RESEARCH ARTICLE

How Should Intelligent Tutoring Systems Sequence Multiple Graphical Representations of Fractions? A Multi-Methods Study

M. A. Rau · V. Aleven · N. Rummel · Z. Pardos

Abstract Providing learners with multiple representations of learning content has been shown to enhance learning outcomes. When multiple representations are presented across consecutive problems, we have to decide in what sequence to present them. Prior research has demonstrated that interleaving *tasks types* (as opposed to blocking them) can foster learning. Do the same advantages apply to interleaving *representations*? We addressed this question using a variety of research methods. First, we conducted a classroom experiment with an intelligent tutoring system for fractions. We compared four practice schedules of multiple graphical representations: blocked, fully interleaved, moderately interleaved, and increasingly interleaved. Based on data from 230 4th and 5th-grade students, we found that interleaved practice leads to better learning outcomes than blocked practice on a number of measures. Second, we conducted a think-aloud study to gain insights into the learning mechanisms underlying the advantage of interleaved practice. Results show that students make connections between representations only when explicitly prompted to do so (and not spontaneously). This finding suggests that reactivation, rather than abstraction, is the main mechanism to account for the advantage of interleaved practice. Third, we used methods derived from Bayesian knowledge tracing to analyze tutor log data from the classroom experiment. Modeling latent

M. A. Rau (✉)
Department of Educational Psychology, University of Wisconsin, Madison, WI, USA
e-mail: marau@wisc.edu

V. Aleven
Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburg, PA, USA

N. Rummel
Institute of Education, Ruhr-Universität Bochum, Bochum, Germany

Z. Pardos
Schools of Education and Information, University of California, Berkeley, CA, USA

measures of students' learning rates, we find higher learning rates for interleaved practice than for blocked practice. This finding extends prior research on practice schedules, which shows that interleaved practice (compared to blocked practice) impairs students' problem-solving performance during the practice phase when using raw performance measures such as error rates. Our findings have implications for the design of multi-representational learning materials and for research on adaptive practice schedules in intelligent tutoring systems.

Introduction

In this paper, we present research conducted with an intelligent tutoring system for fractions learning that focuses on the effects of practice schedules of multiple graphical representations on students' learning. The question of how to sequence multiple graphical representations is interesting for several reasons. First, designers of educational materials frequently face this question, as many domains employ multiple graphical representations across consecutive problems. Second, learning sciences research on the sequence of task types (e.g., of addition and multiplication problems) has frequently demonstrated advantages of interleaved sequences, compared to blocked sequences (e.g., Bahrick et al. 1993; Schmidt and Bjork 1992; de Croock et al. 1998). Yet it remains unclear whether the advantage of interleaving task types also applies to interleaving graphical representations. Thus, the question of how best to interleave multiple graphical representations is also of theoretical interest. Third, problem sequences are highly relevant to intelligent tutoring systems, for example because problem selection methods based on cognitive models often vary in terms of the degree to which they block or interleave problem types (e.g., Koedinger et al. 2011; Koedinger et al. 2013). Finally, conducting research on the effects of sequencing graphical representations in the context of intelligent tutoring systems is particularly interesting because they offer novel opportunities to implement adaptive sequences of multiple graphical representations–adapting, for instance, to the prior knowledge level of a given learner. Although we do not explicitly investigate adaptive sequences in the present paper, our findings provide insights into the potential benefits of adaptively sequencing multiple graphical representations. Thus, the research presented in this paper is of practical and theoretical relevance.

In conducting this research, we combined a number of methodologies. First, we conducted a classroom experiment with the Fractions Tutor, a classroom-proven an intelligent tutoring system for 4th-and 5th-grade fractions learning that we developed as platform for research into the use of multiple graphical representations (Rau et al. 2013c; Rau et al. 2012a). Next, to gain insights into the learning mechanisms that account for the findings of the classroom experiment, we analyzed verbal data from a think-aloud study that we conducted with the version of the Fractions Tutor that was shown to be most effective in the classroom experiment. Finally, we used Bayesian knowledge tracing to analyze the data obtained from the classroom experiment to further clarify open questions about the learning mechanisms. In sum, these different methods complement one another: each method serves the goal to clarify the insights gained by another.

Multiple Graphical Representations of Fractions

In this paper, we present results from a multi-methods program of research to investigate how best to temporally sequence multiple graphical representations within the Fractions Tutor. Graphical representations of learning contents are used not only in fractions instruction, but in many other areas of mathematics (Common Core State Standards Initiative 2010; Kilpatrick et al. 2001; NCTM 1989, 2000, 2006; NMAP 2008) and science (Kozma 2003; National Research Council 2002). Multiple representations are considered to enhance learning in part because different representations emphasize complementary conceptual aspects of the learning material and have differential effects on mental processing (Cox 1999; Cromley et al. 2010; Eilam 2013; Gagatsis and Elia 2004; Gegenfurtner et al. 2011; Goldman 2003; Hinze et al. 2013; Kozma et al. 2000; Larkin and Simon 1987; Reed and Ettinger 1987; Schnotz and Bannert 2003; Schwartz and Black 1996a; Tabachneck et al. 1997; Zhang 1997; Zhang and Norman 1994). Theoretical accounts highlight several beneficial functions of the use of multiple representations in educational materials (Ainsworth 2006; Scaife and Rogers 1996), such as computational offloading (reducing cognitive effort), re-representing (highlighting complementary conceptual aspects), and graphical constraining (mutually constraining interpretations).

In spite of the well-documented promise of learning with multiple representations, it is widely recognized that multiple representations (compared to single representations) do not necessarily enhance student learning. Research on multiple representations shows that they only enhance learning when students adequately understand each individual representation (Ainsworth 2006; Eilam 2013), and when they make connections between representations (Ainsworth 2006; de Jong et al. 1998; Gobert et al. 2011; Gutwill et al. 1999; Özgün-Koca 2008; Rathmell and Leutzinger 1991; Superfine et al. 2009; Uttal 2003; van der Meij 2007). Therefore, one of the challenges curriculum designers face with regard to designing multi-representational learning environments is how best to support students in learning from multiple graphical representations. Unfortunately, we still know little about how best to implement multiple graphical representations in instructional materials, let alone how best to take advantage of the specific opportunities intelligent tutoring systems offer to enhance students' learning with multiple graphical representations.

Fractions are one of many areas in mathematics where multiple graphical representations are used extensively (NMAP; NCTM 2000, 2006). As in many other science and mathematics domains, instructional materials of fractions typically employ different graphical representations. Typically used graphical representations include circle diagrams, rectangles, and number lines (see Fig. 1). Each graphical representation emphasizes a slightly different conceptual viewpoint on fractions (Charalambous and Pitta-Pantazi 2007), as discussed below.
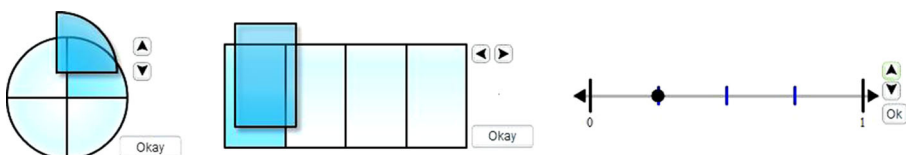
Fig. 1 Interactive circle, rectangle, and number line representations, as used in the fractions tutor

Practice Schedules

When designing instruction that uses multiple graphical representations, curriculum designers must decide how to temporally sequence the different graphical representations. How frequently should the curriculum alternate between graphical representations? Practice schedules are likely to have an impact on students' robust learning of the domain knowledge. In line with Koedinger and colleagues (2012), we define robust learning as the acquisition of knowledge that transfers to novel tasks and lasts over time. In creating practice schedules that involve multiple graphical representations, it matters, most likely, whether the different representations are practiced in a "blocked" manner or are interleaved with practice of other representations. Although instructional materials employ multiple graphical representations in many different ways, including side-by-side use of different representations, our current study focuses on the special case in which each tutor activity involves a single graphical representation, in addition to text and symbols. (As discussed below, this study is a step in a broader program of research that also looks at side-by-side use of representations). Thus, in a blocked practice schedule, consecutive tutor problems may for example involve the following sequence of representations: circle–circle–circle–number line–number line–number line. By contrast, an interleaved sequence might look as follows: circle–number line–circle–number line–circle–number line. Blocked schedules allow students to gain in-depth experience with one graphical representation before switching to a new representation, and may thus enhance students' understanding of individual representations. Interleaved schedules, on the other hand, provide frequent opportunities to compare different graphical representations to one another (every time the student switches from one representation to the other), thus allowing students to make connections between different representations. Research shows that interleaved practice schedules lead to better long-term retention and transfer than blocked practice in a variety of domains including vocabulary learning (Bahrick et al. 1993; Cepeda et al. 2006), motor tasks (Hebert et al. 1996; Immink and Wright 1998; Li and Wright 2000; Meiran 1996; Meiran et al. 2000; Ollis et al. 2005; Schmidt and Bjork 1992; Shea and Morgan 1979; Simon and Bjork 2001), algebra (Rohrer 2008; Rohrer and Taylor 2007; Taylor and Rohrer 2010), troubleshooting (de Croock et al. 1998; Van Merriënboer et al. 2002), and decision-making tasks (Helsdingen et al. 2011). However, interleaved practice schedules often result in lower performance during the acquisition phase (i.e., while students practice).

A limitation of this research is that it has exclusively focused on practice schedules of different task types (for instance, addition–addition–addition–multiplication–multiplication–multiplication, versus addition–multiplication–addition–multiplication–addition–multiplication). Task types differ in terms of the problem-solving procedure (or action sequence, in the case of motor tasks; memory trace in the case of fact retrieval or vocabulary learning) a given problem involves, whereas graphical representations differ in terms of the concepts they invoke. In our prior research (Rau et al. 2013a) we compared interleaving of task types to interleaving of representations (i.e., interleaving task types while blocking graphical representations versus interleaving graphical representations while blocking task types). We found that when the choice is to interleave one while blocking the other, the choice should be made to interleave task types. However, in practice, it is often not necessary to interleave one but block the other; the decision to

interleave can often be made separately for each dimension. Therefore, in the current research, we drop the constraint that only one dimension can be interleaved. Consistent with our prior work, we look at situations in which task types are (moderately) interleaved, and ask what level of interleaving of representations is ideal, in combination with moderately interleaving task types.

The advantage of interleaved practice has been attributed to two kinds of processes that play a role in deep, cognitive processing of the learning material (Rau et al. 2013a). First, interleaved practice schedules require learners to frequently reactivate the knowledge needed to solve each learning task (de Croock et al. 1998; Lee and Magill 1983, 1985): when tasks are presented in an interleaved sequence, the required knowledge has to be retrieved more frequently from long-term memory–it cannot be kept in memory from one task to the next if there is not overlap in requisite knowledge. Retrieval from long-term memory strengthens the association between cues and associated elements in long-term memory, and increases the likelihood that this knowledge can be recalled later on (Anderson 1993; Anderson 2002). Second, interleaving may help students abstract knowledge across different learning tasks (de Croock et al. 1998; Shea and Morgan 1979). When knowledge needed for different learning tasks is simultaneously active in working memory, students can compare the knowledge relevant to the respective learning tasks. While this process may happen consciously or unconsciously, it helps learners to see which task properties are key and which are incidental, thereby directing their attention to aspects relevant to knowledge construction (Bannert 2002; Paas and van Gog 2006; van Merriënboer et al. 2002).

What might these learning processes correspond to when interleaving graphical representations in particular? Frequently switching between different graphical representations may require students to frequently reactivate representation-specific knowledge, such as knowledge regarding the specific conceptual aspects emphasized by the particular graphical representation at hand. Repeated reactivation of representation-specific knowledge may thereby support the ease with which students can retrieve knowledge about individual graphical representations: students may become more fluent at using representation-specific knowledge. Frequently switching between graphical representations may also provide students with more opportunities to make connections between corresponding elements of the different graphical representations, for instance, by relating the numerator presented in a circle to the numerator presented in a number line. This connection-making process might help students to abstract from the different graphical representations a more generic understanding of fractions – regardless of the graphical representation they are depicted as. We note that it reactivation and abstraction processes are not necessarily mutually exclusive: to a certain extent, they may both be active at the same time. Yet, it remains an open question whether one of the two learning processes is more likely than the other to account for the benefits of interleaved practice.

As mentioned, the present study is part of a broader program of research on the use of multiple graphical representations in learning with an intelligent tutoring system. Our prior research showed that multiple graphical representations lead to better learning than a single graphical representation (Rau et al. 2009, 2013c). We further showed that when a choice has to be made to interleave either task types or graphical representations, we should interleave task types (Rau et al. 2013a). However, our prior work did not address the question of whether interleaving graphical

representations of fractions leads to better learning than blocking graphical representations. Another open question regards the learning processes that account for the advantage of interleaved practice: is reactivation or abstraction the most likely mechanism by which interleaved practice leads to better learning than blocked practice? We address these open questions in this paper.

The Fractions Tutor

We conducted our research in the context of the Fractions Tutor: a successful intelligent tutoring system for fractions that uses multiple, interactive graphical representations (e.g., Rau et al. 2013b, c). We have used the Fractions Tutor as a platform in several experiments for investigating research questions about how to use multiple graphical representations to promote robust learning. This research has led to a number of instructional design principles for the use of multiple graphical representations, summarized elsewhere (Rau et al. 2013b, c). The Fractions Tutor has been iteratively updated based on the outcomes of these studies and as a result, embodies these design principles.

