Why Mathematics Education Needs Large-Scale Research

James A. Middleton, Jinfa Cai, and Stephen Hwang

Over the years our community has benefitted greatly from the application of large-scale methods to the discernment of patterns in student mathematics performance, attitudes, and to some degree, policies and practices. In particular, such research has helped us discover differential patterns in socioeconomic, gender, and ethnic groups and point out that, as a system, mathematics curriculum and instruction has hardly been equitable to all students. From the National Center on Education Statistics (in the US), large scale studies such as High School and Beyond, the Longitudinal Study of American Youth, and the National Assessment of Educational Progress came important calls to focus attention on improving instruction for marginalized populations and to increase emphasis on more complex problem solving than had typically been the norm (Dossey & Wu, 2013).

But these studies have been less useful, historically, in helping us design and implement our responses to their call. Mathematics curriculum design has been, typically, an intense form of educational engineering, wherein units or modules are developed and piloted in relatively insular settings, with large-scale field tests held at or near the end of development. Arithmetic materials, for example, have been informed by a large *body* of small to medium *scale* studies of the development of children's mathematical thinking. Algebra, which has many fewer studies of learners' thinking, is even more dependent upon small-scale studies. Towards the end of the 1990s and into the early 2000s, policy devoting more research funding on efficacy studies renewed interest in experimental and quasi-experimental methods, sample size, and generalizability of results (Towne & Shavelson, 2002). The push

J.A. Middleton (✉)
Arizona State University, Tempe, AZ, USA
e-mail: jimbo@asu.edu

J. Cai • S. Hwang
University of Delaware, Ewing Hall 523, Newark, DE 19716, USA
e-mail: jcai@udel.edu; hwangste@udel.edu

J.A. Middleton et al. (eds.), *Large-Scale Studies in Mathematics Education*,
Research in Mathematics Education, DOI 10.1007/978-3-319-07716-1_1

has been to demonstrate the impact of different education interventions on mathematics performance and achievement statistically.

One notable study conducted in this period evaluated the impact of SimCalc, a computer-facilitated system for representing and manipulating functions and coordinating representations with simulation (animations) and real-world data (Roschelle & Shechtman 2013; Roschelle, Tatar, Hedges, & Shechtman 2010). In this set of studies, the authors examined implementation of SimCalc in over 100 schools (150 teachers, 2,500 students) throughout Texas.

This study is a good example of many of the issues facing large-scale research today. For example, the authors took care to select schools from urban as well as rural areas and admitted that urban schools were under-sampled as well as those that served African-American students. In particular, the reality of working with intact classrooms, in schools drawn non-randomly from widely different communities, forced the authors to utilize statistical controls to equilibrate experimental versus control groups across a variety of demographic and attitude variables, to insure that performance differences are meaningfully attributed to the intervention rather than to presage variables.

In addition, fidelity of implementation is an issue impacting the internal validity of a study. The authors had to implement a wide battery of assessments to determine the degree to which SimCalc-as-implemented reflected SimCalc-as-intended. What is noteworthy in this study is the use of multiple indices to understand the implementation of the program as integral to assessing its impact. The authors used a pre-post design to assess student performance and collected teacher knowledge assessments and tests of teacher mathematical knowledge for teaching, teacher attitude questionnaires, teacher logs, teacher interviews, and coordinated this data with demographic data. Such a wide geography of implementation, as well as a wide demography showed that, despite variation in implementation, the structure of the tools themselves constrained student and teacher behavior to be roughly in line with the design intent.

Hierarchical Linear Modeling (HLM) was used to preserve the levels of nested effects (students within classes). Results showed that students who utilized SimCalc in their classes outperformed a control group with effect sizes ranging from .6 to .8 or .9 for complex items (focusing on proportionality for younger students and functions for older students). Even low-complexity items showed significant effect sizes, though lower than those found for complex items (ranging from .1 to .19).

So, this study and others (see Romberg & Shafer, 2008) show that interventions can be developed, theoretically, and analyzed experimentally at a large enough scale to give us some confidence that, if employed elsewhere, there is a good probability the intervention will result in meaningful improvement of teacher practice and student learning.

