WHAT ARE MULTILEVEL QUESTIONS, AND HOW MIGHT WE EXPLORE THEM WITH QUANTITATIVE METHODS?*

Valerie E. Lee
Professor of Education at the University of Michigan, Ann Arbor, MI, USA

Resumo
Sistemas educacionais possuem uma estrutura hierárquica, na qual estudantes estão agrupados em turmas, que se agrupam em escolas. Essas estruturas hierárquicas têm diferentes níveis – o nível do aluno; o nível da turma; o nível da escola. Questões de pesquisa multiníveis buscam investigar relações entre fatores que atuam em distintos níveis. Este artigo discute as principais limitações das tentativas de abordar questões de pesquisa multiníveis com técnicas analíticas de um único nível e introduz algumas características relevantes da modelagem multinível, uma abordagem capaz de lidar adequadamente com questões multiníveis.

Palavras-chave: Modelos multiníveis; métodos quantitativos; sociologia da educação.

Resumen
Los sistemas educacionales contienen una estructura jerárquica en la cual los estudiantes se agrupan en clases y en escuelas. Esas estructuras jerárquicas poseen diferentes niveles: el nivel del alumno, el nivel de la clase y el nivel de la escuela. Cuestiones de pesquisas de múltiplos niveles procuran investigar relaciones entre los factores que actúan en los diferentes niveles. Este artículo enfoca las principales limitaciones de las tentativas de tratar de cuestiones de pesquisa de múltiplos niveles con técnicas analíticas de un único nivel e introduce algunas características relevantes del modelo múltiplo nivel, y una visión capaz de trabajar adecuadamente con cuestiones de múltiplos niveles.

Palabras-clave: Modelos de múltiplo niveles; métodos cuantitativos; sociología de la educación.

Abstract
Educational systems exhibit a hierarchical structure in which students are grouped in classrooms, classrooms in schools. Those hierarchical structures possess different levels – for instance, the students level, the classroom level, the school level. Multilevel research questions are those questions that seek to investigate patterns of relationship between factors acting at different levels. This paper outlines the main pitfalls of trying to investigate multilèvel research questions with single level techniques and introduces some features of multilevel modeling, an approach capable of dealing adequately with multilevel questions.

Keywords: Multilevel models; quantitative methods; sociology of education.

* Visit and seminar sponsored by Pontifical Catholic University of Rio de Janeiro and National Research Council of Brazil.

Introduction

An honor. I want to thank you for inviting me to come talk with you about a research methodology that is hopefully of some interest and use to you. Although I suspect that I was invited to come here to talk about research methodology, I don't consider myself – strictly speaking – a methodologist. Rather, I am a sociologist of education. When quantitative multilevel research methods first became available, about 15 years ago, the "multilevel club" had very few members. Now there are many more of us, many of whom are much more statistical than I am. I do research using these methods and teach others to use them, but I don't develop new methodologies. What has linked me to Brazil for this trip is that in the last few years, two of your countrymen came from Brazil to Ann Arbor, Michigan, to take my course in multilevel research methods – Creso Franco from PUC/Rio in June 2000 and Claudia Travassos, a public health professional in Rio. Having Creso and Claudia in Brazil, but having studied with me in the U.S., makes this visit quite special for me.

The audience. I suspect that there are perhaps three or four different groups in the audience. Group 1 includes people who have engaged in a systematic study of multilevel research methods, and perhaps they have been using these methods in their research for some time. That's the advanced group, and I assume it is pretty small (Creso would surely be in this group). Group 2 is composed of people who may have some knowledge of this methodology, and perhaps they see it as useful for their work. These people want some details but may understand the basics. This group is also small, although more numerous than Group 1. Group 3 is composed of people with a general interest in quantitative research methods and a desire to gain knowledge and skill in many methodologies. Here is one more methodology they might consider, and perhaps they would find a use for it some day. Members of Group 4 were probably encouraged to come by someone, but perhaps haven't spent a lot of time using or studying quantitative research methods. I will try to raise some issues of interest to each group.

Present at the birth. My own experience with multilevel methods is limited in a couple of ways. Although there are now several different multilevel statistical software packages available, my own experience and teaching is confined to only one: called Hierarchical

Linear Models, or HLM. People who are familiar with the different multilevel software packages have generally concluded that HLM is easier to use than the others. Another limitation relates to how I use the methodology in my own research. Mostly I focus on the context of "schools", and I conduct what is known as "school effects research". I hope that some of you may share those interests. But it is important that you understand where I am coming from.