The Fractions Tutor is an example-tracing tutor (Aleven et al. 2009a), a type of Cognitive Tutor (Koedinger and Corbett 2006). Cognitive Tutors have a proven track record in improving students' mathematics achievement (Koedinger and Corbett 2006). Example-tracing tutors are behaviorally similar to Cognitive Tutors, meaning that they provide step-by-step guidance in the form of feedback and on-demand hints. In contrast to Cognitive Tutors, example-tracing tutors rely on generalized examples of correct and incorrect solution paths rather than on a rule-based cognitive model of student behavior. We created the Fractions Tutor with the Cognitive Tutor Authoring Tools (CTAT; Aleven et al. 2009a), designing tutor interfaces separately for each problem type and representation. The design of the interfaces and of the interactions students engage in during problem-solving are based on a number of small-scale user studies that we conducted in our laboratory, on prior classroom experiments (see Rau et al. 2013b for an overview), as well as on Cognitive Task Analysis of the learning domain (Baker et al. 2007; Clark et al. 2007). Furthermore, an experienced mathematics teacher was involved in developing the tutor problems. Across the different classroom experiments, we have iteratively updated and improved the Fractions Tutor based on our findings. The Fractions Tutor covers a comprehensive set of supplementary instructional materials ranging from fraction identification to equivalent fractions and addition and is available to students and teachers on a free website (https://fractions.cs.cmu.edu; Aleven et al. 2009b).

Several other intelligent tutoring systems have been developed that support fractions learning but the Fractions Tutor appears to be unique in that it focuses on conceptual learning with *multiple*, *interactive*, *abstract* graphical representations. Like many other intelligent tutoring systems, the Fractions Tutor includes interactive, virtual manipulatives (Moyer et al. 2002). Research has demonstrated that students can benefit from using virtual manipulatives of fractions (Reimer and Moyer 2005) and that virtual manipulatives can be at least as effective in supporting students' learning as physical manipulatives (Suh et al. 2005). ASSISTments, a system for middle-school math (e.g., Heffernan et al. 2012), includes fractions, but focuses on procedural tasks such as adding fractions. ActiveMath is an intelligent tutoring

system that supports self-regulated learning based on a constructivist approach (e.g., Goguadze et al. 2008). Although ActiveMath includes graphical representations, students do not manipulate them directly. Rather, changes in the representations reflect students' interactions with symbolic fractions. Kong and Kwok (2003) describe an intelligent tutoring system that heavily relies on rectangle representations, but it does not include other graphical representations. In the MFD system ("Mixed numbers, Fractions, and Decimals"; Beck et al. 1997; Arroyo et al. 1999), students interact with various concrete representations of fractions (e.g., sets of dogs and animals, buttons to measure lengths), but it does not include abstract graphical representations. In addition, there are several other interactive learning environments that use multiple, interactive, abstract graphical representations (e.g., Akpinar and Hartley 1996; Reimer and Moyer 2005), but these are not intelligent tutoring systems.

The Fractions Tutor includes several *abstract* and *interactive* graphical representations: circle diagrams, rectangles, and number lines (Fig. 1). Each graphical representation emphasizes certain aspects of different conceptual interpretations of fractions (Charalambous and Pitta-Pantazi 2007). The circle as a part-whole representation depicts fractions as parts of an area that is partitioned into equally-sized pieces. The rectangle is a more elaborate part-whole representation as it can be partitioned vertically and horizontally. At the same time, it does not have a standard shape for the unit, like the circle does. Finally, the number line is considered a measurement representation and thus emphasizes that fractions can be compared in terms of their magnitude, and that they fall between whole numbers. We chose abstract graphical representations based on the notion that they lead to more transferable knowledge because the representation is not tied to a specific scenario (e.g., pizza sharing) (Goldstone et al. 2003; Smith 2003). In addition, abstract representations may be advantageous because they facilitate interpretations of a situation in terms of abstract relations rather than specific attributes (Resnick and Omanson 1987; Schwartz and Black 1996a, b). However, to promote students' understanding of graphical representations based on their prior real-world experiences (e.g., Grady 1998; Heim 2000; Nisbett and Ross 1980), we introduce the abstract graphical representations within real-world contexts and concrete representations (e.g., pizzas, chocolate bars). Thus, our approach to using abstract graphical representations while introducing them with concrete graphical representations corresponds to Goldstone and Son's (2005) approach of "concreteness fading", which was shown to be successful in an experimental study. In our own classroom studies, we found that students enjoy a version of the Fractions Tutor more if it includes problems that introduce the abstract graphical representations in the context of realistic scenarios (Rau et al. 2013b).

The way in which the Fractions Tutor supports students' *interactions* with the graphical representations is based on extensive reviews of education standards (e.g., NCTM 1989, 2006, 2008), interviews and focus groups with teachers, and several iterations of classroom experiments and lab-based user studies (Rau et al. 2013b). The Fractions Tutor includes a variety of multi-step problems and provides step-by-step guidance. It supports various ways for students to interact with the graphical representations: by clicking on fraction

pieces to highlight or select them, by dragging and dropping fraction pieces, and through buttons to change the partitioning of the graphical representations.

The Fractions Tutor covers a comprehensive set of task types including interpreting graphical representations, reconstructing the unit of fraction representations, and improper fractions, described briefly in Table 1. The tutoring system takes a conceptually-focused approach in introducing fractions, as detailed in Table 1. A common theme throughout the Fractions Tutor is the unit of the fraction (i.e., what the fraction is taken of). The concept of the unit is being introduced in the first task types, and revisited during the later task types as students learn about improper fractions. Figure 2 shows an example of a problem in which students make circle representations for two given symbolic fractions and are then prompted to reflect on the relative size of the two fractions. The concept of the unit lays the foundation for introducing improper fractions, by demonstrating that fractions can be larger than one unit (i.e., 1 1/2 is one unit plus 1/2 of that unit).

The present version of the Fractions Tutor builds on our prior research in two important ways. First, each problem included conceptually oriented prompts (see Fig. 2) to help students relate the multiple graphical representations to the symbolic notation of fractions. We found these prompts to be effective in an earlier experimental study (Rau et al. 2009). Second, the Fractions Tutor moderately interleaves task types, building on our earlier finding that interleaving task types leads to better learning than blocking task types (Rau et al. 2013a). We use this moderately interleaved sequence of task types consistently across the different sequences of multiple graphical representations contrasted in the present experiment.

The Fractions Tutor has been demonstrated to lead to significant learning gains across several classroom experiments with over 3,000 students in grades 4–5 (e.g.,

Table 1 Description of task types covered by the fractions tutor

Task type #	Task type description	Description of example task
1	Naming unit fractions and proper fractions	Determine what fraction (unit fractions and proper fractions) a given graphical representation shows (using circle diagrams, rectangles, and number lines)
2	Making representations of unit fractions and proper fractions	Construct a graphical representation for a fraction given symbolically (using circle diagrams, rectangles, and number lines)
3	Reconstructing the unit I	Given a unit fraction, reconstruct the unit of the fraction (using circle diagrams, rectangles, and number lines)
4	Reconstructing the unit II	Given a proper fraction, (a) find the unit fraction, and (b) reconstruct the unit of the fraction (using circle diagrams, rectangles, and number lines)
5	Naming improper fractions	Determine what improper fraction a given graphical representation shows (using circle diagrams, rectangles, and number lines)
6	Making representations of improper fractions	Construct a graphical representation for a fraction given symbolically (using circle diagrams, rectangles, and number lines)

Fig. 2 Making a circle given a symbolic fraction, combined with prompts to compare the two fractions. Reflection prompts are implemented with drop-down menus, shown in the bottom half of each problem

Rau et al. 2012, b). In our most recent classroom experiment with 599 4th- and 5th-graders (this experiment took place after the experiment reported in the current paper–see Rau et al. 2012a), we found that the Fractions Tutor substantially improved students' knowledge of fractions. After 10 h of instruction with the Fractions Tutor, students improved significantly with a medium effect size of $d=0.40$ at the posttest ($p<0.01$). When we administered a delayed posttest a week later, we found that students retained these learning gains with an effect size of $d=0.60$ ($p< 0.01$). These are pre/post effect sizes when the Fractions Tutor was used as supplemental instruction, after the regular fractions instruction had been completed.

Classroom Experiment: Effects of Practice Schedules

The goal of the classroom experiment (also see Rau et al. 2012b) was to evaluate the effect of different practice schedules of graphical representations on students' learning of fractions. We contrasted four conditions that differed only in the degree to which, in the tutor's problem sets, the graphical representations where blocked or interleaved. In accordance with the results from our earlier experiment (Rau et al. 2013a), we consistently used a moderately interleaved practice schedule of task types across all conditions. Furthermore, our goal was to investigate students' learning gains from working with the Fractions Tutor by comparing their performance on equivalent pretests, posttests, and delayed posttests.

Research Hypotheses

Specifically, we contrasted four conditions, all of which worked with multiple graphical representations, but differed with regard to the practice schedule according

to which graphical representations were sequenced. A *blocked practice schedule* switched infrequently between the representations. A *fully interleaved practice schedule* switched maximally frequently between the representations. A *moderately interleaved schedule* of multiple graphical representations switched between representations after every couple of problems. Finally, an *increasingly interleaved schedule* of multiple graphical representations gradually moved from a blocked schedule to a more and more interleaved schedule.

In line with prior research on interleaved practice, we expect that students learn with all four practice schedules but that interleaving graphical representations supports more robust learning than the other schedules, through two possible mechanisms: interleaving may allow students to abstract across multiple graphical representations and to frequently reactivate their knowledge about fractions representations and fractions concepts. Our specific hypotheses are:

Hypothesis 1:   Students significantly improve from pretest to posttest on all measures of robust learning, namely, reproduction with area models, reproduction with number lines, transfer of conceptual knowledge, and transfer of procedural knowledge.

Hypothesis 2:   Students who learn with multiple graphical representations presented in an interleaved fashion will outperform students who learn with multiple graphical representations presented in a blocked fashion on all measures of robust knowledge.

Methods

Experimental Design

Figure 3 illustrates the practice schedules of task types and graphical representations for the four multiple graphical representations conditions. In all conditions, students worked through the same sequence of task types and fraction problems, and switched task types after every six of a total of 108 problems. Each task type was visited three times. We randomly assigned students to one of four conditions. Students in each condition worked with multiple graphical representations, presented according to different practice schedules. In the *blocked* condition, students switched graphical representations after 36 problems. In the *moderate* condition, students switched representations after every six problems (initially offset by three problems so as to not switch representations at the same time as task types[1]). In the *fully interleaved* condition, students switched representations after each problem. In the *increased* condition, the length of the blocks was gradually reduced from twelve problems at the beginning (initially offset by nine problems) to a single problem at the end. To account for possible effects of the order of graphical representations, we randomized the order in which students encountered the graphical representations.

---

[1] We chose not to switch graphical representations at the same time as task types (except, by virtue of its nature, in the fully interleaved condition) so as to make it easier for students to see correspondences between different task types when they use the same graphical representation.
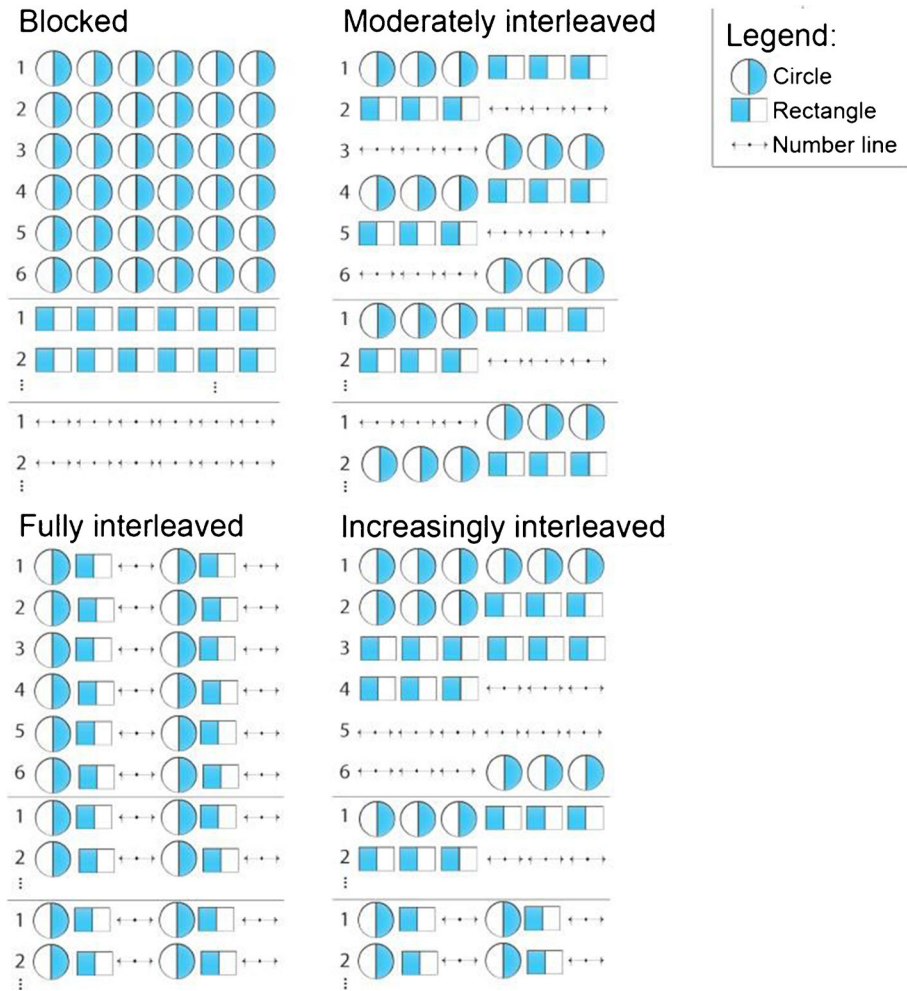
Fig. 3 Practice schedules for the multiple graphical representations conditions. In all conditions, six task types were presented three times. Numbers 1–6 indicate task types, shapes depict representations

Participants

A total of 474 4th- and 5th-grade students from six different schools (31 classes) participated in the study during their regular mathematics instruction. The schools' rankings in the academic year of 2009/2010 were in the top 10 % of 2468 Pennsylvania public schools.[2] In the school year of 2009/2010, 10–30 % of all students in the participating school districts were enrolled in free or reduced-price lunch programs, over 90 % of all students were white, less than 5 % African American. Students were aged 8 to 11 years. All schools were located in Western Pennsylvania.

