Measuring Change in Mathematics Learning with Longitudinal Studies: Conceptualization and Methodological Issues

Jinfa Cai, Yujing Ni, and Stephen Hwang

Learning is about growth and change. Learning is often demonstrated by changes in student achievement from one point in time to another. Therefore, researchers and educators are interested in academic growth as a means to understand the process of student learning. In mathematics education, there has been a growing interest in using longitudinal designs to examine and understand student learning over time. Researchers face a number of issues of measuring change using such designs. In this chapter, we draw on our experience gained from two longitudinal studies of mathematics learning to discuss various issues of measuring change in student learning. We start with a brief introduction of the two studies. Then we discuss the conceptualization and measures of change in mathematics learning. Third, we discuss issues of analyzing and reporting change. Finally, we discuss how to interpret changes in mathematics achievement in longitudinal studies appropriately.

Two Longitudinal Studies Examining Curricular Effect on Student Learning

This chapter draws on two longitudinal projects that studied the effects of curriculum on student learning. The first project was conducted in China and addressed the question, "Has curriculum reform made a difference?" by looking for changes in classroom practice and consequently in student learning. This project (hereafter called the China project) compared the effect of a new, reform-oriented elementary mathematics curriculum to that of the conventional curriculum on classroom

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

Y. Ni
The Chinese University of Hong Kong, Hong Kong, China

J.A. Middleton et al. (eds.), *Large-Scale Studies in Mathematics Education*,
Research in Mathematics Education, DOI 10.1007/978-3-319-07716-1_13

practice and student learning outcomes. The second project—the LieCal project (Longitudinal Investigation of the Effect of Curriculum on Algebra Learning)—was conducted in the USA. This project was designed to investigate both the ways under which a reform curriculum did or did not have an impact on student learning in algebra, and the characteristics of the curricula that led to student achievement gains. Both projects looked into changes in classroom practice by examining the nature of classroom instruction, analyzing cognitive features of the instructional tasks implemented in different classrooms, the characteristics of classroom interactions, and changes in student learning outcomes.

The China project and the LieCal project shared similarities in their designs and data analyses. In particular, both projects addressed a set of common and critical questions about teaching and learning using reform-oriented curricula, including: (1) Does the use of the reform-oriented curriculum affect the quality and nature of classroom teaching; (2) Do students improve at solving problems, as the developers of the reform-oriented curricula claim; (3) Do students sacrifice basic mathematical skills with the reform-oriented curriculum; and (4) To what extent does the use of the reform-oriented curriculum improve learning for all students?

Conceptualizing and Measuring Change in Student Learning

Student learning takes place in various domains; two major domains are cognitive and affective (Krathwohl, 2002), each with multiple factors influencing what is learned, how it is learned, and how it is remembered and used. Here, we will focus on the cognitive domain, and in particular on mathematical thinking, to illustrate the issues of how to conceptualize and measure change in student learning. We will briefly touch on the affective domain afterwards.

Although there is no consensus on what mathematical thinking is, it is widely accepted that there are many aspects of mathematical thinking that warrant examination (Cai, 1995; Ginsburg, 1983; Schoenfeld, 1997; Sternberg & Ben-Zeev, 1996). Studies of mathematics learning over the years have included a focus on identifying those ways that students demonstrate a propensity to "think mathematically" in their actions. For example, Polya found that capable problem solvers employ heuristic reasoning strategies to solve problems (Polya, 1945). Being able to self-generate useful analogies while solving a problem is an example of a heuristic that capable solvers demonstrate as they solve problems. In addition, Krutetskii (1976) found that able students are more likely than less able students to use generalizations in their mathematical problem solving. Other researchers have described and explained mathematical thinking as distinct from the body of mathematical knowledge, focusing on processes such as specializing, conjecturing, generalizing, and convincing (Burton, 1984). More recently, mathematical thinking has been characterized in terms of the learner being able to develop strong understandings in mathematical situations (Kieran & Pirie, 1991) and making connections among concepts and procedures (Hiebert & Carpenter, 1992).

These studies suggest that we need to use multiple measures to assess the mathematical thinking of students. For example, although we know that it is important for students to have algorithmic knowledge to solve many kinds of problems, this does not ensure that they have the conceptual knowledge to solve nonroutine or novel problems (Cai, 1995; Hatano, 1988; Steen, 1999; Sternberg, 1999). Hence, it is crucial that studies of mathematical thinking include tasks that measure students' high-level thinking skills as well as their routine problem-solving skills that involve procedural knowledge. Indeed, as the heart of measuring mathematical performance is the set of tasks on which achievement is to be assessed, it is desirable to use various types of tasks to measure the different facets of students' mathematical thinking and gauge student growth in mathematics learning (Betebenner, 2008; Mislevy, 1995; National Research Council (NRC), 2001).

Recognizing the need to assess mathematical thinking broadly, both the China project and the LieCal project used multiple measures of student achievement. Most of the assessment tasks used in both projects came from Cai's earlier work (1995, 2000), in which he investigated Chinese and US students' mathematical thinking. The design of the achievement measures in each project was guided by the following considerations: (1) a combination of multiple-choice and open-ended assessment tasks should be used to measure students' performance; (2) different cognitive components, specifically, the four components of Mayer's (1987) cognitive model (translation, integration, planning, and computation), should be attended to in the multiple choice tasks; and (3) in responding to open-ended tasks, students should show their solution processes and provide justifications for their answers.

