Unbalanced Sampling Effect on the Power at Level-1 in the Random Coefficient Model

Bonnie J. Steele Colorado Mountain College
Daniel J. Mundfrom Eastern Kentucky University
Jamis Perrett Texas A & M University

Researchers often disregard the potentially negative effects of unbalanced sampling on power estimates when using multilevel models. The purpose of this study was to investigate the effects that unbalanced sampling had on the estimated level-one power in multilevel random coefficient models. Twelve combinations of three effect sizes (0.5, 0.8, and 1.0) and four intraclass correlations (0.2, 0.1, 0.05, and 0.01) were investigated with each of three sampling ratios (0.25:0.75, 0.20:0.80, and 0.15:0.85) and three sample sizes (200, 500, and 800) to compare the effects that the different sampling ratios had on the level-1 power in the random coefficient model. Results indicated that as sampling ratios changed from 0.25:0.75, to incrementally a larger unbalanced sampling ratio of 0.15:0.85, the estimated power was lower in almost every case. This effect was more pronounced for the smaller sample sizes. Fourteen cases displayed differences larger than 5% in aggregate power estimates.

Hierarchical Linear Modeling (HLM) is a derivative or extension of the standard regression model adapted to address the problem of multilevel data, which allows the researcher to confront restrictions previously imposed by single-level analyses (Heck & Thomas, 2000). As a widely utilized technique, HLM has a rich literature containing recommendations regarding the appropriate balanced sample sizes necessary to ensure adequate power in simultaneous variation testing for both within-groups and between-group(s) comparisons (Kreft & De Leeuw, 1998; Raudenbush & Bryk, 2002; Raudenbush & Liu, 2000). For example, the use of HLM with balanced sample sizes is cited in mental health research (Bond, Miller, Krumweid, & Ward, 1988), education (Finn & Achilles, 1990; Mosteller, 1995), and medicine (Haddow, 1991). However, the literature pertaining to unbalanced sample size recommendations in HLM is meager (Raudenbush & Liu). This study builds on the work from previous balanced research perspectives by providing insight into the effect that unbalanced sampling has on the power estimates at level-1 in the random coefficient model with three dissimilar conditions of effect size and four intraclass correlations.

Kraemer and Thiemann (1987) broadly summarized and discussed the effects of sampling in a number of single level models. They found that small differences in sample sizes across groups for single-level analyses may not lower the estimated power of a test, but larger differences become problematic, indicating that unbalanced sampling tends to lower the model's expected power. Larger differences are those with proportional sampling differences that can incrementally differ by as much as 75% or as little as 25%. Such unbalanced sampling occurs as the rule rather than the exception in many educational settings where, for example, several classes are sampled to obtain a sample size of 200. Suppose that, in a particular school, each of 5 sampled classes have 10 students and each of 10 other sampled classes have 15 students. With sampling that is unbalanced to this extent, it would not be unreasonable to expect the same decreasing effect on power in multilevel models as is seen in single-level models.

Raudenbush and Liu (2000) provided a comprehensive summary of expected power estimate calculations based upon parameter estimates, balanced sample sizes, and overall resource expense. Unbalanced sampling, on the other hand, was only minimally addressed as a focus of suggested further research. Likewise, Reise and Duan (2003) suggested that the unbalanced nature of educational data produced design flaws in need of further research to investigate its effects on model efficiency.

Method

Building on the work of Kraemer and Thiemann (1987) and Raudenbush and Liu (2000), three different overall sample sizes (as suggested by Raudenbush and Bryk, 2002) with three unbalanced sampling ratios (as suggested by Kraemer and Thiemann) were investigated in this study, with the focus on the differences in the amount of unbalanced sampling being used to determine if there was a recognizable effect on model power. Other model conditions that were varied included effect size and intraclass correlations (ICC) as possible contributors to decreases in power with multilevel models. The overall resource expense ratio was held to 1.

Sampling Schemes

Three levels of proportionally unbalanced data (75% to 25%, 80% to 20%, and 85% to 15%) were calculated for three different sample sizes of 200, 500, and 800 with each of 12 combinations of effect size and intraclass correlation. Design conditions were limited by the restriction that the number of classes (C_1) times the number of students in each of those classes (S_1) at the first sampling proportion (e.g., 25%) plus the number of classes (C_2) times the number of students in each of those classes (S_2) at the second sampling proportion (e.g., 75%) must equal the desired sample size (N), i.e., $(C_1S_1) + (C_2S_2) = N$, where in the 25%-75% sampling ratio, (C_2S_2) must be 3 times larger than (C_1S_1). This algebraic equation was used to generate 312 possible sampling ratios that would reflect possible classroom scenarios, where in each case the C_1, S_1, C_2, and S_2 values were integers. To be specific, 84 possible sampling combinations were used that corresponded to a total sample size of 200, another 84 possible sampling combinations were used with a total sample size of 500, and 144 possible sampling combinations were used with a total sample size of 800. Whereas Schumacker and Lomax (1996) (as cited in Heck & Thomas, 2000) provided a rule of thumb suggestion that a minimum of 100-150 subjects be included in a study, Heck and Thomas considered anything with $N < 400$ to be a small sample. Sample sizes, sampling schemes, and sampling ratio differences used in this study are displayed in Table 1 where a sample size of $N = 200$ is used as a representative of a small sample, $N = 500$ to represent a moderate sample size, and $N = 800$ to represent a large sample.