We excluded students who missed at least one test day, and who completed less than 67 % of all tutor problems. We had to apply this stringent criterion to ensure that

---

[2] The precise numbers are withheld to preserve anonymity of the participating schools.

students in the blocked condition encountered all three graphical representations (see Fig. 1). This results in a total of $N=230$ ($n=63$ in blocked, $n=53$ in moderate, $n=52$ in fully interleaved, $n=62$ in increased).

Experimental Procedure

Prior to working with the Fractions Tutor, students completed a pretest. The pretest took about 30 min. On the following day, all students started working with the Fractions Tutor. Students accessed the Fractions Tutor from the computer lab at their schools and worked with it for about 5 h as part of their regular math instruction for five to six consecutive school days (depending on the length of the respective school's class periods). All students worked on the Fractions Tutor at their own pace, but the time students spent with the system was held constant across classrooms and across experimental conditions. On the day following the tutoring sessions, students completed the immediate posttest, which took about 30 min. Seven days after the posttest, students completed an equivalent delayed posttest.

Test Instruments

We assessed students' knowledge of fractions at three test times using three equivalent test forms. We randomized the order in which they were administered. The tests included four knowledge types: reproduction with area models (i.e., circles and rectangles), reproduction with number lines, conceptual transfer and procedural transfer. The area model items and number line items covered identifying fractions given a graphical representation, making a graphical representation given a symbolic fraction, and recreating the unit given a graphical representation of both unit fractions and proper fractions. Conceptual transfer items included proportional reasoning questions with and without graphical representations. Procedural transfer items included comparison questions with and without graphical representations. The theoretical structure of the test (i.e., the four knowledge types just mentioned) resulted from a factor analysis performed on the pretest data. All test scales included items adapted from standardized state assessments. The test scales reproduction with area models and reproduction with number lines constitute reproduction items: the test items closely relate to the knowledge covered in the Fractions Tutor. Creating separate scales for area models and number lines seemed reasonable given that number lines are believed to be more challenging for students than area models (Cramer et al. 2008; NMAP 2008).

Results

As mentioned, we analyzed the data of $N=230$ students. There was no significant difference between conditions with respect to the number of students excluded ($\chi^2<1$). There were no significant differences between conditions at pretest for any dependent measure, $ps>0.10$. There was no significant effect for order of multiple graphical representations within the intervention conditions for any dependent measure, $F(5, 285)=1.56$, $ps>0.10$.

    We used a hierarchical linear model (HLM, see Raudenbush and Bryk 2002) with four nested levels to analyze the data in order to take into account for nested sources

of variance, due to the fact that a student's performance can be partially explained by his/her class and school. At level 1, we modeled performance on each of the tests for each student. At level 2, we accounted for differences between students. Level 3 models random differences between classes, and level 4 random differences between schools. The HLM is the outcome of a forwards-inclusion procedure in which we used the Bayesian Information Criterion (BIC) to find whether the inclusion of a variable increased model fit. If the BIC decreased as a consequence of including a variable (indicating better model fit), we kept the variable. If the BIC did not decrease, we did not include the variable. We tested a number of variables, including teacher, sequence of graphical representations, test form sequence, grade level, number of problems completed, total time spent with the tutor, random intercepts and slows for classes and schools. Equation 1 shows the resulting HLM:

$$Y_{ijkl} = \left(\left(\left(\mu + W_l\right) + V_{kl}\right) + \beta_3 * c_j + \beta_4 * p_j + \beta_5 * c_j * p_j + U_{jkl}\right) + \beta_1 * t_i + \beta_2 * c_j * t_i + R_{ijkl}$$

(1)

with

(level 1) $Y_{ijkl} = \varepsilon_{jkl} + \beta_1 * t_i + \beta_2 * c_j * t_i + R_{ijkl}$
(level 2) $\varepsilon_{jkl} = \delta_{kl} + \beta_3 * c_j + \beta_4 * p_j + \beta_5 * c_j * p_j + U_{jkl}$
(level 3) $\delta_{kl} = \gamma_l + V_{kl}$
(level 4) $\gamma_l = \mu + W_l$

with the index i standing for posttest time (i.e., immediate and delayed posttest), j for the student, k for class, and l for the school. The dependent variable Y_{ijkl} is student$_j$'s score on the dependent measures at posttest time t_i (i.e., immediate or delayed posttest), ε_{jkl} is the parameter for the intercept for student$_j$'s score, β_1 is the parameter for the effect of posttest time t_i, β_2 is the effect of the interaction of condition c_j with posttest time t_i, β_3 is the parameter for the effect of condition c_j, β_4 is the parameter for the effect of student$_j$'s performance on the pretest p_j, β_5 is the parameter for an aptitude-treatment interaction between condition c_j and student$_j$'s performance on pretest p_j, δ_{kl} is the parameter for the random intercept for class$_k$, γ_l is the parameter for the random intercept for school$_l$, and μ is the overall average.

Since the HLM described in (1) uses students' pretest scores as a covariate, it does not allow us to analyze whether students in the various conditions improved from pretest to immediate and delayed posttest. To analyze learning gains, we included pretest score in the dependent variable, yielding:

$$Y_{ijkl} = \left(\left(\left(\mu + W_l\right) + V_{kl}\right) + \beta_3 * c_j + U_{jkl}\right) + \beta_1 * t_i + \beta_2 * c_j * t_i + R_{ijkl}$$

(2)

with

(level 1) $Y_{ijkl} = \varepsilon_{jkl} + \beta_1 * t_i + \beta_2 * c_j * t_i + R_{ijkl}$
(level 2) $\varepsilon_{jkl} = \delta_{kl} + \beta_3 * c_j + U_{jkl}$
(level 3) $\delta_{kl} = \gamma_l + V_{kl}$
(level 4) $\gamma_l = \mu + W_l$

with the index i standing for test time (i.e., pretest, immediate, and delayed posttest). The dependent variable Y_{ijkl} is student$_j$'s score on the dependent measures at test time

t_i (i.e., pretest, immediate posttest, or delayed posttest). Excluded from formula (1) were the parameters β_4 for the effect of student$_j$'s performance on the pretest p_j, and the parameter β_5 is for an aptitude-treatment interaction between condition c_j and student$_j$'s performance on pretest p_j.

We used planned contrasts and post-hoc comparisons to clarify results from the HLM analysis, all of which were computed as part of the HLM to clarify results from the HLM analysis. All reported p-values were adjusted using the Bonferroni correction for multiple comparisons. Table 2 shows the means and standard deviations for the dependent measures by condition and test time.

Learning Effects

To investigate hypothesis 1 (that all students significantly improve from pretest to posttest on all measures of robust learning), we analyzed learning gains using the HLM described in formula (2) (which uses pretest as a dependent measure). The main effect of test time (i.e., pretest, immediate posttest, and delayed posttest) was significant for reproduction with number lines, $F(2, 867)=20.09$, $p<0.01$, partial $\eta^2=0.03$, for reproduction with area models, $F(2, 867)=17.54$, $p<0.01$, partial $\eta^2=0.02$, conceptual transfer, $F(2, 867)=38.78$, $p<0.01$, partial $\eta^2=0.03$, and marginally significant for procedural transfer, $F(2, 867)=2.84$, $p<0.10$, partial $\eta^2=0.01$. The interaction between test time and condition was significant for reproduction with area models $F(12, 862)=2.06$, $p<0.05$, partial $\eta^2=0.01$. These results show that students (regardless of condition) benefited from working with the Fractions Tutor on reproduction with number lines, reproduction with area models, procedural and conceptual transfer. On reproduction with area models, students' learning gains depended on the condition.

Table 2 Means and standard deviations (in parentheses) for dependent measures at pretest, immediate posttest, delayed posttest by condition

		Reproduction with area models	Reproduction with number lines	Conceptual transfer	Procedural transfer
Pretest	Blocked	0.56 (0.26)	0.45 (0.28)	0.60 (0.30)	0.51 (0.34)
	Moderately interleaved	0.64 (0.25)	0.50 (0.27)	0.70 (0.27)	0.51 (0.31)
	Fully interleaved	0.59 (0.31)	0.46 (0.25)	0.71 (0.22)	0.60 (0.36)
	Increasingly interleaved	0.64 (0.25)	0.48 (0.26)	0.69 (0.25)	0.55 (0.36)
Immediate posttest	Blocked	0.62 (0.25)	0.51 (0.31)	0.72 (0.30)	0.52 (0.37)
	Moderately interleaved	0.64 (0.26)	0.54 (0.25)	0.77 (0.24)	0.55 (0.36)
	Fully interleaved	0.71 (0.24)	0.58 (0.25)	0.78 (0.23)	0.60 (0.34)
	Increasingly interleaved	0.73 (0.22)	0.58 (0.26)	0.75 (0.27)	0.53 (0.37)
Delayed posttest	Blocked	0.69 (0.25)	0.57 (0.30)	0.71 (0.31)	0.60 (0.37)
	Moderately interleaved	0.71 (0.20)	0.62 (0.25)	0.78 (0.26)	0.65 (0.32)
	Fully interleaved	0.69 (0.25)	0.64 (0.25)	0.83 (0.19)	0.56 (0.37)
	Increasingly interleaved	0.77 (0.20)	0.60 (0.28)	0.77 (0.25)	0.55 (0.32)

Differences between Practice Schedules

To investigate hypothesis 2 (that students who learn with multiple graphical representations presented in an interleaved fashion will outperform students who learn with multiple graphical representations presented in a blocked fashion on all measures of robust knowledge), we computed the HLM presented in formula (1) for the intervention conditions (using pretest as a covariate). There was no significant main effect of practice schedules on any knowledge type, indicating that there was no global effect of practice schedules across immediate and delayed posttests. An interaction between posttest time and condition was marginally significant for reproduction with area models, $F(3, 867)=$ 2.57, $p<0.10$, partial $\eta^2=0.01$, so that the effect of condition was possibly stronger on the immediate posttest than on the delayed posttest, suggesting that the effect of practice schedules on reproduction with area models may be somewhat temporary. The interaction between pretest score and condition was marginally significant for conceptual transfer, $F(3, 219)=2.52$, $p<0.10$, partial $\eta^2=0.02$, suggesting that students with different pretest scores benefit from different practice schedules.

To clarify the interaction between posttest time and condition, we used post-hoc contrasts separately for the immediate and the delayed posttest. To limit the number of comparisons, we only compared the most successful practice schedule against the remaining three practice schedules taken together, as summarized in Table 3. We found some support for a benefit of interleaving multiple graphical representations: the fully interleaved condition significantly outperformed the not-fully-interleaved conditions (i.e., blocked, moderately interleaved, and increasingly interleaved) on conceptual transfer at the delayed posttest. Furthermore, we found a marginally significant advantage for the increasingly interleaved condition over the not-increasingly-interleaved conditions (i.e., blocked, moderately interleaved, and fully interleaved) on reproduction with area models at the immediate and the delayed posttests.

To clarify the interaction between pretest score and condition on conceptual transfer, we computed post-hoc comparisons for students with extremely low or high pretest scores. For students with a pretest score of 15 %, 20 %, and 25 %, we found a

Table 3 Results from post-hoc comparisons on differences between multiple representations conditions at immediate posttest (post) and delayed posttest (delayed) by type of knowledge. "ns" indicates non-significant differences. "–" indicates that no post-hoc comparisons were computed

Effect	Test	Reproduction-area models	Reproduction-number lines	Conceptual transfer	Procedural transfer
Fully interleaved>blocked, moderately interleaved, increasingly interleaved	post	–	ns	ns	–
	delayed	–	ns	$p<0.05$, $d=0.33$	–
Increasingly interleaved > blocked, moderately interleaved, fully interleaved	post	$p<0.10$, $d=0.30$	–	–	–
	delayed	$p<0.10$, $d=0.30$	–	–	–
Moderately interleaved > blocked, fully interleaved, increasingly interleaved	post	–	–	–	ns
	delayed	–	–	–	ns

significant advantage for the fully interleaved over the blocked condition ($ps<0.05$). We found no differences for high prior knowledge students.