Because of their potential for broad content coverage and objective scoring, their highly reliable format, and their low cost, multiple-choice questions were used to assess whether students had learned basic knowledge and skills in mathematics. However, it is relatively difficult to infer students' cognitive processes from their responses to multiple-choice items; such questions are more appropriate for measuring procedural knowledge and basic skills than conceptual understanding. Thus, open-ended tasks were also included to assess student achievement in both projects. The open-ended tasks provided a better window into the thinking and reasoning processes involved in students' problem solving (Cai, 1997). The use of various types of assessment tasks provided the information to address questions such as, "Does the curricular emphasis on conceptual understanding come at the expense of fluency with basic mathematical skills?" For example, the China project showed that both students who received the reform-oriented curriculum and those who did not receive the curriculum had significant improvement in performance on computation and on routine and open-ended problem solving over time. However, the non-reform group showed a faster rate of improvement on the measure of computation. The LieCal project demonstrated that students receiving the reform-oriented CMP curriculum (Connected Mathematics Program, a *Standards*-based curriculum) showed a faster rate of improvement than the students receiving non-CMP curricula on the measures of solving open-ended tasks. However, the two groups did not differ in growth rate on the measure of computation and equation solving.

Research has also shown that changes in learning experiences can lead to changes in feelings towards mathematics, perception of mathematics, and consequently commitment to think mathematically. For example, Schoenfeld (1992) demonstrated how students' beliefs about mathematics could be changed with the experience of being engaged in solving authentic mathematical problems. Reform-oriented mathematics curricula aim not only to help students think mathematically but also to nurture their positive beliefs and attitudes toward learning mathematics. Therefore, the China project administered multiple measures of affective outcomes (interest in learning mathematics, classroom participation and views of what mathematics is about) several times. It was found that, although the students showed significant gains in the three measures of cognitive achievement, their interest in learning mathematics declined from the start of fifth grade to the end of sixth grade for both the reform and non-reform group, with a steeper decline for the non-reform group. This highlights the importance of considering change in students' mathematical learning broadly so that changes can be understood in a broader context of learning. In particular, it highlights the importance of longitudinal analyses so that growth rates can be estimated for key learning variables.

Analyzing and Reporting Change

The major purpose of a longitudinal study is to examine change and the correlates or causes of change over time. Because learning is fundamentally about growth and change, analyzing and reporting change in students' academic achievement is a significant endeavor for the study of learning. However, change is often difficult to document well, given the myriad variables and factors that may influence changes in students' learning. It is even more challenging to identify the causes of a change when change is detected. A sound analysis of longitudinal data relies on a sound study design that includes the use of multiple measures of the same variables over time to help enhance the internal validity of the study (Fisher & Foreit, 2002; Linn, 2007). Given the multifaceted nature of the mathematical thinking that the LieCal and China projects were studying, both projects used three cognitive measures of mathematics achievement (computation, routine problem solving, and complex problem solving) to gain a detailed picture of student growth in mathematics achievement and a possible curricular correlate to the growth.

Within the confines and constraints of non-randomized experimental design, the primary question about change in student achievement that our studies were designed to answer was whether or not there was any meaningful difference in growth rate in mathematics achievement among groups of students using different curricula (Cai, Wang, Moyer, Wang, & Nie, 2011; Ni, Li, Li, & Zhang, 2011).

At the same time, the projects were also designed to address other factors that might affect the students' mathematics achievement growth rate. For example, the LieCal project considered how the conceptual or procedural emphasis of classroom instruction might moderate the curricular influence on the growth rate of students' mathematics achievement. To measure these classroom variables, as the students progressed from sixth through eighth grade, we conducted over 500 lesson observations of over 50 mathematics teachers participating in the project. Each LieCal class was observed four times, during two consecutive lessons in the fall and two in the spring. Trained observers recorded extensive minute-by-minute information about each lesson using a detailed, 28-page observation instrument. The data from these observations were used to characterize key aspects of each lesson, including the degree of conceptual and procedural emphasis of instruction in the CMP and non-CMP classrooms (Moyer, Cai, Wang, & Nie, 2011).

In the China project, each of 60 participating teachers and their classrooms was observed for three lessons on three consecutive days. The videotaped lessons were analyzed in terms of cognitive features of implemented instructional tasks and patterns of classroom discourse. The project found significant differences in instruction between the reform and non-reform classrooms (Li & Ni, 2011). With the measured aspects of classroom instruction, it became possible to examine the relations between curriculum, classroom instruction, and student learning.

In addition, both the LieCal project and the China project attended to elements of the students' sociocultural backgrounds that might influence change in student achievement. Classrooms in the USA have become increasingly ethnically diverse, and there have been persistent concerns about disparities in the mathematics achievement of different ethnic groups. This is particularly true with respect to areas such as algebra and geometry, where success has been shown to help narrow disparities in post-secondary opportunities (Loveless, 2008). Given that middle school mathematics experiences can lay the foundation for students' development of algebraic thinking, the LieCal project explored potential differential effects of reform and traditional curricula on the mathematics performance of students from different ethnic groups (Cai, Wang et al., 2011; Hwang et al., 2015).

The China project took into consideration socioeconomic status (SES) as well. This variable was measured because one purpose of the project was to examine whether achievement gaps between higher SES students and low SES students would decrease or increase in the different aspects of mathematics achievement over time in relation to the different mathematics curricula.