Table 1. Possible Sampling Combinations Equal to Unbalanced Samples of 200, 500, and 800.

Ratio	Classes Trt 1	Subjects Trt 1	Classes Trt 2	Subjects Trt 2
	$N = 200$			
.25-.75	5	10	10	15
	5	10	5	30
.20-.80	2	20	10	16
	4	10	10	16
	2	20	5	32
	4	10	5	32
.15-.85	2	15	10	17
	$N = 500$			
.25-.75	5	25	15	25
.20-.80	4	25	20	20
	10	10	20	20
	4	25	16	25
	10	10	16	25
.15-.85	3	25	17	25
	5	15	17	25
	$N = 800$			
.25-.75	8	25	24	25
	10	20	24	25
	8	25	30	20
	10	20	30	20
.20-.80	10	16	20	32
	10	16	40	16
	5	32	20	32
	5	32	40	16
.15-.85	4	30	17	40
	4	30	20	34
	5	24	17	40
	5	24	20	34

Data Simulation

After the unbalanced sampling schemes with corresponding sample sizes were determined, Step 1 of the simulation began. Ten thousand outcome variables were simulated for each of the 312 different sampling schemes using the SAS PROC IML (see the Appendix). These outcome variables were mechanically constrained to fit within given values for effect size, intraclass correlations, sample sizes, and proportions of unbalanced data. Step 2, performing an HLM analysis on each of the 10,000 iterations of the 312 sampling schemes using SAS PROC MIXED, produced the partitioned level-1 and level-2 power parameters. The 312 possible sampling combinations were grouped and aggregated according to sampling schema to generate 108 estimated level-1 power values.

Upon completion of the simulations, a comparison table was created where the effect of each level of each design characteristic (i.e., sample size, proportion of unbalanced data, effect size, and intraclass correlation) on model power could be investigated on the resultant dependent variable (the calculated estimate of level-1 power) for the simulated unbalanced random coefficient model. Visual comparisons were made.

Results and Conclusions

Simulated data were created following the recommendations of Raudenbush and Liu (2000) for relative magnitudes of effect size, intraclass correlation, total sample size, and proportions of unbalanced data. The 312 possible sampling ratios and 108 level-1 aggregate power estimates from Step 2 are presented in Tables 2A – 4C. Power estimates when N = 200 are in Tables 2A – 2B, with N = 500 in Tables 3A – 3C, and for N = 800 in Tables 4A – 4C.

Table 2A. Aggregate Power for Sample Size of 200 for Sampling Ratios 0.25:0.75 and 0.20:0.80 by Four ICCs & Three Effect Sizes

Ratio	ICC	ES	C1	S1	C2	S2	Power	Average
0.25:0.75	0.2	1.0	5	10	10	15	0.9639	
			5	10	5	30	0.9686	0.9663
		0.8	5	10	10	15	0.8977	
			5	10	5	30	0.8391	0.8684
		0.5	5	10	10	15	0.6643	
			5	10	5	30	0.5889	0.6266
	0.1	1.0	5	10	10	15	0.9945	
			5	10	5	30	0.9939	0.9942
		0.8	5	10	10	15	0.9612	
			5	10	5	30	0.9591	0.9602
		0.5	5	10	10	15	0.7322	
			5	10	5	30	0.7417	0.7370
	0.05	1.0	5	10	10	15	0.9986	
			5	10	5	30	0.9887	0.9937
		0.8	5	10	10	15	0.9838	
			5	10	5	30	0.9819	0.9829
		0.5	5	10	10	15	0.7837	
			5	10	5	30	0.6363	0.7100
	0.01	1.0	5	10	10	15	0.9999	
			5	10	5	30	0.9970	0.9985
		0.8	5	10	10	15	0.9612	
			5	10	5	30	0.9627	0.9620
		0.5	5	10	10	15	0.8483	
			5	10	5	30	0.6604	0.7544
0.20:0.80	0.2	1.0	2	20	10	16	0.8577	
			4	10	10	16	0.9381	
			2	20	5	32	0.8823	0.9066
			4	10	5	32	0.9484	
		0.8	2	20	10	16	0.7725	
			4	10	10	16	0.8614	
			2	20	5	32	0.7950	0.8246
			4	10	5	32	0.8693	
		0.5	2	20	10	16	0.5803	
			4	10	10	16	0.6239	
			2	20	5	32	0.6200	0.6181
			4	10	5	32	0.6482	

