Developmental Psychology
2010, Vol. 46, No. 1, 178–192

0012-1649/10/$12.00 DOI: 10.1037/a0016701

The Developmental Relations Between Conceptual and Procedural Knowledge: A Multimethod Approach

Michael Schneider and Elsbeth Stern
ETH Zurich

Interactions between conceptual and procedural knowledge influence the development of mathematical competencies. However, after decades of research, these interrelations are still under debate, and empirical results are inconclusive. The authors point out a source of these problems. Different kinds of knowledge and competencies only show up intertwined in behavior, making it hard to measure them validly and independently of each other. A multimethod approach was used to investigate the extent of these problems. A total of 289 fifth and sixth graders' conceptual and procedural knowledge about decimal fractions was measured by 4 common hypothetical measures of each kind of knowledge. Study 1 tested whether treatments affected the 2 groups of measures in consistent ways. Study 2 assessed, across 3 measurement points, whether conceptual and procedural knowledge could be modeled as latent factors underlying the measures. The results reveal substantial problems with the validities of the measures, which might have been present but gone undetected in previous studies. A solution to these problems is essential for theoretical and practical progress in the field. The potential of the multimethod approach for this enterprise is discussed.

Keywords: conceptual knowledge, procedural knowledge, measurement, validity, confirmatory factor analyses

Changes in a person's knowledge are among the most powerful mechanisms underlying and facilitating development (Case, 1992; Karmiloff-Smith, 1992; Piaget, 1978; Siegler, 1996; Spelke, 2000). Thus, the question of how a person's changing knowledge can be measured and modeled lies at the heart of developmental psychology. Many developmental psychologists have found it useful to treat knowledge not as a unitary construct but as differentiated into at least two kinds of knowledge: (a) conceptual knowledge, facilitating understanding of abstract principles, and (b) procedural knowledge, assisting in solving concrete problems. For example, some persons might understand the principle of commutativity (i.e., $a + b = b + a$) without applying it correctly to solve a problem. Other persons might apply the principle correctly without understanding why it is correct.

Michael Schneider and Elsbeth Stern, Institute for Behavioral Sciences, ETH Zurich, Zurich, Switzerland.

This article is based on Michael Schneider's doctoral thesis, accepted 2006 at the Technical University of Berlin, Berlin, Germany. Elsbeth Stern was primary advisor of the thesis. Data collection was financed by the Max Planck Institute for Human Development in Berlin, Germany, where the authors held positions at that time.

We thank Helmut Jungermann, Bethany Rittle-Johnson, Lieven Verschaffel, Ilonca Hardy, Roland Grabner, Bert De Smedt, Henrik Saalbach, and Lennart Schalk for helpful comments on earlier versions of this article. Further, we are grateful to David Rush for help with the manuscript preparation and to Hella Beister, Gabriele Cierniak, Barbara Diffendaffer, Anna-Kristina Kalbacher, Jennifer Paetsch, Ruth Wagner, and Sylvia Weber for help with the data collection.

Correspondence concerning this article should be addressed to Michael Schneider, Institute for Behavioral Sciences, ETH Zurich, Universitaetsstrasse 6, CAB G84.2, 8092 Zurich, Switzerland. E-mail: schneider@ifv.gess.ethz.ch

For many decades now, researchers have tried to examine how conceptual and procedural knowledge influence each other during development (cf. Byrnes & Wasik, 1991; Canobi, Reeve, & Pattison, 1998; Dixon & Moore, 1996; Gelman & Gallistel, 1978; Gelman & Meck, 1983; Greeno, Riley, & Gelman, 1984; Hiebert, 1986; Resnick & Ford, 1981; Rittle-Johnson, Siegler, & Alibali, 2001; Sophian, 1997). Key questions concern the naturally occurring order of acquisition of these two kinds of knowledge, their optimal order of acquisition, whether conceptual knowledge causally influences procedural knowledge, and whether procedural knowledge causally influences conceptual knowledge.

Despite this long history of research on the relations between conceptual and procedural knowledge, the conflicting theoretical viewpoints have not converged on a universally agreed upon position but rather have been subject to ongoing debates (Gilmore & Papadatou-Pastou, 2009; LeFevre et al., 2006; Mabbott & Bisanz, 2008; Rittle-Johnson & Star, 2007). The empirical results differ strongly across content domains, studies, and persons (Rittle-Johnson & Siegler, 1998). In the current article, we examine a possible explanation for these difficulties: Different kinds of knowledge only show up intertwined with each other and with other competencies in overt behavior. It is, therefore, not clear to what extent they can be measured validly and partly independently of each other. Tasks used to assess conceptual or procedural knowledge differ between content domains, age groups, and even studies. There are neither established standards for measuring the kinds of knowledge nor set standards for testing the validities of hypothetical measures. We discuss how these validities can be investigated by means of a multimethod approach. In two empirical studies, we demonstrate this approach and show that eight measures commonly used to assess conceptual or procedural knowledge in published studies have insufficient validities.

Characteristics of Conceptual and Procedural Knowledge

The distinction between conceptual and procedural knowledge is used not only in developmental psychology but also in cognitive science (Goldstone & Kersten, 2003; Johnson, 2003), in educational psychology (Baroody & Dowker, 2003; Hiebert & Lefevre, 1986; Rittle-Johnson & Koedinger, 2005), in standards for mathematics teaching (National Council of Teachers of Mathematics, 2000), and in large-scale studies of students' achievement (National Assessment Governing Board & U.S. Department of Education, 2003; OECD, 1999).

Conceptual knowledge is usually viewed as general and abstract knowledge of the core principles and their interrelations in a domain. Accordingly, it is assumed to be stored mentally in some form of relational representation, for example, schemas or semantic networks (Hiebert, 1986), which allow for its flexible transformation through processes of inference and elaboration. It is, therefore, not bound to specific problem types (Baroody, 2003; Hiebert, 1986). We consider declarative knowledge (e.g., Anderson, 1983) as a generic term comprising (general and abstract) conceptual knowledge as well as knowledge about (specific and concrete) instances and events.

Procedural knowledge, in contrast, is usually seen as knowledge of operators and the conditions under which they can be applied to reach certain goals (Anderson, 1993; Baroody, 2003; Rittle-Johnson et al., 2001). It can be automatized to different degrees, depending on the extent of practice. Automatized procedural knowledge can be used with minimal conscious attention and few cognitive resources (cf. Johnson, 2003). This efficiency, however, has the drawback of inflexibility. Because automatized knowledge is only partly open to conscious inspection, it can hardly be verbalized or transformed by higher mental processes. As a consequence, it is often tied to specific problem types (Baroody, 2003; Rittle-Johnson et al., 2001; Singley & Anderson, 1989).

There are four different theoretical viewpoints on the causal interrelations of these kinds of knowledge, each view being supported by some empirical evidence (Baroody, 2003; Haapasalo & Kadjievich, 2000; Rittle-Johnson & Siegler, 1998; Rittle-Johnson et al., 2001). The relations might be unidirectional from conceptual knowledge to procedural knowledge (concepts-first view; e.g., Geary, 1994; Gelman & Williams, 1998; Halford, 1993), unidirectional from procedural knowledge to conceptual knowledge (procedures-first view; e.g., Fuson, 1988; Karmiloff-Smith, 1992; Siegler & Stern, 1998), or bidirectional (iterative model; Rittle-Johnson et al., 2001). Finally, the kinds of knowledge might not be directly causally related (inactivation view; e.g., Resnick, 1982; Resnick & Omanson, 1987).

The Problem of Measuring Knowledge Kinds

As described, pairs of knowledge kinds are postulated to explain empirically found dissociations between solutions of tasks with the same content but different functional characteristics, for example, explaining an abstract principle versus solving a concrete problem. In their literature review, Rittle-Johnson and Siegler (1998) discussed four major types of possible dissociations: (a) different means of the two measures (Canobi, 2004), (b) low or medium-high correlations between the two measures (e.g., Cowan & Renton, 1996), (c) different developmental trends of the measures (Byrnes & Wasik, 1991), and (d) different effects of interventions on the two measures in experimental designs (Rittle-Johnson & Star, 2007). In the literature, these dissociations between measures are often interpreted in terms of conceptual and procedural knowledge.

However, none of these findings shows clearly that two distinguishable kinds of knowledge were assessed. The two measures used in each design might as well have assessed a single knowledge kind. For instance, different means of the two measures could be due to different task difficulties. Low reliabilities of the measures can account for the imperfect correlations. Different developmental trends might be due to developmental differences in background knowledge, such as learning new words that make it easier to explain an abstract principle. The same applies to intervention studies: A certain treatment might have enhanced the background knowledge useful for solving one type of task but not the other one.

The reason why it is so hard to interpret findings obtained with a hypothetical measure of conceptual knowledge and a hypothetical measure of procedural knowledge is that each measure can have four different variance components, which are confounded. First, each measure might indeed reflect differences in the amount of the kind of knowledge it is supposed to measure. Second, each measure reflects unsystematic measurement error. Third, the variance of each measure reflects assessment-specific competencies. For example, if conceptual knowledge is measured by explanations of an abstract principle, understanding of the principle is confoundedly measured with reasoning ability, verbal skills, and vocabulary (cf. Broaders, Cook, Mitchell, & Goldin-Meadow, 2007; Vosniadou, 1994). When procedural knowledge is assessed by routine problems involving the use of diagrams, then, in addition to problem-solving procedures, background knowledge about diagrams and the represented content influences the assessment (Shah & Hoeffner, 2002). Finally, concepts are static memory structures that can only be put into action by procedural knowledge (Goldstone & Kersten, 2003; Medin, 1989), and presumably people can derive new procedures from their concepts (Geary, 1994). Thus, solutions of conceptual assessment tasks might, to some degree, also reflect procedural knowledge, and solutions of procedural assessment tasks might reflect parts of conceptual knowledge (Rittle-Johnson et al., 2001).