Analyzing and Reporting Change Quantitatively

With these purposes in mind, both studies employed a panel design in which a cohort is followed for a period of time and a common set of instruments is administered repeatedly over that period (Ma, 2010). The studies produced data with a hierarchical structure of individual students nested within classes, classes nested within schools, etc. For this type of hierarchically structured data, the technique of hierarchical linear modeling (HLM), and in particular multilevel growth modeling, is appropriate and effective for examining change at both the individual and the group level. This is because this method is able to account for the correlated observations of the different levels due to the clustering effects and thus relax the assumption of independence of observations for the traditional regression analysis (Raudenbush & Bryk, 2002). Therefore, both projects used HLM models to answer their research questions. The HLM analyses revealed that, in the China project, the students showed a faster growth rate in computation and solving routine problems than in solving open-ended problems, and that this trend was more pronounced for the students receiving a conventional curriculum than those receiving a reform curriculum. The LieCal project used four two-level HLM models (one for each outcome measure) with the mean of conceptual emphasis or procedural emphasis across 3 years as a teaching variable together with student ethnicity and curriculum type nested in schools (Cai, Wang et al., 2011). The results of the HLM analysis showed that students who used CMP had a significantly higher growth rate than non-CMP students on open-ended problem-solving and translation tasks while maintaining similar growth rates on computation and equation-solving tasks. Thus, the relatively greater conceptual gains associated with the use of the CMP curriculum did not come at the cost of basic skills.

In addition, to gain a finer-grained picture of the curricular impact and also as a validation of the results of the HLM analyses, Cai, Wang et al. (2011) compared the percentage of students receiving the CMP curriculum who obtained positive gain scores to the percentage of students receiving non-CMP curricula who obtained positive gain scores. These calculations showed the relative sizes of the groups of students whose performance increased on each of the outcome measures whereas the results of the HLM analyses estimated an overall difference in the means of the gain scores between the two groups of students. For example, we found that 89 % of CMP students had positive gains in open-ended problem-solving tasks over the course of the middle grades. This was a statistically significantly larger percentage than for the non-CMP students, of whom 83 % showed gains in open-ended problem solving. With respect to computation, despite the fact that the mean gains were not significantly different between the CMP and non-CMP students, we found that a larger percentage of non-CMP students than of CMP students showed positive gains (78 % vs. 60 %). With respect to equation-solving, however, the two groups were not significantly different either in mean gains or in percentage of students with positive gains (e.g., 50 % of student group A receiving non-CMP curricula obtaining positive gain scores and 70 % of student group B receiving the CMP curriculum doing so) (Cai, Wang et al., 2011).

Using a broad set of measures over time within a study also allows for the collection of information on what trade-offs may be faced with different curricula and about what can be realistically expected in typical classrooms (Brophy & Good, 1986). The China project showed that the non-reform group demonstrated faster growth in proficiency in computation skills from the fifth grade to the sixth grade, and they outperformed the reform group students in the final assessment. Also, the reform group students kept their initial advantage in solving open-ended problems, as they performed better than the non-reform group on the first assessment and the growth rates for the two groups were similar. Nevertheless, given the nature of the design, it could not be concluded that the reform group's better performance on complex problem solving was merely due to the curriculum or to their better initial status. However, the reform group appeared to have achieved a relatively more balanced development in the three measures of mathematics achievement, computation, routine problem solving, and complex problem solving.

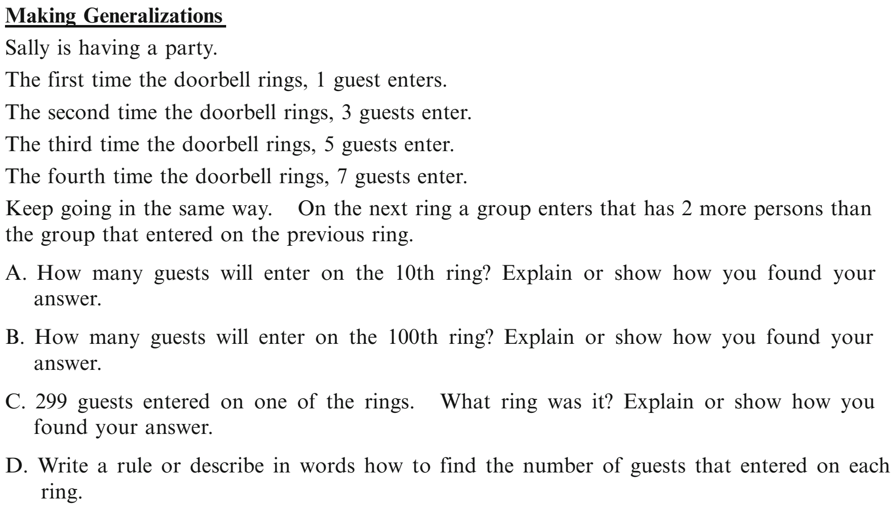
The China project was also concerned with whether or not the different curricula would help reduce achievement gaps between students from different family backgrounds. The project found that the achievement gaps in computation skills between students of high SES backgrounds and those of low SES were narrowed significantly from their fifth grade to sixth grade, but there was no narrowing of the gap in solving open-ended mathematics questions. This was the case for both groups using either a reform curriculum or conventional curriculum. The closing achievement gap in computation but not in solving open-ended mathematics questions suggested that instructional conditions that facilitate mathematical explaining, questioning, exchanging, and problem solving are most valuable for students from low SES families because low SES families are less likely to be able to afford the conditions to facilitate high-order thinking (Ni et al., 2011).

Analyzing and Reporting Change Qualitatively

To deepen analyses of curricular effect on change in student learning it is necessary to look beyond measuring performance differences in terms of mean scores on various types of tasks between groups of students receiving different types of curricula. As useful as such comparisons may be, they do not provide a complete profile of what students who use different curricula can and cannot do. Two students may receive the same score on a task but use very different solution strategies or make very different types of errors. To inform these comparisons of performance on individual tasks, some additional exploration of the thinking and methods that led students to their answers is required.