Table 2B. Aggregate Power for Sample Size of 200, 0.20:0.80 and 0.15:0.85 Sampling Ratios by Four ICCs & Three Effect Sizes

Ratio	ICC	ES	C1	S1	C2	S2	Power	Average
0.20:0.80	0.1	1.0	2	20	10	16	0.9545	0.9704
			4	10	10	16	0.9826	
			2	20	5	32	0.9576	
			4	10	5	32	0.9867	
		0.8	2	20	10	16	0.8526	0.9007
			4	10	10	16	0.9299	
			2	20	5	32	0.8767	
			4	10	5	32	0.9437	
		0.5	2	20	10	16	0.6001	0.6475
			4	10	10	16	0.6669	
			2	20	5	32	0.6317	
			4	10	5	32	0.6912	
	0.05	1.0	2	20	10	16	0.9821	0.9913
			4	10	10	16	0.9969	
			2	20	5	32	0.9899	
			4	10	5	32	0.9962	
		0.8	2	20	10	16	0.9304	0.9503
			4	10	10	16	0.9668	
			2	20	5	32	0.9355	
			4	10	5	32	0.9686	
		0.5	2	20	10	16	0.6585	0.6974
			4	10	10	16	0.7138	
			2	20	5	32	0.6813	
			4	10	5	32	0.7358	
	0.01	1.0	2	20	10	16	0.9982	0.9991
			4	10	10	16	0.9997	
			2	20	5	32	0.9991	
			4	10	5	32	0.9992	
		0.8	2	20	10	16	0.9824	0.9871
			4	10	10	16	0.9905	
			2	20	5	32	0.9854	
			4	10	5	32	0.9902	
		0.5	2	20	10	16	0.7453	0.7680
			4	10	10	16	0.7784	
			2	20	5	32	0.7585	
			4	10	5	32	0.7896	
0.15:0.85	0.2	1.0	2	15	10	17	0.8522	0.8522
		0.8	2	15	10	17	0.7469	0.7469
		0.5	2	15	10	17	0.5471	0.5471
	0.1	1.0	2	15	10	17	0.9357	0.9357
		0.8	2	15	10	17	0.8274	0.8274
		0.5	2	15	10	17	0.5561	0.5561
	0.05	1.0	2	15	10	17	0.9697	0.9697
		0.8	2	15	10	17	0.9063	0.9063
		0.5	2	15	10	17	0.6086	0.6086
	0.01	1.0	2	15	10	17	0.9958	0.9958
		0.8	2	15	10	17	0.9658	0.9658
		0.5	2	15	10	17	0.6737	0.6737

In general, with *N* = 200, adequate average power was achieved with effect sizes equal to 1.0 and 0.8 for all three sampling ratios and all four ICC values (with 3 exceptions—sampling ratio of 0.20:0.80, ICC = 0.2, ES = 0.8; sampling ratio of 0.15:0.85, ICC = 0.2, ES = 0.8; and sampling ratio of 0.15:0.85, ICC = 0.1, ES = 0.8). None of the scenarios with effect size = 0.5 showed adequate average power.

With *N* = 500, adequate average power was achieved for all three ratios of unbalanced sampling, all four ICC values, and all three effect sizes with four exceptions each with effect size = 0.5: sampling ratio of 0.25:0.75, ICC = 0.2; sampling ratio of 0.20:0.80, ICC = 0.2; sampling ratio of 0.15:0.85, ICC = 0.2; and sampling ratio of 0.15:0.85, ICC = 0.1.

Table 3A. Aggregate Power for Sample Size of 500, 0.25:0.75 and 0.20:0.80 Sampling Ratios by Four ICCs & Three Effect Sizes