My introduction to and familiarity with HLM comes through the people who developed the HLM software and wrote the major textbook on it: Steven Raudenbush and Anthony Bryk. Steve and I were in graduate school together at Harvard University in the early 1980s, and Tony Bryk was a young Harvard faculty member who advised both of our dissertations. Steve was a year ahead of me, we shared an office in graduate school, and I used Steve's original (and quite primitive) HLM program in one chapter in my dissertation. Steve is now my faculty colleague at the University of Michigan, after having spent more than a decade on the faculty at Michigan State University. We both continue to work with Tony Bryk, who is now a professor at the University of Chicago. Thus, I am not an impartial consumer/user of multilevel methods. I'm partial to HLM and to school effects studies, and I've learned what I know from Bryk and Raudenbush. For these reasons, you might describe me as "a biased consumer."

Organization of the talk. I have organized my comments to recognize the several levels of familiarity with quantitative methodology and multilevel methods represented in the audience. I've set up my remarks as a series of questions and answers. The questions begin at a pretty basic level, and hopefully progress to a somewhat more developed discussion of multilevel questions, research, and ideas. Here are five questions:

- What are multilevel research questions?
- What are multilevel methods, and why do we need them?
- Can you use single-level methods to explore multilevel questions?
- Which kinds of research questions require multilevel methods?
- What are the drawbacks of using multilevel methods?

Question 1
What Are Multilevel Research Questions?

Nested data. Let me locate the discussion of this question in the context of a hypothetical study – the evaluation of a simple social intervention. Let's say that a team of researchers has been hired to evaluate whether a particular and well defined social intervention program is effective. In this case, the hypothetical program is a course designed to teach secondary school dropouts to use computers over a 3-month period (most likely this course, with a well-defined curriculum, is part of an employment-training program). It is reasonable to assume that the intervention is delivered by teachers in classrooms to groups of students. Thus, our evaluation design should recognize that individual students are "nested" in classes. This nesting raises a central issue for any quantitative research: "What is the unit of analysis?" Here, the intervention is delivered to groups – to classes of young adult students. However, we would expect that the outcome – let's say what we want to measure is "the development of computer proficiency" – would accrue to individuals.

Units of analysis. This design represents the essence of a multilevel question. Why? Because there is more than one unit of analysis – here, the two units are students and classrooms. Any evaluation asks at least three important questions: "Does the program work?" "What do we mean by 'works'?" and, finally, "Works, compared to what?" The first question is the bottom-line evaluation question. The second question pushes evaluators to decide on a dependent variable. In this instance, let's assume that we use a well developed and familiar timed test of computer proficiency (perhaps word processing). Our evaluation team would administer this test to individuals at the end of their 3-month experience in either the intervention or comparison-group classes. Ideally, this variable should also has good statistical properties; it should a normally distributed, continuous, and reliable variable. For the third question, let us assume that we also have access to a reasonable comparison group... perhaps an ordinary computer-training curriculum, also delivered to students in classes. As stated, this represents a standard evaluation design: (1) a well-defined program to evaluate, (2) a control group with which to compare the treatment's effect, and (3) an outcome variable on which to evaluate the program.

What data do we need? To conduct this evaluation, we will need to collect some data about the individual students in these programs. Besides the students' performance on the outcome variable, we might want to know their age, gender, education and skills gained before they left high school, their attendance record in the intervention classes, their familiarity with computers prior to participating in the program, how much time they practiced during the program, and perhaps some measure of their cognitive ability. We would also want to collect some data about the intervention and control-group classes: the number of students in them, how long they lasted (number of hours per day, days per week), qualifications of the teaching staff, something about the computer equipment, how much practice time was available (and whether it was supervised), and of course whether the classes used the intervention curriculum or the comparison curriculum. The point here is important: we want to take into account any information that might provide an alternative explanation for why students in the intervention classes did well or poorly on the outcome measure, compared to those in the control group classes. For this evaluation, then, we would collect data about both the classes and the individuals in them, and carefully administer the instrument to measure the outcome. These are the usual requirements for using multilevel methods like HLM: (1) data on individuals, (2) data on groups, and (3) a normally distributed outcome measure. Here is a larger issue: If we want to know about how participation in groups influences individuals in those groups, we have a multilevel research question. And we need to use the appropriate methodology to analyze our data.