Whatever effect is found with one hypothetical measure of a knowledge kind, one can never be sure which of the four variance components underlies the effect. Therefore, dissociations between one hypothetical measure of a kind of knowledge and one hypothetical measure of another kind of knowledge can principally not yield decisive evidence on the validity of the two assessments. This leads to the key question of this article: What empirical evidence can demonstrate that two measures each validly assess one kind of knowledge?

A Multimethod Approach

For reasons of parsimony, different kinds of knowledge should only be postulated if there is empirical evidence that cannot be more easily explained by alternative constructs. Greeno, Riley, and Gelman (1984) suggested what might constitute such evidence in the context of a study about the relations between children's conceptual principles underlying counting and their procedural counting skills:

> Evidence for understanding of principles always is problematic to some degree. Any single piece of evidence can be explained without recourse to a hypothesis of understanding: performance consistent with the principle could be learned by rote, evaluation could involve simple comparison of example performance and covert performance of a rote procedure, and novel procedures could be generated by trial and error. Even so, a combination of evidence of these various kinds can constitute a compelling argument that principles are understood significantly. (p. 106)

Thus, a kind of knowledge should be postulated if it can efficiently explain comparable effects found simultaneously with several qualitatively different measures (see Anderson, 1983, for a similar argument), thus demonstrating the *convergent validity* (Eid & Diener, 2006) of the measures. Consequently, the assumption of two kinds of knowledge is appropriate if two groups of measures are found, with measures in the same group showing similar effects and measures in different groups varying at least partly independently of each other. This would demonstrate the *divergent validities* (Eid & Diener, 2006) of the two groups of measures. The methodological advantage of using several qualitatively different measures of each kind of knowledge is that it allows disentangling measure-specific from measure-general variance components. If all measures theoretically related to a knowledge kind share a common variance component, the most parsimonious explanation is that this variance component indeed measures the respective kind of knowledge (cf. Campbell & Fiske, 1959).

The Current Studies

We conducted two studies to demonstrate our multimethod approach and to test the validities of four measures of conceptual and procedural knowledge, respectively. In Study 1, we used an experimental design with two treatment groups and a control group. If four measures assess the same knowledge kind (e.g., conceptual knowledge) with a high convergent validity, then an intervention increasing knowledge of this kind should affect all four measures. If these four measures have high divergent validities, they should be affected to a much lesser degree by an intervention increasing a different kind of knowledge (e.g., procedural knowledge). However, the results of this approach do not depend only on the quality of the measures but also on the quality of the treatment interventions. To obtain purer estimates of the validities of our measures, we used a one-group design with three measurement points (i.e., no treatment groups) in Study 2. By confirmatory factor analyses, we tested whether a latent factor underlies the four hypothetical measures of each knowledge kind, which would indicate high convergent validities and whether the two latent factors for conceptual and procedural knowledge covary partly independently of each other, which would indicate high divergent validities of our measures.

This is the first systematic comparison of the convergent and divergent validities of hypothetical measures of conceptual and procedural knowledge reported in the literature. The results can show which of the measures used in past studies are trustworthy, which will assist in understanding the results of these studies and in planning future studies.

To allow for comparisons of the results newly obtained by the multimethod approach with previous findings obtained from single measures, we used the same content area, problem types, and age group as in Rittle-Johnson et al. (2001). In their study with fifth and sixth graders, Rittle-Johnson et al. investigated the adequacy of their iterative model by searching evidence for bidirectional causal relations between conceptual and procedural knowledge. Likewise, we investigated fifth and sixth graders' conceptual knowledge about decimal fractions and their procedural knowledge about locating these fractions on a number line.

To understand the concept of decimal fractions, a child has to know that the value of a digit depends on its position in the digit sequence relative to the decimal point. He or she further must understand that the notational system for numbers has a decimal structure, so that the digits at the different positions indicate units of ten, units of one, units of tenths, units of hundredths, and so forth. By increasing the number of digits after the decimal point, the exactness of a decimal fraction theoretically can be increased limitlessly (Hiebert, 1992; Hiebert & Wearne, 1986; Resnick et al., 1989). The most frequent procedural strategies to locate decimal fractions and other numbers on a number line are the following: (a) counting up from the smallest value on the line; (b) counting down from the largest value on the line; (c) counting up or down from a midpoint value on the line, which is frequently identified by proportional reasoning; and (d) memory recall without counting (Newman & Berger, 1984; Petitto, 1990; Siegler & Opfer, 2003).

In both studies, we used the same four hypothetical measures of conceptual knowledge and four hypothetical measures of procedural knowledge. All eight measures were adapted from published studies in different content domains and age groups, in which they were theoretically justified and used as assessment of either kind of knowledge. In Study 1, we used an experimental pretest–posttest design, and in Study 2, we used a longitudinal design and confirmatory factor analyses to explore the convergent and divergent validities of our measures.

Our knowledge measures are closely related to our characterizations of the two knowledge kinds given above. Therefore, we expected at least acceptable convergent and divergent validities, allowing for subsequent analyses of their causal interrelations despite the concerns expressed above.

Study 1

Rationale and Aims of the Study

We used an experimental pretest–posttest design with two treatment groups and a control group. The conceptual intervention group received a treatment designed to convey mainly conceptual knowledge (and as little procedural knowledge as possible). The procedural intervention group received a treatment designed to convey mainly procedural knowledge (and as little conceptual knowledge as possible). The control group engaged in a nonmathematical activity.

If four measures assess the same kind of knowledge with sufficient convergent validities, then all four measures should indicate an experimentally induced increase in this kind of knowledge.

Likewise, if four measures assess a certain kind of knowledge with sufficient divergent validities, they should not, or should only minimally, be affected by increases in another kind of knowledge. This is the rationale of Study 1. For general arguments in favor of investigating validities of measures by means of experimental designs see Borsboom, Mellenbergh, and van Heerden (2004).

We posed four research questions. First, Is each of the four hypothetical measures of conceptual knowledge more strongly affected by the conceptual intervention than by the intervention of the control group? This would indicate the convergent validity of the measures. Second, Is each of the four hypothetical measures of procedural knowledge increased more strongly by the procedural intervention than by the intervention of the control group? This would indicate the convergent validity of these four measures. Third, Is each of the four hypothetical measures of conceptual knowledge more strongly affected by the conceptual intervention than by the procedural intervention? This would indicate the divergent validities of these measures. Fourth, Is each of the four hypothetical measures of procedural knowledge more strongly affected by the procedural intervention than by the conceptual intervention? This would indicate the divergent validities of these measures.

Method

Participants. The 93 fifth graders were volunteers from 11 elementary schools in Berlin, Germany, and received monetary compensation of 23 Euros (approximately $30) for their participation. The schools were located in middle-class to upper middle class neighborhoods with mainly Caucasian families. Nine participants were excluded from the analyses because of missing data on the pretest or the posttest caused by hardware problems or absence on one measurement point due to illness. The analyses were conducted with data from the remaining 84 children. The conceptual intervention group included 18 girls and 13 boys with a mean age of 11.2 years (SD = 0.4). The procedural intervention group contained 10 girls and 17 boys with a mean age of 11.4 years (SD = 0.9). The control group included 11 girls and 15 boys with a mean age of 11.3 years (SD = 0.5). In Germany, mathematics marks range from 1 (best) to 6 (worst). According to children's self-reported grades from their last school report card, their mean mathematics grade was 2.4 (SD = 0.72) indicating a wide competence range in our sample.

Procedure. The children were tested in small groups at our research institute, each student working individually at a computer, wearing headphones and not seeing each other. The children could solve all tasks at their individual pace and automatically received the prerecorded instructions over the headphones. The experimenter was present and available for questions the whole time. In a first session, the children completed the pretest for about 45 min. In a second session, 5 or 6 days later, the children were randomly assigned to the three experimental groups and participated in their respective intervention. The conceptual intervention group needed an average of 20.6 min (SD = 6.3) to complete their intervention, whereas the procedural intervention group required 13.5 min (SD = 4.4), and the control group required 22.4 min (SD = 10.4). Directly after the treatment, the children completed the posttest. The order of the eight knowledge tests was randomized for each child but kept constant across measurement points. The children also completed a questionnaire about age, grades, and so forth. The interventions used in our study were developed in an unpublished preliminary study with 38 additional participants.

We used the German notation for decimal fractions, that is, a comma instead of the decimal point. However, in this article we report all fractions in the internationally common notation. In accordance with Rittle-Johnson et al. (2001), decimal fractions in the tests and interventions were chosen randomly with equal probabilities from five types: fractions with one, two, or three digits, the latter two having an initial zero after the decimal separator or not. Like Rittle-Johnson et al., we coded solutions of the number line estimation task as correct when they lay within an error interval $\pm 10\%$ of the number line.

Interventions. Study 1 used two intervention groups and a control group.