But also, large-scale studies can help us theoretically, by providing a check against a set of findings drawn from a number of diverse, small-scale exploratory studies. Even a body of data as coherent and long-standing as that for proportional reasoning can be found wanting. In "Exploring the Impact of Knowledge of Multiple Strategies on Students' Learning about Proportions," Vig, Star, Depuis, Lein, and Jitendra (this volume), for example, show us that large-scale data can provide a cross-check on the continued utility of some models developed across many small-scale studies. For

example, they found in a study of implementation of a unit designed to teach proportional reasoning and associated content that cross-multiplication as a wide-ranging strategy may be much less prevalent now than in previous years due to changes in curriculum and instruction. This illustrates the potential from large data to see new things that are impossible to discern on the small scale and to judge the generality of findings of small-scale research in the larger population.

The Institute for Education Sciences and NSF (U.S. Department of Education & National Science Foundation, 2013) present six classes of research in their *Common Guidelines for Education Research and Development*: (1) Foundational Research, (2) Early Stage or Exploratory Research, (3) Design and Development Research; (4) Efficacy, (5) Effectiveness, and (6) Scale-up. From exploring new phenomena in Exploratory research, or development of new theories in Foundational research, to the examination of the effectiveness of interventions across a wide range of demographic, economic, and implementation factors in Scale-Up research, scale is a critical factor to establish the believability of our conceptual models and the potential efficacy of our designed innovations in mathematics education.

What exactly is the scale that would constitute compelling evidence of intervention efficacy? What is the appropriate scale that would convince the field that inequities exist? That those same inequities have been ameliorated significantly? What scale would convince us that a long-standing research finding may no longer be as prevalent? These are unanswered questions that the chapters in this book can help us answer.

What Is Meant by "Large Scale?"

In this chapter, we introduce this book by asking the fundamental question, "What is meant by Large Scale?" In particular, the word "Large" is problematic, as it must be compared with something "small" to be meaningful. Anderson and Postlethwaite (2007), in their comparison of small versus large scale program evaluation research, provide a convenient taxonomy of factors that distinguish issues of scale: (1) Sample size, (2) purpose of the research, (3) generalizability of results, (4) type and complexity of data analysis, and (5) cost. Anderson and Postelthwaite's discussion is limited to studies of program evaluation, but the issues they raise are clearly relevant to curriculum, teaching, learning, and other more basic research foci. We will introduce chapters in this volume utilizing the first four of these issues. Cost is a factor that is determined, in part, by each of the first four and will be woven into our discussion as appropriate.

Sample Size

At first pass, we can define "Large" in terms of the sheer size of the sample(s) being examined. Chapters in this book address samples on the order of 2,500 participants, to three orders of magnitude greater for international data sets. Small, therefore,

would include the undertaking just held up as an exemplar for large scale, the work of Rochelle and colleagues. The scale of such studies, in terms of sample size, therefore, must be tempered with the kinds of methods used. Complex methods that employ multiple measures, including qualitative approaches such as interviews and observation, can be considered "Large" with samples in the hundreds, as opposed to relatively "simple" studies that may employ only a single measure.

Thomas, Heck, and Bauer (2005) report that many large-scale surveys must develop a complex method for determining the sampling frame so that important subpopulations with characteristics of interest (SES, ethnicity, grade level, for example) will be insured representation. A simple random sample, in many cases, will not yield enough members of the target subpopulation to generate adequate confidence intervals. In "Longitudinally Investigating the Impact of Curricula and Classroom Emphases on the Algebra Learning of Students of Different Ethnicities," Hwang, Cai, Shih, Moyer, and Wang (this volume) illustrate this issue clearly. Their work on curriculum implementation required a sample large enough disaggregate results for important demographic groups. Their research shows that while achievement gaps tended to lessen for middle school students engaged in reform-oriented, NSF-sponsored curricula, the performance gap between White students and African-American students remained robust to the intervention. These results show demonstrably that curriculum implementation is not uniform for all ethnic groups, and that we have much work to do to create tasks and sequences that *do* address the cultural and learning needs of all students.