The use of open-ended assessment tasks makes it possible not only to measure students' higher-order thinking skills and conceptual understanding, but also to analyze students' solution strategies, representations, and mathematical justifications (Cai, 1997). The strategies that students employ and the ways that they represent their solutions can provide insight into their mathematical ideas and thinking processes. For example, in the LieCal project, we supplemented our analysis of the correctness of answers with a longitudinal analysis of the changes in students' strategies over time (Cai, Moyer, Wang, & Nie, 2011). Figure 1 shows the doorbell problem, an open-ended task used in the LieCal assessments. In this problem, students were asked to generalize from the given pattern of doorbell rings.

Student performance on this task were analyzed longitudinally over the course of 3 years and found that, in general, both CMP and non-CMP students increased their generalization abilities over the middle school years and that CMP students developed, on average, greater generalization abilities than non-CMP students. More specifically, the success rate for each question improved over time for both CMP and



Making Generalizations

Sally is having a party.

The first time the doorbell rings, 1 guest enters.

The second time the doorbell rings, 3 guests enter.

The third time the doorbell rings, 5 guests enter.

The fourth time the doorbell rings, 7 guests enter.

Keep going in the same way. On the next ring a group enters that has 2 more persons than the group that entered on the previous ring.

A. How many guests will enter on the 10th ring? Explain or show how you found your answer.

B. How many guests will enter on the 100th ring? Explain or show how you found your answer.

C. 299 guests entered on one of the rings. What ring was it? Explain or show how you found your answer.

D. Write a rule or describe in words how to find the number of guests that entered on each ring.

Fig. 1 The doorbell problem used in the LieCal open-ended assessment

non-CMP students, but the CMP students' success rate increased significantly more than that of the non-CMP students on questions A and C in the doorbell problem over the course of the middle grades (Cai, Moyer et al., 2011).

By examining the students' solution strategies on this open-ended task, we obtained further data to inform and confirm this finding. We coded the solution strategies for each of these questions into two categories: abstract and concrete. Students who chose an abstract strategy generally formulated an algebraic representation of the relationship between the ring number and the number of guests entering at that ring (e.g., the number of guests who enter on a particular ring of the doorbell equals two times that ring number minus one). These students then were able to use their generalized rule (e.g., to determine the ring number at which 299 guests entered). In contrast, those who used a concrete strategy made a table or a list or noticed that each time the doorbell rang two more guests entered than on the previous ring and so added 2's sequentially to find an answer.

Looking at the changes over time in the solution strategies students employed to solve the doorbell problem, we found that both CMP and non-CMP students increased their use of abstract strategies over the middle grades. Indeed, in the fall of 2005, only one CMP student and none of the non-CMP students used an abstract strategy to correctly answer question A, but in the spring of 2008, nearly 9 % of the CMP students and 9 % of the non-CMP students used abstract strategies to correctly answer question A. Similarly, nearly 20 % of the CMP students and 19 % of non-CMP students used an abstract strategy to correctly answer question B by the spring of 2008. Although only a small proportion of the CMP and non-CMP students used abstract strategies to correctly answer question C in the spring of 2008, the rate of increase for the CMP students who used abstract strategies from the fall of 2005 to the spring of 2008 was significantly greater than that for non-CMP students ($z=2.58$, $p<.01$).

Thus, these results provided additional detail that informed our conclusion that both CMP and non-CMP students increased their generalization abilities over the middle school years, but that on average, the CMP students developed their generalization ability more fully than did non-CMP students.

The China project did a similar qualitative analysis of the solution strategies that students employed to solve open-ended mathematics questions. A similar observation was obtained that the students receiving the new curriculum were more likely to use a more generalized strategy (e.g., algebraic or arithmetic representation) to solve open-ended questions such as the doorbell problem than the students receiving the conventional curriculum (Ni, Li, Cai, & Hau, 2009). The advantage of using the more generalized strategy became evident in students' solutions to the part of the doorbell problem where 299 guests enter.

Analyzing and Reporting Change Beyond the Grade Band

Generally speaking, mathematics curricula are designed to address the needs of students within a particular grade band, whether it be the elementary, middle, or secondary grades. Analyses of curricular effect, however, should not be limited to the grades in which students encounter the curriculum. Indeed, students' experiences with mathematics curricula can set them up for success or failure in their future mathematics classes. Thus, it is important for longitudinal curriculum analyses to follow students beyond the grade band in which they experience a curriculum to gauge the long-term effects of the curriculum.

The LieCal project initially measured curricular effect on students' learning of algebra while they were still in middle school. The middle school results suggested a potential parallel with findings from studies of Problem-Based Learning (PBL) in medical education (Hmelo-Silver, 2004; Vernon & Blake, 1993). Specifically, medical students who were trained using PBL approaches performed better than non-PBL (e.g., lecturing) students on clinical components in which conceptual understanding and problem solving ability were assessed, but performed as well as non-PBL students on measure of factual knowledge. When the medical students were assessed again 6 months to a few years later, the PBL students were found to perform better than their counterparts on clinical components and measures of factual knowledge (Vernon & Blake, 1993).