Conceptual intervention group. The conceptual intervention consisted of two parts. During the first part, the children heard a 4-min long verbal explanation of the most important characteristics of decimal fractions: the meaning of the comma, the role of place values in the decimal system, and the fact that decimal fractions denote parts of whole units. Thus, the intervention aimed to convey content identified by Resnick et al. (1989) and Hiebert (1992) as crucial for the understanding of decimal fractions. During the audio instruction, the children saw several decimal fractions on the screen illustrating the points explained in the text.

After the audio explanation, the children solved three tasks designed to stimulate active elaborations and reflections of the explanations they had heard before. The children saw successively (a) a decimal and a photo of one and a half apples, (b) a decimal and a fraction with a fraction bar, and (c) a decimal and a bar chart on the computer screen. In each instance, the children were asked to write down one similarity and one difference of the two different number representations.

Procedural knowledge builds with practice. The children in this group did not practice the number-line estimation task we used to assess procedural knowledge. Therefore, we expected the conceptual intervention group to acquire more conceptual knowledge than procedural knowledge.

Procedural intervention group. The procedural intervention group saw five different decimal fractions (0.3, 0.71, 0.04, 0.492, and 0.082) one by one on the computer screen, together with their respective position on a number line ranging from 0 to 1. The students were asked to memorize the positions of the numbers. Their knowledge of the positions of the five numbers was then tested by the computer program. Learning phase and testing phase alternated until a child could correctly reproduce the positions of all five numbers. In the second part of the intervention, the children solved 80 trials of the number line estimation task to derive problem-solving procedures from the five examples and to automatize these procedures. In each trial, the children indicated the position of a given decimal fraction on a number line ranging from 0 to 1. We used the same types of decimal fractions as in the hypothetical measures of procedural knowledge. These were randomly selected. Prior research has firmly established that people can derive procedural knowledge from memorized example solutions through practice (Anderson & Fincham, 1994; Logan, 1988; VanLehn, 1986). The children in this group practiced the number-line estimation task. We used the same task to assess procedural knowledge in our tests. Thus, we expected increases in procedural knowledge in the procedural intervention group. In contrast, the members of this group received no feedback during the intervention so that they could not test hypotheses about relations between decimal fractions and their positions on the number line. Therefore, we expected no or only minimal increases in conceptual knowledge.

Control group. The control children completed a 634-word long cloze about the characteristics of elephants and corrected spelling mistakes that we had purposely built into the text. We chose a task without mathematical content to minimize the probability of negative transfer from the intervention to the tests.

Hypothetical measures of conceptual knowledge. We measured conceptual knowledge using tasks that demanded a general understanding of decimal fractions but did not require the actual placement of a decimal fraction on the number line. The instructions for all four measures emphasized that the children should optimize answer accuracy, not solution times. For all four measures, the percentage of correctly solved trials or achieved points was computed per child. The following measures were used.

Evaluation. Each participant read eight different verbal descriptions of problem-solving strategies for the routine problems from the intervention (see Appendix A). Four of them were correct, and the child had to evaluate each strategy as *rather good* or *rather bad* by clicking a respectively labeled button. This measure is based on the idea that one has to reflect about the general attributes of a procedure to judge its adequacy. The measure has been used in many studies (e.g., Gelman & Meck, 1983; Rittle-Johnson & Alibali, 1999; Siegler & Crowley, 1994) to assess conceptual knowledge.

Representation. The idea behind the representation measure used by Byrnes and Wasik (1991) as well as Hecht, Close, and Santisi (2003) is that one can translate between types of magnitude representations only if one understands both notational systems. In each of the 20 trials, the children saw a decimal fraction, together with four pie charts. A part of each pie chart was shaded gray. The children were asked to click on the pie chart where the proportion of the gray area as measured against the whole area corresponded to the decimal fraction.

Comparison. One needs to understand the ordinal relations between numbers to judge which of two numbers has the higher magnitude. For example, Stafylidou and Vosniadou (2004) and Rittle-Johnson et al. (2001) used tasks involving magnitude comparisons to assess conceptual knowledge. We presented the children with 20 pairs of decimal fractions. The children were asked to click on the number with the higher value.

Explanation. Verbal explanations of general principles are a frequently used measure of conceptual knowledge research on science and mathematics learning (e.g., Siegler & Stern, 1998; Vamvakoussi & Vosniadou, 2004). The advantage is that children can be asked questions of an abstract, general, or even hypothetical nature. The disadvantage is that, under some circumstances, children cannot verbalize their conceptual understanding comprehensively (Goldin-Meadow, Alibali, & Church, 1993). We asked the children to write down the answers to four questions about general properties of decimal fractions. Translated from German, the questions were the following: (a) "What does it say about the number, when it contains a comma?" (b) "Why does it make sense to measure the quantity of fuel sold at gas stations in decimal fractions?" (c) "In what everyday situations is it better to use whole numbers instead of decimal fractions?" (d) "The longer a whole number the higher its value. Is this also true for decimal fractions? Please explain." The answers were independently coded by two trained raters. Children earned 2 points for a fully correct answer, 1 point for a partly correct answer, and 0 points for a wrong or missing answer. The raters discussed and unified diverging judgments together. The sum scores of the two raters were correlated, with r = .90 for the pretest and r = .85 for the posttest.

Hypothetical measures of procedural knowledge. Procedural knowledge is tied to routine problems. Therefore, all four measures used in our study required the location of given decimal fractions on a number line, as practiced in the procedural intervention group. The four measures assessed four alternative aspects of problem-solving behavior. We used a different task surface for each measure to ensure that comparable effects found with different hypothetical measures of the same knowledge kind could not be attributed to identical perceptual or motor characteristics of the tasks.

Accuracy. Procedural knowledge is a necessary condition for solving problems. Therefore, the percentage of correctly solved routine problems has been used by a wide variety of authors (e.g., Byrnes & Wasik, 1991; Canobi et al., 1998; Rittle-Johnson & Alibali, 1999) to assess procedural knowledge. In our test, the children completed 20 trials of the number line estimation task. The children located the value of a decimal fraction on the number line by moving a lever to this position with the mouse. They could readjust the position as often as they wanted and then confirm their final answer by clicking a button. The instructions stated that the correctness, not the speed of their solutions, mattered.

Speed. Together with accuracy, problem-solving speed is among the most frequently used measures of procedural knowledge (Canobi, 2004; Torbeyns, Verschaffel, & Ghesquiére, 2005). It is closely related to the degree of automatization of procedural knowledge (cf. Anderson et al., 2004). For a given problem-solving strategy, solution times decrease as a logarithmic function of practice (Rickard, 1997; Ritter & Schooler, 2001). The participants solved 20 trials of the number-line estimation task by clicking at the correct position with the mouse. The students were asked to work as quickly as possible without sacrificing accuracy. The times from the presentation of a new trial until the mouse click were recorded in milliseconds. Values more than three standard deviations above or below the individual mean of a child were excluded as outliers. Times of error trials were also discarded. All analyses were carried out with the—more symmetrically distributed—natural logarithm of the solution time, but we report all times in seconds to aid interpretation. Although we used both accuracy and speed as measures, we expected no distortion of the results as a result of a speed–accuracy trade-off. Accuracy and speed were measured on different trials. The task surface and instructions for the accuracy trials were designed to optimize accuracy. The task surface and instructions for the speed trials were designed to optimize speed. A person with well-established routine procedures can be expected to solve problems both more accurately and more quickly than a novice who must construct and try out a new approach for reaching a solution.

Asymmetry. Procedural knowledge is goal-directed and, thus, asymmetric. Increases in procedural knowledge therefore decrease solution times for the practiced direction of a task (i.e., translating a number into a position on the number line) more than they decrease solution times for the unpracticed direction (i.e., translating a position on the number line into a number). This makes asymmetry of access a hypothetical measure of procedural knowledge (e.g., Anderson & Fincham, 1994; Anderson, Fincham, & Douglass, 1997; Pennington, Nicolich, & Rahm, 1995; Rabinowitz & Goldberg, 1995). Thus far, asymmetry of access has only been demonstrated with adults. However, explanations of the phenomenon in the literature do not relate to specific age groups or content domains but to the nature of procedural knowledge in general. Therefore, we expect that the asymmetry can be found in children and that it reflects their procedural knowledge. The children in our study solved 20 trials of the practiced direction (Asymmetry I), then 40 trials of the unpracticed direction (Asymmetry II), and finally 20 trials of the practiced direction (Asymmetry I) again. The answer alternatives were presented in a multiple-choice format. The solution times for Asymmetry I (practiced direction) and Asymmetry II (unpracticed direction) were recorded and cleaned as was done for the measurement of speed. The final values for asymmetry were computed as Asymmetry I minus Asymmetry II, so that higher values indicated more procedural knowledge.

Dual-task costs. In studies with adults, dual-task costs have been shown to be negatively related to the extent of a person's practice of a task (e.g., Schumacher, Seymour, Glass, Kieras, & Meyer, 2001), because individuals with better procedural knowledge need less cognitive resources for solving a task and, thus, can better solve a second task simultaneously (Anderson, Taatgen, & Byrne, 2005; Salvucci & Taatgen, 2008). For the measurement of the dual-task costs, the children solved two different types of tasks in an ABBA design. In the 40 trials of the single-task condition, the children saw a decimal fraction on the screen and clicked on one of four arrows that indicated potential positions of the number on a number line. We recorded the solution rate for the number-line estimation task (Dual-Task Costs I). In the dual-task condition, after solving 10 practice trials, the children again solved 40 number-line estimation trials but simultaneously counted given names they heard on a headphone. After every 10 tasks, children entered the scores for the names into a window before they could proceed. Again, we recorded the solution rate for the number-line estimation task (Dual-Task Costs II). If the rate on a block of 10 trials dropped below 60%, a computer-generated message asked the child to try harder. We computed the measure of dual-task costs as 100 × (Dual-Task Costs I minus Dual-Task Costs II)/Dual-Task Costs I. We used solution rates instead of solution times, because the former had more plausible intercorrelations with the other measures and more plausible mean changes over time in Studies 1 and 2.

