Bootstrapping within the Multilevel/Hierarchical Linear Modeling Framework: A Primer for Use with SAS and SPLUS

J. Kyle Roberts
University of North Texas

Xitao Fan
University of Virginia

Nested data structure obtained from a cluster sampling design often calls for hierarchical linear modeling (HLM) analysis. Such data structure warrants some special considerations when the bootstrap technique is applied. This paper presents some discussions and examples for applying the bootstrap method within the framework of hierarchical linear modeling. A two-level dataset (about 900 students nested under 20 schools) extracted from the High School and Beyond (HSB) was used for illustration. Bootstrap resampling was implemented in both SAS and S-PLUS, and a hierarchical linear model with one Level-1 predictor (student SES), and one Level-2 predictor (type of schools, Catholic or public) was applied.

In quantitative research in education and psychology, over-reliance on statistical significance testing has been called into question. Several issues have been raised concerning the use of statistical significance testing in research practice including sample size, the meaningfulness of the traditional null hypothesis, and questions involving the validity of theoretical assumptions underlying parametric statistical inferences (e.g., Carver, 1978; Shaver, 1993; Thompson, 1993). As a result of these and other concerns, researchers are increasingly turning to empirically-grounded resampling procedures in quantitative analyses.

Applauded as one of the newest breakthroughs in statistics (Kotz & Johnson, 1992), the bootstrap is often considered the best-known resampling method. The importance of bootstrapping as a versatile analytic approach with which to conduct data analysis has been widely recognized not only by those in the area of statistics, but also by quantitative researchers in education, psychology, and social and behavioral sciences in general.

Statistical inference (e.g., in a t test, rejection of the null hypothesis that two populations have equal means) is usually made based on the sampling distributions of a statistical estimator. For parametric statistics, the derivation of such sampling distributions is typically based on a set of theoretical assumptions. The bootstrap method attempts to estimate these sampling distributions empirically, using information drawn from the sample of observations used to estimate the statistical model in the first place (Diaconis & Efron, 1983; Efron, 1979). In doing so, the bootstrap approach avoids some of the pitfalls of traditional statistical significance testing. As discussed by Lunneborg (2000):

> Until inexpensive computing power made replicate data analysis practical, the drawing of statistical inferences from a set of data almost always required that we accept an idealized model for the origin of those data. Such models can be either inappropriate or inadequate for the data in our study. Resampling techniques allow us to base the analysis of a study solely on the design of that study, rather than on a poorly-fitting model. (p. xi)

The bootstrap method has found a variety of research applications in social and behavioral sciences. For example, the bootstrap method has been applied in sociological research (e.g., Stine, 1989), and in research for psychological measurement issues such as differential test predictive validity (e.g., Fan & Mathews, 1994) and item bias (e.g., Harris & Kolen, 1989). Application of bootstrapping has also involved many different statistical techniques, including correlation analysis (e.g., Mendoza, Hart, & Powell, 1991; Rasmussen, 1987), regression analysis (e.g., Fan & Jacoby, 1995), descriptive discriminant analysis (e.g., Dalgleish, 1994; Thompson, 1992), canonical correlation analysis (e.g., Fan & Wang, 1996; Thompson, 1995), factor analysis (e.g., Lambert, Wildt, & Durand, 1991; Thompson, 1988), and structural equation modeling (e.g., Bollen & Stine, 1990; Yung & Bentler, 1996).

Although bootstrap was proposed as a versatile tool for non-parametric statistical inference (Efron, 1985), Thompson (1993) has also advocated the use of bootstrapping as a descriptive tool and an internal replication mechanism for assessing the stability and replicability of sample results of an individual study. This descriptive use of bootstrap is meaningful when our interest is not about statistical inference, but rather, about understanding how stable the results may be across repeated sampling.

Bootstrapping is a computing-intensive data resampling strategy, and easy access to powerful computing facilities makes bootstrapping an attractive and viable procedure for research practitioners. Unfortunately, although the logic of bootstrapping is conceptually straightforward, bootstrapping has yet

to enjoy widespread application in many areas of research and for some statistical techniques. Because bootstrapping is not typically implemented as an automated option in the major commercial statistical software packages (e.g., SAS, SPSS), researchers who desire to use this approach usually have to deal with programming for performing bootstrap resampling. This can be a daunting endeavor for many who do not have the skills, knowledge, or interest required to carry out such a task. Consequently, this appears to be a major obstacle for implementing bootstrapping in substantive research. Some methodologists have sensed the need for programs to perform bootstrapping; as a result, some special programs have been published for bootstrap application in different analytic techniques, such as regression analysis (Fan & Jacoby, 1995) and factor analysis (Thompson, 1988). But overall, bootstrapping remains procedurally difficult for most research practitioners.