As an alternative test for hypothesis 2, we used post-hoc comparisons within the HLM described in formula (2) in order to investigate whether students' learning gains differ between conditions (using pretest as a dependent variable). Specifically, we computed post-hoc comparisons that contrasted students' scores at the immediate posttest and the delayed posttest, compared to the pretest. Tables 4 and 5 provides a summary of these post-hoc comparisons. Generally, we found significant learning gains at the delayed posttest for most conditions on reproduction with area models, reproduction with number lines, and conceptual transfer. On procedural transfer, only the moderate condition showed significant learning gains at the delayed posttest. The learning gains are most consistent for the fully interleaved condition: we found significant learning gains on all measures but procedural transfer at the immediate and delayed posttest.

Discussion

The results from the classroom experiment are generally in line with hypothesis 1 (that students significantly improve from pretest to posttest on all measures of robust learning). We found that students across conditions significantly improved from pretest to posttest on reproduction with number lines, area models, and on conceptual transfer, albeit with small effect sizes. Altogether, the learning gains were most consistent for the fully interleaved condition. As Tables 4 and 5 illustrates, the analysis of learning gains by condition shows learning gains for most conditions on reproduction with area models, reproduction with number lines, and conceptual transfer. However, only the moderately interleaved condition showed significant gains on procedural transfer, albeit only at the delayed posttest. The lack of learning gains on procedural transfer may reflect the fact that the Fractions Tutor focuses on conceptual learning of fractions more so than on procedural learning. Thus, altogether, we can conclude that students learn from the Fractions Tutor, especially when they work with the fully interleaved version.

Table 4 Improvement of test scores at immediate posttest (post) over pretest (pre) and delayed posttest (delayed) over pretest by knowledge types and conditions. "ns" indicates non-significant differences

Condition	Effect	Reproduction with area models	Reproduction with number lines	Conceptual transfer	Procedural transfer
Blocked	post>pre	ns	ns	$p<0.05$, $d=0.42$	ns
	delayed>pre	$p<0.05$, $d=0.52$	$p<0.01$, $d=0.39$	$p<0.05$, $d=0.39$	ns
Moderately interleaved	post>pre	ns	ns	$p<0.05$, $d=0.29$	ns
	delayed>pre	ns	$p<0.01$, $d=0.50$	$p<0.05$, $d=0.30$	$p<0.05$, $d=0.45$
Fully interleaved	post>pre	$p<0.05$, $d=0.45$	$p<0.01$, $d=0.51$	$p<0.01$, $d=0.34$	ns
	delayed>pre	$p<0.05$, $d=0.38$	$p<0.01$, $d=0.75$	$p<0.01$, $d=0.60$	ns
Increasingly interleaved	post>pre	$p<0.05$, $d=0.38$	$p<0.01$, $d=0.43$	ns	ns
	delayed>pre	$p<0.05$, $d=0.55$	$p<0.01$, $d=0.46$	ns	ns

Table 5 Number of surface connections and conceptual connections by implicit and explicit prompts averaged across students

	Implicit prompts	Explicit prompts	Overall
Surface	4.17	2.33	2.94
Conceptual	0.58	1.63	1.28
Overall	2.38	1.98	

The results provide qualified support for hypothesis 2 (that students who learn with multiple graphical representations presented in an interleaved fashion will outperform students who learn with multiple graphical representations presented in a blocked fashion on all measures of robust knowledge). We found a significant advantage of the fully interleaved condition (compared to the other conditions, see Table 3) only on conceptual transfer at the delayed posttest. We found a marginally significant advantage of the increasingly interleaved condition (compared to the other conditions, see Table 3) only on reproduction with area models. Yet, there was a significant interaction of condition with pretest, so that the fully interleaved condition showed significantly better performance on the posttests than the blocked condition for students with low prior knowledge. This finding was consistent regardless of which cut-off value was used to identify low prior knowledge students. Further support for hypothesis 2 comes from the analysis of learning gains by condition. As Table 3 illustrates, only the fully interleaved condition shows consistent learning gains on all dependent measures (except for procedural transfer, on which we found no learning gains, with one exception). Furthermore, the blocked condition never outperformed any of the interleaved conditions (see Tables 2 and 3). Thus, we can carefully conclude that there is an advantage of interleaving graphical representations over blocking them, especially for students with low prior knowledge.

The finding that the effect of practice schedules on students' learning outcomes depends on their prior knowledge is particularly interesting. While students with low prior knowledge benefit from fully interleaved practice, we found no effect of practice schedules for students with high prior knowledge. This finding might indicate that students with high prior knowledge are equipped to abstract across different graphical representations even when they are presented across a longer period of the learning sequence (as in the blocked condition). They might also have less of a need to frequently reactivate knowledge about the specific representations and the conceptual aspects they highlight because this type of knowledge is more accessible to them than to low prior knowledge students.

The finding that the increasingly interleaved condition (which gradually moves from a blocked sequence to a more and more interleaved sequence) is most effective on reproduction with area models but not on reproduction with number lines might be attributed to the relative difficulty of number lines, compared to area models. Area models are considered to be relatively intuitive and easy to understand (Cramer 2001; Lamon 1999), whereas number lines tend to be more difficult and less intuitive (Siegler et al. 2010; NMAP 2008). To a very limited extent, our finding thus supports the notion that allowing students to gain in-depth experience with one representation before introducing another representation (i.e., increasingly interleaved practice)

helps students improve their understanding of a graphical representation that is easy to learn. Early in the learning sequence, students might benefit from a blocked schedule because it allows them to apply one graphical representation across a sequence of different task types. This procedure might allow students to gain deeper understanding of the graphical representation. However, keeping a blocked practice schedule across the entire learning sequence (as in the blocked condition) does not enhance students' learning. A blocked practice schedule is only effective in the early learning sequence, provided that later on, students switch increasingly frequently between representations. That procedure may help students to consolidate their understanding of area models by allowing them to reactivate their understanding of area models frequently, every time they switch to a new representation. How might we explain that increasingly interleaved practice does not lead to an advantage (compared to other conditions) in learning about the number line? Students tend to have little prior knowledge about number lines (Siegler et al. 2010; NMAP 2008). It is possible that practice schedules do not have an impact on students' learning of a more difficult graphical representation. It is also possible that a different pace of moving from a blocked to a more and more interleaved practice schedule might have been more successful than the practice schedules we implemented. In fact, given that the effect of practice schedules appears to depend on students' prior knowledge, and the difficulty of a graphical representation, it is possible that students may have benefited more from a schedule that moves less rapidly from a more blocked to an increasingly interleaved schedule.

Although our interpretations regarding how the effectiveness of different practice schedules relates to students' prior knowledge and the particular target knowledge (i.e., conceptual transfer versus reproduction with area models and number lines) are speculative, they highlight an open question that is particularly interesting with respect to intelligent tutoring systems. If indeed, the effectiveness of a given practice schedule depends on a learner's level of prior knowledge and the type of target knowledge, intelligent tutoring systems might be used to take advantage of these effects. To take into account the hypothesized interaction between practice schedule and prior knowledge, the intelligent tutoring system might (1) initially select a practice schedule that is appropriate for the given student's level of prior knowledge, and (2) monitor the student's acquisition of knowledge throughout the learning process to adapt the practice schedule accordingly. To take into account the hypothesized interaction between practice schedule and type of target knowledge, the intelligent tutoring system might (3) initially prioritize on a particularly important type of knowledge, such as conceptual knowledge, to select the appropriate practice schedule, and (4) use mastery learning to detect when the student has mastered that target knowledge, to switch to a different practice schedule that is more appropriate to a secondary type of target knowledge (e.g., reproduction with area models). Future research should investigate a potential three-way interaction between practice schedules, prior knowledge (perhaps even speed of learning), and type of target knowledge, as well as implications for the use of adaptive practice schedules in intelligent tutoring systems.

In conclusion, the findings from the classroom experiment provide (albeit limited) support for the notion that instructional materials should provide interleaved practice with multiple graphical representations in order to promote students' robust learning, in particular if the goal is to promote the acquisition of conceptual knowledge that

transfers to novel tasks. We further found that the multiple-representations version of the Fractions Tutor leads to significant learning gains (in particular when multiple graphical representations are presented according to an interleaved practice schedule) on most measures of robust learning, and that these learning gains persist over at least 1 week after students' work with the tutoring system.

Think-Aloud Study: Underlying Mechanisms

To gain insights into the cognitive processes underlying the advantage of interleaved practice schedule (as identified in the classroom experiment), we additionally conducted a small think-aloud study (also see Rau et al. 2012b). The goal of the think-aloud study was to investigate the role of specific mechanisms that underlie the advantage of interleaved practice, namely, whether repeated reactivation or abstraction are more likely to account for the advantage of the interleaved practice schedule. As mentioned, one possible mechanism is repeated reactivation (de Croock et al. 1998; Lee and Magill 1983; Sweller 1990), which might help students to become more fluent in using representation-specific knowledge. Another possible mechanism is abstraction (de Croock et al. 1998; Shea and Morgan 1979), which might help students to make connections between graphical representations when they are presented in an interleaved fashion.

In order to gain further insight into these cognitive processes underlying the benefits of an interleaved practice schedule, we conducted a small-scale think-aloud study with six students who worked on the fully interleaved version of the Fractions Tutor. The fully interleaved condition was selected for this analysis because it was the most successful condition for two of four measures (see Table 2), and because the learning gains were most consistent for the fully interleaved condition (see Table 3). The goal of the think-aloud study was to gather information that might help us distinguish between the two alternative explanations just described. Thus, we wanted to investigate what kinds of spontaneous connections students make between graphical representations when working with the interleaved version of the Fractions Tutor, and whether students who fail to make spontaneous comparisons can be prompted to do so. If the mechanism underlying the advantage of interleaved practice consists mainly in abstraction of fractions knowledge across multiple graphical representations, we would expect to see evidence of spontaneous connection making. If, however, the main mechanism is repeated reactivation of representation-specific knowledge, we may not expect students to make many spontaneous connections between graphical representations. We also investigated whether students are able to make connections between consecutively presented graphical representations when prompted to do so.

Methods

Six 5th-grade students participated in the think-aloud study. The think-aloud study was conducted in our laboratory and included three sessions. During the first session, students took the same pretest that was used in the classroom experiment reported above. The pretest took about 30 min to complete. During the second session, students worked for 1 h on a subset of problems taken from the interleaved version

of the tutoring system while being prompted to think aloud, following the procedure described in Ericsson and Simon (1984). In the third session, students worked with similar tutor problems for 1 h while being prompted to relate the different graphical representations to one another. We varied the type of prompts based on a within-subjects design: the prompt questions were either implicit (i.e., without directly prompting comparisons between the representations; e.g. "How is this problem the same as the last two you did?" or "How is this problem different from the last one you did?"), or explicit (i.e., directly referring to aspects that the different representations share; e.g., "What is the unit in the circle / rectangle / number line?" or "How are the rectangle and the circle and the number line the same / different?"). All students received two implicit prompts and four explicit prompts, in a fixed sequence.

Students' utterances were recorded and transcribed. We combined top-down and bottom-up approaches in developing a coding scheme: the experimenters identified types of connections that students might make prior to the think-aloud study, and then refined the coding scheme after viewing the transcripts from the think-aloud study. Connections between graphical representations were coded as surface connections if they either referred to the color of the representation, the shape of the representation, or the action performed on the representation (e.g., dragging and dropping). For example, when asked "how is the circle like the rectangle?" a student's response "you have to drag something into a diagram of the unit" would be coded as a surface connection. Connections were coded as conceptual if they referred to the corresponding features of the representations (i.e., numerator, denominator, unit), or the magnitude represented. For instance, when asked: "how is the number line like the circle?" for improper fractions, a student's answer "they both have one whole unit plus a fraction of another unit that's the same" would be coded as a conceptual connection.

Results

The results from the pretest indicate that all students had a good understanding of fractions. During the spontaneous comparison phase of the think-aloud study, we found only five instances of connections. These five connections were uttered by five of the six students. All five connections were surface connections. In the prompted session, we found 138 instances of prompted connection making. Tables 4 and 5 summarizes the average number of connections coded as surface and conceptual connections for implicit and explicit prompts. Given the small number of students, a statistical test on the types of connections in response to implicit and explicit prompts is not warranted. Tables 4 and 5 suggests, however, that students generated substantially more surface connections than conceptual connections. We can also see that the implicit prompts yielded most of the surface connections, but almost none of the conceptual connections. Explicit prompts seem to have yielded more of the conceptual connections, compared to the implicit prompts.

Discussion

The observations from the think-aloud study show that students tend not to spontaneously make connections between multiple graphical representations: we found only five spontaneous connections, and all of them were surface connections. However,

students are able to make these connections when prompted to do so. In particular, explicit prompts are well-suited to enhance conceptual connections.

It is important not over-interpret the generalizability of these observations, as the think-aloud study was conducted with only a small number of students. Yet, our results do not provide any indication that the advantage of interleaved practice might stem from spontaneous connection-making activities between multiple graphical representations. Thus, it seems that students' benefit from interleaved practice with multiple graphical representations does not stem from conscious abstraction across the different representations. Rather, interleaved practice may be attributed to requiring students to repeatedly reactivate knowledge about the specific graphical representations. The fact that students were able to make connections when prompted to do so demonstrates that the lack of spontaneous connection-making activities is not an artifact of the think-aloud method being an unsuitable metric for detecting students' connection-making processes.