Table 1
Means and Standard Deviations of the Knowledge Measures at the Two Measurement Points for the Treatment Groups

Treatment group and measure	Time 1		Time 2	
	M	*SD*	*M*	*SD*
Hypothetical measures of conceptual knowledge				
Evaluation				
Conceptual intervention	61.7	14.8	69.8	20.4
Procedural intervention	56.5	20.9	69.0	20.0
Control	55.8	20.7	66.4	21.4
Representation				
Conceptual intervention	49.2	15.0	56.5	17.9
Procedural intervention	55.4	21.6	66.7	20.5
Control	49.6	15.9	56.5	16.7
Comparison				
Conceptual intervention	74.0	14.5	77.1	18.4
Procedural intervention	78.7	14.6	86.9	11.9
Control	75.4	14.6	79.8	13.3
Explanation				
Conceptual intervention	16.5	21.5	27.0	18.9
Procedural intervention	20.4	21.1	21.8	27.0
Control	12.0	15.6	15.4	15.5
Hypothetical measures of procedural knowledge				
Accuracy				
Conceptual intervention	54.8	20.3	66.1	20.4
Procedural intervention	56.5	21.0	83.5	18.4
Control	51.2	17.3	65.0	21.4
Speed				
Conceptual intervention	4.1	1.7	3.3	1.2
Procedural intervention	3.7	1.1	2.6	0.6
Control	3.4	1.2	2.8	0.5
Asymmetry				
Conceptual intervention	−0.4	0.8	0.1	0.6
Procedural intervention	−0.2	0.8	0.3	0.5
Control	−0.2	1.0	0.0	0.4
Dual-task costs				
Conceptual intervention	11.5	17.5	11.9	17.9
Procedural intervention	9.1	25.8	7.9	10.8
Control	10.2	17.4	13.6	14.0

Note. Speed and the asymmetry measures are reported in seconds; all other measures are reported in percentages.

Results

The means and standard deviations for the measures at pretest and posttest are given in Table 1 for the three treatment groups. We expected to find interaction effects between the variables of measurement point (pretest, posttest) and treatment group (conceptual intervention group, procedural intervention group, control group) on each measure. As shown in Table 2, we found these interaction effects only for the measures of explanation and accuracy. The measure of explanation indicated significantly bigger knowledge gains for the conceptual intervention group (Cohen's d = 0.37) than for the control group (d = 0.15) or the procedural intervention group (d = 0.04). The measure of accuracy indicates significantly stronger gains of the procedural intervention group (d = 0.97) as compared with the conceptual intervention group (d = 0.39) and the control group (d = 0.50). Thus, the data found with explanation and accuracy are in line with our hypotheses, whereas the data of all the other measures are not.

The first column of Table 3 shows Cronbach's alphas as indicators of the internal consistencies of each knowledge measure. The coefficients for evaluation and explanation are low, whereas the other values are good or acceptable. Because the children had to evaluate qualitatively different procedures and explain relatively independent properties of decimal fractions, the low coefficients are likely due to the heterogeneity of the items rather than to low reliabilities. The internal consistencies of the reaction time (RT) measures are not given, because RTs of error trials were excluded. Internal consistencies, therefore, could only have been computed for the 6 to 8 persons with no error trials on these measures.

The rightmost three columns of Table 3 give the results of the analyses of variance (ANOVAs) predicting children's scores on a measure by the position of this measure in the test sequence (one

Table 2
Influences of Measurement Point and Treatment on the Knowledge Measures

Measure	Measurement point		Measurement point × Treatment group	
	Partial η^2	p	Partial η^2	p
Hypothetical measures of conceptual knowledge				
Evaluation	.21	.000	.01	.705
Representation	.19	.000	.01	.598
Comparison	.15	.000	.03	.287
Explanation	.11	.002	.07	.045
Hypothetical measures of procedural knowledge				
Accuracy	.49	.000	.13	.003
Speed	.30	.000	.03	.300
Asymmetry	.21	.000	.01	.717
Dual-task costs	.00	.717	.01	.737

to eight). Significant test-order effects were found for evaluation, comparison, and Dual-Task Costs I.

Discussion

The results indicate low convergent validities of our assessments. We found predicted treatment effects for the measures of explanation and accuracy, that is, for one hypothetical measure of each knowledge kind but not for the remaining six measures. The fact that different effects were found with hypothetical assessments of the same knowledge kind demonstrates that our assessments were not pure measures of conceptual and procedural knowledge but additionally or fully reflected other competencies.

The analysis of test-order effects further supports this hypothesis. If four measures assess the same knowledge kind validly and with similar sensitivities, then all four of them should indicate a test-order effect on this knowledge kind. However, we found test-order effects only for some hypothetical measures of each knowledge kind and not for others. Therefore, the effects were not caused by increases in conceptual or procedural knowledge but rather by assessment-specific variance components. The internal consistencies of the measures, which were just medium high, confirm the assumption that each single measure—to some extent—reflected error variance or different competence dimensions.

Our findings illustrate the importance of a multimethod approach to the measurement of conceptual and procedural knowledge. If we had measured conceptual and procedural knowledge only by explanation and accuracy, we would have incorrectly concluded that our treatments had worked the intended way. If we had used a pair of the other measures, we would wrongly have concluded that the treatment had not been effective at all.

A question that remains open is whether the effects found here can be generalized over different treatments. If our measures have only partial validities, they could have been affected by stronger or broader treatments in a more consistent way than they were affected by our treatment. In this case, the unexpected findings in Study 1 may have been due in part to a deficient treatment rather than to low validities of our measures alone. However, it is not known how conceptual and procedural knowledge can be influenced consistently and partly independently of each other. As yet, four different publications (Byrnes & Wasik, 1991; Hiebert & Wearne, 1996; Rittle-Johnson & Alibali, 1999; Rittle-Johnson et al., 2001) have reported experiments with treatments hypothesized to increase conceptual knowledge more strongly than procedural knowledge or vice versa. In none of the experiments did we find different effects of the treatments on the amounts of conceptual or procedural knowledge despite the fact that the studies were carried out expertly, yielded other plausible results, and comprised a variety of very different treatments ranging from 10-min long verbal instructions (Rittle-Johnson et al., 2001) to different 3-year curricula (Hiebert & Wearne, 1996). For this reason, we reinvestigated the relations among our eight knowledge measures in Study 2 without the use of treatment interventions.

Study 2

Aims

In Study 2, we used the same eight knowledge measures as in Study 1 but in a longitudinal design with three measurement points. We explored the convergent and divergent validities of our measures by means of confirmatory factor analyses. If the four hypothetical measures of the same kind of knowledge really assess the same construct, we should find evidence for an underlying latent factor, related to all four of them, at all measurement points. In addition, if the eight measures assess two kinds of knowledge with sufficient divergent validities, we should find, at each measurement point, two interrelated latent factors instead of a single latent factor (cf. Bryant, Christie, & Rendu, 1999). If latent factors are found at each measurement point, we can then explore the causal relations between the latent factors by means of a cross-lagged panel design (Burkholder & Harlow, 2003). Finally, we can explore the longitudinal interrelations of our knowledge kinds to determine whether these relations are measure specific or measure general.

Table 3
Internal Consistencies and Test Order Effects of the Knowledge Measures at Time 1

Measure	Cronbach's α	Test order effects η^2	p
Hypothetical measures of conceptual knowledge			
Evaluation	.27	.27	.001
Representation	.72	.06	.643
Comparison	.73	.18	.029
Explanation	.49	.15	.088
Hypothetical measures of procedural knowledge			
Accuracy	.77	.12	.201
Speed		.09	.422
Asymmetry I		.12	.184
Asymmetry II		.06	.705
Asymmetry		.05	.764
Dual-task costs I	.92	.20	.013
Dual-task costs II	.93	.13	.131
Dual-task costs		.08	.497

We posed the following research questions. First, Do different hypothetical measures of conceptual knowledge assess the same construct? Second, Do different hypothetical measures of procedural knowledge assess the same construct? And third, Do hypothetical measures of conceptual knowledge and of procedural knowledge assess two interrelated, but distinguishable constructs instead of only one construct?

Method

Participants. We tested 231 fifth-grade and sixth-grade volunteers from 10 primary schools in Berlin, Germany. The schools were located in middle-class to upper middle class neighborhoods populated mainly by Caucasian families. None of the children had participated in Study 1. The participants were tested on 2 consecutive days (Time 1 and Time 2). At a third measurement point (Time 3), about 4 months after the second one, 213 of the participants were tested again. Eight of the 213 participants were excluded from the analyses, because they did not complete all tests, they obviously violated the instructions, or their data were lost as a result of hardware problems. All of the following analyses were conducted on the data of the remaining 205 children (mean age = 11.3 years, SD = 0.7). There were no missing data on any of the knowledge tests. The sample was about half female (51%) and half male (49%). About 47% were fifth graders, with the rest being sixth graders. The children's mean Mathematics grades from the last report card was 2.5 (SD = 0.9), falling in the midrange between 1 (best) and 6 (worst). Because in Berlin the general mathematical properties of decimal fractions are usually not taught before the end of sixth grade (Senatsverwaltung für Bildung Jugend und Sport Berlin et al., 2004), the majority of our participants had no extensive school instruction on this topic. However, they could have had some prior conceptual and procedural knowledge as a result of the usual first-grade to fourth-grade lessons on diagrams as well as on distances and prices, which are often decimal fractions.