Multilevel modeling is an area where bootstrapping has not yet enjoyed much application. As is the case for other statistical techniques, bootstrapping within multilevel modeling may serve two main purposes: making non-parametric inferences about parameter estimates and correcting potential bias in parameter estimation. This non-parametric approach can be especially helpful in samples where assumptions about data may have been violated (e.g., data non-normality, Bryk & Raudenbush, 1992), or in samples where the number of Level 1 observations (e.g., individuals) may be small within each Level 2 unit (e.g., schools).

This paper provides some heuristic examples of implementing bootstrap analysis for hierarchical linear modeling (HLM). Although there has been little application of bootstrapping in hierarchical linear modeling, it is hoped that the demonstration of the use of these methods will encourage future researchers to utilize these techniques. Procedures for conducting bootstrapping analyses with both SAS and S-PLUS are presented with heuristic datasets.

Bootstrap Approach

Bootstrap as the most popular resampling method is mainly used for estimating the sampling distribution of a statistic of interest for which parametric alternatives either do not exist, or the validity of the parametric alternatives are in question (e.g., violated assumptions). The basic bootstrap method typically has three straightforward steps:

1. select *B* independent bootstrap samples, each consisting of n observations drawn with replacement from the original sample,
2. obtain the statistic of interest from each bootstrap sample, and
3. evaluate the sampling distribution of the bootstrapped statistic of interest by
    a) estimating the standard error of the statistic of interest by the sample standard deviation of the *B* bootstrap replications, or
    b) using exact percentiles (e.g., 97.5%; 2.5%) for constructing empirical confidence intervals.

Approach a) above assumes distribution normality of the bootstrapped statistic, and parametric confidence intervals can be constructed through the use of the estimated standard error. Approach b), however, does not assume distribution normality of the bootstrapped statistic, and the resultant confidence intervals are non-parametric in nature.

Although the bootstrapping method as described above is procedurally straightforward, its application in hierarchically nested data structure such as those used in hierarchical linear modeling may warrant some special considerations. Typical bootstrapping involves sampling individual observations with replacement, and there is no consideration for the nested data structure in HLM, (e.g., individual students (Level 1 units) are nested under schools (Level 2 units)). Because of this nested data structure, potentially, there can be different resampling approaches for hierarchically nested data. From a sample data with two levels (e.g., Level 1: students, and Level 2: schools), with k schools, and each with ni students, and the total sample size of N [$N = \mathbf{\Sigma} n_i$, i = 1, 2, 3 . . j, k], the following bootstrap sampling approaches may potentially be applied:

1. draw a bootstrap sample of *N* students with replacement, and totally ignore the nested data structure;
2. draw a bootstrap sample of ni students with replacement from each and every school in the sample data, and the bootstrap sample has sample size of *N*;
3. bootstrap *k* schools with replacement while selecting all n_i students in each bootstrapped *k* school;
4. first, drawn a bootstrap sample of *k* schools with replacement; from each sampled school, draw a bootstrap sample of n_i students with replacement.

The first two approaches will provide a consistent sample size of *N* for each bootstrap iteration. But the third and fourth approach will not provide a consistent sample size of *N* for each bootstrap iteration, unless $n_i = n$ for each Level 2 unit (i.e., each school contains the same number of students in the original sample).

Theoretically, both Level 2 and Level 1 units should be considered as randomly drawn from the population. In other words, in clustered sampling design, Level 2 units (schools) are randomly drawn first. Level 1 units (students) are then randomly drawn from the school. In this sense, the fourth approach of bootstrap sampling for hierarchically nested data described above makes good sense. In practice, however, the fourth approach will typically not provide a consistent bootstrap sample size of *N*, because hierarchically nested sample data typically do not have equal sample size within each Level 2 unit. Without a consistent sample size of *N*, it would not be possible to construct an empirical sampling distribution for a statistical estimator of interest because the sampling distribution is always associated with a specific sample size. For this reason, we only used the first two bootstrap sampling approaches in our examples.