Ratio	ICC	ES	C1	S1	C2	S2	Power	Average
0.25:0.75	0.2	1.0	5	25	15	25	0.9943	0.9943
		0.8					0.9598	0.9598
		0.5					0.7886	0.7886
	0.1	1.0	5	25	15	25	0.9998	0.9998
		0.8					0.9967	0.9967
		0.5					0.8733	0.8733
	0.05	1.0	5	25	15	25	1.0	1.0
		0.8					0.9996	0.9996
		0.5					0.9522	0.9522
	0.01	1.0	5	25	15	25	1.0	1.0
		0.8					1.0	1.0
		0.5					0.9929	0.9929
0.20:0.80	0.2	1.0	4	25	20	20	0.9745	0.9885
			10	10	20	20	0.9997	
			4	25	16	25	0.9804	
			10	10	16	25	0.9994	
		0.8	4	25	20	20	0.9193	0.9569
			10	10	20	20	0.9884	
			4	25	16	25	0.9338	
			10	10	16	25	0.9861	
		0.5	4	25	20	20	0.7351	0.8016
			10	10	20	20	0.8619	
			4	25	16	25	0.7512	
			10	10	16	25	0.8581	
	0.1	1.0	4	25	20	20	0.9985	0.9991
			10	10	20	20	1.0	
			4	25	16	25	0.9979	
			10	10	16	25	1.0	
		0.8	4	25	20	20	0.9802	0.9910
			10	10	20	20	0.9997	
			4	25	16	25	0.9854	
			10	10	16	25	0.9985	
		0.5	4	25	20	20	0.8234	0.8840
			10	10	20	20	0.9346	
			4	25	16	25	0.8427	
			10	10	16	25	0.9351	

Table 3B. Aggregate Power for *N* = 500, 0.20:0.80 Sampling Ratio by 2 ICCs & 3 Effect Sizes

Ratio	ICC	ES	C1	S1	C2	S2	Power	Average
0.20:0.80	0.05	1.0	4	25	20	20	1.0	1.0
			10	10	20	20	1.0	
			4	25	16	25	0.9999	
			10	10	16	25	1.0	
		0.8	4	25	20	20	0.9988	0.9991
			10	10	20	20	0.9999	
			4	25	16	25	0.9980	
			10	10	16	25	0.9998	
		0.5	4	25	20	20	0.9051	0.9388
			10	10	20	20	0.9663	
			4	25	16	25	0.9144	
			10	10	16	25	0.9694	
	0.01	1.0	4	25	20	20	1.0	1.0
			10	10	20	20	1.0	
			4	25	16	25	1.0	
			10	10	16	25	1.0	
		0.8	4	25	20	20	0.9999	1.0
			10	10	20	20	1.0	
			4	25	16	25	1.0	
			10	10	16	25	1.0	
		0.5	4	25	20	20	0.9803	0.9859
			10	10	20	20	0.9892	
			4	25	16	25	0.9835	
			10	10	16	25	0.9904	

Table 3C. Aggregate Power for *N* = 500, 0.15:0.85 Sampling Ratio by 4 ICCs & 3 Effect Sizes

Ratio	ICC	ES	C1	S1	C2	S2	Power	Average
0.15:0.85	0.2	1.0	3	25	17	25	0.9510	0.9660
			5	15	17	25	0.9810	
		0.8	3	25	17	25	0.8891	0.9328
			5	15	17	25	0.9764	
		0.5	3	25	17	25	0.6901	0.7154
			5	15	17	25	0.7406	
	0.1	1.0	3	25	17	25	0.9920	0.9956
			5	15	17	25	0.9992	
		0.8	3	25	17	25	0.9565	0.9725
			5	15	17	25	0.9884	
		0.5	3	25	17	25	0.7529	0.7878
			5	15	17	25	0.8227	
	0.05	1.0	3	25	17	25	0.9999	1.0
			5	15	17	25	1.0	
		0.8	3	25	17	25	0.9930	0.9957
			5	15	17	25	0.9984	
		0.5	3	25	17	25	0.8474	0.8725
			5	15	17	25	0.8975	
	0.01	1.0	3	25	17	25	1.0	1.0
			5	15	17	25	1.0	
		0.8	3	25	17	25	0.9999	1.0
			5	15	17	25	1.0	
		0.5	3	25	17	25	0.9470	0.9557

Table 4A. Aggregate Power for *N* = 800, 0.25:0.75 Sampling Ratio by 4 ICCs and 3Effect Sizes.

Ratio	ICC	ES	C1	S1	C2	S2	Power	Average
0.25:0.75	0.2	1.0	8	25	24	25	0.9997	0.9994
			10	20	24	25	0.9993	
			8	25	30	20	0.9986	
			10	20	30	20	1.0	
		0.8	8	25	24	25	0.9922	0.9932
			10	20	24	25	0.9936	
			8	25	30	20	0.9898	
			10	20	30	20	0.9973	
		0.5	8	25	24	25	0.8927	0.9017
			10	20	24	25	0.9134	
			8	25	30	20	0.8880	
			10	20	30	20	0.9125	
	0.1	1.0	8	25	24	25	1.0	1.0
			10	20	24	25	1.0	
			8	25	30	20	1.0	
			10	20	30	20	1.0	
		0.8	8	25	24	25	1.0	0.9999
			10	20	24	25	1.0	
			8	25	30	20	0.9997	
			10	20	30	20	1.0	
		0.5	8	25	24	25	0.9679	0.9703
			10	20	24	25	0.9733	
			8	25	30	20	0.9593	
			10	20	30	20	0.9807	
	0.05	1.0	8	25	24	25	1.0	1.0
			10	20	24	25	1.0	
			8	25	30	20	1.0	
			10	20	30	20	1.0	
		0.8	8	25	24	25	1.0	1.0
			10	20	24	25	1.0	
			8	25	30	20	1.0	
			10	20	30	20	1.0	
		0.5	8	25	24	25	0.9929	0.9938
			10	20	24	25	0.9959	
			8	25	30	20	0.9904	
			10	20	30	20	0.9961	
	0.01	1.0	8	25	24	25	1.0	1.0
			10	20	24	25	1.0	
			8	25	30	20	1.0	
			10	20	30	20	1.0	
		0.8	8	25	24	25	1.0	1.0
			10	20	24	25	1.0	
			8	25	30	20	1.0	
			10	20	30	20	1.0	
		0.5	8	25	24	25	1.0	0.9999
			10	20	24	25	0.9999	
			8	25	30	20	0.9998	
			10	20	30	20	1.0	