Likewise, in "A Randomized Trial of Lesson Study with Mathematical Resource Kits: Analysis of Impact on Teachers' Beliefs and Learning Community," Lewis and Perry (this volume) performed a randomized control trial of lesson study implementation, examining the impact of a set of support materials that provide imbedded professional development, and a structure for neophytes to implement lesson study on fractions with fidelity. The authors took great pains in their sampling frame, to establish the equivalence of control versus treatment groups. Their results show that such support improves teachers' knowledge of fractions, their fidelity of implementation of lesson study, and subsequent student performance on fractions. Their use of multiple methods, across multiple levels (students, teachers, teacher lesson study groups) also highlights how studies across diverse geography and demography can explore the efficacy of locally organized professional development (with nationally designed support) versus larger policy-level organization.

For government agencies, these sampling issues are often addressed through multi-stage cluster sampling, often including oversampling of under-represented groups. These strategies have implications for the calculation of standard errors in subsequent analyses, particularly if within-group variation is smaller than cross-group variation. "A Review of Three Large-Scale Datasets Critiquing Item Design, Data Collection, and the Usefulness of Claims," "Using NAEP to Analyze Eighth-Grade Students' Ability to Reason Algebraically," and "Homework and Mathematics Learning: What Can We Learn from the TIMSS Series Studies in the Last Two Decades?" (Orletsky, Middleton & Sloane, this volume; Kloosterman et al., this

volume; and Zhu, this volume) each deal with these complex sampling issues, both practically as the authors implement studies that must take sample weighting into account and by methodological critique of secondary databases.

Researchers without the financial wherewithal of government agencies often must resort to other methods for insuring the sample of a study effectively represents some general population. Matching participants across experimental units on a variety of important covariates is a statistical method for making the case that experimental units are *functionally* equivalent prior to an intervention, and therefore, that any differences found after the study are due solely to the intervention. Such methods do not account for *all* preexisting variation in groups; some systematic variation is inevitably unaccounted for. Lakin and Lai (2012), for example, show that the generalizability of standardized tests can be much lower than for non-ELLs. Their study showed that ELL students would have had to respond to more than twice as many mathematics items and more than three times as many verbal items for the instrument to show the same precision as non-ELL students. Care must be taken, then, to not underestimate the standard error of measurements for subpopulations.

In "A Lesson for the Common Core Standards Era from the NCTM Standards Era: The Importance of Considering School-level Buy-in When Implementing and Evaluating Standards Based Instructional Materials, Kramer, Cai, & Merlino (this volume) performed a quasi-experiment examining the impact of school-level attitudes and support on the efficacy of two NSF-supported middle school curricula. Using data from a Local Systemic Change project, they assessed "Will to Reform," a survey-proxy for fidelity of implementation, roughly defined as teacher buy-in to the curriculum, and principal support for the curriculum. Carefully matching, statistically, schools implementing either *Connected Mathematics Project* or *Mathematics in Context*, they found that choice of material did not matter so much as the degree to which schools supported the curricula, teachers showed buy-in to the methods, and principals supported teachers' reform. The ability to match schools across several potential nuisance factors (such as prior mathematics and reading scores, demographics, SES) *requires* a large enough sample to provide adequate variability across all matching factors.

Regardless of the techniques used, the point is to reduce the overall systematic variation between experimental units enough to claim that the residual variation has a relatively minor effect. Careful choice of covariates is critical to make this claim, in addition to randomization or other equilibration techniques.

Purpose of the Study

Small-scale studies tend to be used towards the beginning of a research program: To explore new phenomena for which existing measures are not yet developed. Many focus on developing measures, drafting tasks for curriculum and assessment, or for exploring new teaching practices. Large-scale studies, in contrast, tend to be

employed after such methods or instruments have been piloted and their use justified, and the phenomena to which they apply have been adequately defined. Anderson and Postlethwaite (2007) define the purpose of large-scale studies as describing a system as a whole and the role of parts within it. But the complexity of the system and the type of understanding to be gained from the study greatly impact how large the scale must be.

When examining a relatively simple system (say, performance on a test of proportional reasoning), the relatively low cost of administering a single measure, relative to the necessary power for detecting a particular effect, makes a "large" scale smaller, proportionally, than when examining the interaction between a set of variables. In general, the more variables one is interested in, the larger the scale one must invest in. But this is even more crucial if the *interaction* among variables is under study. Even relatively simple factorial designs to test interaction effects require a polynomially increasing sample size as the number of interactions increases. When ordered Longitudinally, concepts assessed in year 1 of a study, for example, do not ordinarily have a one-to-one relationship with concepts in subsequent year. Thus, the combinatorial complexity of human learning requires a huge sample size if the researcher is interested in mapping the potential trajectories learners may travel across a domain (Confrey & Maloney, this volume; Hwang, et al., this volume; Lewis & Perry, this volume).