Data Source

Bryk and Raudenbush (1992) used a dataset from the national High School and Beyond (HSB) database to illustrate the application of HLM. The same dataset is also used by Singer (1998) in her illustration of using SAS for fitting HLM models. For bootstrapping illustrations in this paper, we used a dataset of 20 schools randomly selected from the dataset of 160 schools as used in Bryk and Raudenbush (1992) and Singer (1998).

Table 1 presents the basic descriptive information for the variables used in our HLM bootstrapping example. The student level predictor SES is centered with mean of zero. The variable SECTOR is dummy coded, with Catholic schools coded as 1 and public schools coded as 0. So the mean of SECTOR (0.35) indicates that, of the 20 schools in this dataset, seven (35%) are Catholic schools, and the remaining 13 are public schools.

Table 2 presents the descriptive information for math achievement scores for the 20 schools, and the sample sizes of the 20 schools. The total sample size for this data is 914, with sample size for individual schools ranging from 25 to 66, and the average sample size across the 20 schools of 45.7. It is noticed that there appears to be some noticeable variation among the school averages of math achievement score. This suggests that some school-level variable may potentially be useful in accounting for the variation among the school means of the math achievement score.

A conditional two-level model, with SES as the Level 1 (student level) predictor, and SECTOR as the Level 2 (school level) predictor, was fitted to the data, as shown below (Y: math achievement score; notations as used in Bryk and Raudenbash, 1992):

$$\text{Level 1:} \quad Y_{ij} = \beta_{0j} + \beta_{1j}\,(\text{SES}) + r_{ij}\,, \quad \text{and}$$

$$\text{Level 2:} \quad \beta_{0j} = \gamma_{00} + \gamma_{01}\,(\text{SECTOR}) + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}\,(\text{SECTOR}) + u_{1j}\,.$$

To provide information about how much variation in the math achievement score is within and between schools in this dataset, a one-way ANOVA model with random effects was fitted the data. The one-way ANOVA model with random effect takes the following form:

$$\text{Level 1:} \quad Y_{ij} = \beta_{0j} + r_{ij}\,, \quad \text{and}$$

$$\text{Level 2:} \quad \beta_{0j} = \gamma_{00} + u_{0j}$$

Table 1. Descriptive Statistics for the Dataset Used

Variable		
Student-Level (1)	Mean	SD
Math Achievement (Y)	13.00	7.20
SES	0.00	0.65
School Level (2)		
Sector	0.35	0.49

Table 2. Descriptive Statistics of Math Achievement Scores for the 20 Schools

School ID	*N*	Mean	SD
1317	48	13.18	5.46
1374	28	9.73	8.36
1461	33	16.84	6.95
1477	62	14.23	7.15
2458	57	13.99	5.85
2629	57	14.91	5.17
2768	25	10.89	7.29
2771	55	11.84	6.80
3427	49	19.72	3.54
3716	41	10.37	8.48
3838	54	16.06	5.10
3967	52	12.04	6.89
4383	25	11.47	7.45
5619	66	15.42	7.28
5762	37	4.32	4.99
6291	35	10.11	6.59
6897	49	15.10	6.65
7697	32	15.72	6.62
7890	51	8.34	6.25
8946	58	10.38	6.52

Table 3. Results of One-Way ANOVA Model and the HLM Model

One-Way ANOVA			
Fixed Effects	Coefficients		Description
γ_{00}	12.76		overall mean math score
Random Effect	Variance Component		
School Mean, u_0	11.09		variation of school means
Level-1 Residual	41.86		
Intra-Class Correlation $\rho = 11.09/(11.09 + 41.86) = .21$			
HLM Model: (Level 1 Predictor: SES; Level 2 Predictor: SECTOR)			
	Coefficients		
Fixed Effects	SAS	S-PLUS	Description
γ_{00}	11.33	11.33	mean math score for public schools
γ_{01}	4.02	4.02	SECTOR main effect
γ_{10}	3.35	3.34	SES main effect
γ_{11}	-1.77	-1.77	**a** SECTOR effect on SES slope
	Variance Component		
Random Effects	SAS	S-PLUS	
School Mean, u_0	7.64	6.78	variation of intercept (τ_{00})
SES-Math Slope, u_1	2.54	2.07	variation of slope (τ_{11})
Covariance (u_0,u_1)	2.16	1.91	covariation of u_0 and u_1 (τ_{01})
Level-1 Residual	37.72	37.72	Var (r_{ij})

a In Catholic schools (coded as 1 on SECTOR), the student performance on Math is less related to SES (Catholic schools are more equitable). See Chapter 4 in Bryk and Raudenbush (1992) for discussion related to this issue.

