Sample Size Requirements in HLM: An Empirical Study

The Relationship Between the Sample Sizes at Each Level of a Hierarchical Model and the Precision of the Outcome Model

Sabine Meinck and Caroline Vandenplas

IEA Data Processing and Research Center, Hamburg, Germany

This report was funded by the National Center for Education Statistics under Contract No. ED-08-CO-0117 with the International Association for the Evaluation of Educational Achievement (IEA). Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

IERI Monograph Series
Issues and Methodologies in Large-Scale Assessments

Special Issue 1
Sample Size Requirements in HLM: An Empirical Study
The Relationship Between the Sample Sizes at Each Level of a Hierarchical Model and the Precision of the Outcome Model

Sabine Meinck and Caroline Vandenplas

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

Main Text

Appendix

Acknowledgements

The authors would like to express their warmest thanks to several people who helped them to complete this project. We thank Dirk Hastedt for encouraging us to take on this study. We thank Heiko Sibberns, Leslie Rutkowski, Wolfram Schulz, Eugenio Gonzalez, and the anonymous reviewer for their thorough evaluation of the paper and their valuable feedback. We are also grateful for the help of our colleagues in the Sampling Unit and the Research and Analysis Unit at the IEA Data Processing and Research Center, who helped us with various details throughout the study. Without the generous help of these individuals, this investigation would not have been possible. Finally, we thank the National Center for Education Statistics (NCES), U.S. Department of Education, Institute of Education Sciences, for funding this research project.

Abstract

This study focused on the properties of data collected in large-scale assessments (LSA) in order to explore the relationships between sample sizes at different levels of clustered data and the sampling precision of the results derived from hierarchical linear models (HLM). A Monte Carlo simulation study was used in order to explore various population and sample conditions. The varied conditions were sample sizes of and within clusters, intraclass correlation coefficients, covariance distribution, use of sampling weights, and model complexity. As expected, the precision of all explored parameters increased as sample sizes increased. The dependency took a nonlinear format—a general observation that held true for all settings. The magnitude of the increase, and whether the effect became more pronounced as sample size increased on either of the hierarchical levels, could depend, however, on all explored sample and population conditions and could also vary across the different model parameters. In conclusion, the results showed that required sample sizes depend heavily on the parameter of interest. In particular, sampling precision differed widely for fixed model parameters versus variance estimates. For certain model parameters, the effect of how the covariance was distributed between the hierarchical levels appeared to be even more pronounced than the effect of varying sample sizes. The inclusion of sampling weights in the model decreased the sampling precision of all explored parameters consistently by approximately 10%. The model complexity had an influence on the sampling precision of all observed parameters except the residual variance. The influence thus varied according to the parameter of interest as well as the considered case of covariance distribution.

1. Introduction

Beyond controversy is the premise that education is an important factor influencing the development of national economies worldwide (Brown & Lauder, 1996; Decker, Rice, & Moore, 1997). National assessments exploring the quality and outcomes of education systems have consequently become popular in recent decades, while accretive levels of globalization have led to education increasingly being viewed from within a broader context (Dale, 2000; Suárez-Orozco & Qin-Hilliard, 2004). These developments have heightened interest in international comparative studies of education, many of which include large-scale assessments (LSA). The increasing number of educational surveys conducted by the International Association for the Evaluation of Educational Achievement (IEA) and the Organisation for Economic Co-operation and Development (OECD) are evidence of this growing interest.[1]

When analyzing data collected in large-scale educational surveys, researchers still tend to use (or to suggest the use of) simple linear regression models (Foy & Olson, 2009; Olson, Martin, & Mullis, 2008). While the application of these models is appropriate for certain types of analyses or data structures, limitations regarding their usefulness become apparent when the data have a nested structure, that is, follow specific hierarchies (Aitkin, Anderson, & Hinde, 1981; Robinson, 1950). Simple linear regression models do not consider the effects of multiple factors on different levels of the hierarchy or on their interactions. These limitations can be avoided by using hierarchical linear modeling (HLM) (e.g., Bryk & Raudenbush, 1992; Hox, 1995; Snijders & Bosker, 1999). HLM takes the multilevel structure of a comparison problem into account and allows predictors to be introduced at different levels, thereby making it possible to study the effect of the variables at the specific level in which they occur.