Thus, the LieCal project subsequently followed 1,000 of the CMP and non-CMP students into high school to investigate the hypothesis that the superior conceptual understanding and problem solving abilities gained by CMP students in middle school might result in better performance on delayed assessments of procedural skill, conceptual understanding, and problem solving. We used measures of open-ended problem solving in the ninth grade, basic mathematical skills (on the state test) in the tenth grade, and problem solving and posing in the 11th grade to probe the long-term effects of the CMP and non-CMP curricula that the students had used in middle school. On all three measures, we found that the use of the CMP curriculum in middle school had positive effects, not only on students' middle school performance, but also on their high school performance (Cai, Moyer, & Wang, 2013).

More specifically, we found that, controlling for middle school achievement, the ninth grade, former CMP students performed as well as or significantly better than the non-CMP students on open-ended mathematics problems. On the tenth grade state standardized test of basic mathematical skills, we found that the CMP students had a significantly higher scaled mean score than the non-CMP students (Cai, Moyer, & Wang, 2013). This result held for a series of analyses of covariance controlling for the students' sixth grade baseline scores on LieCal multiple choice and open-ended tasks as well as for their sixth, seventh, and eighth grade state standardized mathematics test scores. Similarly, on problem-posing tasks administered in the 11th grade, we examined the performance of groups of CMP and non-CMP students who had performed similarly on their sixth grade baseline examinations (Cai, Moyer, Wang, Hwang, et al., 2013). We found that the CMP students were more likely to pose problems that correctly reflected the mathematical conditions of the given problem situation than the comparable non-CMP students. Moreover, a detailed analysis of the students' problem-solving performance and strategy use showed that the CMP students appeared to have greater success algebraically abstracting the relationship in the problem-solving task (Cai, Silber, Hwang, Nie, Moyer, & Wang, 2014). Together, these results point to the longer-term effects of curriculum and thus highlight the importance of analyzing and reporting change beyond the immediate grade band in which a curriculum is implemented.

Interpreting Change in Mathematics Achievement

Interpreting change in mathematics achievement means identifying the causes that may be responsible for the observed change. This is an extremely important task for advancing knowledge of how educational inputs are related to educational outputs and thus to inform educational practice. It is also an extremely difficult task to accomplish. Below we describe our approach to interpreting change in our longitudinal studies and the lessons we have learned in the process (Cai, Ni, & Lester, 2011). In particular, we focus on the importance of establishing equivalent groups of students in comparative curricular studies and on the need for a conceptual model that informs an initial hypothesis.

Equivalence of Student Sample Groups

Both the LieCal and China studies were designed to investigate curricular influence on change in student learning outcomes by comparing two curricula. To infer any causal links between a curriculum and observed change in student learning outcomes in this type of comparative study, it is of paramount importance to set up equivalent groups of students to receive the curricula (NRC, 2004). However, it is often challenging to implement random assignment of students to one or the other curriculum because of administrative and ethical constraints. When this is not possible, it is wise to collect as much information as possible about the student sample and consider how any observed change in student achievement may be associated with characteristics of the student sample in addition to the curriculum factor. The LieCal project randomly selected reform curriculum schools, and was able to obtain information on the prior achievement of the students to create statistically comparable groups by selecting comparable non-reform schools. However, this was not possible in the Chinese project. The researchers could not equate the groups statistically because they lacked prior achievement data. This resulted in a high degree of uncertainty about the observed changes in student achievement being due to the different curricula the students had received. The problem might have been mitigated if the Chinese project had, for example, administered an intelligence test and used it as a control variable in the analyses. However, a problem would still have remained because intelligence test scores are only moderately correlated with school achievement. This underscores the importance of obtaining adequate information about student populations prior to the beginning of a comparative study.

Initial Conceptual Model

One must have a theory or hypothesis, regardless how rudimentary it may be at first, to design a curriculum study that can test how curricular influence is related to classroom instruction and, in turn, to students' mathematics achievement (Christie & Fierro, 2010; NRC, 2004; Weiss, 1998). In the LieCal project, we used the conceptual model shown in Fig. 2 of the relations among curriculum, teaching and learning to frame our investigation of the factors or processes that likely caused the observed changes in students' mathematics achievement (e.g., Cai & Moyer, 2006). We considered that curriculum materials including curriculum standards, textbooks, and teacher manuals would affect the kinds of learning tasks that the teachers selected and implemented and the types of classroom discourse that the teachers engaged in with their students. The nature of the learning tasks and classroom discourse implemented in the classroom would in turn affect learning processes and learning outcomes for students.

It would be ideal to test the entire set of relations described in Fig. 2 simultaneously and conclusively. However, this is almost impossible to implement technically. Among other issues, one major obstacle is that a measurement model involving so many variables would produce a covariance matrix so complicated that it would be impossible to make a sensible estimation of the parameters concerned (Ni, Li, Cai, & Hau, in press; Raudenbush & Bryk, 2002). This complication is made even more acute by the difficulty in reliably measuring the variables.

