2. Literature Review

2.1 The Concept of Hierarchical Models and Their Use in Educational Research

Inevitably, individuals interact with their social contexts. Individuals' characteristics can thus be influenced by factors attributed to the group they belong to. For instance, students in schools without a gymnasium may not be as athletic as students in schools that have one. Also, features of groups are often driven by the individuals they contain, which means that these individuals are influenced, in turn, by the "emerged" additive feature of the group to which they belong. For example, students coming from high socioeconomic backgrounds may be more likely to attend private schools than students from lower socioeconomic backgrounds. Inversely, this characteristic of students can be descriptive of private schools. In this case, one feature (socioeconomic background) influences individuals in two dimensions (or two levels of a hierarchy), that is, the individual level and at group level. Both may influence, for example, the mathematics achievement of the students. Finally, interactions between variables on both (or even more) levels are possible and may also influence any dependent variable, for instance, achievement.

Simple linear regression models have been used—and still are used—to analyze LSA data. But these models have weaknesses. One is the underlying assumption that individuals answer independently of the cluster they belong to (Burstein, 1980; Rogosa, 1978). Another is the assumption that, in terms of magnitude and direction, relationships within each group are the same as those across groups. Ignoring the nested structure of the data can lead to aggregation bias, ecological fallacy (Cronbach, 1967; Robinson, 1950), and misestimates of the precision (Aitkin et al., 1981; Knapp, 1977). Apart from these technicalities, most linear models do not allow for analyzing the group effect on the individuals or the different effects of an explanatory variable that is group dependent. It is possible to illustrate this problem within the context of the introductory example above by investigating how much of the variability of the achievement scores in the full population can be explained by introducing socioeconomic status (SES) as a group-level effect.

To overcome the constraints of the simple regression models, researchers developed a model that takes the hierarchical structure of the data into account (Aitkin & Longford, 1986; De Leeuw & Kreft, 1986; Goldstein, 1986, Raudenbush & Bryk, 1986). This model, known as the hierarchical linear model (HLM),[4] allows analysts to investigate effects, relationships, and variability at multiple levels. It also permits different intercepts and coefficients at the various levels, thus allowing the model to fit the actual data structure more accurately (Hox, 1995, 1998; Raudenbush, 1988; Thomas & Heck, 2001).

In order to gain a very brief mathematical introduction to these features, consider a basic HLM model for any dependent micro- (or individual-) level variable Y_{ij} of the *i*th individual in group *j* with one micro-level explanatory variable x_{ij} and one macro- (or group-) level explanatory variable z_j. This can be described as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + R_{ij}$$

$$\text{where } \begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01} z_j + U_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11} z_j + U_{1j} \end{cases} \quad \text{and} \quad \begin{cases} R_{ij} \sim N(0, \sigma^2) \\ \begin{bmatrix} U_{0j} \\ U_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix} \right). \end{cases}$$

In this equation, β_{0j} is the random intercept, β_{1j} is the random slope, and R_{ij} is the micro-level error term. Furthermore, γ_{00} is the mean intercept, γ_{01}, γ_{11}, and γ_{10} are the mean slopes at the macro-, cross-, and micro-levels, and U_{0j} and U_{1j} are the macro-level residuals. The three equations can also be combined and written in a linear form as

$$Y_{ij} = \gamma_{00} + \gamma_{01} z_j \gamma_{10} x_{ij} + \gamma_{11} z_j x_{ij} + U_{0j} + U_{1j} + R_{ij}.$$

A main feature of HLM is that parameters (i.e., intercepts and slopes) can be specified as being *fixed* or *random* at all levels, and the error variance/covariance matrix can take different structures. Also, if the theoretical framework of a research hypothesis suggests it, more than one predictor at each level can be introduced; or, aside from the plain effects of the predictors, interaction terms can be included. Finally, models for more than two levels can be formed (e.g., students nested within teachers, teachers nested within schools). Note, however, that the interpretation complexity of any multilevel model is closely related to the model complexity.[5] Because these matters are not the scope of this report, we do not offer further detailed explanations, but instead refer interested readers to Bryk and Raudenbush (1992), Hox (1995), Goldstein (1996), and Snijders and Bosker (1999), all of whom provide excellent introductions to the topic.