Table 4B. Aggregate Power for *N* = 800, 0.20:0.80 Sampling Ratio by 4 ICCs and 3 Effect Sizes

Ratio	ICC	ES	C1	S1	C2	S2	Power	Average
0.20:0.80	0.2	1.0	10	16	20	32	0.9990	0.9964
			10	16	40	16	1.0	
			5	32	20	32	0.9948	
			5	32	40	16	0.9917	
		0.8	10	16	20	32	0.9945	0.9762
			10	16	40	16	0.9930	
			5	32	20	32	0.9633	
			5	32	40	16	0.9541	
		0.5	10	16	20	32	0.9046	0.8555
			10	16	40	16	0.8914	
			5	32	20	32	0.8260	
			5	32	40	16	0.8000	
	0.1	1.0	10	16	20	32	1.0	0.9999
			10	16	40	16	1.0	
			5	32	20	32	0.9997	
			5	32	40	16	0.9999	
		0.8	10	16	20	32	0.9996	0.9983
			10	16	40	16	0.9997	
			5	32	20	32	0.9975	
			5	32	40	16	0.9962	
		0.5	10	16	20	32	0.9677	0.9294
			10	16	40	16	0.9650	
			5	32	20	32	0.8983	
			5	32	40	16	0.8867	
	0.05	1.0	10	16	20	32	1.0	1.0
			10	16	40	16	1.0	
			5	32	20	32	1.0	
			5	32	40	16	1.0	
		0.8	10	16	20	32	1.0	1.0
			10	16	40	16	1.0	
			5	32	20	32	0.9999	
			5	32	40	16	0.9999	
		0.5	10	16	20	32	0.9926	0.9787
			10	16	40	16	0.9904	
			5	32	20	32	0.9666	
			5	32	40	16	0.9651	
	0.01	1	10	16	20	32	1.0	1.0
			10	16	40	16	1.0	
			5	32	20	32	1.0	
			5	32	40	16	1.0	
		0.8	10	16	20	32	1.0	1.0
			10	16	40	16	1.0	
			5	32	20	32	1.0	
			5	32	40	16	1.0	
		0.5	10	16	20	32	0.9996	0.9989
			10	16	40	16	0.9994	
			5	32	20	32	0.9983	
			5	32	40	16	0.9984	

Table 4C. Aggregate Power for *N* = 800, 0.15:0.85 Sampling Ratio by 4 ICCs and 3 Effect Sizes

Ratio	ICC	ES	C1	S1	C2	S2	Power	Average
0.15:0.85	0.2	1.0	4	30	17	40	0.9839	0.9886
			4	30	20	34	0.9899	
			5	24	17	40	0.9903	
			5	24	20	34	0.9903	
		0.8	4	30	17	40	0.9474	0.9539
			4	30	20	34	0.9424	
			5	24	17	40	0.9618	
			5	24	20	34	0.9640	
		0.5	4	30	17	40	0.7846	0.7969
			4	30	20	34	0.7814	
			5	24	17	40	0.8150	
			5	24	20	34	0.8064	
	0.1	1.0	4	30	17	40	0.9988	0.9992
			4	30	20	34	0.9987	
			5	24	17	40	0.9999	
			5	24	20	34	0.9993	
		0.8	4	30	17	40	0.9888	0.9919
			4	30	20	34	0.9923	
			5	24	17	40	0.9920	
			5	24	20	34	0.9946	
		0.5	4	30	17	40	0.8621	0.8799
			4	30	20	34	0.8689	
			5	24	17	40	0.8919	
			5	24	20	34	0.8966	
	0.05	1.0	4	30	17	40	1.0	1.0
			4	30	20	34	1.0	
			5	24	17	40	1.0	
			5	24	20	34	1.0	
		0.8	4	30	17	40	0.9994	0.9996
			4	30	20	34	0.9997	
			5	24	17	40	0.9996	
			5	24	20	34	0.9998	
		0.5	4	30	17	40	0.9370	0.9472
			4	30	20	34	0.9401	
			5	24	17	40	0.9582	
			5	24	20	34	0.9536	
	0.01	1.0	4	30	17	40	1.0	1.0
			4	30	20	34	1.0	
			5	24	17	40	1.0	
			5	24	20	34	1.0	
		0.8	4	30	17	40	1.0	1.0
			4	30	20	34	1.0	
			5	24	17	40	1.0	
			5	24	20	34	1.0	
		0.5	4	30	17	40	0.9913	0.9941
			4	30	20	34	0.9936	
			5	24	17	40	0.9959	
			5	24	20	34	0.9954	