Facing this challenge in our projects, we used the problem-solving heuristic of "divide-and-conquer" to address our research questions. After having observed the changes in students' mathematics achievement and their association with the type of

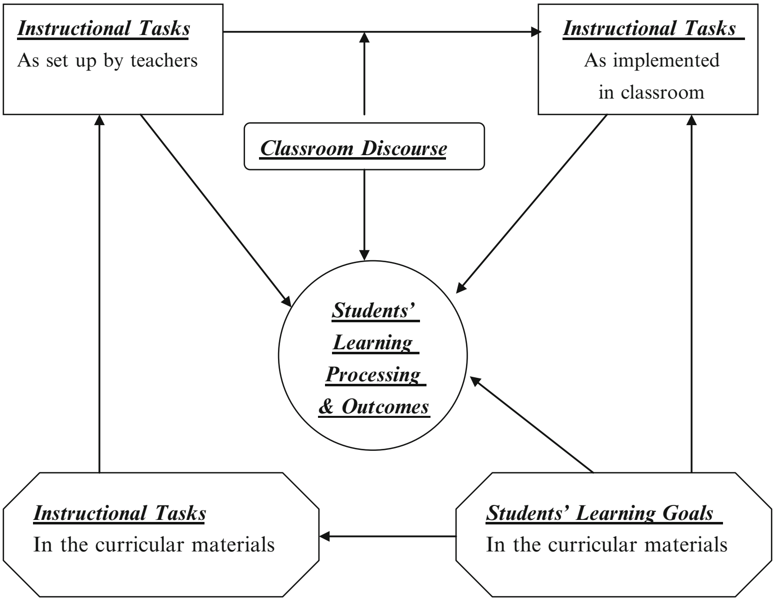

Instructional Tasks
As set up by teachers
Instructional Tasks
As implemented in classroom
Classroom Discourse
Students' Learning Processing & Outcomes
Instructional Tasks
In the curricular materials
Students' Learning Goals
In the curricular materials

Fig. 2 Framework used in the two projects (Cai, 2007; Cai & Moyer, 2006; Ni et al., in press)

curriculum being implemented, the LieCal project used HLM analyses to investigate whether the conceptual or procedural emphasis of classroom instruction moderated the curricular influence on the achievement gains of the students. However, these variables did not show any meaningful influence. We then looked into the effect of the cognitive demand of instructional tasks. Using the classification scheme of Stein and Lane (1996), the instructional tasks actually used in the CMP and non-CMP classrooms were classified into four increasingly demanding categories of cognition: memorization, procedures without connections, procedures with connections, and doing mathematics. We found that the distributions of types of instructional tasks in the CMP and non-CMP classrooms were significantly different, with CMP teachers implementing a higher percentage of cognitively demanding tasks (procedures with connections and doing mathematics) than non-CMP teachers (Cai, 2014). In contrast, non-CMP teachers implemented a significantly higher percentage of tasks with low cognitive demand (memorization or procedures without connections). Moreover, we found that this variable was a significant predictor of achievement gains in the students receiving either curriculum.

Similarly, following the conceptual framework in Fig. 2, the China project examined the relationships of the cognitive features of instructional tasks (high cognitive demand, multiple representations, and multiple solution-strategies) to teacher–student classroom discourse on the one hand (Ni, Zhou, Li, & Li, 2014) and to students' mathematics achievement gains on the other hand in the Chinese mathematics classrooms (Ni, Zhou, Li, & Li, 2012). The results showed that high cognitive demand tasks were associated with teachers' high-order questions, which in turn led to students' highly participating responses. However, teachers tended to be more authoritative in evaluating student responses when they used high cognitive demand tasks or high-order questions. It was unexpected that teachers tended to ask low-order Yes or No questions when they elicited multiple solution methods from students for an instructional task. It appeared that the teachers just wanted students to talk more but did not press students to be accountable for their answers when pursuing multiple solution methods. Concerning the effects of the cognitive features of instructional tasks on student learning, the China project found that the cognitive features did not predict achievement gain on any of the cognitive learning outcomes (computation, routine problem solving, and complex problem solving). However, high cognitive demand of instructional tasks was shown to positively predict affective outcomes including students' expressed interest in learning mathematics, classroom participation, and a more dynamic view about mathematics. In turn, the indicators of students' positive attitude towards learning mathematics were significantly associated with their cognitive learning outcomes. These results illustrated the richness, complexity, and uncertainty of the links from the written curriculum to the implemented curriculum in classrooms and then to the achieved curriculum as shown in changes in student learning.

Our experience with the two projects indicates that a conceptual framework, such as the one in Fig. 2, is a necessary tool for planning and executing a quality longitudinal study of students' mathematics learning in relation to curricula and classroom instruction.

Conclusion

The LieCal project and the China project provide opportunities for us to consider the challenges in conducting high-quality longitudinal research into student learning. It is clear that the constructs we are interested in measuring are broad, requiring both careful definitions and well-chosen measures to address properly. If we wish to measure growth and change in students' academic achievement, it is necessary to use a variety of measures that address multiple facets of that growth and change. To characterize the effects of curriculum on student learning, diverse measures of conceptual understanding, procedural skill, problem-solving and problem-posing abilities, and interest and attitude toward learning mathematics are all useful tools.

In addition, the contexts and structures within which students learn guarantee that the data we collect will be complex. The methods of analyses we choose must therefore be suitable for the structure of the data and be sufficiently robust to take into consideration the many influences on student learning. Social and socioeconomic factors, the nature of classroom instruction, and many other factors can influence student learning, and thus the design of studies that include these factors must be carefully considered. Of course, no study design, however solid it may be, can address all of the potential influences. As we have done in planning the LieCal and China projects, researchers must use their conceptual models and hypotheses strategically to choose what to address and how, given the constraints of experimental design and ethical considerations.

As we consider the results from these two projects, we look forward to continued longitudinal research that seeks to conceptualize, measure, analyze, and interpret change in student learning. We conclude with a final note on the role of experimental studies and our expectations for them. It is important to note that the analyses done by both the LieCal project and the China project about the relations between classroom processes and gains in student mathematics achievement were descriptive in nature. Therefore, experimental studies are yet required to test and prove a causal link of the classroom processes to student learning outcomes. However, these correlational findings were derived from naturalistic situations in which the classrooms differed with respect to factors such as teachers' allocation of time to academic activities, classroom organization, and student backgrounds. The patterns of association observed in these situations do provide meaningful results that can guide further experimental studies and classroom practice (Brophy & Good, 1986).