Furthermore, the observation that students generate a substantial number of conceptual comparisons between the graphical representations when explicitly prompted suggests that students might benefit from receiving such explicit prompts as part of a future version of the Fractions Tutor. Indeed, the literature on learning with multiple representations demonstrates the importance of making connections between multiple representations (Ainsworth 2006; Cook et al. 2007; Even 1998; Gutwill et al. 1999; Özgün-Koca 2008; Plötzner et al. 2001; Plötzner et al. 2008; Schnotz and Bannert 2003; Schwonke et al. 2008; Schwonke and Renkl 2010).

Bayesian Knowledge Tracing: Differences during the Acquisition Phase

Another goal of our research was to investigate whether we can detect advantages of interleaved practice using data obtained while students practiced with the Fractions Tutor in the classroom experiment (i.e., acquisition-phase data). Analyzing student performance during the acquisition phase (i.e., while students learn) is particularly interesting when investigating the effects of practice schedules: a common finding is that interleaved practice schedules lead to better long-term retention and to better transfer than blocked schedules, but they often lead to worse performance during the acquisition phase (Battig 1972; de Croock et al. 1998; Helsdingen et al. 2011; Pashler et al. 2007; Rohrer and Taylor 2007; Schmidt and Bjork 1992; Schneider 1985; Simon and Bjork 2001; Van Merriënboer et al. 2002). Therefore, it is often believed that the advantage of interleaved practice over blocked practice is not apparent during the acquisition phase, but can only be detected with long-term retention tests and transfer tests administered *after* the acquisition phase. However, it may be that educational data mining techniques focused on *latent* student variables *during* the acquisition phase may have something to offer over previous investigations, none of which used such techniques, to the best of our knowledge.

We use Bayesian knowledge tracing (Corbett and Anderson 1995) based on the tutor log data to investigate whether "machine-learned" learning rate estimates constitute a more suitable metric to detect the effects of practice schedules on students' learning during the acquisition phase (also see Rau and Pardos 2012). Knowledge tracing tracks student knowledge over time using a two state Hidden Markov Model assumption of learning. It uses correct and incorrect responses in students' problem-solving attempts to infer the probability of a student knowing the skill underlying the problem-solving step

at hand. This method has been used to investigate learning differences between conditions during the acquisition phase (Pardos et al. 2011).

More specifically, we investigate whether learning rate estimates, based on knowledge tracing, can detect the advantage of interleaved practice even during the acquisition phase. We chose Bayesian knowledge tracing and not Performance Factors Analysis (which is often used to predict students' performance in intelligent tutoring systems research and educational data mining; Pavlik et al. 2009), because the latter focuses on prediction of performance and does not include the notion of ability change over time. In other words, Performance Factors Analysis would not allow us to model our variable of interest: learning rates. To summarize, our analysis investigates whether knowledge tracing provides a suitable metric for detecting the effects of an intervention that is known not be accessible through simpler metrics such as performance during the acquisition phase.
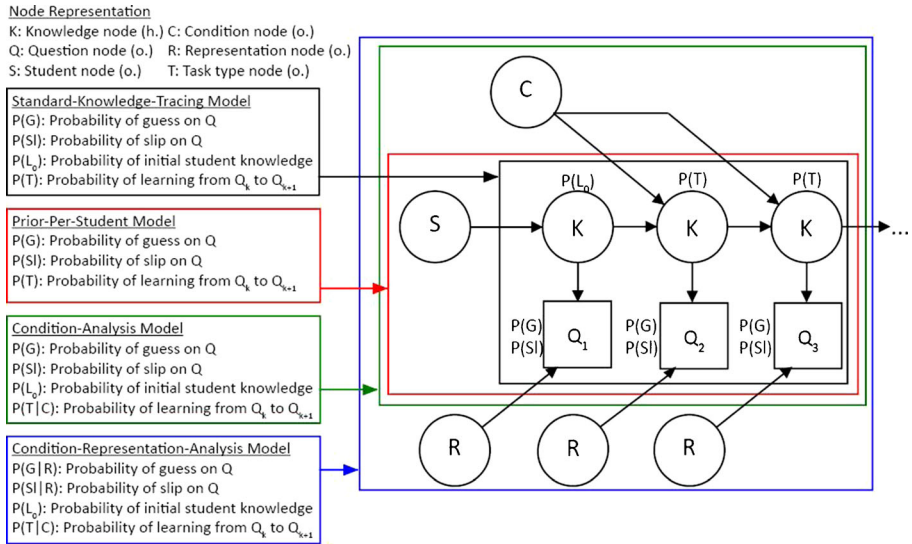
Bayesian Models

We combined our Bayesian model with several other extensions to knowledge tracing to each of the four conditions of the experimental study to investigate differences in estimated learning rates between the conditions in the Fractions Tutor. Specifically, we evaluated four Bayesian models based on the Fractions Tutor log data. Two of the models were created for the purpose of analyzing the learning rates of the conditions in the experiment while the other two were used as baseline models to gauge the relative predictive performance of the new models. None of the tested models included a knowledge component model, so each step in the tutor is treated as a knowledge component.

Learning Analysis Models

We employed two models that served as benchmarks for model fit and designed two novel models for evaluating learning differences among the experiment conditions. We compared the resulting four Bayesian models all of which were based around knowledge tracing. Figure 4 provides an overview of the different models that we compared. The Standard-Knowledge-Tracing model and the Prior-Per-Student model correspond to our two benchmark models. The *Standard-Knowledge-Tracing model* includes only knowledge tracing without taking students' prior knowledge (S) (Pardos and Heffernan 2010), experimental condition (C), or fraction representation (R) into account. The *Prior-Per-Student model* (Pardos and Heffernan 2010) includes the individual students' prior knowledge (S). Both the Standard-Knowledge-Tracking model and the Prior-Per-Student model assume that there is a probability that a student will transition from the unlearned to the learned knowledge state at each opportunity regardless of the particular problem just encountered or practice schedule of the student.

The Condition-Analysis model and the Condition-Representation-Analysis model are analogous to hypothesis 2 of the classroom experiment described above, namely, that condition (i.e., different practice schedules are a significant predictor of students' learning rates). Specifically, we hypothesize that within each given task type (described in Table 1), the fully interleaved condition will show higher learning rates than the blocked condition. Thus, we depart from the simplifying assumption of a single learning rate per skill and instead fit a separate learning rate for each of the four practice schedules implemented in the Fractions Tutor. To do so, we adapted

Fig. 4 Overview of the four different Bayesian Networks tested, with observed (o.) and hidden (h.) nodes

modeling techniques from prior work that evaluated the learning value of different forms of tutoring in (non-experiment) log data of an intelligent tutor (Pardos et al. 2010). Furthermore, we use techniques from KT-IDEM (Pardos and Heffernan 2011) to model different guess and slips for problems depending on the representation used in the tutor problem. This procedure allows us to estimate four different learning rates per task type, each corresponding to the particular condition (i.e., blocked practice, fully interleaved, moderately interleaved, or increasingly interleaved) assigned to the student–as opposed to using a single learning rate per task type, independent of condition. The *Condition-Analysis model* includes students' prior knowledge and models the effect of experimental condition (C). In addition, the *Condition-Representation-Analysis model* takes into account that different representations of fractions are expected to result in different degrees of difficulty in solving the tutor problem (Charalambous and Pitta-Pantazi 2007). Thus, the Condition-Representation-Analysis model incorporates students' prior knowledge (S), condition (C), and the graphical representation encountered by each student in each problem (R).

Model Fitting Procedure

In order to determine model fit by task type, we analyzed the log data by task type. For the evaluation of predictive performance, reported in the next section, a 5-fold cross-validation at the student level was used. For the reporting of learning rates by practice schedule, all data was used to train the model.

The parameters in all four models were fit using the Expectation Maximization algorithm implemented in Kevin Murphy's Bayes Net Toolbox (Murphy 2001). For the Condition-Representation-Analysis Model the number of parameters fit per task was 12 (2 prior+4 learn rate+3 guess+3 slip). Probabilities of knowledge are fixed at 1 if the skill was already known, $P(L_{n-1})=1$, to represent a zero chance of forgetting, an assumption made in standard knowledge tracing. If a student was previously (at

learning opportunity $n–1$) in the unlearned state, the probability that he/she will now (at opportunity n) have transitioned to the learned state is:

$$P(L_n) = P(L_{n-1}) + ((1-P(L_{n-1}))*P(T|C_S)), \tag{3}$$

where $P(L_{n-1})$ is the probability of a student already knowing the skill, is the condition assigned to a student (i.e., blocked, fully interleaved, moderately interleaved, increasingly interleaved), and T is the given task type (see Table 1).

Evaluation Results

To evaluate the predictive accuracy of each of the student models mentioned above, we conducted a 5-fold cross-validation at the student level. By cross-validating at the student level we can have greater confidence that the resulting models and their assumptions about learning will generalize to new groups of students. The metrics used to evaluate the model are root mean squared error (RMSE) and area under the curve (AUC). Lower RMSE equals better prediction accuracy. For AUC, a score of 0.50 represents a model that is predicting no better than chance. An AUC of 1 is a perfect prediction.

As shown in Table 6, the Standard-Knowledge-Tracing model has an overall RMSE of 0.3445, the Prior-Per-Student model has an RMSE of 0.3469, the Condition-Analysis model has an RMSE of 0.3466, and the Condition-Representation-Analysis model has the lowest RMSE with 0.3427 as well as the best AUC. The fact that the model fit indices altogether are relatively low might be attributed to the fact that we did not include a knowledge component model, but instead treated each step in the tutor as a separate knowledge component. We conclude that the Bayesian network that includes students' prior knowledge (S), experimental condition (C), and representations used for a certain problem (R) provides the best model fit.

Table 7 provides a summary of students' performance on the Fractions Tutor problems during the acquisition phase, based on the overall first-attempt correct steps students made during practice with the Fractions Tutor. A repeated measures ANOVA with students' performance on each task type as dependent measure and practice schedule as independent factor showed that students' performance during the acquisition phase did not significantly differ between practice schedules ($F<1$). Planned contrasts between the blocked condition and each of the interleaved condition did not yield significant differences in students' performance ($ts<1$). Table 8 shows the learning rates obtained from the Condition-Representation-Analysis model for each

Table 6 Summary of the cross-validated prediction results of the four tested models using RMSE and AUC metrics

	Model	RMSE	AUC
1	Condition-Representation-Analysis Model	0.3427	0.6528
2	Standard-Knowledge-Tracing Model	0.3445	0.6181
3	Condition-Analysis Model	0.3466	0.5509
4	Prior-Per-Student Model	0.3469	0.5604

Table 7 Average number of correct first attempts by task type and practice schedule (standard deviation in brackets). Higher numbers indicate higher performance during the acquisition phase

Task type	Blocked	Moderate	Interleaved	Increased
1	0.85 (0.08)	0.86 (0.06)	0.86 (0.06)	0.87 (0.05)
2	0.88 (0.07)	0.89 (0.07)	0.89 (0.05)	0.89 (0.05)
3	0.91 (0.05)	0.88 (0.07)	0.88 (0.06)	0.87 (0.07)
4	0.87 (0.06)	0.84 (0.07)	0.81 (0.07)	0.84 (0.06)
5	0.83 (0.10)	0.86 (0.06)	0.83 (0.08)	0.83 (0.07)
6	0.88 (0.10)	0.90 (0.06)	0.89 (0.07)	0.89 (0.07)
Overall	0.87 (0.08)	0.87 (0.07)	0.86 (0.07)	0.87 (0.06)

condition for each of the task types that the Fractions Tutor covered. Overall, the learning rate estimates align with the results obtained from the posttest data: the fully interleaved condition demonstrates higher learning rates overall than the other conditions. Examining the learning rates by task type provides more specific insights on the nature of the differences between conditions in learning rates. For all but the fourth task type (naming improper fractions), the fully interleaved condition demonstrates a higher learning rate than the blocked condition. To test whether these differences are statistically significant, we employed the binomial test used in Pardos et al. (2010). The advantage of fully interleaved practice over blocked practice was statistically significant for task types 1, 2 and 3 (ps<0.05) and moderately significant for task type 5 ($p<0.10$). The fully interleaved condition achieved the highest overall learning rate, which was twice that of any other condition. This advantage is remarkable, given that performance, as established by the average number of errors made during the acquisition phase, did not differ between conditions. Learning rates of the increased condition fall between the blocked and fully interleaved conditions on most task types 1, 2, and 5, as might be expected. However, the increasingly interleaved condition shows very low learning rates on task types 3 and 4; these are task types that required students to reconstruct the unit of a fraction, a particularly challenging topic that is typically not part of school curricula for fractions.

Table 8 Learning rates by task type and practice schedule from the Condition-Representation Analysis Model. Higher numbers indicate higher learning rates during the acquisition phase. * indicates significant differences between conditions, (*) indicates marginally significant differences

Task type	Blocked	Moderate	Interleaved	Increased
1*	0.0061	0.0061	0.0080	0.0072
2*	0.0019	0.0032	0.0065	0.0036
3*	0.0149	0.0059	0.0337	0.0030
4	0.0037	0.0022	0.0035	0.0014
5(*)	0.0108	0.0220	0.0124	0.0130
6	0.0043	0.0107	0.0078	0.0090
Overall	0.0062	0.0056	0.0120	0.0062

Discussion

The findings from the Bayesian knowledge tracing analysis support and augment the findings from the classroom study in several ways. The finding that the Condition-Representation-Analysis model provides the best fit to the log data is in line with the overall finding from the classroom experiment that practice schedules of multiple graphical representations matter. The differences between conditions on learning rate estimates provide further support for hypothesis 2 in the classroom experiment, that students who learn with multiple graphical representations presented in an interleaved fashion will outperform students who learn with multiple graphical representations presented in a blocked fashion.