4 The HLM is also known as the "multilevel model" (Hox, 1995; Snijders & Bosker, 1999), "variance component model" (Longford, 1993), and "random coefficient model" (De Leeuw & Kreft, 1986).

5 Complexity in the sense used here means increasing the numbers of predictors, introducing more than two hierarchical levels, or considering random instead of fixed slopes.

Data collected in educational LSA generally follow a hierarchical structure. The reason is that such studies usually apply two-stage cluster sampling designs. This specific sampling design implements two selection steps:

1. Clusters are selected from an exhaustive list of clusters (in educational studies, clusters are generally schools);
2. Individuals are selected from inside each cluster sampled in the first stage.[6]

Although these cluster samples have one important disadvantage—a considerable decrease in the precision of the sample (Cochran, 1977; Kish, 1965; Lohr, 1999)—other important reasons make this approach the preferred option. First, it often reduces the costs of the study because it is less expensive to test, for instance, one whole class in each of 150 schools than just one student in 400 schools located countrywide.[7] Second, a simple random sample requires a complete list of all individuals in the target population (e.g., all Grade 4 students in a country), which is usually not available.

Finally, a main research interest with respect to educational LSA involves investigating how group-level variables influence individual-level variables and cross-level interactions, that is, the interaction between variables measured at different levels of the hierarchy. For example, if the relationship between mathematics achievement and the SES of a student differs in terms of the averaged SES of the schools, there is a cross-level interaction.

Although the benefits of using HLM for data analyses have rarely been a critical consideration for assessment designers,[8] many educational researchers have taken advantage of the benefits of hierarchical modeling when endeavoring to best accommodate the existing data structure (e.g., Anderson, Milford, & Ross, 2009; Baker, Goesling, & Letendre, 2002; Braun, Jenkins, & Grigg, 2006; Cheong, Fotiu, & Raudenbush, 2001; Desimone, Smith, Baker, & Ueno, 2005; Green, Camilli, & Elmore, 2006; Koretz, McCaffrey, & Sullivan, 2001; Lamb & Fullarton, 2001; Lubienski & Lubienski, 2006; Ma & McIntyre, 2005; Pong & Pallas, 2001; Rumberger, 1995; Wang, 1998; Wenglinsky, 2002).

6 Various international LSA incorporate yet a further sampling step: within schools, classes are selected, and within the selected classes, all students are selected or a subsample of students is selected (as in, e.g., TIMSS and the Progress in Reading Literacy Study [PIRLS]).

7 Both designs are comparably efficient (assuming a moderate intraclass correlation coefficient of about 0.3), although the total sample sizes deviate by factor 10.

8 Most LSA in education have international comparisons of population estimates and trend measurement as main focuses. Study designs are mainly driven by these focuses.

2.2 Precision of the Estimates in Multilevel Models for Complex Sample Survey Data

One goal that researchers try to meet when designing a survey is to achieve a certain target level of precision for estimates of the population parameters so that they can ensure that the estimate—calculated using sample data—reflects the actual value in the population within specific margins of error.[9] Researchers may also want to detect a difference between certain groups, expose the effects of covariates, and allege interactions between different independent variables—all activities leading to conclusions that can be made only within certain confidence levels. This happens because the inference pertaining to the population is based on data collected from a sample. A measure that can be used to determine the precision of any sample estimate is the standard (or sampling) error, which allows researchers to calculate confidence intervals.

In general, the sampling error is a monotonic decreasing function of the sample size (Snijders & Bosker, 1999), and it is further affected by population variance. If complex sampling designs are applied, additional factors influence the sampling error. First, data collected from clusters are not independent. For example, students within a class are more alike than students from different classes because all members of the former group receive the same tuition from a teacher. A measure that illustrates this effect is the intraclass correlation coefficient (ICC). It displays the ratio of the between-group variability to the total variability and ranges from 0 to 1 (Kish, 1965). During estimation of sampling error for complex samples (assuming simple random sampling), sampling error estimates become downwardly biased as ICCs increase. To overcome this obstacle, sampling errors are estimated using repeated replication methods such as Jackknife Repeated Replication or Balanced Repeated Replication. This use is very common in educational LSA (e.g., Olson et al., 2008. Organisation for Economic Co-operation and Development [OECD], 2006, 2009).