Procedure. As in Study 1, the students were tested in small groups at our research institute, each student working individually at a computer, wearing headphones and not seeing the others. On the first day (Time 1), the children completed the knowledge tests and received the first half of an intervention. The intervention was the same for all children and served to activate and increase their knowledge. One day later (Time 2), the children received the second half of the intervention and completed the knowledge tests as well as a short questionnaire about personal data. About 4 month later (Time 3), the students solved eight intervention problems as a reminder of the study content and then completed the knowledge tests again. Each of the three data collection sessions was about 90 min in duration. The children were volunteers and received a small monetary compensation. Prior to the main study, the adequacy of the materials and instructions were iteratively optimized in an unpublished preliminary study with 19 additional fifth and sixth graders.

Intervention. We used the "catch-the-monster game" adapted from the study of Rittle-Johnson et al. (2001). In each of the 160 trials, a decimal fraction was presented to a child on the computer screen, together with a number line ranging from 0 to 1. Only the positions of the 0 and the 1 numerals were marked by ticks and labeled with numbers. The children were asked to click on the position on the number line that corresponded to the value of the decimal fraction. They were told that, by doing so, they could catch a monster hiding at this position. Each time a child entered an answer, the picture of a monster appeared at the correct position of the decimal fraction on the line, thus providing feedback. In 8 trials, the children additionally were asked to write down a self-explanation for the correct answer after the feedback. In every 10th trial, unlabeled ticks appeared at the position of the tenths on the line to help the children grasp the decimal structure of the presented numbers and the line. On the basis of the findings of Rittle-Johnson et al. (2001), we expected increases from Time 1 to Time 2 to occur in all of our eight knowledge measures.

Measures. The same eight knowledge measures as in Study 1 were used. Again their order was randomized for each person but kept constant across measurement points. The only changes made for Study 2 consisted of slight modifications of the wording of some of the instructions. The answers on the explanation measure were coded in the same way and by the same two raters as in Study 1. The interrater reliabilities computed as Pearson correlations were .88, .87, and.92 for the three respective measurement points.

Data analyses. We used the MPlus program (Muthén & Muthén, 1998–2007) to analyze the covariance structure of our data (see Appendix B). We chose the maximum-likelihood estimator MLM (Muthén & Muthén, 1998–2007, pp. 482–485) based on the Satorra-Bentler scaled chi-square statistic (Nevitt & Hancock, 2004; Satorra, 2000; Satorra & Bentler, 1999) for the analyses, because it is robust to the non-normality of distributions. The estimator can only be used without missing data. We specified the factor metrics by fixing the unstandardized factor loadings of the evaluation measure to the value 1 in all models and by additionally fixing the unstandardized loadings of the accuracy measure to the value 1 in all models with two latent factors. We allowed the residuals of speed and asymmetry to correlate, because they were the only RT measures among the eight knowledge tests. All manifest measures were z-standardized prior to the analyses to account for their different metrics.

Results

Description of the manifest measures. Means and standard deviations of the eight knowledge measures are given in Table 4 together with the significance of the mean changes across measurement points. Most of the solution rates were clearly above a chance level. As hypothesized, all measures showed significant knowledge increases from Time 1 to Time 2. The solution rates of the four hypothetical measures of conceptual knowledge and asymmetry of access increased, whereas solution times and dual-task costs decreased. Changes over the 4 months between Times 2 and 3 were less systematic.

Table 5 displays the internal consistencies of the accuracy scores as indicated by their Cronbach's alphas. Most of them indicated good consistencies. Only the coefficients of the measures of evaluation and explanation were below .7. This replicates the pattern of internal consistencies found in Study 1. Table 5 further displays test-order effects at Time 1, computed as in Study 1. Significant effects occurred for five measures.

Convergent validities. We specified separate models for the two kinds of knowledge at the three measurement points, respectively. In each of the six models, the four respective measures were

Table 4
Means and Standard Deviations of the Eight Knowledge Measures Together With the Significances of Changes Over the Three Measurement Points

	Time 1		Time 2		Time 3	
Measure	*M*	*SD*	*M*	*SD*	*M*	*SD*
Hypothetical measures of conceptual knowledge						
Evaluation	60.2	21.0	66.8	21.6	76.2	18.8
Representation	51.2	16.9	69.6	19.1	69.9	17.8
Comparison	75.1	15.7	83.7	12.7	86.9	17.5
Explanation	16.5	20.6	21.6	23.2	31.9	25.9
Hypothetical measures of procedural knowledge						
Accuracy	56.8	21.3	88.1	16.3	86.3	17.5
Speed	3.9	2.8	2.7	0.8	2.6	0.8
Asymmetry I	3.8	1.4	2.8	0.8	2.6	0.7
Asymmetry II	3.4	1.3	3.0	0.8	2.7	0.8
Asymmetry	−0.4	1.1	0.2	0.7	0.1	0.6
Dual-task costs I	60.3	22.5	90.9	13.6	89.3	15.8
Dual-task costs II	51.9	24.9	85.2	19.5	85.4	19.7
Dual-task costs	15.3	19.3	7.0	14.7	5.4	12.6

Note. Speed and asymmetry measures are reported in seconds; all other measures are reported in percentages.

specified as loading on a latent factor. The fit indices of these models are given in Table 6. A comparative fit index (*CFI*) above .95, a weighted root mean square residual (*WRMR*) below 1, and a root mean square error of approximation (*RMSEA*) below 0.05 indicate a good fit (Hu & Bentler, 1999; Yu, 2002). This was the case for both kinds of knowledge at Times 1 and 2. For Time 3, one or two indices indicated a slightly nonoptimal fit of the models, whereas the others still indicated a good fit.

Each factor explained only 14% to 49% of the variance of its respective indicators. The absolute values of the factor loadings (see Table 7) ranged from .04 to .83. More than half of the factor loadings had an absolute value smaller than .60, which indicated insufficient convergent validities of our measures.

The proportions of explained variance were higher for conceptual knowledge (all ≥38%) than for procedural knowledge (all ≤28%). The loadings of the hypothetical measures of conceptual knowledge were significantly related to their latent factors, whereas the hypothetical measures of procedural knowledge only showed such relations at Time 2. This was mirrored by the factor loadings, which mostly were higher for conceptual knowledge than for procedural knowledge.

All significant factor loadings exhibited the expected signs, with the exception of speed at Time 2. This loading indicates that children with lower procedural knowledge also had lower solution times. We observed an explanation for this during the data collection: Children without the procedural knowledge necessary to solve the task quickly clicked at random positions, whereas children with more procedural knowledge needed more time to find and enter the correct answer. This also explains why the measure was not significantly related to the latent factor at Times 1 or 3. In all, the knowledge measures clearly were not mutually independent but exhibited insufficient convergent validities, particularly for procedural knowledge. All subsequently reported analyses should be considered exploratory and thus should be interpreted with caution, because our measures had insufficient validities with respect to measuring conceptual and procedural knowledge, contrary to their characterization in the introduction.

Divergent validities. We specified, for each measurement point, a one-factor model (i.e., all eight measures loading on a single latent factor) and a two-factor model (i.e., the hypothetical measures of each kind of knowledge loading on a latent factor, respectively, in which the two factors are allowed to correlate). As shown in Table 8, the fit indices of the one-factor model and the two-factor models indicate almost equally good fits at each measurement point. The reason for this consists in the high intercorrelations of the latent factors in the two-factor models: The Pear-

Table 5
Internal Consistencies and Test-Order Effects of the Knowledge Measures

	Cronbach's α			Test-order effects at Time 1	
Measure	Time 1	Time 2	Time 3	η^2	*p*
Hypothetical measures of conceptual knowledge					
Evaluation	.44	.54	.47	.08	.015
Representation	.70	.80	.76	.06	.069
Comparison	.77	.75	.80	.01	.901
Explanation	.52	.59	.54	.05	.158
Hypothetical measures of procedural knowledge					
Accuracy	.82	.85	.85	.09	.013
Speed				.02	.772
Asymmetry I				.10	.003
Asymmetry II				.05	.243
Asymmetry				.05	.244
Dual-task costs I	.92	.92	.93	.14	.000
Dual-task costs II	.93	.94	.95	.09	.012
Dual-task costs				.03	.454

Table 6
Fit Indices of the Models for the Analyses of the Convergent Validities

Measurement point	$\chi^2(1)$	*p*	*CFI*	*WRMR*	*RMSEA*	R^2
			Hypothetical measures of conceptual knowledge			
Time 1	0.09	.769	1.000	0.065	0.000	.38
Time 2	0.10	.753	1.000	0.073	0.000	.39
Time 3	1.70	.192	.997	0.198	0.058	.49
			Hypothetical measures of procedural knowledge			
Time 1	0.36	.551	1.000	0.222	0.000	.14
Time 2	1.12	.290	.997	0.269	0.024	.28
Time 3	2.43	.119	.933	0.456	0.083	.25

Note. *CFI* = comparative fit index; *WRMR* = weighted root mean square residual; *RMSEA* = root mean square error of approximation.

son correlations between the factors estimating conceptual knowledge and procedural knowledge were .93 at Time 1, .95 at Time 2, and .97 at Time 3. However, these correlations and the approximately equally good fits of the alternative models should not be overinterpreted. The latent factor values were estimated on grounds of measures that overlap only to a limited extent, making it hard to say what construct is actually reflected by each latent factor. The low convergent validities of the measures might, therefore, have led to an overestimation of the factor intercorrelations. This is further corroborated by the standardized covariances of our manifest measures given in Appendix B. All four measures of conceptual knowledge as well as accuracy, that is, one measure of procedural knowledge, were significantly intercorrelated, whereas the remaining three measures of procedural knowledge were almost independent of conceptual knowledge and of each other. This explains why the one-factor model fits the data as well as a two-factor model. For the time being, we must leave open the question of whether two kinds of knowledge or a single kind of knowledge were assessed by our measures.