With N = 800, adequate average power was achieved for all three ratios of unbalanced sampling, all four ICC values, and all three effect sizes with only one exception: sampling ratio of 0.15:0.85, ICC = 0.2 and ES = 0.5. Overall, with effect size set at 1.0 or 0.8, the only scenarios for which adequate average power was not achieved was with the small sample size of N = 200. With N = 500 or 800, the only scenarios for which adequate average power was not achieved all had effect size = 0.5 and ICC values equal to 0.2 or 0.1.

Aggregate data in Table 5 represent summaries for sample sizes of N = 200, 500, and 800 by each of the three sampling ratios, the four ICC values, and the three effect sizes for a total of 108 average power estimates. These results indicate that increasing the width of the sampling ratios has the effect of lowering the estimated power in most cases. For each sample size in each column, the aggregate estimated power decreased as the width of the sampling ratio increased. This effect is more pronounced for the smaller sample size of N = 200.

Five of these 108 aggregated power estimates (4.6%) exhibited an exception to the decreasing pattern, where contrary to every other estimate, power showed a slight increase. In every case, this exception occurs in the sampling ratio of 0.20:0.80 three times with N = 200 and twice with N = 500. These exceptions are shaded in Table 5.

Plausible explanations for these differences come from the work of Kreft and De Leeuw (1998) and Raudenbush and Liu (2000) where each researcher determined that using a larger number of groups have a greater positive effect on estimated power than having more subjects within groups. The five aggregate power estimates that are exceptions come from samples in which the group size is small. It also appears to generally be the case that even when intraclass correlations and effect sizes are high, power estimates are not compromised at the lower sample size. A wider sampling ratio, or greater unbalanced sampling, has the most pronounced effect on power and presents the greatest threat to research results.

Table 5. Comparisons of Aggregate Power Estimates for All Variables

Comparison of Power for ICC = 0.2 and Effect Size = 1.0, 0.8, & 0.5

	ICC = 0.2 & ES = 1.0			ICC = 0.2 & ES = 0.8			ICC = 0.2 & ES = 0.5		
Ratio	200	500	800	200	500	800	200	500	800
0.25 : 0.75	0.9663	0.9943	0.9994	0.8684	0.9598	0.9932	0.6266	0.7886	0.9017
0.20 : 0.80	0.9066	0.9885	0.9964	0.8246	0.9569	0.9762	0.6181	0.8016	0.8555
0.15 : 0.85	0.8522	0.9660	0.9886	0.7469	0.9328	0.9539	0.5471	0.7154	0.7969

Comparison of Power for ICC = 0.1 and Effect Size = 1, 0.8, & 0.5

	ICC = 0.1 & ES 1.0			ICC = 0.1 & ES = 0.8			ICC = 0.1 & ES = 0.5		
Ratio	200	500	800	200	500	800	200	500	800
0.25 : 0.75	0.9942	0.9998	1.0000	0.9602	0.9967	0.9999	0.7370	0.8733	0.9703
0.20 : 0.80	0.9704	0.9991	0.9999	0.9007	0.9910	0.9983	0.6475	0.8840	0.9294
0.15 : 0.85	0.9357	0.9956	0.9992	0.8274	0.9725	0.9919	0.5561	0.7878	0.8799

Comparison of Power for ICC = 0.05 and Effect Size = 1, 0.8, & 0.5

	ICC = 0.05 & ES = 1.0			ICC = 0.05 & ES = 0.8			ICC = 0.05 & ES = 0.5		
Ratio	200	500	800	200	500	800	200	500	800
0.25 : 0.75	0.9937	1.0000	1.0000	0.9829	0.9996	1.0000	0.7100	0.9522	0.9938
0.20 : 0.80	0.9913	1.0000	1.0000	0.9503	0.9991	1.0000	0.6974	0.9388	0.9787
0.15 : 0.85	0.9697	1.0000	1.0000	0.9063	0.9957	0.9996	0.6086	0.8725	0.9472