Table 3 presents the results of fitting the two different models to this dataset. The first one is an unconditional model, or the one-way ANOVA model with random effect, and the second one is the HLM model we used for later bootstrapping illustration. From the one-way ANOVA model with random effect, the intraclass correlation was obtained to be 0.21, suggesting that 21% of the variance in the math scores is between-school variation, while the remaining 79% variation is within schools. This indicates that the HLM model is warranted for this dataset. If it turned out that only a negligible proportion of the total variance is between-school variation, HLM would not be as useful. Since the nested bootstrap will be illustrated with two software packages, SAS and S-PLUS, results from each of these packages will be presented in the HLM model in the following tables.

Further comparisons between the two models show that a) for the between-school variation, 31% of the variance [(11.09 – 7.64)/11.09] is accounted for by the school-level predictor SECTOR, and b) 10% of within-school variation is accounted for by the student-level predictor SES [(41.86 – 37.72)/41.86]. The results of the HLM model indicate that, within Catholic schools, the relationship between math achievement and SES is weaker than that within public schools (γ_{11} = -1.77). More specifically, for

public schools, the average regression slope between math score and SES is 3.35. For Catholic schools, the average regression slope between math score and SES is 1.58 (3.35-1.77), suggesting that Catholic schools appeared to be more equitable with regard to student SES level (Bryk & Raudenbush, 1992, Chapter 4).

Method

As has been noted previously, performing a nested bootstrap within the HLM framework is not just a "point and click" procedure in any software package. Although some programs, such as MLwiN, provide a method of performing the bootstrap with hierarchically structured data, this method is based on residuals bootstrap, which redistributes the residuals at each appropriate level (see bootstrap #4 above) rather than nesting the bootstrap within Level 2 units.

The nested bootstrap utilizes a nested looping structure within both the SAS and S-PLUS architecture. Inside the inner loop, a dataset is being created from the original dataset by extracting the data, one school (or Level 2 unit) at a time. The programs will search through the data and find the first appearing school and then extract all other pieces of data that have the same value for the school variable. In the case of the HSB dataset, 20 total schools were selected. The dataset, after being split into the 20 schools, is bootstrapped across the data contained in each school such that the number of people in the original school equals the number of people in the now bootstrapped school. The iterative process can be described as follows:

1. Select all data in school k.
2. Bootstrap data in school k such that $n_i = n_i'$.
3. Repeat process for next k school.
4. Append data from school $k + 1$ to school k.
5. Repeat steps 3 and 4 until all schools have been selected.

After this inner loop has created the bootstrapped dataset, the HLM analysis is conducted in an outer loop and the desired components are extracted. This outer loop then reverts back to the original dataset and the entire process is begun again. In the outer loop, the extracted components are appended to the previously extracted components across all bootstrapped samples. In the case of this paper, we chose 2000 bootstrap samples. It should be noted that these two programs are computer intensive and require between 2 and 3 hours of processing time for 2000 iterations on a Pentium III 600 MHz with 128meg RAM.

For comparison purposes, we also chose to include in the analysis a typical (non-nested) bootstrap. In this method, student scores were bootstrapped regardless of which school they appeared in (see bootstrap method #1 above). It is conceivable that in this method, within a single bootstrap, one school may contain no student estimates for a given bootstrap sample. This analysis was only conducted in SAS and as such, should be compared against the original SAS estimates.

Results

Criteria were set for which pieces of information that should be extracted from the HLM analysis from the specified model. Since this was a model with two levels and random effects at the second level, four fixed effects, two random effects, the covariance between the random effects, and the Level 1 residual were extracted in each bootstrapped sample. The results of these bootstrapped fixed and random effects can be seen in Tables 4 and 5.

In Tables 4 and 5, columns labeled "Original Data SAS" and "Original Data S-PLUS" correspond to the results from the original sample analysis in Table 3. These were included for comparison purposes. Results from the SAS nested bootstrap program, the S-PLUS nested bootstrap program, and from the non-nested bootstrap are also included in these tables.