In contrast, when the number of potentially interacting variables is high, the analysis is fine-grained (such as interviews of individual learning trajectories or observation of classroom interactions), and the purpose of the study is to create a new model of the phenomenon, smaller sample sizes may be needed to distinguish subtle differences in effects of tasks, questioning techniques, or other relevant factors. Middleton et al. (this volume), in "A Longitudinal Study of the Development of Rational Number Concepts and Strategies in the Middle Grades," show that, with only about 100 students, intense interview and observation techniques over several years can be considered large scale due to the purpose of the study as modeling student development of rational number understanding. The authors found that, contrary to their initial hypotheses, students' understanding grew less complex over time due to key biases in models used to teach fractions and ratios.

In "Engineering [for] Effectiveness in Mathematics Education: Intervention at the Instructional Core in an Era of Common Core Standards," Confrey and Maloney (this volume) provide a reconciliation of these extremes. They make the case that findings across *many* such studies can highlight the interplay among factors central to what they term the "instructional core"—the complex system bounded by curriculum, assessment, and instruction. The scale here is defined as the extent of the common efforts across studies. They call for the creation of collaborative efforts among large-scale development and implementation projects and the development of technologically-facilitated systems of data collection and sharing to facilitate analysis of this complex system *as* a complex system, using modern analytics.

Generalizability and Transportability of Results

Generalizability is a valued outcome of most large-scale studies. We report data not just as a description of the local, individual participants and their behavior, but as a model for *other* participants. When we test the efficacy of a teacher professional development program, for example, we are reporting our belief that the effects found can be replicated, under similar conditions, in some population of teachers. For primarily quantitative data, generalizability is established by the sampling frame—the methods by which the author makes the case that the sample represents the population of interest—the operational definition of the measure, and the appropriateness of the analyses. Standard errors are used to find the probability that a measure adequately reflects the typical behavior of the population. For such studies, size really does matter: The sample size is inversely proportional to the standard error. The issues of the complexity of sampling frames and the analyses mentioned above are largely important due to their impact on generalizability.

For other studies, those that use more qualitative methods, or those that cannot make random assignment to conditions, generalizability is difficult to impossible to establish statistically. Instead, a concept from design research becomes useful: Transportability of results. Transportability has to do with the functionality of the innovation being studied. Curricula, for example, may have different ways of being applied depending on teacher knowledge, available technology, state and local level standards, and so on. How robust the curriculum is, and how adaptable it is when transported from one situation to another, is a critical consideration for studies of applicability (Lamberg & Middleton, 2009).

In "Challenges in Conducting Large-Scale Studies of Curricular Effectiveness: Data Collection and Analyses in the COSMIC Project," Tarr and Soria (this volume) address both of these issues adroitly in their multi-level study of the impact of curriculum type on student achievement. The authors had to take multiple measures of prior achievement from state-level tests, convert them to z-scores, then map the state z-scores to NAEP scores to model student achievement as a result of reform-oriented curricula versus more traditional curricula. Effects of teachers, due to lack of observational data, were modeled using paper and pencil scales of teacher beliefs as proxies. Moreover, because so many teacher variables had potential impact on student achievement, potentially obscuring the impact of curriculum type, the authors reduced these dimensions using Principle Components Analysis. In this study of 4,600 students across 135 teachers, the sheer number of variables measured, and their potential interactions necessitated a large scale to have enough power to detect any effect curriculum might have had. Through iterative multi-level models, reducing the dimensionality of the system each iteration, they found that curriculum DOES matter, but prior achievement, opportunity to learn, and teacher effects mediate curriculum significantly.

The scale and sampling frame for this study establishes good generalizability of the results in a statistical sense. However, the iterative methods used in this chapter

allowed the authors to show that many key variables impact the transportability of different curricula from one situation to another. Curriculum matters, but not to the exclusion of factors of implementation.

Type and Complexity of Data Analysis

"Large" is also determined, to a great extent, by the methods used to answer the research question. Observational methods, for example, because of their inherent cost in terms of time and analytic complexity, may constitute only dozens of records, depending on whether or not single units are observed multiple times, or whether multiple units are observed once or twice.