Comparison of Power for ICC = 0.01 and Effect Size = 1, 0.8, & 0.5

	ICC = 0.01 & ES = 1.0			ICC = 0.01 & ES = 0.8			ICC = 0.01 & ES = 0.5		
Ratio	200	500	800	200	500	800	200	500	800
0.25 : 0.75	0.9984	1.0000	1.0000	0.9620	1.0000	1.0000	0.7544	0.9929	0.9999
0.20 : 0.80	0.9990	1.0000	1.0000	0.9871	1.0000	1.0000	0.7680	0.9859	0.9989
0.15 : 0.85	0.9958	1.0000	1.0000	0.9658	1.0000	1.0000	0.6737	0.9557	0.9941

Aggregate data presented in Table 6 exhibits level-1 power estimates for sample sizes of 200, 500, and 800 with three sampling ratios, four intraclass correlations, and three effect sizes for a total of 108 mean power estimates. Cohen (1988) suggests aggregate power estimates at or above .80 that possess adequate magnitude to ensure research integrity. Being slightly more conservative, this study considered power estimates that were less than .85 to possess inadequate magnitude to ensure research integrity.

Counting the number of estimates that fell below the selected value, the results indicated that increasing the width of the three sampling ratios has the effect of lowering the estimated power in most cases. For each sample size in the last three columns of Table 6, the aggregate estimated power reduced as the breadth of the three sample ratios increased. For example, within the column of sample sizes = 200, 42% of the power estimates were below .85.

Table 6. Aggregated Estimated Power

Levels of Proportionally Unbalanced Data	Intraclass Correlations of Level 2 Units	Effect Size	Aggregate Level-1 Power $N = 200$	Aggregate Level-1 Power $N = 500$	Aggregate Level-1 Power $N = 800$
0.25 : 0.75	0.2	1	0.9663	0.9943	0.9994
		0.8	0.8684	0.9598	0.9932
		0.5	0.6266	0.7886	0.9017
	0.1	1	0.9942	0.9998	1.0
		0.8	0.9602	0.9967	0.9999
		0.5	0.7370	0.8733	0.9703
	0.05	1	0.9937	1.0	1.0
		0.8	0.9829	0.9996	1.0
		0.5	0.7100	0.9522	0.9938
	0.01	1	0.9985	1.0	1.0
		0.8	0.9620	1.0	1.0
		0.5	0.7544	0.9929	0.9999
0.20 : 0.80	0.2	1	0.9066	0.9885	0.9964
		0.8	0.8246	0.9569	0.9762
		0.5	0.6181	0.8016	0.8555
	0.1	1	0.9704	0.9991	0.9999
		0.8	0.9007	0.9910	0.9983
		0.5	0.6475	0.8840	0.9294
	0.05	1	0.9913	1.0	1.0
		0.8	0.9503	0.9991	1.0
		0.5	0.6974	0.9388	0.9787
	0.01	1	0.9991	1.0	1.0
		0.8	0.9871	1.0	1.0
		0.5	0.7680	0.9859	0.9989
0.15 : 0.85	0.2	1	0.8522	0.9660	0.9886
		0.8	0.7469	0.9328	0.9539
		0.5	0.5471	0.7154	0.7969
	0.1	1	0.9357	0.9956	0.9992
		0.8	0.8274	0.9725	0.9919
		0.5	0.5561	0.7878	0.8799
	0.05	1	0.9697	1.0	1.0
		0.8	0.9063	0.9957	0.9996
		0.5	0.6086	0.8725	0.9472
	0.01	1	0.9958	1.0	1.0
		0.8	0.9658	1.0	1.0
		0.5	0.6737	0.9557	0.9941

Table 7. Number & Percentage of Power Estimates ≤ 0.85 for Unbalanced Sample Schemes

Levels	*N* = 200	*N* = 500	*N* = 800
0.25 : 0.75	4 (33%)	1(8%)	0 (0%)
0.20 : 0.80	5 (42%)	1(8%)	0 (0%)
0.15 : 0.85	6 (50%)	2 (17%)	1 (8%)

Additionally, comparing samples of 200, where the unbalanced levels are measured at 0.25:0.75, 33% of the power estimates fell below .85. At 0.20:0.80, 42% of the power estimates fell below .85 and at 0.15:0.85, 50% of the power estimates fell below .85. This effect is more pronounced for the smaller sample size of 200. The majority of the scenarios with sample sizes equal to 500 and 800 produced smaller comparative power differences (see Table 7). In the 108 possible power estimates, five (4.6%) exceptions to the decreasing pattern are seen where contrary to every other estimate, the power increases slightly.