In first looking at the results from Table 4, it can be seen that the γ_{00} and γ_{01} fixed effects had bootstrapped sample estimates that differ only slightly from the original estimate. This was not the case, however, with the bootstrapped estimates for g10 in the S-PLUS bootstrapped estimate and for γ_{11} in both the SAS and S-PLUS bootstrapped estimate. In these later estimates, it can be seen that the estimate 95%

Table 4 Results of the HLM Bootstrap of the HSB Data for the Fixed Effects

γ_{00} (mean math score for public schools)					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	11.33	11.33	11.33	11.33	11.33
Minimum		10.11		10.23	10.37
Maximum		12.24		12.29	12.23
SD		0.29		0.29	0.29
SEM		0.01		0.01	0.01
LCL Mean		11.32		11.31	11.32
UCL Mean		11.34		11.34	11.35
Skewness		-0.03		-0.04	-0.12
Kurtosis		0.04		0.09	-0.10
γ_{01} (SECTOR main effect)					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	4.02	4.01	4.02	4.03	4.01
Minimum		2.66		2.79	2.78
Maximum		5.25		5.44	5.34
SD		0.40		0.40	0.41
SEM		0.01		0.01	0.01
LCL Mean		3.99		4.01	4.00
UCL Mean		4.02		4.04	4.03
Skewness		0.03		0.08	0.03
Kurtosis		-0.14		0.01	-0.12
γ_{10} (CSES main effect)					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	3.35	3.36	3.34	3.37	3.37
Minimum		2.08		2.08	1.78
Maximum		4.80		4.91	4.72
SD		0.42		0.42	0.42
SEM		0.01		0.01	0.01
LCL Mean		3.34		3.35	3.35
UCL Mean		3.38		3.39	3.39
Skewness		0.03		0.06	0.01
Kurtosis		-0.07		0.04	-0.09
γ11 (SECTOR effect on CSES slope)					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	-1.77	-1.80	-1.77	-1.83	-1.81
Minimum		-3.75		-3.70	-3.72
Maximum		0.55		-0.01	-0.03
SD		0.58		0.59	0.60
SEM		0.01		0.01	0.03
LCL Mean		-1.83		-1.85	-1.84
UCL Mean		-1.78		-1.80	-1.78
Skewness		0.01		-0.02	-0.06
Kurtosis		0.08		-0.16	-0.08

Table 5 Results of the HLM Bootstrap of the HSB Data for the Random Effects

Variation of Regression Intercept u_0 (τ_{00})					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	7.64	8.61	6.78	7.61	8.55
Minimum		4.56		3.82	5.43
Maximum		13.30		12.66	12.68
SD		1.26		1.12	1.23
SEM		0.03		0.25	0.03
LCL Mean		8.56		7.56	8.50
UCL Mean		8.67		7.66	8.61
Skewness		0.14		0.25	0.22
Kurtosis		0.12		0.21	-0.11
Variation of Regression Slope u_1 (τ_{11})					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	2.54	4.43	2.07	3.68	4.43
Minimum		0.43		0.01	0.72
Maximum		11.67		8.79	11.69
SD		1.56		1.36	1.56
SEM		0.04		0.03	0.04
LCL Mean		4.36		3.63	4.37
UCL Mean		4.50		3.74	4.50
Skewness		0.56		0.37	0.45
Kurtosis		0.66		0.11	0.27
Covariation Between u_0 and u_1 (τ_{00})					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	2.16	2.33	1.91	2.07	2.34
Minimum		-1.74		-1.75	-1.37
Maximum		6.38		5.61	6.83
SD		1.15		1.02	1.11
SEM		0.03		0.02	0.03
LCL Mean		2.27		2.02	2.29
UCL Mean		2.38		2.11	2.40
Skewness		-0.05		0.02	-0.00
Kurtosis		0.17		-0.01	0.01
Level-1 Residual (Var (r_{ij}))					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	37.72	36.18	37.72	36.09	36.16
Minimum		31.22		30.70	30.56
Maximum		41.78		42.37	40.87
SD		1.53		1.53	1.55
SEM		0.03		0.03	0.03
LCL Mean		36.11		36.02	36.10
UCL Mean		36.24		36.16	36.23
Skewness		0.05		-0.02	0.06
Kurtosis		-0.01		0.13	-0.11

confidence intervals did not capture the original data estimates. In instances of the fixed effects estimates, it would then be proper to default to the bootstrapped estimate rather than assume the original estimate.

In Table 5 we can see the differences between the original estimates of the variance components and the bootstrapped estimates. Upon looking at these results, we can see that in every bootstrapped estimate across both the SAS and S-PLUS results, the original estimate of the variance component is not captured by the 95% confidence intervals around the nested bootstrapped estimate. This is especially troubling since on many occasions, model specification issues are often based on these estimates.