Shih, Ing, and Tarr (this volume), in "Addressing Measurement Issues in Two Large-Scale Mathematics Classroom Observation Protocols," for example, critique two different observational protocols, designed to view the same classroom phenomena, regarding how they account for, and treat as parameters sources of error variation. Their analysis highlights the need to run comparative analyses of reliability across competing or even seemingly complementary methods. One issue appears to be particularly important: Protocols aiming to determine general features of practice may tend to ignore or gloss over important differences in content and curriculum, which are the *central* features of other protocols, while those protocols focusing on the within-effects different tasks and curricula may report results that do not generalize across those factors. They also provide methodological insight by showing that utilization of multiple raters may improve reliability of observational protocols more effectively than increasing the number of items on a scale.

Like observation, face-to-face interview methods, all things being equal, will not allow samples as large as phone or online interviews. In the world of survey methods, the ability to use computerized (including online) collection methods enables larger sample sizes and more complex methods of assigning items to individuals. These methods, of course, both depend on, and interact with, the kinds of research questions being asked. As Shih et al. show, questions about the generalizability of a known finding requires more data than questions about the possible ways in which teachers might implement a particular concept in their class (also see Lewis et al., this volume).

In "Turning to Online Courses to Expand Access: A Rigorous Study of the Impact of Online Algebra I for Eighth Graders," Jessica Heppen and her colleagues (Heppen, Clements, & Walters, this volume) provide an excellent example of how the unit of analysis, coupled with the research question, influences what we consider "large." They report an efficacy study of providing online access to Algebra I to rural eighth-grade schools, which, heretofore had limited access to the content (some of the surveyed schools only had four eighth graders, presumably making staffing and curriculum adoption impractical and/or cost-prohibitive). In their study, the unit of analysis is *schools*. Schools are the appropriate unit for studying curriculum access, as individual students are typically nested within available curriculum,

and typically schools adopt a single set of materials (see also Kramer, Cai, & Merlino, this volume). Thirty five schools receiving online access to Algebra 1 were compared to 33 control schools. The authors report that providing such access can improve eighth-grade performance as well as improve the probability of subsequent advanced mathematics coursetaking as students move to high school.

Characteristics of the Measurement

Size may matter, but what is being "measured" matters as well. It is clear, for example, from the high degree of variability and low goodness of fit for participant scores in mathematics assessments, that a large amount of any person's score is error of measurement. Any effect, therefore, includes not only true differences in the variable of interest, but also a whole host of spurious effects (Shadish, Cook, & Campbell, 2002).

Seltiz (1976) discusses the different components that make up a typical effect in social research. These effects include: (1) Stable characteristics other than those intended to be measured (such as the person's motivation in mathematics impacting their effort on a test of performance); (2) Relatively transient factors such as health or fatigue; (3) Variation in the assessment situation, for example, taking a test in a testing center versus the classroom or interviewing a teacher in her room versus in the researcher's lab; (4) Variation in administration (different instructions given or tools made available); (5) Inadequate sampling of items; (6) Lack of clarity of measuring instruments; and (7) Variation due to mechanical factors, such as marking an incorrect box on a multiple choice test, of incorrect coding of an interview item.

Multiple-methods and mixed methods (e.g., Mujtaba, Reiss, Rodd, & Simon, this volume) provide both statistical confidence and qualitative depiction of typical or expected attitudes, practices, or student behaviors in context and help the researcher understand when one or more of these factors may play an important role in measurement.

Error of Measurement

Inadequate or inconsistent sampling of items from the conceptual domain under study reduces the degree to which we can have confidence in the results of any assessment utilizing those items. In "Using NAEP to Analyze Eighth-Grade Students' Ability to Reason Algebraically," Kloosterman et al. (this volume) perform a secondary analysis of NAEP items, classifying them by their mathematical content, and then analyzing student performance for that content *longitudinally*. Their study represents a heroic effort just getting access to, and classifying NAEP items, given the proprietary nature to which specific item content and wording is guarded by the National Center for Education Statistics. Their results show that US eighth students' performance on NAEP, both overall and for algebra-specific content, has improved steadily from 1990 to 2011. Analysis of different items, however,

shows consistent difficulty in content associated with proportional reasoning and on equations and inequalities utilizing proportional concepts. As a nationally representative sample, their works illustrate how huge-scale data can simultaneously provide us with information regarding how we are improving (or not), in mathematics instruction, but also provide specific critique on areas where we may still be falling short despite overall improvement.