Reference to an extreme example illustrates the meaning of the ICC and its effect on precision. Imagine that all students within different classes are identical, but that students from different classes differ from one another (ICC = 1). We will not obtain any further information about the population if we sample more students within the selected classes. In other words, the precision will not increase as sample sizes within clusters increase.

Multilevel models reduce the impact of ICC on the precision of the parameter and sampling error estimates. Maas and Hox (2005) report, for example, that starting with ICCs larger than 0.1 produces biased estimated parameters and sampling errors only when fewer than 30 clusters are sampled. Nevertheless, intraclass correlation

9 In the literature, most authors use the term "standard error" instead of "sampling error." In many circumstances, both terms have identical meaning. However, LSA often use the plausible value technique for (at least) their main outcome variables (see Von Davier, Gonzalez, & Mislevy, 2009). In these instances, the standard error captures two sources of variation—sampling error and measurement error. The measurement error is not a focus of this research. To avoid confusion, we consistently use the term "sampling error" throughout this report.

continues to be specified as one important factor influencing the quality of estimation (Asparouhov & Muthén, 2006; Asparouhov et al., 2006) and should therefore be taken into account.

Although Scherbaum and Ferreter (2009) report that the estimation of ICCs a priori (i.e., before the actual survey is done) is difficult, this is not true for most educational LSA. This is because excellent data sources are available for many participating countries from which to estimate ICCs reliably. These sources include databases from previous cycles of a survey, or surveys with similar subjects or similar target populations. Many of these databases are publicly available.[10] Note, however, that ICCs vary from one variable to the next and may vary across survey cycles.

In summary, sampling errors within multilevel models are no longer simple monotonic functions of the total sample size. As a general rule, the higher the ICC, the less the increase in precision if the sample size within clusters is increased. We review this aspect in more detail in Section 2.3.

As we have already mentioned, educational LSA require implementation of complex sampling designs. Weights reflect multiple sampling steps, selection probabilities, and non-response at each sampling stage. The use of sampling weights for estimating population parameters is a well-established procedure (see, for example, Cochran, 1977). If the probabilities of selection are ignored, the parameter estimates can be substantially biased. In most cases, the use of weighted data also affects sampling errors. Despite these occurrences, the use of sampling weights in HLM analysis has only recently been addressed in the literature.

Among those who have discussed the biased parameter estimates that occur when standard multilevel modeling without weights is used are Korn and Graubard (2003), Longford (1996), Pfeffermann et al. (1998), and Rabe-Hesketh and Skrondal (2006). Asparouhov et al. (2006) provide a discussion of the different methods of normalizing sampling weights and their impact on parameter estimation. They also offer guidelines on how to scale weights under specific conditions. Chantala et al. (2006) provide programs in Stata and SAS that allow computation of correctly scaled weights for multilevel modeling of complex survey data. Zaccarin and Donati (2008) evaluate the influence of different choices of sampling weights in HLM on PISA results.[11]

Pfeffermann, Moura, and Silva (2006) suggest a model-based approach instead of probability weighting under informative sampling designs. They found that their approach outperformed probability weighting under certain conditions in a simulation study but admitted that the latter approach is far easier to implement and needs significantly less computational power.

Finally, the inclusion of covariates at either level can influence the precision of multilevel models. This is because of their potential to reduce the between-group variance (Raudenbush, 1997; Reise & Duan, 2003).

10 For example, all databases from previous cycles of TIMSS and PIRLS can be downloaded at http://timssandpirls.bc.edu/ or from www.iea.nl, together with all technical documentation and user guides.

11 Programme for International Student Assessment, conducted by the OECD: http://www.pisa.oecd.org/

2.3 Sample Size Requirements and HLM—Knowledge at Hand

A general problem associated with applying any method designed to define optimal sample sizes[12] is that the sample sizes optimal for, say, the estimation of a population parameter might not be optimal for the test of, for example, a cross-level interaction effect. As Snijders and Bosker (1999) aptly point out, the fact that optimality depends on one's objectives is a general problem of life that cannot be solved by reference to a textbook.

