 simulated data sets for each combination of parameter specifications.