For studies that assess the structure of variables in a network model or that employ advanced regression methods, the critical relationship between the number of items used to measure a construct and its reliability becomes extremely important. Even if each item is an excellent measure of its individual construct, the degree to which the items, together, predict some larger class of understandings can be eroded through their incorporation into a subscale. This increases the error of estimate of the latent variable.

Ebby and Sirinides (this volume) studied the interaction among several key variables, heretofore studied separately, in "Conceptualizing Teachers' Capacity for Learning Trajectory-Oriented Formative Assessment in Mathematics". They report on the development of an instrument to measure several aspects of teachers' Mathematical Knowledge for Teaching, including their assessment of the validity of the mathematics students used to solve problems, their assessment of students' mathematical thinking, and their orientation towards thinking of students' work in a learning trajectory. Fourteen hundred teachers were assessed by 15 different raters in this study! Using structural equation modeling (SEM), the authors found that teachers utilize their assessment of the validity of the mathematics to help them diagnose students' mathematical thinking. Their understanding of children's mathematical thinking, in turn, impacts their understanding of the students' learning trajectory. Together, these three variables significantly impact teachers' instructional decision making.

Complexity of the Measure

Assessments that measure multiple constructs versus a single one run into the tendency to under-sample the domain for each sub-construct, increase fatigue due the length of the administration of the assessment, and subsequently increase the number of mechanical errors recorded. Mujtaba et al. (this volume) clearly illustrate this in "Methodological issues in mathematics education research when exploring issues around participation and engagement". The authors studied motivational variables and their individual and collective impact on students' intended choice of mathematics in post-compulsory education. Multi-level modeling allowed the authors to account for school-based variation, to focus analyses on individual determinants of future course choice. What scale afforded the authors was an opportunity to examine *multiple* variables in concert, without sacrificing predictive validity of any variable apart from the others. Intrinsic motivation in mathematics, beliefs about extrinsic material gain from studying mathematics and advice all were shown to be significant contributors to students' decisions.

P = "Publish"

Large-scale studies are prone to errors due to "fishing." Because, particularly for secondary data analysis, researchers have access to so many variables at once, the tendency to run analyses without clear hypotheses or theoretical justification is almost too easy. The probability values of these results may be very low, due to the effect of large sample size on the standard error of measurement. The literature is currently full of findings of dubious utility, because the probability that a correlation is zero due to random chance may be very small. But how large is the correlation? For experimental research, an effect may have low probability of occurrence by random chance. But how large is the effect? Large-scale studies, because of the relative stability that large samples provide for estimates, can give us indication of the size of effect of an intervention, and therefore its potential practical significance.

Orletsky et al. (this volume), in "A Review of Three Large-Scale Datasets Critiquing Item Design, Data Collection, and the Usefulness of Claims," compare and contrast three large-scale longitudinal studies (ELS, NAEP, & TIMSS), examining the potential threats to validity that are probable when performing secondary data analysis. In particular, because of the relationship between sample size and standard error of estimate, the tendency for large-scale "findings" to have low *p-values* may yield *many* spurious results. Heppen et al. (this volume) utilize the narrow standard errors of large-scale research methodologically in a clever and unique way by hypothesizing that a *lack of* statistically significant side effects of an intervention may be considered supportive evidence for its efficacy. When combined with *significant* performance outcomes, large sample sizes enable researchers to examine unintended consequences of interventions statistically.

Zhu (this volume), in "Homework and Mathematics Learning: What Can We Learn from the TIMSS Series Studies in the Last Two Decades?," utilized the TIMSS database to compare the mathematics homework practices of five east Asian nations with three Western nations. Overall, though there were key differences from nation to nation, homework practices were found to be highly similar. Well over 90 % of teachers surveyed assigned homework. Homework varied from about ½ h per day (US, Japan, England), to about 45 min per day (Singapore). Most homework consisted of worksheet problems. One key finding shows that across all the studied nations, the prevalence of classroom discussion of homework problems has steadily increased from 1995 to 2011. Without large samples capable of being disaggregated by nation, the stability of these findings would have been near impossible to establish.