Over the past 15 or so years, several research projects, many of which are simulation studies, have endeavored to address the issue. Only a few studies have examined the impact of various factors on statistical precision and sample sizes in hierarchical models as well as their interactions. We review the most important of these studies below.

Snijders and Bosker (1993) developed approximation formulas to calculate optimal sample sizes on two-level designs for fixed regression coefficients. They evaluated their work as being valid for sample sizes with more than 10 units on both levels. Applying their formulas to an example, a consideration of budget constraints, they showed that if small sampling errors of regression coefficients are to be achieved, then higher sample sizes at the macro-level are always preferable to increasing the sample sizes within clusters. Sampling errors increase if the number of sampled clusters decreases. This situation holds true if the total sample size is kept constant, and even when the total sample size increases. Snijders and Bosker's example also makes clear that the sampling errors of the regression coefficient of a macro-level effect are much more sensitive to sample sizes than are interaction effects between two different macro-level variables.

Afshartous (1995) addressed the topic of estimation bias in hierarchical modeling due to small samples. He showed that necessary sample sizes of micro- and macro-level units respectively vary depending on whether the interest is mainly in obtaining accurate and reliable estimates for variance components or for fixed effects. He found, in a specified setting, that 320 schools were needed in order to obtain unbiased estimates of variance components, whereas as few as 40 schools appeared to suffice for estimation of regression coefficients. However, Afshartous admitted that this effect might depend on the type of fixed effect being studied (e.g., intercept or slope). Also, Afshartous used only one specific dataset for his research (NELS[13]) and analyzed clearly delimited subsamples of the base dataset.

In a very thorough study, Mok (1995) investigated samples of students derived from a real dataset pertaining to 50 schools. She set a fixed total sample size, let the number of schools and students within schools vary, and then considered a variety of estimators, including regression coefficients, variances, and covariances. In agreement with other authors, she found that designs using more schools and fewer students are more

12 Here, "optimal sample size" refers to a sample size that will meet certain precision requirements.

13 National Education Longitudinal Survey, U.S. Department of Education.

efficient than designs that allocate sample sizes the other way around. Based on her review of simulation studies, Kreft (1996) offered a 30/30 rule of thumb, leading to a minimum total sample size of 900, no matter what type of effect is studied. Bell, Morgan, Schoeneberger, Loudermilk, Kromrey, and Ferron (2010) have since argued against this viewpoint, claiming that this commonly cited rule would likely not yield high levels of statistical power for the fixed effects at both levels of the model.

Raudenbush (1997) made clear the fact that inclusions of covariates have an impact on the optimal design. Covariates are non-negligible because they explain substantive parts of the variance of the dependent variable. According to Raudenbush, the explanatory power of the covariate at each level becomes highly relevant for choosing optimal sample sizes. Raudenbush also focused in his paper on the efficiency of cluster randomized trials and considered cost implications. Snijders (2006) added to this aspect by observing that the reduction in sampling error depends on the intraclass correlation of the dependent variable and on the within-group and the between-group residual correlation between the dependent variable and the covariate.

Moerbeek, Van Breukelen, and Berger (2000) have also described how to allocate sample sizes to the macro- and micro-level in a cluster-randomized trial. The authors considered different treatments and budget constraints, and aimed for specified levels of power with regard to treatment effects. In another article, these authors again investigated this topic, but this time their focus was on binary outcome variables (Moerbeek, Van Breukelen, & Berger, 2001).

Cohen (1998) implemented an approach similar to that of Snijders and Bosker's (1993). He reported that the estimation of micro-level variances requires larger samples within clusters (and hence fewer clusters, assuming a fixed cost budget) than does estimation of traditional quantities, such as means, totals, and ratios.

Hox (1995) provided another rule of thumb. He advocated sample sizes of 50 clusters and 20 individuals per cluster as appropriate for multilevel modeling.

Maas and Hox (2005) carried out a simulation study with varying numbers of clusters ($N = 30, 50, 100$), varying cluster sizes ($n = 5, 30, 50$), and varying intraclass correlations (ICC = 0.1, 0.2, 0.3) in order to explore the effect of these variations on parameter estimates and estimates of their sampling errors. The authors found that the regression coefficients and variance components were all estimated with negligible bias (using restricted maximum likelihood as the estimation method). Also, sampling errors for regression coefficients were estimated correctly. However, the authors stated that sampling error estimates of macro-level variances were downwardly biased when the number of clusters was substantially lower than 100 (i.e., 50 or 30 in their study).