Question 2
What Are Multilevel Methods and Why Do We Need Them?

No need to choose a single unit of analysis. Before multilevel methods were widely available and people were trained to use them (which wasn't very long ago), researchers had to address questions like the one I invented for this evaluation example at only one unit of analysis: either at the level of the individual (most common) or at the level of the group. But either decision ignores the essential "nesting" of the phenomenon under study; in this example, the nesting of students in classes. If we had data on enough classes, we might have decided to use the class as the unit of analysis, and conduct our study using average (or aggregate) data on individuals in each class. At that point, we would be discarding a lot of information about differences between the individuals in the same class.

The idea of independence. Please forgive me for getting a bit more technical here. Staying with the computer-training example, we need to think about an appropriate dependent variable – I suggested some measure of computer proficiency (typically a timed test of skills – maybe word processing). We would measure this skill on a large number of students, who experienced some kind of training delivered in different classes (some using the curriculum we want to evaluate, some with a generic curriculum). Remember that the outcome variable is a normally distributed continuous variable. As you may recognize, the major single-level method we might have wanted to analyze the data for this study would be ordinary least squares regression – we have a continuous outcome and predictor variables that are both continuous and dummy-coded. However, one of the major assumptions of regression – one that should not be violated – is that the cases are independent of one another. But if we analyzed our data for this evaluation at the individual level, we would – because of the design – violate that assumption. We can't assume that individual students are independent of one another, particularly when they're grouped in classes for instruction on the computer. Our outcome measure, computer proficiency, is in a form that is appropriate for using both regression and HLM.

The ICC. Enter the notion of "the intra-class correlation" (ICC). This simple idea is very important for understanding and using multilevel methods. Whenever we have a multilevel data structure and a multilevel research question, we need to ask a simple but important question: "What proportion of the total variance in the dependent variable lies systematically between groups?" Computing the ICC allows us to answer that question. It also allows us to test whether the independence assumption is valid, and whether regression would be an appropriate technique to use. In a two-level HLM (the most common type), we assume that variance is either between groups or between individuals in those groups. When we compute the ICC, we partition that variance into those two categories: within groups and between groups.

If the ICC – the between-group proportion of variance – is very low, say below ten percent), then our assumption of independence is probably valid. However, if a substantial proportion of the variance in the outcome is between groups, then multilevel methods are not only suggested, they really are required. Computing the ICC is quite easy with HLM. In fact, the first step in any multilevel analysis is to evaluate the intra-class correlation for the dependent variable. Remember, we are talking about partitioning the variance into its between-group and within-group portions. If your research question and data suggest the value of using multilevel methods, but your ICC is quite low, it also indicates that the success you will have in finding multilevel effects will probably be quite limited. In this work (beyond our hypothetical example), besides assuming that our dependent variable is continuous and normally distributed, we'll also assume that it is measured with reasonable reliability, something you also learn about when you use HLM to compute the ICC. If any of these assumptions don't hold (especially if the outcome variable is not very reliable), you've got a problem finding effects with any quantitative analysis method – including HLM.

Question 3
Can You Address Multilevel Questions with Single-Level Methods?

I think that from what you've heard me say so far, you might know my answer: no. If you have multilevel data and multilevel questions, then single-level analyses won't do. You're violating important assumptions. On the other hand, if you find that the ICC is quite low, even if the data were collected with a multilevel research design, you're on firmer ground in using single-level methods like regression or analysis of variance. But it would mean that you should reformulate your research questions, so that they are no longer multilevel. But if your research requires a multilevel research question (such as our example above did – looking for classroom-level effects on individual outcomes), you're in trouble. A low ICC suggests that there isn't much between-group variance in your outcome and that you won't be able to find group-level effects.

Four problems. For those of you who are even a little familiar with quantitative methods, you may recognize other problems you might encounter when using single-level methods with multilevel data. Let spell out four common problems. The first, which we call "aggregation bias," occurs when we try to capture a phenomenon at one level by measuring it at another. For example, we in education find that it is sometimes quite difficult to obtain measures of individual children's socioeconomic status, or SES.