For example, consider the intraclass correlation from the original dataset and nested bootstrap (where $ICC = \tau_{00}/[\tau_{00} + r_{ij}]$). In the case of the present dataset, the ICC for the original data in the SAS equation would be .168, suggesting that only 16.8% of the variance in the math scores is between-school variation, while the remaining 83.2% variation is within schools. We can contrast these results to the nested bootstrap sample where the ICC is .192. This is critical when we consider that Kreft and de Leeuw (1998) and Roberts (2004) have defined an ICC of .20 or greater as a large effect. These differences in the variance estimates might lead a researcher to interpret a fixed effect as a small or medium effect when in fact the effect is quite large (or vice versa). This could prove problematic when basing modeling decisions on the interpretation of variance estimates alone.

Discussion

One question that might be brought to attention from the results presented in Tables 4 and 5 is whether or not the effort justifies the ends. In this analysis, the results from the individual bootstrap and the nested bootstrap yield similar results. Although this has proven true in this case, it does not hold that the two types of resampling will yield similar results across all hierarchical datasets. Consider when a dataset (unlike the present dataset) has few Level 1 units inside each Level 2 unit. In this case, the individual bootstrap would be much more likely to obtain bootstrap estimates in which entire Level 2 units are ignored, whereas the nested bootstrap will always include the same n for each Level 2 unit in every analysis. As was previously noted, the nested bootstrap will prove especially useful when *N* for the entire dataset is very small.

While the present paper has only identified certain components of the hierarchical linear model to bootstrap, it can be seen that bootstrapping other components of the model could help to answer some of the problems associated with assumptions in HLM. For example, we might wish to test the mutual independence of all residuals by testing to see if they are normally distributed and have zero means given the explanatory variables across bootstrap samples. Furthermore, we might also want to test each bootstrapped sample for heteroscedasticity and in cases where heteroscedasticity is high across bootstrap samples, apply a Box-Cox transformation to the dataset and then reapply the bootstrap (Snijders & Bosker, 1999).

Although it has been the primary purpose of this paper to discuss and illustrate the nested bootstrap in hierarchical linear and multilevel modeling, further applications of this type of analysis could be utilized beyond the topics presented currently. For example, this type of nested bootstrap could prove vitally useful in Generalizability theory studies where actual items are bootstrapped rather than just individuals. This nested bootstrap might further be utilized in ANOVA type studies where researchers are concerned about the robustness of variance estimates within levels of a given way.

This type of resampling can also encourage researchers to think seriously about resampling designs in other types of analysis. For example, consider if we were to apply a jackknife resampling design to the present study. Since HLM type analyses require such large sample sizes, we are unlikely to see much of a difference in our parameter estimates. Consider, however, if we were to apply a nested jackknife to the data where actual schools are jackknifed rather than individuals. In this case, a researcher could easily note the potential contribution (or lack of contribution) for each school in the dataset. This type of analysis could also be applied to Generalizability theory where items (or some other facet) are jackknifed rather than individuals.

This type of resampling could be further applied in a nested jackstrap (a combination of the nested bootstrap and nested jackknife). In this type of analysis (in a school-effects model), schools would first be jackknifed and then the nested bootstrap would be applied to each jackknifed sample. One might consider that this type of analysis could conceivably run on a single computer for a couple of days, but

the results could help solve some sampling issues that a researcher might be facing. It is hoped that the presentation of this paper will encourage researchers to consider more complex resampling techniques that are more appropriate to the type of data and type of analysis that they might run.

References

Bollen, K. A., & Stine, R. (1993). Bootstrapping goodness-of-fit measures in structural equation models. In K. A. Bollen & J. S. Long, (Eds.), *Testing structural equation models* (pp. 111-135). Newbury Park, CA: Sage.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: SAGE publications.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378-399.

Dalgleish, L. I. (1994). Discriminant analysis: Statistical inference using the jackknife and bootstrap procedures. *Psychological Bulletin*, *116*, 498-508.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, May, 116-130.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1-26.

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika*, *72*, 45-48.

Fan, X., & Jacoby, W. R. (1995). BOOTSREG: A SAS matrix language program for bootstrapping linear regression models. *Educational and Psychological Measurement*, *55*, 764-768.

Fan, X., & Mathews, T. A. (1994, April). *Using bootstrap procedures to assess the issue of predictive bias in college GPA prediction for ethnic groups*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No: ED 372 117)

Fan, X. & Wang, L. (1996). Comparability of jackknife and bootstrap results: An investigation for a case of canonical analysis. *Journal of Experimental Education*, *64*, 173-189.