Predictive relations. The longitudinal relations between the latent factors modeling conceptual and procedural knowledge could not be investigated because the validities were too low. To determine whether different pairs of hypothetical measures of conceptual and procedural knowledge lead to comparable results concerning the causal interrelations of the knowledge kinds (as should be the case if they have good validities), we specified cross-lagged panel models for all possible pairings of measures of conceptual and procedural knowledge. In each model, one hypothetical measure of conceptual knowledge and one hypothetical measure of procedural knowledge at Time 3 are regressed on the same two measures at Time 2, and these two measures are regressed on the same two measures at Time 1. At each measurement point, the measures are allowed to intercorrelate.

The results of the 16 models are displayed in Table 9. The results obtained with the different pairs of single measures were not homogeneous. For example, the absolute values of the correlations were all below .1 for evaluation and speed but were all above .5 for representation and accuracy. The intercorrelations of the single measures were generally much smaller than the intercorrelations of the latent factors. Some pairs of measures indicated a stronger influence of conceptual on procedural

Table 7
Standardized Factor Loadings of the Models for the Analyses of the Convergent Validity

	Time 1		Time 2		Time 3	
Measure	Standardized factor loading	*p*	Standardized factor loading	*p*	Standardized factor loading	*p*
			Hypothetical measures of conceptual knowledge			
Evaluation	.64	—[a]	.49	—[a]	.60	—[a]
Representation	.67	.000	.75	.000	.83	.000
Comparison	.68	.000	.77	.000	.77	.000
Explanation	.42	.000	.34	.000	.57	.000
			Hypothetical measures of procedural knowledge			
Accuracy	.56	—[a]	.62	—[a]	.66	—[a]
Speed	−.04	.829	.36	.014	.25	.078
Asymmetry	.44	.188	.54	.001	.44	.050
Dual-task costs	−.24	.125	−.54	.001	−.54	.051

[a] Coefficient could not be computed, because the metric of the latent factor was fixed to the metric of this indicator.

Table 8
Fit Indices of the Models for the Analysis of the Divergent Validities

Measurement point/Model	Maximum likelihood method estimated χ^2	*df*	*p*	*CFI*	*WRMR*	*RMSEA*	Akaike information criterion
Time 1							
One factor	39.66	25	.032	.949	1.528	0.053	4,430
Two factors	36.91	23	.033	.952	1.567	0.054	4,431
Time 2							
One factor	57.65	25	.000	.895	1.656	0.080	4,347
Two factors	59.46	23	.000	.882	1.645	0.088	4,350
Time 3							
One factor	46.41	25	.006	.940	1.519	0.065	4,238
Two factors	46.57	23	.003	.934	1.595	0.071	4,240

Note. *CFI* = comparative fit index; *WRMR* = weighted root mean square residual; *RMSEA* = root mean square error of approximation.

knowledge (e.g., evaluation and dual-task costs), others showed a stronger influence of procedural on conceptual knowledge (e.g., evaluation and accuracy), and still others indicated bidirectional influences (e.g., representation and accuracy).

Discussion

The results of Study 2 confirm and differentiate the findings from Study 1. Each latent factor explained less than 50% of the pooled variance of its indicators, showing that the single measures reflected the influences of other constructs to a stronger degree than they reflected conceptual or procedural knowledge. The latent factors for conceptual knowledge explained more variance than did the latent factors for procedural knowledge at all three measurement points. The factor structure of the hypothetical measures of procedural knowledge changed across measurement points. Only at Time 2, directly after the intervention, were all indicators significantly related to the latent factor. However, the loading of the speed measure on the factor had an unexpected sign at this measurement point. The low convergent validities of all measures are further reflected by the fact that, again, test-order effects occurred only for some of the measures and by the fact that the longitudinal relations between a single hypothetical measure of each kind of knowledge indicated qualitatively different causal interrelations of the measures.

General Discussion

The Insufficient Validity of Commonly Used Measures of Conceptual and Procedural Knowledge

Many studies have sought to investigate the developmental relations between conceptual and procedural knowledge. As

Table 9
Results of the Crossed-Lagged Panel Models With Pairs of Single Measures

Hypothetical measures of conceptual knowledge (C) and procedural knowledge (P)	Correlation coefficient *r*			Regression weight β			
	C_1 with P_1	C_2 with P_2	C_3 with P_3	P_2 on C_1	C_2 on P_1	P_3 on C_2	C_3 on P_2
Evaluation							
Accuracy	.42	.36	.54	.06	.23	.12	.27
Speed	−.06	.00	.01	−.07	−.02	.03	.02
Asymmetry	.06	.07	.23	.09	.04	.12	.12
Dual-task costs	−.12	−.23	−.36	−.26	.06	−.16	−.12
Representation							
Accuracy	.57	.63	.66	.21	.21	.25	.29
Speed	.01	.12	.13	.01	−.01	.06	.02
Asymmetry	.17	.31	.26	.09	.12	.24	.07
Dual-task costs	−.23	−.34	−.33	−.15	−.08	.03	−.16
Comparison							
Accuracy	.48	.64	.63	.14	.21	.13	.21
Speed	−.05	.15	−.02	.00	.05	.04	−.08
Asymmetry	.12	.17	.18	−.02	.10	.08	.05
Dual-task costs	−.21	−.30	−.22	−.16	−.13	−.08	.00
Explanation							
Accuracy	.28	.27	.51	.17	.14	.04	.30
Speed	−.12	−.08	.08	−.07	.01	−.01	.12
Asymmetry	.12	.19	.25	.15	.05	.20	.13
Dual-task costs	−.04	−.22	−.33	−.18	.01	−.11	−.20

Note. Subscripts indicate the measurement point.

shown by our results, these studies can be criticized for trying to investigate relations of conceptual and procedural knowledge before it is known how these constructs can be measured validly and at least partly independently of each other. Previous studies in the field postulated, rather than empirically tested, the validities of their measures.

For the first time in the literature, we discuss here how the validities of hypothetical measures of conceptual or procedural knowledge can be empirically tested. A multimethod approach helps to examine the convergent and divergent validities of assessments, thus disentangling measure-specific from measure-general variance components. Only variance components common to several qualitatively different hypothetical measures of the same kind of knowledge can be considered valid indicators of this kind of knowledge.

The results of our two empirical studies consistently indicated severe problems with the validities of measures commonly used to assess conceptual or procedural knowledge. Measures designed to assess the same kind of knowledge proved to be inhomogeneous: Only some of them were affected by a treatment (in Study 1). Moreover, the four measures shared only 14% to 49% common variance—depending on the kind of knowledge and measurement point. In the case of procedural knowledge, the assessments were not always significantly related to a commonly underlying latent factor. Pairs of a single hypothetical measure of each kind of knowledge had qualitatively different predictive interrelations (in Study 2). Test-order effects occurred selectively for only some of the measures supposedly assessing the same construct (in both studies). These findings demonstrate low convergent validities of the eight measures. This problem made it impossible for us to yield trustworthy estimates of the divergent validities and the longitudinal interrelations of the knowledge kinds in our study. In summation, the empirical results showed that previous research has critically underestimated the problem of measuring conceptual and procedural knowledge validly and partly independently of each other.

In their review of the findings concerning the interrelations of conceptual and procedural knowledge, Rittle-Johnson and Siegler (1998) pointed out that findings are inhomogeneous across studies and suggested domain differences between the observed phenomena as one cause. Our results suggest a second, additional cause: The choice of measures of conceptual and procedural knowledge determines the obtained empirical results. Clear evidence for this comes from the analyses of the longitudinal relations between our measures in Study 2. Depending on which of the 16 pairs of single measures of conceptual and procedural knowledge is chosen, the results support either the concepts-first view, the procedures-first view, the iterative model, or the independence view, although these viewpoints are mutually exclusive.