Children's school records often don't contain information on family income or parents' education; U.S. parents would typically find it intrusive to be asked to provide such information. However, U.S. schools have a program whereby poor children are provided with free or reduced-cost school meals.

Thus, the school might be willing to provide a proxy measure of SES at the school level – the proportion of low-income children in the school (usually measured the proportion of students receive free or reduced lunches). Schools may not be willing to tell you which children were eligible. In this circumstance, that measure at one level – for example, school average SES – doesn't indicate the SES of any individual student at the school. Quite simply, aggregated variables measure group characteristics, which almost always mean something different than they would mean for individuals. A second major problem is "mis-estimated standard errors". We talked about this earlier. If we use individual-level analysis for an outcome that really has a considerable part of its variance between groups, then the standard errors associated with significance testing of predictors would be wrong. A third problem is technically labeled "heterogeneity of regression slopes." I'll talk more about this in a minute. Here, let me simply say that if we assume that the relationship between a dependent variable and the independent variable among individuals is the same within all groups, we might not be right. Moreover, the different relationships across groups might really be interesting to explore (a really powerful feature of HLM).

The fourth problem you encounter when using single-level methods for multilevel questions is that group-level effects are systematically underestimated this way. If it is group-level effects you care about (and if your research question is multilevel, that's what you do care about), you would be less likely to find effects. With our evaluation example, we would be unlikely to find differences between the treatment and control groups in computer proficiency if we did not investigate these effects in the multilevel format (here, individual students nested in classrooms).

Perhaps some of you are familiar with the "Coleman Report". In the early 1960s, sociologist James Coleman was commissioned by the

U.S. Congress to conduct a large-scale study of U.S. schools – with the aim of documenting inequalities in school resources that characterized schools that served students of different races. At the same time, Americans believed that social interventions – financed by the federal government – could correct social inequalities, in this case in schools serving mostly Black or mostly White students.

Coleman and his colleagues collected data on thousands of schools and hundreds of thousands of school children. This was the largest social research study ever undertaken at that time, and the researchers made use of computers to analyze their data for the first time. The conclusions from the Coleman Report, which appeared in 1966, were upsetting and startling.

- They found that school resource differences really didn't make much difference in explaining the substantial race differences in student achievement and learning.
- Instead, Coleman and his colleagues concluded that the most important explanatory factor of achievement was children's home background, particularly their socioeconomic status (SES).

These findings caused a furor in the social policy world. "Schools don't count?" "How could this be?" We now recognize that Coleman and his colleagues were, quite simply, asking multilevel questions, but addressing them with single-level methods. The Coleman Report encountered this fourth problem. They didn't find the "school effects" they were looking for because they didn't use the right methods to analyze their data.

Of course, you can't blame them – there were no multilevel methods in the 1960s. Analysis of variance and regression were "advanced methods" back then. And James Coleman, who died about four years ago, went on to a very distinguished career as a sociologist. In fact, it was Coleman who convinced Tony Bryk to leave Harvard and come to the University of Chicago, mainly to develop the HLM methodology. A positive outcome from the Coleman Report is the basis of my own work. Many people were determined to "prove Coleman wrong", to demonstrate that "schools do count". I'm one of those people.

Question 4
Which Types of Research Questions Require Multilevel Methods?

Type 1 question: group effects. There are three types of research questions that are not only appropriate for investigation with multilevel methods, but they should be explored that way. One is the type we've been talking about and that Coleman was interested in: questions that focus on identifying group-level effects on individuals. The type of research I usually do, called "school effects" studies, ask how characteristics of schools – their size, whether they are public or private, how they are organized academically (e.g., tracking structure), their social organizations (e.g., how people relate to one another – influence how much students learn in the schools they attend. This is the first and most common type of multilevel question: estimating how characteristics of groups influence the individuals who are group members. My program evaluation example, as described so far, would be of this type. Did the students in the intervention classrooms become more proficient with computers than those in the control classrooms?

