Unifying Computer-Based Assessment Across Conceptual Instruction, Problem-Solving, and Digital Games

William L. Miller[1], Ryan S. Baker[2], Lisa M. Rossi[3]

[1]*Reasoning Mind, Houston, TX*

[2]*Teachers College, Columbia University, New York, NY*

[3]*Georgia Institute of Technology, Atlanta, GA*

neal.miller@reasoningmind.org, baker2@exchange.tc.columbia.edu, lrossi@gatech.edu

A**bstract.** As students work through online learning systems such as the Reasoning Mind blended learning system, they often are not confined to working within a single educational activity; instead, they work through various different activities such as conceptual instruction, problem-solving items, and fluency-building games. However, most work on assessing student knowledge using methods such as Bayesian Knowledge Tracing has focused only on modeling learning in only one context or activity, even when the same skill is encountered in multiple different activities. We investigate ways in which student learning can be modeled across activities, towards understanding the relationship between different activities and which approaches are relatively more successful at integrating information across activities. However, we find that integrating data across activities does not improve predictive power relative to using data from just one activity. This suggests that seemingly identical skills in different activities may actually be cognitively different for students.

Keywords: Bayesian Knowledge Tracing; educational activities

1. Introduction

Increasingly, online learning environments such as blended learning environments assess students as they are learning, including problem-solving environments (Razzaq et al., 2007; Koedinger & Corbett, 2006), simulation microworlds (Gobert et al., 2013; Quellmalz, Timms, and Schneider 2009), action-style games (Lomas et al. 2012), and game-based virtual environments (Baker and Clarke-Midura 2013; McQuiggan et al. 2008). These assessments are becoming increasingly effective, and may even be reaching asymptotic performance for specific types of assessment, such as assessing a single skill in online problem-solving (see Pardos et al., 2011). Some online assessments are becoming sufficiently high-quality to be able to predict external measures such as grade point average (Wüstenberg, Greiff, and Funke, 2012) and standardized exam performance (Pardos et al., 2013).

However, current online assessment is typically limited to the context of one learning environment. As a student practices a skill, it is considered in isolation within the current learning environment. The possibility that the student is gaining proficiency at the skill in other contexts, in between practice opportunities with the current environment, is sometimes acknowledged but not explicitly modeled. This limitation was not serious a decade ago, when students typically used a single online learning environment in a concentrated fashion; but as students increasingly study the same topic in different activities and contexts, this limitation may be slowing the adaptivity of online learning systems. Recent work has created single infrastructure with the capacity for bringing information about a single student's learning in multiple online learning systems together (Archibald & Poltrack, 2011), but the algorithmic methods to actually integrate that information together, and for one system to effectively utilize evidence from a different online learning system, remain lacking.

Indeed, the entire goal of most modern learning environments is not just to produce learning and improved performance just within that system, but to produce robust learning that is retained over time and generalizes to new contexts (Koedinger, Perfetti, and Corbett 2012). There is evidence that the skills that are assessed in many of these environments are not just specific to a single learning environment but can generalize outside of them. Many of the skills currently being assessed are cross-curricular (Greiff and Funke 2009) -- that is to say skills

that are relevant and useful across multiple disciplines or situations (e.g., complex problem solving). For instance, the science inquiry skills measured by Gobert et al. (2013), Baker and Clarke-Midura (2013), and Molnár, Greiff, and Csapó (2013) manifest in a wide range of learning environments and real-world situations.

Some researchers have explicitly looked at whether skills transfer between two contexts. For instance Baker, Gowda, and Corbett (2011) developed an automated detector that can predict, from learning of a skill within an online tutor, how successful the student will be at learning the next skill from a paper curriculum. This detector utilized information on the meta-cognitive behaviors students demonstrated during learning, and performed significantly better than skill assessment alone. Also, Schwonke and colleagues (2009) found a significant relationship between procedural transfer performance and principle-based self-explanations, as well as goal-operator combinations, suggesting that different learning approaches (principle-oriented or solution-oriented) produce different patterns in transfer performance.

Despite the fact that skill can transfer outside of learning environments, and many students use multiple different kinds of learning environments, thus far the development of skill in multiple contexts, environments, or activities is not explicitly modeled in existing online learning environments. In fact, this has not yet been done as far as we know, even within a single learning system that has different types of learning activities, although there has been work to model transfer between different versions of the same activity where different information is present as well as the core information (Maass & Pavlik, 2013).

Within this paper, we study how to most effectively model student learning that is being acquired in multiple activities in an interleaved fashion. In specific, we look at the learning of over a hundred skills, occurring within three activities, of fundamentally different design but involving the same knowledge components (e.g. skills and concepts). Our goal is to answer the research question, "What formalism best accounts for the learning of the same skill occurring across different activities?", or more concretely, "Will a student's learning of a knowledge component in one context during authentic learning translate to superior performance on future applications of that knowledge component in other contexts?"

It is known that learning does indeed transfer between activities in many cases. Gick and Holyoak (1980), for example, observed that college students were able to apply problem-solving strategies generated in prior tasks to a novel, semantically unrelated task even where no prompting of transfer was given. Chen (1996) reported equivalent transfer of problem solving approaches in children aged 5 to 8, with a combination of superficial (i.e. solution-irrelevant), structural (i.e. solution-relevant), and procedural (i.e. solution-relevant, process features) links between target and source used by participants to facilitate transfer. Similarly, Gentner, Loewenstein, and Thompson (2003) observed that college students were able to transfer negotiation strategies learned from prior business cases to new instances, and where multiple solutions were possible, solutions that were consistent with approaches learned earlier were more likely to be selected.