Level of Data Analysis

Many of the chapters in this volume address this issue explicitly, so we do not go into depth here. Suffice it to say that learning studies where students can be randomly assigned to experimental conditions require fewer records than nested

designs. HLM and other multi-level methods are only valuable if the appropriate number of participants is sampled *at each level*. Within group variation then becomes an issue, depending on the heterogeneity of students within classrooms, or classrooms within schools. The more within group variation, the more groups will be needed to establish the group effect (Hwang et al., this volume; Kramer et al., this volume; Lewis & Perry, this volume).

Summary

This monograph is timely in that the field of mathematics education is becoming more diverse in its methods, and the need to investigate the efficacy of policies, tools, and interventions on mathematics teaching and learning is becoming more and more acute. In particular, the diverse ways in which students from a variety of backgrounds and with a variety of interests can become more powerful, mathematically, is still an open question. While examples can be provided with investigations of a few students in a few classrooms, the generality of those examples across the tremendous diversity of conditions of implementation in the world must be established with studies of a scale large enough to detect and estimate the probabilities of interventions' effectiveness with populations of interest disaggregated.

The chapters in this book show that large scale studies can be both illuminative—uncovering patterns not yet seen in the literature, and critical—changing how we think about teaching, learning, policy, and practice. The authors examine topics as diverse as motivation, curriculum development, teacher professional development, equity, and comparative education. Organizationally, we divide the chapters into four thematic sections:

Section I: Curriculum Implementation
Section II: Teachers and Instruction
Section III: Learning and Dispositions
Section IV: Methodology

But, it must be noted that most of this work crosses lines of teaching, learning, policy, and practice. The studies in this book also cross the boundaries of the six types of research discussed in the IES/NSF *Common Guidelines for Education Research and Development* (2013). We have selected these authors because their research and commentary are complex, illuminating problems to look out for, methodologically, as well as insight for how to better create robust, generalizable information for the improvement of curriculum, teaching, and learning. We anticipate this volume will help researchers navigate this terrain, whether engaging in designing and conducting efficacy research on the one hand, or analyzing secondary data on the other.

References

Anderson, L. W., & Postlethwaite, T. N. (2007). *Program evaluation: Large-scale and small-scale studies*. International Academy of Education: International Institute for Education Planning. www.unesco.org/iiep/PDF/Edpol8.pdf.

Dossey, J. A., & Wu, M. L. (2013). Implications of international studies for national and local policy in mathematics education. In *Third international handbook of mathematics education* (pp. 1009–1042). New York: Springer.

Institute for Education Sciences, National Science Foundation. (2013). *Common guidelines for education research and development.* Washington, DC: IES. http://ies.ed.gov/pdf/CommonGuidelines.pdf.

Lakin, J. M., & Lai, E. R. (2012). Multigroup generalizability analysis of verbal, quantitative, and nonverbal ability tests for culturally and linguistically diverse students. *Educational and Psychological Measurement, 72*(1), 139–158.

Lamberg, T., & Middleton, J. A. (2009). Design research perspectives on transitioning from individual microgenetic interviews in a laboratory setting to a whole class teaching experiment. *Educational Researcher, 38*(4), 233–245.

Romberg, T. A., & Shafer, M. C. (2008). *The impact of reform instruction on student mathematics achievement: An example of a summative evaluation of a standards-based curriculum*. London: Routledge.

Roschelle, J., & Shechtman, N. (2013). SimCalc at scale: Three studies examine the integration of technology, curriculum, and professional development for advancing middle school mathematics. In S. P. Hegedus & J. Roschelle (Eds.), *The SimCalc vision and contributions* (pp. 125–143). Dordrecht, The Netherlands: Springer.

Roschelle, J., Tatar, D., Hedges, L., & Shechtman, N. (2010). *Two perspectives on the generalizability of lessons from scaling up SimCalc*. Paper presented at the Society for Research on Educational Effectiveness, Washington, DC.

Seltiz, C. (1976). *Research methods in social relations*. New York: Holt, Rinehart and Winston.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.

Thomas, S. L., Heck, R. H., & Bauer, K. W. (2005). Weighting and adjusting for design effects in secondary data analyses. *New Directions for Institutional Research, 2005*(127), 51–72.

Towne, L., & Shavelson, R. J. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academies Press.