The learning rates model a latent factor for students' gains in knowledge, separate from problem difficulty induced by the graphical representation, which is accounted for by conditioning the guess-and-slip parameters on the graphical representation used in each step (in the Condition-Representation-Analysis model). This procedure allows us to assess students' learning from different practice schedules more accurately than pure performance measures do: in a way, Bayesian knowledge tracing allows us to "tease apart" the effects of practice schedules on learning (captured by the learning rate estimates) and on problem difficulty (captured by the guess-and-slip parameters).

The literature on practice schedules shows that interleaved sequences often impair performance during the acquisition phase (e.g., de Croock et al. 1998). It is assumed that temporal variation between consecutive problems interferes with immediate performance since students have to use a new problem-solving procedure each time they encounter a new task. This interference leads to higher processing demands and lower performance during the acquisition phase, but results in better long-term retention and transfer performance later on. In the light of this literature, one might expect that higher learning gains in the interleaved condition become apparent only in the posttest data, but not during the acquisition phase because they might be "masked" by impaired performance due to contextual interference. Our data does not show that interleaved practice schedules result in lower performance during the acquisition phase. It is possible that in tutored problem solving, the performance differences may be less pronounced than in untutored problem solving. Furthermore, building on our prior work (Rau et al. 2013a), we interleaved task types in all conditions so that even in the condition that blocked graphical representations, another aspect of the tutor problems (namely task types) were interleaved. This consistent degree of interleaving in all conditions may have diminished the expected differences between conditions in performance during the acquisition phase.

Although we do not find lower performance in the interleaved conditions, our findings are in line with the overall notion that performance measures are not suitable for detecting differences between practice schedules during the acquisition phase. Rather than investigating differences between directly observed behaviors, Bayesian knowledge tracing models "machine-learn" a latent variable, namely the probability that a student transitions from the unlearned state to the learned state. These learning rate estimates appear to be a more suitable metric to detect advantages of interleaved practice even during the acquisition phase. In other words, "naïve" methods such as performance during the acquisition phase are not suitable to detect differences in students' learning from different practice schedules. Bayesian knowledge tracing

analyses allow detecting learning gains that may be too subtle to detect during the acquisition phase when relying on student performance only.

Why might we not have found significant differences between conditions on learning rate estimates for all topics? There were no differences on task types 4 and 6. Interestingly, task type 4 (reconstructing the unit for proper fractions) strongly builds on task type 3 (reconstructing the unit for unit fractions). Likewise, task type 6 (making improper fractions) strongly builds on task type 5 (naming improper fractions). Following our argument that the effect of interleaved practice might depend on students' prior knowledge, it might be possible that task type 3 equipped students with substantial "prior knowledge" to task type 4 (and task type 5 to task type 6, respectively) so that the advantage of interleaved practice was diminished as a result. The surprisingly low learning rates for the increasingly interleaved condition on the particularly challenging but unfamiliar task types 3 and 4 may also reflect a possibly complex interaction between practice schedule, prior knowledge, and task type difficulty. As noted earlier, this explanation is highly speculative. However, this finding might illustrate yet again that much is to be gained by investigating more thoroughly the interaction between practice schedules of graphical representations and students' prior knowledge.

General Discussion

Taken together, our analysis of the learning outcomes from the classroom study, the think-aloud study, and the Bayesian knowledge tracing analysis yield interesting insights that are both of theoretical and practical significance. From a practical perspective, our results provide qualified evidence that interleaving graphical representations leads to better learning than blocking graphical representations. The analysis of the learning outcomes shows a significant advantage of interleaved practice only on transfer of conceptual knowledge at the delayed posttest, and a marginally significant advantage of the increasingly interleaved condition on reproduction with area models. The advantage of the fully interleaved condition over the blocked condition was particularly true for students with low prior knowledge. Furthermore, the blocked condition never outperformed any of the interleaved conditions. Finally, the learning gains from pretest to the (immediate and delayed) posttests were most consistent for the fully interleaved condition.

The Bayesian knowledge tracing analysis provides further support for this practical recommendation. The results show that a model that includes practice schedules as a predictor fit the data best. This finding is in line with our interpretation of the results on learning outcomes, namely, that practice schedules affect students' learning. Furthermore, the Bayesian knowledge tracing analysis shows that interleaving graphical representations leads to better learning than blocking graphical representations.

In sum, based on the results from the learning outcomes and from the tutor log data, we cautiously recommend that designers of learning materials provide multiple graphical representations in an interleaved rather than in a blocked sequence, in particular if learners have low prior knowledge, and if the goal is to promote conceptual transfer. Given that graphical representations are used across many educational technologies in science and mathematics domains, our findings provide guidance for instructional designers of a wide range of instructional materials.

Our results also provide novel insights from a theoretical perspective. We extend the literature on interleaved practice, which has mostly focused on the effects of task types in a variety of domains. In particular, we provide evidence that the advantage of interleaved practice generalizes to sequences of multiple graphical representations.

The small-scale think-aloud study suggests that interleaved practice does not enhance students' learning by the mechanism of abstraction. Students who worked with the fully interleaved version of the Fractions Tutor did not spontaneously make connections between representations or abstract across them. This observation suggests that repeated reactivation of representation-specific knowledge, and not abstraction, is the main mechanism that accounts for the advantage of interleaved practice with multiple graphical representations. When students work with interleaved graphical representations, they have to reactivate the knowledge relevant to using that graphical representation to solve fractions problems more often than when working with blocked practice schedules of graphical representations. The process of loading representation-specific knowledge components from long-term memory into working memory increases the strength of the association between the graphical representation and that knowledge component, which in turn improves the chances that a student will be able to retrieve the knowledge component later on. However, these considerations are somewhat speculative, given that we did not directly assess reactivation processes in the classroom experiment. Future work should thus investigate whether indeed the advantage of interleaving representations results from repeated reactivation of knowledge about specific representations.

Understanding which of the proposed mechanisms is most likely to account for the advantage of interleaved practice is not only interesting from a theoretical standpoint but also has important practical implications as to which scenarios we can expect our findings to generalize to. If reactivation is the major accountable mechanism, we expect that interleaving graphical representations will lead to better learning than blocking them, *provided that* the representations are sufficiently dissimilar in terms of some critical conceptual aspect. In other words, there has to be some critical knowledge that is not shared between representations and that (consequently) is being reactivated and thereby strengthened every time students switch between representations. Reactivation may even occur if the different representations are maximally dissimilar–they might even be about a completely different topic (although that might not be a wise design decision for pedagogical reasons or learning-efficiency considerations).

If abstraction is the major accountable mechanism, we expect that interleaving graphical representations will lead to better learning than blocking them, *provided that* the representations are sufficiently dissimilar (so that there is some difference to abstract across) *and* sufficiently similar, so that there is some common conceptual commonality that students can abstract from the different representations. For abstraction to occur, students need to hold relevant components of knowledge (which are shared by both representations) in working memory at the same time. If the number of shared knowledge components exceeds working memory capacities (i.e., if the representations are too similar), abstraction might be jeopardized, especially if students lack prior knowledge to select conceptually relevant aspects to attend to. If the number of shared knowledge components is too small (i.e., if the representations are too dissimilar), abstraction will fail because there's not enough shared information to abstract.

These considerations regarding the mechanisms that account for the advantage of interleaved practice have implications for the generalizability of our findings to learning materials. In our classroom experiment, we employed only three graphical representations. Do our findings apply to learning materials that include more than three graphical representations? If reactivation is the main learning mechanism, we would expect that the advantage of interleaved over blocked practice does not depend on the number of graphical representations involved. If, however, abstraction is the main learning mechanism, we might expect that whether or not interleaved practice leads to better learning than blocked practice crucially depends on the type of information shared between consecutively presented graphical representations. As long as students can abstract the target concepts from consecutive graphical representations, we expect that interleaved practice will lead to better learning than blocked practice.

We note again that both learning mechanisms might be at work, as they are not necessarily mutually exclusive. Although we did not observe explicit abstraction in our think-aloud study, it is possible that (1) abstraction does occur, but it did not in our sample of six students, (2) abstraction does occur but remains unconscious, or (3) that students would benefit even more from interleaved practice if they were also explicitly prompted to abstract across representations. While we cannot make claims regarding arguments (1) and (2), the findings from the third phase of the think-aloud study, that students make connections between representations when prompted to do so, is in line with argument (3). Given the extensive literature that shows that students benefit from receiving support for connection making between text and diagrams (Bodemer and Faust 2006; Bodemer et al. 2004; Plötzner et al. 2001) and between symbols and graphs (van der Meij and de Jong 2006), it is likely that students' would benefit from explicit support for connection making between different graphical representations. We investigate this question in subsequent work (Rau et al. 2012a; Rau et al. 2013d).

Our analysis of the tutor log data using Bayesian knowledge tracing also provides interesting theoretical insights. The results from the log data analysis are in line with the interpretation that interleaved practice mainly enhances students' learning via repeated reactivation. Repeated reactivation of representation-specific knowledge may support the ease with which students can retrieve knowledge about a given graphical representation. The higher learning rates that we found for the interleaved condition (compared to the blocked condition) indicates that students become more accurate at solving fractions problems, all of which include graphical representations. This is what we would expect if students acquire representational fluency: the ability to use graphical representations as tools to solve domain-relevant tasks. Based on this finding, we hypothesize that interleaved practice with graphical representations enhances students' learning of fractions by promoting their representational fluency (through repeated reactivation of representation-specific knowledge).

Finally, the Bayesian knowledge tracing analysis extends the literature on interleaved practice by showing that the advantage of interleaved practice cannot only be detected based on long-term retention and transfer assessments (de Croock et al. 1998), but also based on "machine-learned" latent variable of students' learning rates, inferred from students' problem-solving behaviors. To the best of our knowledge, the present study is the first to empirically establish advantages of interleaved practice over blocked practice using data from the acquisition phase. We demonstrate that methods of educational data mining provide unique opportunities to gain deeper insights into educational

psychology questions in a way that is not possible using "naïve" methods of looking at performance data alone.

Our findings also extend our own prior work on learning with multiple graphical representations. We showed that multiple graphical representations lead to better learning than a single graphical representation (Rau et al. 2009). Further, we showed that, when faced with a choice to interleave one dimension, task types *or* representations, we should interleave task types rather than representations. Our current work extends this finding by showing that interleaving both dimensions, task types and graphical representations, leads to the best learning gains. Building on the observations in the small-scale think-aloud study, our subsequent work (Rau et al. 2012a) shows that students' learning can additionally be enhanced by providing explicit support for connection making between graphical representations.

Future Research Directions

There are several open questions that we might consider in future research. One particularly interesting question concerns the generalizability of our findings to other domains. We have argued that most STEM domains use multiple graphical representations to emphasize different conceptual aspects of the domain knowledge, such as chemistry (Kozma et al. 2000; Kozma and Russell 2005; Stieff et al. 2011; Zhang and Linn 2011), biology (Cook et al. 2007; Simons and Keil 1995), physics (Larkin and Simon 1987; Lewalter 2003; Urban-Woldron 2009; van der Meij and de Jong 2006), engineering (Nathan et al. 2011; Walkington et al. 2011), and programming (Kordaki 2010). Let us consider chemistry as one example. Chemistry uses a variety of graphical representations of molecules (Kozma et al. 2000; Kozma & Russell 2005; Stieff et al. 2011; Zhang and Linn 2011): Electrostatic Potential Map (EPM) representations make the concept of electron density and molecular dipoles easily accessible, but make it more difficult to perceive the details of the chemical structure and molecular geometry. On the other hand, ball-and-stick figures show the complete chemical structure of a molecule and provide information on the geometry (i.e., the spatial arrangement of the molecule's atom), but they do not depict electron density. To predict the reactivity of the molecule, both electron density and molecular geometry are important factors. Thus, both graphical representations share a common concept (i.e., chemical molecules) but emphasize complementary aspects of the concept (electron density versus molecular geometry). Based on our research, we would expect the best learning gains (e.g., in the ability to predict reactivity of given molecules) if these different graphical representations were presented to students in an interleaved rather than in a blocked fashion, so that students can frequently reactivate their knowledge about molecule surfaces and molecule structure. To a (possibly) lesser extent, students may also abstract conceptual understanding of what constitutes a chemical molecule from molecule surface features and molecule structure features. However, in light of our observations in the small-scale think-aloud study, we might further hypothesize that students will require additional, more explicit support to engage in connection-making activities that allow them to abstract conceptual knowledge about chemical molecules from these graphical representations. As this example illustrates, we often employ multiple, conceptually complementary graphical representations to support students' learning of domain knowledge in STEM domains.