Harris, D. J., & Kolen, M. J. (1989). Examining the stability of Angoff's delta item bias statistic using bootstrap. *Educational and Psychological Measurement*, *49*, 81-87.

Kotz, S., & Johnson, N. L. (1992). *Breakthroughs in statistics: Volumes 1 and 2*. New York: Springer-Verlag.

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

Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1991). Approximating confidence interval for factor loadings. *Multivariate Behavioral Research*, *26*, 421-434.

Lunneborg, C. E. (2000). *Data analysis by resampling: Concepts and applications*. Pacific Grove, CA: Duxbury.

Mendoza, J. L., Hart, D. E., & Powell, A. (1991). A bootstrap confidence interval based on a correlation corrected for range restriction. *Multivariate Behavioral Research*, *26*, 255-269.

Rasmussen, J. L. (1987). Estimating the correlation coefficients: Bootstrap and parametric approaches. *Psychological Bulletin*, *101*, 136-139.

Roberts, J. K. (2004). An introductory primer on multilevel and hierarchical linear modeling. *Learning Disabilities: A Contemporary Journal*, *2*(1), 30-38.

Shaver, J.P. (1993). What significance testing is, and what it isn't. *Journal of Experimental Education*, *61*, 293-316.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *24*, 323-355.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.

Stine, R. A. (1989). An introduction to bootstrap methods: Examples and ideas. *Sociological Methods and Research*, *8*, 243-291.

Thompson, B. (1988). Program FACSTRAP: A program that computes bootstrap estimates of factor structure. *Educational and Psychological Measurement*, *48*, 681-686.

Thompson, B. (1992). DISCSTRA: A computer program that computes bootstrap resampling estimates of descriptive discriminant analysis function structure coefficients and group centroids. *Educational and Psychological Measurement*, *52*, 905-911.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, *61*, 361-377.

Thompson, B. (1995). Exploring the replicability of a study's results: Bootstrap statistics for the multivariate case. *Educational and Psychological Measurement*, *55*, 84-94.

Yung, Y., & Bentler, P. M. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 195-226). Mahwah, New Jersey: Lawrence Erlbaum.

Send correspondence to: J. Kyle Roberts, Ph.D.
P.O. Box 311335
University of North Texas
Denton, TX 76203-1335
Email: kroberts@unt.edu