Implications

The cost of utilizing larger sampling techniques to ensure model adequacy may not meet the challenges of today's dwindling budgets. "Doing more with less" would be the preferred method despite the mixed messages inferred from previous research. For example, Bassari (1988) estimates detection of cross-level effects with sufficient power needed at least 30 groups with 30 participants per group or a total sample of 900. Kreft and De Leeuw (1998) (as cited in Heck & Thomas, 2000) found groups as low as 20 were sufficient to determine cross-level effects (i.e., with a total sample size of 600). The results of the present study can help to update educational researchers concerning the recommended sample sizes needed to achieve adequate power when utilizing unbalanced sampling in multilevel models.

References

Bassiri, D. (1988). *Large and small sample properties of maximum likelihood estimates for the hierarchical linear model*. Unpublished doctoral dissertation, Michigan State University.

Bond, G., Miller, L., Drumweid, R., & Ward, R. (1988). Assertive cases management in three CMHs: A controlled study. *Hospital and Community Psychiatry, 9,* 411-418.

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

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27*, 557-577.

Haddow, J. (1991). Cotinine-assisted intervention in pregnancy to reduce smoking and low birthweight delivery. *British Journal of Obstetrics and Gynecology, 98*, 859-865.

Heck, R. H., & Thomas, S. L. (2000). *An introduction to multilevel modeling techniques*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research.* Newbury Park, CA: Sage Publications, Inc.

Kreft, I., & De Leeuw, J., (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage Publications, Inc.

Mosteller, F. (1995). Optimal design in psychological research. *Psychological Methods, 2*, 3-19.

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

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*, 199-213.

Reise, S. P., & Duan, N. (2003). Design issues in multilevel studies. In S. Reise & N. Duan (Eds.), *Multilevel modeling methodological advances, issues, and applications* (pp. 285 – 298). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Schumacker, R E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum Associates.

Yeates, K. O., & Taylor, H. G. (1997). Predicting premorbid neuropsychological functioning following pediatric traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology, 19,* 825-.

Send correspondence to: Bonnie J. Steele
Colorado Mountain College
Email: bsteele@coloradomtn.edu

APPENDIX
SAS Program for Simulating Outcome Variables and Estimating Power

The following is the SAS code for running the simulation that produces 10,000 outcome variables for each designated case then takes these outcome variables through SAS PROC MIXED procedure to estimate power for each set of sampling ratios. Note: Intraclass correlations and effect sizes must be defined as fractions, not decimals or the program will cease to run.

```
/** Generate unbalanced two-level data **/
%let icc=1/100; *intraclass correlation coefficient;
%let g1=2; *number of classes in treatment group 1;
%let g2=10; *number of classes in treatment group 2;
%let n1=20; *number of subjects/class in treatment group 1;
%let n2=16; *number of subjects/class in treatment group 2;
%let ti=2; *number of treatments.  DO NOT CHANGE THIS VALUE.;
%let es=1; *effect size;
%let se=1; *standard deviation of individuals (level 2);
%let iter=10000; *this is the number of times you want the simulation to
iterate;

*Note: standard deviation of classes is determined computationally by
 the standard deviation of individuals as well as the effect size.;

title;
data tests;
probf=1;
delete;
run;

/** Generate Data **/
%macro datagen;
   ods select none;
   proc iml;
   icc=&icc; g1=&g1; n1=&n1; g2=&g2; n2=&n2; ti=&ti; se=&se; es=&es;
   mu=j(ti,1,1);
   mu[ti]=mu[ti]+es*se;
   se=1;
   sd=sqrt((icc/(1-icc))*se*se);
   y={0 0 0 0};
   CREATE datagen From y [colname={trt,class,student,y}];
   j=1;
      do k=1 to g1;
         z=normal(0);
         do i=1 to n1;
            w=rannor(0);
            y[1]=j;y[2]=k;y[3]=i;
            y[4]=mu[j]+sd*z+se*w;
            APPEND FROM y;
         end;
      end;
   j=2;
      do k=1 to g2;
         z=normal(0);
         do i=1 to n2;
            w=normal(0);
            y[1]=j;y[2]=k;y[3]=i;
            y[4]=mu[j]+sd*z+se*w;
            APPEND FROM y;
         end;
      end;

   close datagen;
   quit;
```

```
   proc mixed data=datagen;
      class class;
      model y=trt;
      random class;
    ods output tests3=tests3;
   run;quit;

   data tests;
   set tests tests3;
   run;

   ods select all;
%mend datagen;
%macro iterate;
   options nonotes nodate nonumber;ods results off;
   %do i=1 %to &iter;
   %datagen;
   %end;
   options notes;ods results on;

   %if &es+0=0 %then %do;
      title 'This is the simulated value of alpha.';
   %end;
   %if &es+0^=0 %then %do;
      title 'This is the simulated value of power.';
   %end;

data prop;
   set tests;
   rejects=probf<.05;
run;

proc means data=prop mean;
   var rejects;
run;
title;
%mend iterate;
%iterate;
```