Of course, not every experimental study using random assignment will produce causal links between a set of assumed factors and the observed outcomes. Conversely, it is always questionable for a non-randomized study to draw such causal links. Indeed, caution is always appropriate when interpreting the results of any single study. Consequently, consistency and replication of findings is the key to the generalization of any finding. A good example of this is the evaluation of the federally funded early childhood programs in the USA (Heckman, Doyle, Harmon, & Tremblay, 2009; Reynolds, 2000). On the one hand, the implementation of early childhood education varied in different states and communities. This made generalization of any particular finding about its effectiveness difficult. On the other hand, the assemblage of evaluations of programs that were carried out in diverse situations provided an excellent opportunity to examine whether or not a given finding about the effects of the programs could be observed across different circumstances. Converging evidence was obtained that indicated that the cognitive advantages for the children participating in the programs tended to disappear approximately 3 years after leaving the programs. However, those children who participated did benefit in terms of increased likelihood of retention in grade school, high school graduation, college education, and employment. The conclusions that arose from the convergence of consistent findings and the replication of those findings across diverse contexts have subsequently contributed to well-informed educational policy and practice for early childhood education. Similar concerted efforts are required to examine the robustness of findings about the influences of curricular and classroom variables on gains in student mathematics achievement in different circumstances and with different methods.

Acknowledgments Preparation of this article was supported by grants from the National Science Foundation (ESI-0454739 and DRL-1008536) and Research Grant Council of HKSAR, China (CERG-462405; CERG-449807) and the National Center for School Curriculum and Textbook Development, Ministry of Education of People's Republic of China. Any opinions expressed herein are those of the authors and do not necessarily represent the views of the funding agencies.

References

Betebenner, D. W. (2008). Toward a normative understanding of student growth. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 155–170). New York: Taylor & Francis.

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 328–366). New York: Macmillan.

Burton, L. (1984). Mathematical thinking: The struggle for meaning. *Journal for Research in Mathematics Education, 15*, 35–49.

Cai, J. (1995). A cognitive analysis of U.S. and Chinese students' mathematical performance on tasks involving computation, simple problem solving, and complex problem solving. *Journal for Research in Mathematics Education monograph series 7*. Reston, VA: National Council of Teachers of Mathematics.

Cai, J. (1997). Beyond computation and correctness: Contributions of open-ended tasks in examining students' mathematical performance. *Educational Measurement Issues and Practice, 16*(1), 5–11.

Cai, J. (2000). Mathematical thinking involved in U.S. and Chinese students' solving process-constrained and process-open problems. *Mathematical Thinking and Learning, 2*, 309–340.

Cai, J. (2007). *Empirical investigations of U.S. and Chinese students' learning of mathematics: Insights and recommendations*. Beijing, China: Educational Sciences Publishing House.

Cai, J. (2014). Searching for evidence of curricular effect on the teaching and learning of mathematics: Some insights from the LieCal project. *Mathematics Education Research Journal*. doi:10.1007/s13394-014-0122-y.

Cai, J., & Moyer, J. C. (2006). *A conceptual framework for studying curricular effects on students' learning: Conceptualization and design in the LieCal Project*. Paper presented at the annual meeting of the International Group of Psychology of Mathematics Education, Prague, Czech Republic: Charles University in Prague.

Cai, J., Moyer, J. C., Wang, N., & Nie, B. (2011). Examining students' algebraic thinking in a curricular context: A longitudinal study. In J. Cai & E. Knuth (Eds.), *Early algebraization: A global dialogue from multiple perspectives* (pp. 161–186). New York: Springer.

Cai, J., Ni, Y. J., & Lester, F. (2011). Curricular effect on the teaching and learning of mathematics: Findings from two longitudinal studies in China and the United States. *International Journal of Educational Research, 50*(2), 63–143.

Cai, J., Wang, N., Moyer, J. C., Wang, C., & Nie, B. (2011). Longitudinal investigation of the curricular effect: An analysis of student learning outcomes from the LieCal project in the United States. *International Journal of Educational Research, 50*, 117–136.

Cai, J., Moyer, J. C., & Wang, N. (2013a). Longitudinal investigation of the effect of middle school curriculum on learning in high school. In A. M. Lindmeier & A. Heinze (Eds.), *Proceedings of the 37th Conference of the International Group for the Psychology of Mathematics Education* (pp. 137–144). Kiel, Germany: PME.

Cai, J., Moyer, J. C., Wang, N., Hwang, S., Nie, B., & Garber, T. (2013b). Mathematical problem posing as a measure of curricular effect on students' learning. *Educational Studies in Mathematics, 83*, 57–69.

Cai, J., Silber, S., Hwang, S., Nie, B., Moyer, J. C., & Wang, N. (2014). Problem-solving strategies as a measure of longitudinal curricular effects on student learning. In P. Liljedahl, C. Nicol, S. Oesterle, & D. Allan (Eds.), *Proceedings of the 38th Conference of the International Group for the Psychology of Mathematics Education and the 36th Conference of the North American Chapter of the Psychology of Mathematics Education* (Vol. 2, pp. 233–240). Vancouver, BC, Canada: PME.

Christie, C. A., & Fierro, L. A. (2010). Program evaluation. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (Vol. 3, pp. 706–712). Oxford: Elsevier.

Fisher, A., & Foreit, J. (2002). *Designing HIV/AIDS intervention studies: an operations research handbook*. Washington, DC: Population Council.