Snijders (2005) took a more general approach when addressing the topic. He pointed out that the sample size at the micro-level (i.e., the total sample size) matters if the effect of a micro-level variable is of main interest, and (vice versa) that the sample size on the macro-level is more important when testing a main effect of a macro-level variable. He concluded that, in most instances, a sample with more macro-level

units will be more informative than a sample where the within-cluster sample size is enlarged but fewer clusters are selected. He also explained that small cluster sizes are unproblematic when testing regression coefficients but have a negative impact on test power when testing random slope variances at the macro-level. Snijders gives, in line with suggestions made during an earlier work (Snijders & Bosker, 1999), some formulas that can be used to obtain insight into the design aspects that are most influential on power and sampling errors. Both sources indicate that the formulas will give only very rough estimates of the required sample sizes if several correlated explanatory variables, some of which will have random slopes, are to be introduced in the model.

Okumura (2007) presented a new simulation-based approach to determine optimal sample sizes for HLM that lead to desired levels of statistical power and mean ranges of confidence intervals. Specifically, his method acknowledges uncertainty in parameter values, given the posterior distribution for the unknown parameters. Okumura cited, as disadvantages of his approach, the fact that the method takes much more computational time than existing techniques and that it is very difficult to adapt computer programs to meet specific model conditions.

Finally, various computer programs are available that enable users to conduct power estimations under specific conditions. The two programs that serve modules closest to the object of our interest are PinT[14] and OD.[15]

PinT (*Power in Two-level designs*) calculates sampling errors of regression coefficients in two-level designs as a function of fixed total-sample sizes. It also takes into account cost constraints. According to Snijders and Bosker (1999), the greatest difficulty in using this software is that means, variances, and covariances of all explanatory variables and random effects have to be specified. Furthermore, the program uses relatively rough, large sample approximations to obtain sampling errors.

The other program, OD (*Optimal Design*), calculates power and optimal sample sizes for testing treatment effects and variance components in multisite and cluster-randomized trials with balanced two-group designs, and in repeated-measurement designs (Raudenbush, Spybrook, Liu, & Congdon, 2005). Because LSA are generally observational surveys rather than experimental ones, this program is another that does not fully fit the needs of sample-size calculations for these assessments.

14 With manual available for free download at http://stat.gamma.rug.nl/multilevel.htm#progPINT

15 With manual available for free download at http://sitemaker.umich.edu/group-based/optimal_design_software

2.4 Cost Implications of Sample Size Considerations

All decisions pertaining to sample sizes have cost implications. Many authors have therefore addressed this issue in their research and tried to optimize sample size to accommodate budget constraints (e.g., Cohen, 1998; Moerbeek et al., 2000, 2001; Mok, 1995; Snijders & Bosker, 1993, 1999).

Cost reductions are often obvious and significant when fewer macro-level units need to be selected for a survey because, in most cases, it is more expensive to survey one more cluster than one more individual within each cluster. However, reducing micro-level sample sizes could also have significant cost implications in certain circumstances. This effect will typically show up in surveys that do not have predetermined micro-level sample sizes.[16] These surveys are often those carried out with non-student target populations. Inservice teachers in ICCS[17] and TALIS[18] and future teachers in their final year of training and their educators in TEDS-M[19] provide examples of these populations. With these surveys, the decision to select, for instance, 15 or 20 teachers from a total of 150 schools does indeed matter because the need for high participation rates often makes necessary considerable engagement with personnel or the setting of incentives, such as payments. Faced with limited budgets, researchers need to focus on securing optimal survey designs that have minimum cost implications.

16 In student surveys, full classrooms are often surveyed. In these cases, the within-cluster sample size is predetermined.

17 International Civic and Citizenship Education Survey, conducted by IEA: http://iccs.acer.edu.au/

18 Teaching and Learning International Survey, conducted by OECD: http://www.oecd.org/edu/talis

19 Teacher Education and Development Study in Mathematics, conducted by IEA: http://teds.educ.msu.edu/