Another interesting question regards the potential benefits of adaptive interleaved practice schedules. Our finding that interleaved practice schedules are particularly effective for students with low prior knowledge suggests that, as students acquire more and more robust knowledge through practice, the choice of practice schedules has a diminished impact on their learning. Furthermore, our finding that interleaved practice enhances conceptual transfer and reproduction with area models, but not procedural transfer and reproduction with number lines, suggests that the effectiveness of a practice schedule may depend on the knowledge that is targeted. It is even possible that there is a three-way interaction between practice schedules, prior knowledge, and target knowledge. Our study is limited in that we implemented only four fixed practice schedules. Future research should investigate the effectiveness of other possible practice schedules in enhancing different aspects of domain knowledge for particular types of learners. Intelligent tutoring systems offer unique possibilities in adapting practice schedules to the students' prior knowledge level and learn rate. Yet, we still know too little about what types of adaptive practice schedules might be most effective. A further interesting question for future research might be to what extent the higher effectiveness of adaptive problem sequences (compared to fixed problem sequences) can be traced back to the fact that adaptive problem sequences are often more highly interleaved than fixed problem sequences. By demonstrating that practice schedules do matter, and that their effectiveness depends on students' prior knowledge, our research provides a first step in this direction.

Conclusions

In sum, our research demonstrates how multiple methodological approaches can complement one another to investigate different aspects of a research question. To investigate *which* practice schedule works best, we conducted a controlled classroom experiment and analyzed the learning outcomes. This is a common approach in many fields related to intelligent tutoring systems research (including instructional design research, learning sciences research, educational psychology research, etc.). Yet, analyzing the learning outcome data does not answer the question of *why* we find differences between conditions. This question is not only of theoretical relevance, it also has practical implications regarding possible scenarios that we can expect our findings to generalize to (as discussed above). To address this open question about which learning mechanism is most likely to account for differences between conditions, we conducted a think-aloud study. Although we cannot rule out that abstraction may occur concurrently with reactivation processes, our think-aloud study does not provide evidence of students *explicitly* engaging in abstraction processes. Our observations in the think-aloud study suggest that reactivation, rather than abstraction, is more likely to account for the advantage of interleaved practice we found in the classroom experiment. To further augment these observations, we used Bayesian knowledge tracing analysis to investigate differences between conditions in learning rates during the acquisition phase. Our finding that latent measures of learning replicate the advantage of interleaved practice that we found in the classroom experiment is in line with our interpretation of the think-aloud study that reactivation is a likely learning mechanism to account for the advantage of interleaved practice (although other interpretations are possible, as argued above).

Taken together, our work illustrates how the use of a variety of methodologies can complement one another to answer the questions of *what* works, and *why* does it work.

Within the research community on intelligent tutoring systems, there are many other examples illustrating what is gained by this multi-methods approach. For example, Li and colleagues (2012) use SimStudent, a machine-learning agent that learns skills from demonstrated solutions, to investigate whether positive or negative feedback accounts for the effects of interleaved versus blocked practice. Their results show that interleaved practice schedules increase the amount of negative feedback the simulated student receives. They conclude that negative, rather than positive feedback may account for the differences between practice schedules. Pavlik and colleagues (2013) use additive factors modeling to investigate forgetting and spacing effects in an experiment of interleaved versus blocked practice in musical training. Work by Rummel and colleagues (2012) illustrates how the analysis of verbal data and log data contribute complementary perspectives in our understanding of the mechanisms by which scripted collaborative learning with intelligent tutoring systems enhances students' robust learning. Like our multi-methods study, these other examples illustrate one of the key benefits of conducting interdisciplinary research. Intelligent tutoring systems research can crucially benefit from the insights we gain from multi-methods approaches about mechanisms that underlie effects of instructional design. Furthermore, we can use this knowledge to further our research on developing interventions that make use of the advantage that intelligent tutoring systems offer, for instance, to provide adaptive practice schedules.

In conclusion, our research was done in the context of a successful intelligent tutoring system that focuses on conceptual learning with multiple interactive, abstract graphical representations. We extend the literature on learning with multiple representations by demonstrating that interleaved practice of graphical representations promotes students' learning. It seems more likely that the mechanism underlying the advantage of interleaved practice is repeated reactivation of representation-specific knowledge rather than abstraction. Further, we demonstrate that learning rate estimates based on Bayesian knowledge tracing are an appropriate metric for detecting advantages of interleaved practice even during the acquisition phase. Our findings lead to instructional design recommendations that developers of intelligent tutoring systems can draw upon when designing instructional materials that present multiple graphical representations across consecutive tasks. Although our findings are subject to further investigation, we recommend that learning materials should provide multiple graphical representations in an interleaved fashion, rather than in a blocked fashion, especially if the goal is to promote students' acquisition of conceptual knowledge that can transfer to new tasks and if students have little prior knowledge of the domain.

References

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*(3), 183–198. doi:10.1016/j.learninstruc.2006.03.001.

Aleven, V., McLaren, B., Sewall, J., & Koedinger, K. (2009a). A new paradigm for intelligent tutoring systems: example-tracing tutors. *International Journal of Artificial Intelligence in Education, 19*(2), 105–154.

Aleven, V., McLaren, B. M., & Sewall, J. (2009b). Scaling up programming by demonstration for intelligent tutoring systems development: an open-access website for middle-school mathematics learning. *IEEE Transactions on Learning Technologies, 2*(2), 64–78. doi:10.1109/TLT.2009.22.

Akpinar, Y., & Hartley, J. R. (1996). Designing interactive learning environments. *Journal of Computer Assisted Learning, 12*(1), 33–46.

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

Anderson, J. R. (2002). *ACT: A simple theory of complex cognition*. Cambridge, MA, US: MIT Press.

Arroyo, I., Beck, J., Schultz, K., & Woolf, B. P. (1999). Piagetian psychology in intelligent tutoring systems, *Proceedings of the Ninth International Conference on Artificial Intelligence in Education, 1999*.

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of Foreign-Language Vocabulary and the Spacing Effect. *Psychological Science, 4*(5), 316–321.

Baker, R. S., Corbett, A., & Koedinger, K. (2007). The difficulty factors approach to the design of lessons in intelligent tutor curricula. *International Journal of Artificial Intelligence in Education, 17*(4), 105–154.

Bannert, M. (2002). Managing cognitive load-recent trends in cognitive load theory. *Learning and instruction, 12*(1), 139–146.

Battig, W. F. (1972). Intratask interference as a source of facilitation in transfer and retention. In R. F. Thompson & J. F. Vos (Eds.), *Topics in learning and performance* (pp. 131–159). New York: Academic.

Beck, J., Stern, M., & Woolf, B. P. (1997). Using the student model to control problem difficulty. In A. Jameson, C. Paris, & C. Tasso (Eds.), *User modeling: proceedings of the sixth international conference, UM97*. Vienna: Springer Wien New York.

Bodemer, D., & Faust, U. (2006). External and mental referencing of multiple representations. *Computers in Human Behavior, 22*(1), 27–42.

Bodemer, D., Ploetzner, R., Feuerlein, I., & Spada, H. (2004). The Active Integration of Information during Learning with Dynamic and Interactive Visualisations. *Learning and Instruction, 14*(3), 325–341.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380.

Charalambous, C. Y., & Pitta-Pantazi, D. (2007). Drawing on a theoretical model to study students' understandings of fractions. *Educational Studies in Mathematics, 64*(3), 293–316. doi:10.1007/s10649-006-9036-2.

Clark, R. E., Feldon, D. F., Van Merrienboër, J. J. G., Yates, K. A., & Early, S. (2007). Cognitive task analysis. In J. M. Spector, M. D. Merrill, J. J. G. Van Merrienboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology*. Mahwah: Lawrence Erlbaum Associates.

Common Core State Standards Initiative. (2010). Common core state standards for mathematics. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Cook, M., Wiebe, E. N., & Carter, G. (2007). The influence of prior knowledge on viewing and interpreting graphics with macroscopic and molecular representations. *Science Education, 92*(5), 848–867.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253–278.

Cox, R. (1999). Representation construction, externalised cognition and individual differences. *Learning and instruction, 9*(4), 343–363.

Cramer, K. (2001). Using models to build an understanding of functions. *Mathematics teaching in the middle school, 6*(5), 310–318.

Cramer, K., Wyberg, T., & Leavitt, S. (2008). The role of representations in fraction addition and subtraction. *Mathematics teaching in the middle school, 13*(8).

Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010). Cognitive activities in complex science text and diagrams. *Contemporary Educational Psychology, 35*(1), 59–74.

de Croock, M. B. M., Van Merrienboër, J. J. G., & Paas, F. G. W. C. (1998). High versus low contextual interference in simulation-based training of troubleshooting skills: effects on transfer performance and invested mental effort. *Computers in Human Behavior, 14*(2), 249–267.

de Jong, T., Ainsworth, S. E., Dobson, M., Van der Meij, J., Levonen, J., Reimann, P., et al. (1998). Acquiring knowledge in science and mathematics: The use of multiple representations in technology-based learning environments. In M. W. Van Someren, W. Reimers, H. P. A. Boshuizen & T. de Jong (Eds.), Learning with Multiple Representations. Oxford.

Eilam, B. (2013). Possible Constraints of Visualization in Biology: Challenges in Learning with Multiple Representations *Multiple Representations in Biological Education* (pp. 55–73): Springer Netherlands.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: verbal reports as data*. Cambridge: MIT Press.

Even, R. (1998). Factors involved in linking representations of functions. *The Journal of Mathematical Behavior, 17*(1), 105–121.

Gagatsis, A., & Elia, I. (2004). *The effects of different modes of representation on mathematical problem solving*. Paper presented at the 28th Conference of the International Group for the Psychology of Mathematics Education.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review, 23*(4), 523–552.

Gobert, J. D., & Koedinger, K. R. (2011). Using model-tracing to conduct performance assessment of students' inquiry skills within Technology. Paper presented at the SREE Fall 11 Conference.

Goguadze, G., Melis, E., DFKI. (2008). Feedback in active math exercises, International Conference on Mathematics Education (ICME).

Goldman, S. R. (2003). Learning in complex domains: When and why do multiple representations help? *Learning and instruction, 13*(2), 239–244.

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *Journal of the Learning Sciences, 14*(1), 69–110.

Goldstone, R. L., Steyvers, M., & Rogosky, B. J. (2003). Conceptual interrelatedness and caricatures. *Memory & Cognition, 31*(2), 169–180.

Grady, S. M. (1998). *Virtual reality: computers mimic the physical world*. New York: Facts on File.

Gutwill, J. P., Frederiksen, J. R., & White, B. Y. (1999). Making Their Own Connections: Students' Understanding of Multiple Models in Basic Electricity. *Cognition and Instruction, 17*(3), 249–282.

Hebert, E. P., Landin, D., & Solmon, M. A. (1996). Practice schedule effects on the performance and learning of low- and high-skilled students: An applied study. *Research Quarterly for Exercise and Sport, 67*(1), 52–58.

Heffernan, N. T., Heffernan, C., Deceoteu, M., & Militello, M. (2012). Effective and meaningful use of educational technology: three cases from the classroom. In C. Dede & J. Richards (Eds.), *Digital teaching platforms* (pp. 88–102). New York: Teacher's College Press.

Heim, M. (2000). *Virtual realism*. New York: Oxford University Press.

Helsdingen, A., van Gog, T., & Van Merrienboër, J. J. G. (2011). The effects of practice schedule and critical thinking prompts on learning and transfer of a complex judgement task. *Journal of Educational Psychology, 103*(2), 383–398.

Hinze, S. R., Rapp, D. N., Williamson, V. M., Shultz, M. J., Deslongchamps, G., & Williamson, K. C. (2013). Beyond ball-and-stick: Students' processing of novel STEM visualizations. *Learning and Instruction, 26*, 12–21.

Immink, M. A., & Wright, D. L. (1998). Contextual interference: A response planning account. *The Quarterly Journal of Experimental Psychology Section A, 51*(4), 735–754.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: helping children learn mathematics*. Washington: National Academy Press.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). Knowledge-learning-instruction framework: bridging the science-practice chasm to enhance robust student learning. Cognitive Science, 36(5), 757–798.

Koedinger, K. R., & Corbett, A. (2006). *Cognitive tutors: technology bringing learning sciences to the classroom*. New York: Cambridge University Press.

Koedinger, K.R., Pavlik Jr., P.I., Stamper, J.C., Nixon, T. & Ritter, S. (2011). Avoiding problem selection thrashing with conjunctive knowledge tracing. In *Proceedings of the Fourth International Conference on Educational Data Mining* (pp. 91–100).

Koedinger, K. R., Stamper, J. C., McLaughlin, E. A., & Nixon, T. (2013). Using data-driven discovery of better student models to improve student learning. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial intelligence in education* (pp. 421–430). Berlin: Springer.

Kong, S. C., & Kwok, L. F. (2003). A graphical partitioning model for learning common fraction: designing affordances on a web-supported learning environment. *Computers and Education, 40*, 137–155.

Kordaki, M. (2010). A drawing and multi-representational computer environment for beginners' learning of programming using C: Design and pilot formative evaluation. *Computers & Education, 54*(69–87).

Kozma, R. (2003). The material features of multiple representations and their cognitive and social affordances for science understanding. *Learning and Instruction, 13*, 205–226.

Kozma, R., & Russell, J. (2005). Students becoming chemists: Developing representational competence Visualization in science education (pp. 121–145).

Kozma, R., Chin, E., Russell, J., & Marx, N. (2000). The roles of representations and tools in the chemistry laboratory and their implications for chemistry learning. *The Journal of the Learning Sciences, 9*(2), 105–143.