HLM is usually excellently suited for analyzing data collected in educational surveys. The education systems with students embedded in classes, classes embedded in schools, schools in districts, and districts in countries display the data structure for which HLM techniques were developed. In addition, general sampling strategies in international LSA generally imply the same hierarchical approach (see, for example, Martin, Mullis, & Kennedy, 2007; Olson et al., 2008).

1 http://www.iea.nl; http://www.oecd.org/edu

The first stage of the approach involves, in each participating country, selecting a sample of schools and stratifying them according to certain organizational criteria (e.g., public versus private, or regions comprising different strata). The second stage sees classes and/or students sampled from within each participating school. The hierarchical data structure also opens a window into broadly defined concepts of student achievement in relation to some correlates of learning, such as socioeconomic (SES) background and school resources.

Given these advantages, it is not surprising that more and more researchers want to employ HLM analysis in this field of research. However, this desire has to be taken into account when developing the general study design of an educational assessment. Researchers need to be aware at this time of an important problem associated with designing studies suitable for multilevel modeling, namely the required sample sizes at the different levels of the hierarchy (see, for example, Maas & Hox, 2005; Scherbaum & Ferreter, 2009; Snijders & Bosker, 1999).

In recent years, a number of researchers have tried to address the problem by conducting (mostly) simulation studies with certain conditions in order to produce rules of thumb or even software that enable users to determine the optimal survey design. However, the literature available on the subject tends to be highly technical, hard to apply, and not easily procured. Most importantly, existing simulation studies are based on assumptions that do not fully apply to data collected in educational LSA, either because they fail to or only partially address the features typical of these datasets.

But what are the characteristics of typical LSA survey designs? In general, minimum sample sizes in LSA are predetermined by multiple factors, such as the requested precision of population estimates, the number of items and the item rotation design (connected to the need to have minimum response numbers per item), minimum cell assignments in cross tables, and so on. For example, most IEA surveys specify a minimum sample size of 150 schools to ensure that certain precision requirements are met. To give another example, the item rotation design applied in studies such as TIMSS[2] calls for a sample size of at least 4,000 tested students per education system because each tested student takes only one-seventh of the whole assessment (Olson et al., 2008). In this second case, the total student sample size is dictated by the item rotation design while the total cluster (school) sample size is dictated by the precision requirements and the design effect. Furthermore, cluster sampling of classes often dictates within-cluster sample sizes of about 20 to 30 individuals per cluster.

In addition, data originating from complex surveys carry weights that reflect the multiple selection probabilities of each unit, adjusted for non-response. Although general sampling designs usually aim for self-weighted samples (e.g., Joncas, 2008),[3] estimation weights always vary due to stratification, practical constraints associated with implementation of the ideal sampling design, and non-response adjustments,

2 Trends in International Mathematics and Science Study, conducted by IEA: http://timss.bc.edu/

3 Samples that lead to equal selection probabilities of the units of interest are called self-weighted samples.

a situation that can lead to increased sampling variance. Since the development of multilevel analysis techniques, the need to consider sampling weights when engaged in multilevel modeling, as well as the influence of that modeling on estimates, has attracted attention (albeit limited) in the literature (see, for example, Asparouhov, Muthén, & Muthén, 2006; Chantala, Blanchette, & Suchindran, 2006; Korn & Graubard, 1995; Pfeffermann, Skinner, Holmes, Goldstein, & Rabash, 1998; Rabe-Hesketh & Skrondal, 2006; Stapleton, 2002; Zaccarin & Donati, 2008). However, no mention seems to have been made in this body of work of relationships between sampling weights, the statistical precision of the models, and required sample sizes.

All these constraints suggest the desirability of an evaluation of the sample sizes required to achieve a predetermined level of precision when applying multilevel modeling oriented toward the specific structure of data collected from educational large-scale assessments. Our aim, therefore, in this paper is to extend knowledge about the association between sample sizes and precision of the estimates under varying population and sample conditions and relative to model complexity.