On a more general level, our results suggest that it is easier by far for children to build sound knowledge about decimal fractions than it is to build such knowledge about fractions with a numerator and a denominator (e.g. 3/5). Studies from many countries have found consistently over several decades that fractions with a numerator and denominator are very hard to understand for students and even some adults (e.g., Hope & Owens, 1987; Stafylidou & Vosniadou, 2004; U.S. Department of Education, 2008). In contrast, and consistent with these prior findings, the fifth and sixth graders in our study exhibited a relatively good understanding of decimal fractions. At the first measurement points of both studies, solution rates were already above the chance level, and children's knowledge further increased significantly after that. At the last measurement point of the longitudinal design in Study 2, the solution rates were as high as 87% for fraction comparison and 86% for number-line estimation. This constitutes further evidence for the hypothesis that the main problem in the development of students' knowledge about fractions is not the understanding of rational numbers or of the non-whole quantities to which they refer (Gallistel & Gelman, 2000; Mix, Levine, & Huttenlocher, 1999). Instead, the specific notational form of fractions with a numerator and a denominator seems to make it hard for students to see how both components together denote a single magnitude (cf. Ni & Zhou, 2005).

Issues for Future Research on Conceptual and Procedural Knowledge

A possible explanation for the low convergent validities found in our study is that our measures each tapped a different facet of children's knowledge about decimal fractions (e.g., understanding of place values, ordinal relations, density property of rational numbers). Children's knowledge stems from a variety of sources (everyday life experience, books, teachers, peers, self-explanations, etc.). Oftentimes, children fail to see how the pieces of conceptual or procedural knowledge acquired in these superficially different situations are connected on the level of scientific concepts and theories (diSessa, Gillespie, & Esterly, 2004; Schneider & Stern, 2009). Future studies will be needed to test the degree to which knowledge fragmentation accounts for the low intercorrelations between our measures.

Thus far, studies on the interrelations of conceptual and procedural knowledge have defined and treated both kinds of knowledge as homogeneous (i.e., one-dimensional) constructs. In the case where knowledge fragmentation is identified as cause of the low convergent validities, future studies should use more cautious definitions that acknowledge that both kinds of knowledge can be fragmented or integrated to different degrees.

The exact degree of the generalizability of our findings across studies and designs is not clear, because we conducted the first multimethod study on conceptual and procedural knowledge. Further multimethod studies should test the generalizability of our results across content domains, measures, and designs. For example, we used the number-line estimation task with decimal fractions to assess procedural knowledge. For a person who conceptually understands decimal fraction, however, this task is easy to solve. Future studies in procedurally more demanding content domains, for instance, multistep equation solving, could possibly find stronger dissociations between measures of conceptual or procedural knowledge.

There are also different ways to model data from multimethod studies. We selected some of the most basic models available. When future studies obtain data with a higher quality, they should start to capitalize on the advantages of more complex models. For instance, we modeled each measure as a sum score of its respective item scores. A more sophisticated alternative is to split the items of each measure into two groups and compute the sum score of each group. This would allow specification of a model with the pairs of sum scores as manifest variables, the measures as first-order factors, and the kinds of knowledge as second-order factors. In this way, unsystematic measurement error on each measure could be estimated separately from error variance due to low convergent validities of the measures.

Future studies should also try to optimize the measurement of procedural knowledge by controlling for children's different solution strategies instead of just registering solution times. Attempts should be made to gauge the number of strategies, strategy preferences, efficiency of strategy execution, and adaptation of strategy choices to problem characteristics separately (Lemaire & Siegler, 1995).

Finally, the multimethod approach requires that each person be tested with several measures at each measurement point. Thus, test-order effects and retest effects can increase error variance. Our results show that these effects are only significant for some measures and have only small to medium high effect sizes. We controlled for test-order effects by counterbalancing the order of assessments across participants. Future studies should seek to develop designs and apply statistical modeling techniques (e.g., Moses, Yang, & Wilson, 2007) that reduce or control for test-order and retest effects to an even greater extent than we did.

These points illustrate how complex and demanding multimethod studies generally are. For this reason, we do not argue that every future study on conceptual and procedural knowledge should be conducted as a multimethod study. In particular, cognitive processes in the range of seconds and below cannot be investigated by means of tests that take 45 min to complete. The most promising strategy for future research seems to be a separation of labor. Some large multimethod studies should seek to find valid measures of the kinds of knowledge, establish standards for testing, and perhaps even assist in the development of standardized tests. Other studies can then apply the measures with established validities for efficient in-depth investigations of the underlying cognitive processes.

Other areas of psychology have already greatly benefited from this division of labor. Psychometric studies with multiple measures have led to the construction of valid tests of intelligence or personality, which could then economically be used in subsequent studies. In their *Handbook of Multimethod Measurement in Psychology*, Eid and Diener (2006) provided a comprehensive review of the history, methodology, and examples of multimethod studies. Caroll (1993) reviewed and reanalyzed a huge number of factor-analytic studies of human abilities. These overviews show how much other areas of psychology have profited from a multimethod approach as well as the lack of such studies in research on conceptual and procedural knowledge.

Further multimethod studies are also important from the viewpoint of ecological validity. In experimental psychology, knowledge kinds are investigated under controlled laboratory conditions with measures specifically designed for the study. Notwithstanding the fact that this line of research has yielded fascinating results, it is unclear to what extent the obtained results are relevant for the explanation of learning processes in everyday life, for instance, in schools and universities, where a large number of factors influence what is learned and how it is learned. By focusing on broad patterns of knowledge, the multimethod approach might be more suitable than classical experimental approaches for investigating such real-life learning situations in field studies.

Our theoretical and methodological arguments do not apply only to conceptual and procedural knowledge but also to kinds of knowledge in general. Theorists have proposed numerous other pairs of knowledge kinds, for example, "knowing that" and "knowing how" (Ryle, 1949), competence and performance (Chomsky, 1965), structures and procedures of the mind (Inhelder & Piaget, 1980), declarative and procedural knowledge (Anderson, 1983), and explicit and implicit knowledge (Schacter, 1987). Researchers are far from understanding how these kinds of knowledge relate to each other and how they shape development. Valid empirical measures are an indispensable precondition for scientifically investigating these questions rather than merely speculating about them.

References

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. R. (1993). *Rules of the mind.* Hillsdale, NJ: Erlbaum.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglas, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111,* 1036–1060.

Anderson, J. R., & Fincham, J. M. (1994). Acquisition of procedural skills from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1322–1340.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 932–945.

Anderson, J. R., Taatgen, N., & Byrne, M. D. (2005). Learning to achieve perfect time sharing: Architectural implications of Hazeltine, Teague, and Ivry (2002). *Journal of Experimental Psychology: Human Perception and Performance, 31,* 749–761.

Baroody, A. J. (2003). The development of adaptive expertise and flexibility: The integration of conceptual and procedural knowledge. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise* (pp. 1–33). Mahwah, NJ: Erlbaum.

Baroody, A. J., & Dowker, A. (Eds.). (2003). *The development of arithmetic concepts and skills: Constructing adaptive expertise.* Mahwah, NJ: Erlbaum.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

Broaders, S. C., Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making children gesture brings out implicit knowledge and leads to learning. *Journal of Experimental Psychology: General, 136,* 539–550.

Bryant, P., Christie, C., & Rendu, A. (1999). Children's understanding of the inverse relation between addition and subtraction: Inversion, identity, and decomposition. *Journal of Experimental Child Psychology, 74,* 194–212.

Burkholder, G. J., & Harlow, L. L. (2003). An illustration of a longitudinal cross-lagged panel design for larger structural equation models. *Structural Equation Modeling, 10,* 465–486.

Byrnes, J. P., & Wasik, B. A. (1991). Role of conceptual knowledge in mathematical procedural learning. *Developmental Psychology, 27,* 777–786.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Canobi, K. H. (2004). Individual differences in children's addition and subtraction knowledge. *Cognitive Development, 19,* 81–93.

Canobi, K. H., Reeve, R. A., & Pattison, P. E. (1998). The role of conceptual understanding in children's addition problem solving. *Developmental Psychology, 34,* 882–891.

Caroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Case, R. (1992). *The mind's staircase: Exploring the conceptual underpinnings of children's thought and knowledge.* Hillsdale, NJ: Erlbaum.

Chomsky, N. (1965). *Aspects of a theory of syntax.* Cambridge, MA: MIT Press.

Cowan, R., & Renton, M. (1996). Do they know what they are doing? Children's use of economical addition strategies and knowledge of commutativity. *Educational Psychology, 16,* 407–420.

diSessa, A. A., Gillespie, N. M., & Esterly, J. B. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science, 28,* 843–900.

Dixon, J. A., & Moore, C. F. (1996). The development of intuitive principles in choosing mathematical strategies. *Developmental Psychology, 32,* 241–253.

Eid, M., & Diener, E. (Eds.). (2006). *Handbook of multimethod measurement in psychology.* Washington, DC: American Psychological Association.

Fuson, K. C. (1988). *Children's counting and concept of number.* New York: Springer.

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences, 4,* 59–65.

Geary, D. C. (1994). *Children's mathematical development: Research and practical implications.* Washington, DC: American Psychological Association.

Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number.* Cambridge, MA: Harvard University Press.

Gelman, R., & Meck, E. (1983). Preschoolers' counting: Principles before skill. *Cognition, 13,* 343–359.

Gelman, R., & Williams, E. M. (1998). Enabling constraints for cognitive development and learning: Domain specificity and epigenesis. In D. Kuhn & R. S. Siegler (Eds.), *Handbook of child psychology: Cognition, perception, and language* (5th ed., Vol. 2, pp. 575–630). New York: Wiley.

Gilmore, C. K., & Papadatou-Pastou, M. (2009). Patterns of individual differences in conceptual understanding and arithmetical skill: A meta-analysis. *Mathematical Thinking and Learning, 11,* 25–40.