Lamon, S. J. (Ed.). (1999). *Teaching fractions and ratios for understanding*. Mahwah: Lawrence Erlbaum Associates.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science: A Multidisciplinary Journal, 11*(1), 65–100.

Lee, T. D., & Magill, R. A. (1983). The locus of contextual interference in motor-skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(4), 730–746.

Lee, T. D., & Magill, R. A. (1985). Can forgetting facilitate skill acquisition? In D. Goodman, R. B. Wilberg, & I. M. Franks (Eds.), *Differing perspectives in motor learning, memory, and control* (p. 3e22). Amsterdam: Elsevier.

Lewalter, D. (2003). Cognitive Strategies for Learning from Static and Dynamic Visuals. *Learning and Instruction, 13*(2), 177–189.

Li, N., Cohen, W. W., & Koedinger, K. R. (2012). Problem order implications for learning transfer. In S. Cerri, W. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (Vol. 7315, pp. 185–194). Berlin: Springer.

Li, Y. H., & Wright, D. L. (2000). An assessment of the attention demands during random- and blocked-practice schedules. Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology, 53(2), 591–606.

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology-Learning Memory and Cognition, 22*(6), 1423–1442.

Meiran, N., Chorev, Z., & Sapir, A. (2000). Component processes in task switching. *Cognitive Psychology, 41*(3), 211–253.

Moyer, P., Bolyard, J., & Spikell, M. A. (2002). What are virtual manipulatives? *Teaching children mathematics, 8*, 372–377.

Murphy, K. (2001). The bayes net toolbox for matlab. *Computing science and statistics, 33*(2), 1024–1034.

Nathan, M. J., Atwood, A. K., Prevost, A., Phelps, L. A., & Tran, N. A. (2011). How Professional Development in Project Lead the Way Changes High School STEM Teachers' Beliefs about Engineering Education. *Journal of Pre-College Engineering Education Research (J-PEER), 1*(1), 15–29.

National Research Council. (2002). Scientific research in education. In R. J. Shavelson & L. Towne (Eds.), *Committee on scientific principles for educational research*. Washington: National Academy Press.

NCTM. (1989). Curriculum and evaluation standards for school mathematics: national council of teachers of mathematics

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

NCTM. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: a quest for coherence*. Reston: National Council of Teachers of Mathematics.

Nisbett, R. E., & Ross, L. (1980). *Human inference: strategies and shortcomings of social judgment*. NJ: Prentice Hall.

NMAP (2008). Foundations for success: report of the national mathematics advisory board panel: U.S. government printing office.

Ollis, S., Button, C., & Fairweather, M. (2005). The influence of professional expertise and task complexity upon the potency of the contextual interference effect. *Acta Psychologica, 18*(3), 229–244

Özgün-Koca, S. A. (2008). Ninth grade students studying the movement of fish to learn about linear relationships: The use of video-based analysis software in mathematics classrooms. *The Mathematics Educator, 18*(1), 15–25.

Pardos, Z. A., Dailey, M. D., & Heffernan, N. T. (2010a). Learning what works in ITS from non-traditional randomized controlled trial data. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Proceedings of the 10th International Conference on ITS2010* (Vol. 6094, pp. 41–50). Berlin: Springer.

Pardos, Z. A., Dailey, M., & Heffernan, N. (2011a). Learning what works in ITS from non-traditional randomized controlled trial data. *The International Journal of Artificial Intelligence in Education, 21*(1), 47–63.

Pardos, Z., & Heffernan, N. T. (2010). Navigating the parameter space of Bayesian knowledge tracing models: visualizations of the convergence of the expectation maximization algorithm. In S. Baker, A. Merceron & P. I. Pavlik (Eds.), Proceedings of the 3rd international conference on educational data mining (pp. 161–170).

Pardos, Z. A., & Heffernan, N. T. (2011). KT-IDEM: Introducing item difficulty to the knowledge tracing model. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero & J. Stamper (Eds.), Proceedings of the 4th International Conference on Educational Data Mining.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. P. (2007). Enhancing learning and retarding forgetting: choices and consequences. *Online Submission Description, 19*.

Pavlik, P. I., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis—A new alternative to knowledge tracing. In V. Dimitrova, R. Mizoguchi, & B. du Boulay (Eds.), *Proceedings of the 14th international conference on artificial intelligence in education* (pp. 531–538). Amsterdam: Ios Press.

Pavlik, P., Hua, H., Williams, J., & Bidelman, G. M. (2013). Modeling and optimizing forgetting and spacing effects during musical interval training. In S. K. D'Mello, R. A. Calvo & A. Olney (Eds.), Proceedings of the 6th international conference on educational data mining (EDM 2013) (pp. 145–152): International Educational Data Mining Society.

Plötzner, R., Bodemer, D., & Feuerlein, I. (2001). *Facilitating the mental integration of multiple sources of information in multimedia learning environments*. Paper presented at the World Conference on Educational Multimedia, Hypermedia & Telecommunications.

Plötzner, R., Bodemer, D., & Neudert, S. (2008). Successful and less successful use of dynamic visualizations in instructional texts. In R. K. Lowe & W. Schnotz (Eds.), *Learning with Animation. Research implications for design*. New York: Cambridge University Press.

Rathmell, E. C., & Leutzinger, L. P. (1991). Number representations and relationships. *The Arithmetic Teacher, 38*(7), 20–23.

Rau, M. A., Aleven, V., & Rummel, N. (2009). Intelligent tutoring systems with multiple representations and self-explanation prompts support learning of fractions. In V. Dimitrova, R. Mizoguchi, & B. du Boulay (Eds.), *Proceedings of the 14th international conference on artificial intelligence in education* (pp. 441–448). Amsterdam: Ios Press.

Rau, M. A., Aleven, V., Rummel, N., & Rohrbach, S. (2012a). Sense making alone doesn't do it: Fluency matters too! ITS Support for robust learning with multiple representations. In S. Cerri, W. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent tutoring systems* (Vol. 7315, pp. 174–184). Berlin: Springer.

Rau, M. A., Rummel, N., Aleven, V., Pacilio, L., & Tunc-Pekkan, Z. (2012b). How to schedule multiple graphical representations? A classroom experiment with an intelligent tutoring system for fractions. In J. V. Aalst, K. Thompson, M. J. Jacobson, & P. Reimann (Eds.), *The future of learning: Proceedings of the 10th international conference of the learning sciences (ICLS 2012)* (Vol. 1, pp. 64–71). Sydney: ISLS.

Rau, M. A., Aleven, V., & Rummel, N. (2013a). Interleaved practice in multi-dimensional learning tasks: which dimension should we interleave? *Learning and Instruction, 23*, 98–114.

Rau, M. A., Aleven, V., Rummel, N., & Rohrbach, S. (2013b). Why interactive learning environments can have it all: Resolving design conflicts between conflicting goals. Paper submitted to CHI 2013.

Rau, M. A., Aleven, V., & Rummel, N. (2013c). How to use multiple graphical representations to support conceptual learning? Research-based principles in the Fractions Tutor. In H. C. Lane, K. Yacef, J. Mostow & P. Pavlik (Eds.), *Artificial Intelligence in Education* (pp. 762–765). Berlin Heidelberg: Springer.

Rau, M., Aleven, V., & Rummel, N. (2013d). Complementary effects of sense-making and fluency-building support for connection making: A matter of sequence? In H. C. Lane, K. Yacef, J. Mostow & P. Pavlik (Eds.), Artificial Intelligence in Education (pp. 329-338). Berlin Heidelberg: Springer. (pdf)

Rau, M. A., & Pardos, Z. A. (2012). Investigating practice schedules of multiple fraction representations using knowledge tracing based learning analysis techniques. In K. Yacef, O. Zaïane, H. Hershkovitz, M. Yudelson & J. Stamper (Eds.), *Proceedings of the 5th International Conference on Educational Data Mining* (pp. 168–171).

Raudenbush, S. W., & Bryk, A. S. (Eds.). (2002). *Hierarchical linear models: applications and data analysis methods* (2nd ed.). Newbury Park: Sage Publications.

Reed, S. K., & Ettinger, M. (1987). Usefulness of tables for solving word problems. *Cognition and Instruction, 4*(1), 43–59.

Reimer, K., & Moyer, P. S. (2005). Third-graders learn about fractions using virtual manipulatives: a classroom study. *Journal of Computers in Mathematics and Science Teaching, 24*(1), 5–25.

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: Lawrence Erlbaum Associates, Inc.

Rohrer, D. (2008). The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education, 40*(1), 4–17.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science, 35*, 481–498.

Rummel, N., Mullins, D., & Spada, H. (2012). Scripted collaborative learning with the cognitive tutor algebra. *International Journal of Computer Supported Collaborative Learning, 7*(2), 307–339. doi:10.1007/s11412-012-9146-z.

Scaife, M., & Rogers, Y. (1996). External cognition: how do graphical representations work? *International Journal of Human-Computer Studies, 45*(2), 185–213.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207–217.

Schneider, W. (1985). Training high-performance skills—fallacies and guidelines. *Human Factors, 27*(3), 285–300.

Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction, 13*(2), 141–156.

Schwartz, D. L., & Black, J. B. (1996a). Analog imagery in mental model reasoning: depictive models. *Cognitive Psychology, 30*(2), 154–219.

Schwartz, D. L., & Black, J. B. (1996b). Shuttling between depictive models and abstract rules: induction and fallback. *Cognitive Science: A Multidisciplinary Journal, 20*(4), 457–497.

Schwonke, R., Ertelt, A., & Renkl, A. (2008). Fostering the translation between external representations. Does it enhance learning with an intelligent tutoring program? In J. Zumbach, N. Schwartz, T. Seufert & L. Kester (Eds.), *Beyond knowledge: The legacy of competence* (pp. 117–119): Springer Netherlands.

Schwonke, R., & Renkl, A. (2010). *How do proficient learners construct mental representations of different but related external representations?* Paper presented at the EARLI SIG 2.

Shea, C. H., & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory, 5*(2), 179–187.

Siegler, R. S., Carpenter, T., Fennell, F., Geary, D., Lewis, J., Okamoto, Y., et al. (2010). *Developing effective fractions instruction: a practice guide*. Washington: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Simon, D. A., & Bjork, R. A. (2001). Metacognition in motor learning. *Journal of Experimental Psychology, 27*(4), 907–912.

Simons, D. J., & Keil, F. C. (1995). An abstract to concrete shift in the development of biological thought: The insides story. *Cognition, 56*(2), 129–163.

Smith, L. B. (2003). Learning to recognize objects. *Psychological Science, 14*(3), 244–250.

Stieff, M., Hegarty, M., & Deslongchamps, G. (2011). Identifying representational competence with multi-representational displays. *Cognition and Instruction, 29*(1), 123–145.

Suh, J., Moyer, P. S., & Heo, J.-J. (2005). Examining technology uses in the classroom: Developing fraction sense using virtual manipulative concept tutorials. Journal of Interactive Online Learning, 3(4)

Superfine, A. C., Canty, R. S., & Marshall, A. M. (2009). Translation between external representation systems in mathematics: All-or-none or skill conglomerate? *The Journal of Mathematical Behavior, 28*(4), 217–236

Sweller, J. (1990). Cognitive load as a factor in the structuring of technical material. *Journal of Experimental Psychology, 119*(2), 176e192. doi:10.1037//0096-3445.119.2.176.

Tabachneck, H. J. M., Leonardo, A. M., & Simon, H. A. (1997). CaMeRa: A computational model of multiple representations. *Cognitive Science, 21*(3), 305–350.

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology, 24*(837–848).

Urban-Woldron, J. (2009). Interactive simulations for the effective learning of physics. *Journal of Computers in Mathematics and Science Teaching, 28*(2), 163–176.

Uttal, D. H. (2003). On the relation between play and symbolic thought: The case of mathematics manipulatives. In O. Saracho & B. Spodek (Eds.), *Contemporary Perspectives in Early Childhood* (pp. 97–114): Information Age Press.

van der Meij, J. (2007). *Support for learning with multiple representations designing simulation-based learning environments*. Universiteit Twente, Twente, Netherlands.

van der Meij, J., & De Jong, T. (2003). *Learning with multiple representations*. Paper presented at the EARLI conference.

van der Meij, J., & de Jong, T. (2006). Supporting Students' Learning with Multiple Representations in a Dynamic Simulation-Based Learning Environment. *Learning and Instruction, 16*(3), 199–212.

Van Merriënboer, J. J. G., Schuurman, J. G., de Croock, M. B. M., & Paas, F. G. W. C. (2002). Redirecting learners' attention during training: effects on cognitive load, transfer test performance and training efficiency. *Learning and Instruction, 12*(1), 11–37.

Walkington, C., Nathan, M., Wolfgram, M., Alibali, M., & Srisurichan, R. (2011). Bridges and barriers to constructing conceptual cohesion across modalities and temporalities: Challenges of STEM integration in the precollege engineering classroom. In S. Publishers (Ed.), *Engineering in PreCollege Settings: Research into Practice*.

Zhang, J. (1997). The nature of external representations in problem solving. *Cognitive Science, 21*(2), 179–217.

Zhang, Z. H., & Linn, M. C. (2011). Can generating representations enhance learning with dynamic visualizations? *Journal of Research in Science Teaching*, 48(10), 1177–1198.

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. Cognitive Science: *A Multidisciplinary Journal, 18*(1), 87–122.