Type 2 question: slopes as outcomes. The second type of question focuses on what we call heterogeneity of regression slopes," which I mentioned earlier. Let me suggest a simple example, one that I care very much about. We know that U.S. children from families of different social class backgrounds attend the same school – almost any school enrolls children from many different kinds of families (probably also the case in Brazil). With HLM, we can compute the relationship (or slope) between SES and an outcome, let's say achievement, in any school. In almost all schools, this relationship is positive: more advantaged students achieve at higher levels than their less-advantaged schoolmates. With HLM, we can use this relationship in each school as an outcome. If we made the mistake of using single-level methods, we would be assuming that the relationship between SES and achievement was the same in each school we were studying. However, with HLM we can estimate this regression slope separately in each school, and then trying to identify the characteristics of schools that decrease this relationship. Why decrease? Because I believe that social equity is a good thing, and so trying to decrease the SES/achievement slope would foster social equity. The measure – the relationship between social class and achievement – can be investigated as an outcome in HLM. It is quite straightforward to look for characteristics of schools associated with social equity in outcomes.

We might also investigate the slopes-as-outcomes phenomenon in the hypothetical evaluation of computer proficiency for school dropouts. Perhaps this particular intervention also claims to be particularly effective for students who left school at an early age. Thus, we might investigate a within-classroom slope: the relationship between the age of school leaving and computer proficiency. This slope is probably positive in all classrooms (student who stay in school longer may be more computer-proficient). But it may be the case that the intervention we are evaluating is negatively related to this slope – that is, students' computer proficiency is less strongly related to their dropout age in the intervention classrooms. This would be a second positive finding about the intervention, because it might induce not only computer proficiency, in average, but also induce equity by giving a special boost to early school-leavers.

Type 3 question: measuring change over time. The third main type of research question we can explore with multilevel methods involves change over time in some outcome. In school-based research, we know that students learn something every year (so the change is positive), but there are other instances where change over time is important. Perhaps we might see that depression decreases over time for mental patients who are receiving some sort of therapy. Perhaps we would like to see children's aggressive behavior decrease over time, based on some kind of intervention. Research that focuses on measuring change has always been difficult, but with HLM this type of research question is easily explored. In Bahia state, a project where researchers are assessing achievement at several time points during the same school year on the same children, would be a good example of this type of research question. The nesting here is a bit different, in that multiple measures of the same phenomenon, recorded at several time points, are "nested" in individuals. Here, we would ask how characteristics of individuals (e.g., their gender, their race, their SES, their training, their ability, their history of mental problems, their treatment) are associated with "growth" or "change" in the outcome under study. Of course, this growth may be positive or negative. You might also include a further unit of analysis, or level, where characteristics of schools could be linked to these "growth curves".

Let's return to the evaluation example. Earlier I described this as a 3-month intervention. However, perhaps in several classrooms the program goes on for 12 months, and students' computer proficiency is measured every two months. We might then be able to estimate a growth curve for all students (those in the treatment and control classrooms), and investigate whether more able student gain proficiency faster, whether male and female students learn at the same rate, to determine how the age of dropping out of school influences the students' computer proficiency learning rate. This might not be an evaluation question, per se, but it might be (i.e., the intervention students might learn more quickly at first and then level off thereafter).

Let me review. Although there are other types of multilevel research questions, "the big three" are: (1) group level effects on individual outcomes, (2) equity slopes as outcomes, and (3) measuring change over time in individuals. These encompass a wide range of multilevel questions in social science research, and HLM works to address each type of question. These are also the sorts of research questions that other research methods can answer only with difficulty or not at all.

Question 5
Are There Drawbacks of Using Multilevel Methods?

So far, I've described to you important advantages in quantitative social science research that are associated with using multilevel methods, especially HLM. Hopefully, I have convinced you of the value of this methodology, at least in theory. The next questions you might ask yourself include, "Do I really need to use these methods"? "How can I learn to do this"? "What data do I need to use these methods"? "Will my colleagues understand what I am talking about"? "Can we accomplish the same advantages with other methodologies that people are already familiar with"? Because I teach courses in research methodology in general and multilevel methods in particular, some of these questions are always posed to me by my students.