Goldin-Meadow, S., Alibali, M. W., & Church, R. B. (1993). Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review, 100,* 279–297.

Goldstone, R. L., & Kersten, A. (2003). Concepts and categorization. In I. B. Weiner (Ed.), *Handbook of psychology* (Vol. 4, pp. 599–621). Hoboken, NJ: Wiley.

Greeno, J. G., Riley, M. S., & Gelman, R. (1984). Conceptual competence and children's counting. *Cognitive Psychology, 16,* 94–143.

Haapasalo, L., & Kadjievich, D. (2000). Two types of mathematical knowledge and their relation. *Journal für Mathematik-Didaktik, 21,* 139–157.

Halford, G. S. (1993). *Children's understanding: The development of mental models.* Hillsdale, NJ: Erlbaum.

Hecht, S. A., Close, L., & Santisi, M. (2003). Sources of individual differences in fraction skills. *Journal of Experimental Child Psychology, 86,* 277–302.

Hiebert, J. (Ed.). (1986). *Conceptual and procedural knowledge: The case of mathematics.* Hillsdale, NJ: Erlbaum.

Hiebert, J. (1992). Mathematical, cognitive, and instructional analyses of decimal fractions. In G. Leinhardt, R. Putnam & R. Hattrup (Eds.), *Analysis of mathematics for mathematics teaching* (pp. 283–322). Hillsdale, NJ: Erlbaum.

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 1–27). Hillsdale, NJ: Erlbaum.

Hiebert, J., & Wearne, D. (1986). Procedures over concepts: The acquisition of decimal number knowledge. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 199–223). Hillsdale, NJ: Erlbaum.

Hiebert, J., & Wearne, D. (1996). Instruction, understanding, and skill in multidigit addition and subtraction. *Cognition and Instruction, 14,* 251–283.

Hope, J. A., & Owens, D. T. (1987). An analysis of the difficulty of learning fractions. *Focus on Learning Problems in Mathematics, 9,* 25–40.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Inhelder, B., & Piaget, J. (1980). Procedures and structures. In D. R. Olson (Ed.), *The social foundations of language and thought* (pp. 19–27). New York: Norton.

Johnson, A. (2003). Procedural memory and skill acquisition. In I. B. Weiner (Ed.), *Handbook of psychology* (pp. 499–523). Hoboken, NJ: Wiley.

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science* (2nd ed.). Cambridge, MA: MIT Press.

LeFevre, J.-A., Smith-Chant, B. L., Fast, L., Skwarchuk, S.-L., Sargla, E., Arnup, J. S., et al. (2006). What counts as knowing? The development of conceptual and procedural knowledge of counting from kindergarten through Grade 2. *Journal of Experimental Child Psychology, 93,* 285–303.

Lemaire, P., & Siegler, R. S. (1995). Four aspects of strategic change: Contributions to children's learning of multiplication. *Journal of Experimental Psychology: General, 124,* 83–97.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492–527.

Mabbott, D. J., & Bisanz, J. (2008). Computational skills, working memory, and conceptual knowledge in older children with mathematics learning disabilities. *Journal of Learning Disabilities, 41,* 15–28.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44,* 1469–1481.

Mix, K. S., Levine, S. C., & Huttenlocher, J. (1999). Early fraction calculation ability. *Developmental Psychology, 35,* 164–174.

Moses, T., Yang, W.-L., & Wilson, C. (2007). Using kernel equating to assess item order effects on test scores. *Journal of Educational Measurement, 44,* 157–178.

Muthén, L. K., & Muthén, B. O. (1998–2007). *MPlus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

National Assessment Governing Board & U.S. Department of Education. (2003). *Mathematics framework for the 2003 national assessment of educational progress.* Retrieved June 1, 2008, from http://www.nagb.org/pubs/math_framework/toc.html

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics: Standards 2000.* Reston, VA: National Council of Teachers of Mathematics.

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research, 39,* 439–478.

Newman, R. S., & Berger, C. F. (1984). Children's numerical estimation: Flexibility in the use of counting. *Journal of Educational Psychology, 76,* 55–64.

Ni, Y., & Zhou, Y.-D. (2005). Teaching and learning fraction and rational numbers: The origins and implications of the whole number bias. *Educational Psychologist, 40,* 27–52.

OECD. (1999). *Measuring student knowledge and skills: A new framework for* *assessment.* Retrieved June 1, 2008, from http://www.pisa.oecd.org/document/4/0,2340,en_32252351_32236159_33668932_1_1_1_1,00.html

Pennington, N., Nicolich, R., & Rahm, J. (1995). Transfer of training between cognitive subskills: Is knowledge really use specific? *Cognitive Psychology, 28,* 175–224.

Petitto, A. L. (1990). Development of numberline and measurement concepts. *Cognition and Instruction, 7,* 55–78.

Piaget, J. (1978). *Success and understanding.* Cambridge, MA: Harvard University Press.

Rabinowitz, M., & Goldberg, N. (1995). Evaluating the structure–process hypothesis. In F. Weinert & W. Schneider (Eds.), *Memory performance and competencies: Issues in growth and development* (pp. 225–242). Hillsdale, NJ: Erlbaum.

Resnick, L. B. (1982). Syntax and semantics in learning to subtract. In T. P. Carpenter, J. M. Moser & T. A. Romburg (Eds.), *Addition and subtraction: A cognitive perspective* (pp. 136–155). Hillsdale, NJ: Erlbaum.

Resnick, L. B., & Ford, W. W. (1981). *The psychology of mathematics for instruction.* Hillsdale, NJ: Erlbaum.

Resnick, L. B., Nesher, P., Leonard, F., Magone, M., Omanson, S., & Peled, I. (1989). Conceptual bases of arithmetic errors: The case of decimal fractions. *Journal for Research in Mathematics Education, 20,* 8–27.

Resnick, L. B., & Omanson, S. F. (1987). Learning to understand arithmetic. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 3, pp. 41–95). Hillsdale, NJ: Erlbaum.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General, 126,* 288–311.

Ritter, F. E., & Schooler, L. J. (2001). The learning curve. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences* (pp. 8602–8605). Amsterdam: Pergamon Press.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology, 91,* 175–189.

Rittle-Johnson, B., & Koedinger, K. R. (2005). Designing knowledge scaffolds to support mathematical problem solving. *Cognition and Instruction, 23,* 313–349.

Rittle-Johnson, B., & Siegler, R. S. (1998). The relation between conceptual and procedural knowledge in learning mathematics: A review. In C. Donlan (Ed.), *The development of mathematical skills* (pp. 75–328). Hove, United Kingdom: Psychology Press.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93,* 346–362.

Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology, 99,* 561–574.

Ryle, G. (1949). *The concept of mind.* London: Hutchinson House.

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review, 115,* 101–130.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis: A Festschrift for Heinz Neudecker* (pp. 233–247). London: Kluwer.

Satorra, A., & Bentler, P. M. (1999). A scaled difference chi-square test statistic for moment structure analysis. Retrieved February 1, 2005, from http://ssrn.com/abstract=199064

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 501–518.

Schneider, M., & Stern, E. (in press). The inverse relation of addition and subtraction: A knowledge integration perspective. *Mathematical Thinking and Learning.*

Schumacher, E. H., Seymour, T. L., Glass, J. M., Kieras, D. E., & Meyer, D. E. (2001). Virtually perfect time sharing in dual-task performance: Uncorking the central attentional bottleneck. *Psychological Science, 121,* 101–108.

Senatsverwaltung für Bildung Jugend und Sport Berlin et al. (2004). *Rahmenlehrplan Grundschule: Mathematik* [Elementary school curriculum: Mathematics]. Retrieved June 1, 2008, from http://www.berlin.de/sen/bildung/schulorganisation/lehrplaene/

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review, 14,* 47–69.

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking.* New York: Oxford University Press.

Siegler, R. S., & Crowley, K. (1994). Constraints on learning in nonprivileged domains. *Cognitive Psychology, 27,* 194–226.

Siegler, R. S., & Opfer, J. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science, 14,* 237–243.

Siegler, R. S., & Stern, E. (1998). Conscious and unconscious strategy discoveries: A microgenetic analysis. *Journal of Experimental Psychology: General, 127,* 377–397.

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill.* Cambridge, MA: Harvard University Press.

Sophian, C. (1997). Beyond competence: The significance of performance for conceptual development. *Cognitive Development, 12,* 281–303.

Spelke, E. S. (2000). Core knowledge. *American Psychologist, 55,* 1233–1243.

Stafylidou, S., & Vosniadou, S. (2004). The development of students' understanding of the numerical value of fractions. *Learning and Instruction, 14,* 503–518.

Torbeyns, J., Verschaffel, L., & Ghesquiére, P. (2005). Simple addition strategies in a first-grade class with multiple strategy instruction. *Cognition and Instruction, 23,* 1–21.

U.S. Department of Education. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington, DC: Author.

Vamvakoussi, X., & Vosniadou, S. (2004). Understanding the structure of the set of rational numbers: A conceptual change approach. *Learning and Instruction, 14,* 453–467.

VanLehn, K. (1986). Arithmetic procedures are induces from examples. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics.* Hillsdale, NJ: Erlbaum.

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction, 4,* 45–69.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Unpublished doctoral dissertation, University of California, Los Angeles, CA.

Received August 8, 2008
Revision received April 27, 2009
Accepted May 5, 2009 ■