Appendix A

SAS Code for the Nested Bootstrap

```
*** SAS PROGRAM FOR BOOTSTRAPPING INDIVIDUALS WITHIN EACH SCHOOL ***;

LIBNAME BTHSB20 'C:\HLM Bootstrap';

DATA HSB20; INFILE 'C:\HLM Bootstrap\HSB20.TXT';
  INPUT SCHID MATH SECTOR CSES;

     *** direct the SAS log to a disk file to avoid SAS LOG Window becoming full;
PROC PRINTTO LOG='C:\HLM Bootstrap\LOGFILE.TMP';
 RUN;

%MACRO BTRAP;           *** start of bootstrap macro 'BTRAP';
%DO BTRAP=1 %TO 2000;   *** 2000 bootstrapped samples, about 4.5 sec. each iteration;
%DO A=1 %TO 20;         *** select each school sequentially;
DATA D1; SET HSB20;     *** (20 schools in the data set, unequal N in each school);
   IF SCHID=&A;

     *** sampling with replacement within each selected school;
     *** bootstrapped sample size equal to the original sample size in each school;
     *** bootstrapped sample within each school is named BTDATA_n;
 DATA BTDATA;
  DROP I;
  DO I=1 TO N;
    IOBS=INT(RANUNI(0)*N) + 1;
    SET D1 POINT=IOBS NOBS=N;
  OUTPUT;
  END;
STOP;
      *** assign a unique random number for later combining data sets;
DATA BTDATA_&A;
  SET BTDATA; UNIQUE=RANNOR(0);

%IF &A=1 %THEN %DO;
      DATA BTDATA_ALL; SET BTDATA_&A;
    %END;
    %IF &A>1 %THEN %DO;
       PROC SORT DATA=BTDATA_ALL;   BY UNIQUE; RUN;
      PROC SORT DATA=BTDATA_&A;    BY UNIQUE; RUN;

          *** combining bootstrapped samples from each school;
      DATA BTDATA_ALL;
      UPDATE BTDATA_ALL BTDATA_&A;  BY UNIQUE; RUN;
    %END;
%END;
```

```
   *** direct PROC MIXED output to a file on disk;
   *** avoids potential problem of SAS Output Window becoming too full;
  FILENAME MIXEDOUT 'C:\HLM Bootstrap\MIXEDFILE';
PROC PRINTTO PRINT=MIXEDOUT NEW;
RUN;

PROC MIXED data=btdata_all NOCLPRINT COVTEST NOITPRINT;
  CLASS SCHID;
  MODEL MATH = SECTOR CSES SECTOR*CSES/SOLUTION DDFM=BW NOTEST;
  RANDOM INTERCEPT CSES/TYPE=UN SUB=SCHID;
     ODS OUTPUT COVPARMS=CP;      *** output random cov. terms to a SAS-DATA-SET;
     ODS OUTPUT SolutionF=FIXED;  *** output fixed effects to a SAS-DATA-SET;
  RUN;

    *** re-direct the output to SAS output window;
PROC PRINTTO PRINT=PRINT; RUN;

DATA COV; SET CP;
  KEEP CovParm Estimate;
PROC TRANSPOSE DATA=COV OUT=COVOUT LET; RUN;

      *** obtain the variances/covariance of random effects;
      *** Use Bryk and Raudenbush notations;
DATA COVOUT; SET COVOUT;
  DROP _NAME_; RENAME COL1=U0 COL2=U01 COL3=U1 COL4=R;

DATA COEFF; SET FIXED;
  KEEP EFFECT ESTIMATE;
PROC TRANSPOSE DATA=COEFF OUT=COEFF LET; RUN;

     *** obtain the model parameter estimates of the fixed effects;
     *** Use Bryk and Raudenbush notations;
DATA COEFF; SET COEFF;
  DROP _NAME_; RENAME COL1=GA00 COL2=GA01 COL3=GA10 COL4=GA11;

     *** combine the two data sets to have one observation for each bootstrapped
sample;
DATA BOTH; MERGE COVOUT COEFF;

      *** append estimates from each bootstrap iteration;
      *** to a permanent SAS dataset on disk: HSB20_L1ONLY;
PROC APPEND BASE=BTHSB20.HSB20_L1ONLY FORCE; RUN;
%END;                       *** end bootstrap iterations;
%MEND BTRAP;                *** end of bootstrap macro;
%BTRAP;                     *** execute the BTRAP macro;


/*
    *** read in the data of bootstrapped results;
    *** (2000 observations from 2000 bootstrap iterations);

DATA TEMP;
  SET BTHSB20.HSB20_L1ONLY;

    *** obtain some basic descriptive statistics for;
    *** the bootstrapped distributions of the estimates;

PROC means n mean std skew kurtosis min max maxdec=3;
  title 'descriptive statistics of HLM model - HSB data';
  title2 'bootstrap individuals within each school';
RUN;
*/
```

Appendix B

S-PLUS Code for the Nested Bootstrap

```
### Nested Bootstrap for HLM

### The following code is for performing a nested bootstrap within
### the HLM framework using lme in S-PLUS.

### Identify the number of schools or groups here
schools<-c(20)

### In the split command, identify the dataset and then the grouping variable
abc<-split(Hsb20, Hsb20$schid)

### Identify the number of bootstrap samples to be drawn
nboot<-2000

### In this matrix, the number of columns must equal number
### of components to be extracted
results.out<-matrix(0, ncol=8, nrow=nboot)
for (j in 1:nboot){

abc.total<-abc[[1]][1,]
for(i in 1:schools){
  data.index <- sample(nrow(abc[[i]]), size =
                 nrow(abc[[i]]), replace = T)
  temp<-abc[[i]][data.index,]
  abc.total<-rbind(abc.total,temp)}

  abc.total<-abc.total[2:nrow(abc.total),]
  final.data<-data.frame(abc.total)

### Define the lme model here but note that the dataset in this
### case is final.data and not your original dataset
  model.out<-menuLme(fixed = math~sector*cses, data = final.data,
        random = ~  cses  |  schid, method = "ML")

### Define which components you want to extract from lme here
results.out[j,]<-c(VarCorr(model.out)[1], VarCorr(model.out)[8],
            VarCorr(model.out)[2], VarCorr(model.out)[3],
            model.out$coefficients$fixed[1],
            model.out$coefficients$fixed[2], model.out$coefficients$fixed[3],
            model.out$coefficients$fixed[4])}
```