Interpretation of results. I am going to be candid with you. Although the HLM software is relatively easy to use, interpreting the output is not. At various academic conferences in the U.S., 2-day workshops on HLM are often available. These workshops are very valuable for introducing researchers to the methodology, but they seldom become skilled in their use during that time. My course is a regular semester-long course, but it is packed into four weeks. As my students would tell you, it takes some time (and much effort) to learn to use the methodology well. As they might also tell you, HLM is a "data-hungry" methodology. To estimate group-level effects on individuals, you need to have data on many individuals in each group, and you also need data on many groups. So small samples create problems for HLM. You also need to have data available at the units of analysis you are studying. For the evaluation example, you would need to collect complete data on all the participants in both the treatment and control group classrooms, and you would need substantial numbers of classrooms as well (with good data on all of them). This attention to collecting good data on many individuals and groups is troublesome, costly, and absolutely necessary. Of course, in the evaluation example, attrition could be quite a problem, as well.

Is HLM really necessary? Whether of not HLM is absolutely necessary to address the multilevel research questions you wish to explore can be addressed with a simple HLM, where you compute the intra-class correlation (ICC). If the proportion of variance in your outcome is quite low (say, below 10 percent), then the phenomenon you are exploring probably isn't multilevel (that is, you won't find group-level effects on individuals). Similarly, if your dependent variable is measured unreliably, you are unlikely to find any effects at all. However, to answer multilevel research questions, there is absolutely no substitute for multilevel methods.

Will colleagues understand your multilevel research? What about your colleagues? It is likely that many of your professional colleagues are interested in the substance of your research but not necessarily skilled in research methodology. Ten or 15 years ago, research addressing multilevel questions but using single-level analysis to explore them were published quite regularly in U.S. academic journals. However, it is now quite unlikely that this will happen. This methodology is now well enough known among U.S. academic researchers that a manuscript would probably not make it through the review process if the question were multilevel and the methods used to address it were not. I'm not sure about Brazilian journals, but this methodology has caught on fast elsewhere. I have also heard that many researchers from outside the U.S. are anxious to publish their work in

U.S. journals. I was the keynote speaker at a summer 2000 meeting of European sociologists of education in Holland. I was very impressed by the widespread knowledge and use of multilevel methods among this group, and many of them have published their work in U.S. journals. They were obviously receiving excellent training in multilevel methods at their universities, as well (that is, they're not all learning it in the U.S.). So I'm guessing that if the "multilevel bug" has not struck yet in Brazil and other Latin American countries, it will arrive soon. My presence here today may represent one example of that bug.

Let me encourage you to make your work approachable by your non-technical colleagues. Even if you use these methods, your write-ups don't need to be overly technical – in fact, I would argue against that. In my own work, I have tried hard to make my research understandable by a much wider audience than other quantitative researchers. There is a real danger when you learn something new to let it dominate your writing. Of course, you do need to explain your methods. However, it is your findings that should be highlighted, along with the validity of those findings, and I hope your writing will be accessible even to your non-multilevel colleagues.

Final Comment

My intention here today has been to introduce multilevel research methods to those who aren't familiar with them, or perhaps to expand the interest of those who have heard of them before. I've also tried to give you a general idea of the sorts of research questions that are particularly appropriate to address with multilevel methods and to be honest about the data demands for using them. In other talks while I am here in Brazil I am presenting a bit more detail about how these methods work in practice, using some of my own research as examples. As I mentioned earlier, my "research context" is education, so the examples I discuss are from this field. My special "context of interest" is the school. Much of my research has focused on adolescents and high schools, so many examples center on students of these ages and the schools that serve them.

However, my current quantitative research makes use of the U.S. Department of Education's newest data collection effort, the Early Childhood Longitudinal Study (called ECLS). The data collection began as children entered kindergarten (in 1998), and follows the same students through fifth grade. The first four waves of ECLS focus on about 20,000 children in almost 1,000 schools at the beginning and end of kindergarten and first grade. Although much early childhood research is very new for me, many of the questions are the same. So far we have evaluated the efficacy of particular policies (e.g., full-day vs. half-day kindergarten) and instructional approaches (e.g., we have evaluated the effects of phonics instruction on literacy learning), and investigated the social and academic differences that distinguish children in their first formal school experience (investigating how social background is linked to school quality).

At the end of this talk, let me also mention that I've also been doing quite a lot of qualitative research in high schools for the last three or four years. Although I love quantitative studies, and teach students to use quantitative methods, I also recognize that there are all sorts of research questions in education that lend themselves better to a closer study in a smaller number of "contexts".