Ginsburg, H. P. (Ed.). (1983). *The development of mathematical thinking*. New York: Academic.

Hatano, G. (1988). Social and motivational bases for mathematical understanding. In G. B. Saxe & M. Gearhart (Eds.), *Children's mathematics* (pp. 55–70). San Francisco: Jossey Bass.

Heckman, J., Doyle, O., Harmon, C., & Tremblay, R. (2009). Investing in early human development: Timing and economic efficiency. *Economics and Human Biology, 7*, 1–6.

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 65–97). New York: Macmillan.

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16*, 235–266.

Hwang, S., Cai, J., Shih, J., Moyer, J. C., Wang, N., & Nie, B. (2015). Longitudinally investigating the impact of curricula and classroom emphases on the algebra learning of students of different ethnicities. In J. A. Middleton, J. Cai, & S. Hwang (Eds.), *Large-scale studies in mathematics education*. New York: Springer.

Kieran, T., & Pirie, S. B. (1991). Recursion and the mathematical experience. In L. P. Steffe (Ed.), *Epistemological foundations of mathematical experience* (pp. 78–101). New York: Springer.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice, 41*, 212–218.

Krutetskii, V. A. (1976). *The psychology of mathematical abilities in schoolchildren*. Chicago: University of Chicago Press.

Li, Q., & Ni, Y. J. (2011). Impact of curriculum reform: Evidence of change in classroom practice in the mainland China. *International Journal of Educational Research, 50*, 71–86.

Linn, R. L. (2007). Performance standards: What is proficient performance? In C. E. Sleeter (Ed.), *Facing accountability in education: Democracy and equity at risk* (pp. 112–131). New York: Teachers College Press.

Loveless, T. (2008). The misplaced math student. *The 2008 Brown Center Report on American Education: How well are American students learning?* Washington, DC: Brookings.

Ma, X. (2010). Longitudinal evaluation design. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (Vol. 3, pp. 756–764). Oxford, England: Elsevier.

Mayer, R. E. (1987). *Educational psychology: A cognitive approach*. Boston: Little & Brown.

Mislevy, R. J. (1995). What can we learn from international assessments? *Educational Evaluation and Policy Analysis, 17*(4), 419–437.

Moyer, J. C., Cai, J., Wang, N., & Nie, B. (2011). Impact of curriculum reform: Evidence of change in classroom practice in the United States. *International Journal of Educational Research, 50*, 87–99.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

National Research Council. (2004). *On evaluating curriculum effectiveness: Judging the quality of K-12 mathematics evaluations*. Washington, DC: The National Academies Press.

Ni, Y. J., Li, Q., Cai, J., & Hau, K. T. (2009). *Has curriculum reform made a difference? Looking for change in classroom practice*. Hong Kong, China: The Chinese University of Hong Kong.

Ni, Y. J., Li, Q., Cai, J., & Hau, K. T. (in press). Has curriculum reform made a difference in classroom? An evaluation of the new mathematics curriculum in the Mainland China. In B. Sriraman, J. Cai, K-H. Lee, F. Fan, Y. Shimuzu, C. S. Lim, K. Subramanium (Eds.). *The first sourcebook on Asian research in mathematics education: China, Korea, Singapore, Japan, Malaysia and India*. Charlotte, NC: Information Age.

Ni, Y., Li, Q., Li, X., & Zhang, Z.-H. (2011). Influence of curriculum reform: an analysis of student mathematics achievement in mainland China. *International Journal of Educational Research, 50*, 100–116.

Ni, Y. J., Zhou, D., Li, Q., & Li, X. (2012). *To feel it to better learn it: Effect of instructional tasks on mathematics learning outcomes in Chinese primary students.* Paper presented at the third meeting of the EARLI SIG 18 Educational Effectiveness, Zurich, Switzerland, 29–31 August 2012.

Ni, Y. J., Zhou, D. H., Li, X., & Li, Q. (2014). Relations of instructional tasks to teacher-student discourse in mathematics classrooms of Chinese primary schools. *Cognition and Instruction, 32*, 2–43.

Polya, G. (1945). *How to solve it*. Princeton, NJ: Princeton University Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models*. Newbury Park, CA: Sage.

Reynolds, A. J. (2000). *Success in early intervention: The Chicago Child-Parent Centers*. Lincoln, NE: University of Nebraska Press.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334–371). New York: Macmillan.

Schoenfeld, A. H. (Ed.). (1997). *Mathematical thinking and problem solving*. Mahwah, NJ: Erlbaum.

Steen, L. A. (1999). Twenty questions about mathematical reasoning. In L. V. Stiff & F. R. Curcio (Eds.), *Mathematical reasoning in grades K-12 (1999 Yearbook of the National Council of Teachers of Mathematics)*. NCTM: Reston, VA.

Stein, M. K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation, 2*, 50–80.

Sternberg, R. J. (1999). The nature of mathematical reasoning. In L. V. Stiff & F. R. Curcio (Eds.), *Developing mathematical reasoning in grades K-12* (pp. 37–44). Reston, VA: National Council of Teachers of Mathematics.

Sternberg, R. J., & Ben-Zeev, T. (Eds.). (1996). *The nature of mathematical thinking*. Hillsdale, NJ: Erlbaum.

Vernon, D. T., & Blake, R. L. (1993). Does problem-based learning work? A meta-analysis of evaluative research. *Academic Medicine, 68*, 550–563.

Weiss, C. H. (1998). *Evaluation: Methods for studying programs and policies* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.