While such studies successfully identified transfer in smaller-scale, laboratory settings, more recent research has revealed transfer of cognitive skills in larger-scale and more externally valid educational environments (i.e. authentic classroom learning). For example, Rittle-Johnson and Star (2007) examined transfer in the classroom, confirming transference of mathematical skills as well as defining conditions under which transfer is most successful. Where seventh-graders were taught to compare and contrast different solution approaches they were able to apply learned strategies to novel problems. Students who were instructed to learn through repeated study of the same solution method across different questions were also able to solve novel problems but with less accuracy.

Conditions of transfer have been further studied in middle-school equation solving. Although all students demonstrated transfer of learned skills to solving of problems with novel features, students were more likely to be able to transfer knowledge between context when they learned to compare different solution methods, as opposed to comparing similar solution methods and/or different problem types. This activity led to better-developed conceptual knowledge and fluency, and in turn to enhanced transfer (Rittle-Johnson & Star, 2009). Consistent results have been found in middle school physics students, with superior transfer observed where learning was geared towards developing conceptual knowledge (Schwartz, Chase, Oppezzo, & Chin, 2011), and second-graders learning arithmetic where emphasis of fluency/flexibility of application also supported transfer (Blöte, Van der Burg, & Klein, 2001). More thorough conceptual

knowledge of a domain and greater capacity to abstract and selectively apply learning to novel and varied situations, are thus integral components of transfer. In addition, more thoughtful meta-cognitive behaviors during learning can also support a student in developing learning that will transfer (Baker, Gowda, & Corbett, 2011). These findings have contributed to the Pittsburgh Science of Learning Center theoretical framework, which argues that conceptual understanding, fluency, and successful meta-cognition lead to generally more robust learning, which transfers, is associated with more successful future learning, and which is retained over time (Koedinger, Corbett, & Perfetti, 2011).

However, the research question of interest for this paper is not whether learning transfers between activities (as there is a lengthy literature, much more extensive than the papers reviewed here, attesting that it can do so in many contexts), or what cognitive mechanisms lead to this transfer, but specifically what modeling formalism best accounts for learning occurring in multiple activities in an interleaved fashion.

We study the issue of whether transferring information about student skill across activities leads to better prediction of performance within the context of a standard algorithm for student modeling, Bayesian Knowledge Tracing (Corbett and Anderson 1995), to track student development of specific cognitive skills across all three activities. In recent years, Corbett and Anderson's Bayesian Knowledge Tracing model (Corbett and Anderson 2008) has been used to model student knowledge of a variety of types in a variety of systems, including tutors for mathematics (Koedinger 2002), computer programming (Corbett and Anderson 2008), scientific inquiry skill (Sao Pedro et al., 2013), and reading skill (Beck and Chang 2007). BKT is statistically equivalent to both a Hidden Markov Model, and to a two-node Dynamic Bayesian network (Reye 2004). Bayesian Knowledge Tracing (BKT) keeps a running assessment of the probability that a student currently knows each skill, continually updating the estimate based on student behavior. BKT is known to be effective, comparable to any other algorithm currently in active use (Pardos et al., 2011). BKT can predict future student performance within a given activity (Corbett & Anderson, 1995; Baker et al., 2011; Pavlik et al., 2011), can predict post-test use of the same skill in other modalities (Corbett & Anderson, 1995; Corbett & Bhatnagar, 1997; Sao Pedro et al., 2013), and when applied to data from an entire year of mathematics

instruction, BKT can predict standardized exam scores (Feng, Heffernan, & Koedinger, year; Pardos et al., 2013) and future college attendance (San Pedro et al., 2013). Other algorithms have in some cases been used to model strength of knowledge for fluency (e.g. Pavlik & Anderson, 2008), but this achieves a substantially different goal than BKT, which measures whether a skill or concept is known at all. In addition, Cognitive Mastery Learning built on top of Bayesian Knowledge Tracing has been shown to significantly improve student learning (Corbett, 2001). In one of the few examples of the use of BKT or similar models to study learning across contexts, Sao Pedro et al. (2013) used BKT to model science inquiry skills such as designing a controlled experiment across two different scientific concepts taught in the same fashion. We extend this work by exploring transfer in a broader way, extending BKT to track student skill across three different types of activities that teach the same skills in substantially different fashions, and examine which of several possible model variants is most effective for doing this. In Bayesian Knowledge Tracing, below, we discuss the model variants that will be analyzed, and make an explicit hypothesis for which model variant will be most effective.

By studying how to most effectively track student skill across multiple activities, we can provide more accurate and complete assessment to teachers. In addition, better assessment of this sort will lead to more efficient optimization of student learning experiences, since better knowledge estimates lead to more targeted amounts of practice on each skill (cf. Cen, Koedinger, and Junker 2007).

2. Reasoning Mind

The work presented in this paper is conducted in the context of Reasoning Mind Genie 2, a blended learning mathematics curriculum for elementary and middle school students (current offerings cover the second through the sixth grades), which is implemented within classrooms with a teacher present. Reasoning Mind combines extensive professional development, a rigorous curriculum drawing from successful curricular design in Russia, and a game-like, internet-based interface. Reasoning Mind is designed to be used by students working independently (although in practice, as in most blended learning curricula, there is some degree of collaboration – cf. Schofield, 1995). The students' mathematics teacher was present in the classroom. The primary role of

the teacher in a Reasoning Mind classroom is performing interventions; teachers conduct pre-planned interventions based upon the data provided by the system, as well as assisting struggling students.

Student learning in Reasoning Mind takes place in "RM City," a virtual city where students engage in learning activities in different "buildings." The primary mode of study for students is "Guided Study," wherein they are guided by a pedagogical agent named "Genie" through a series of learning objectives. It is used by approximately 100,000 students a year, primarily in the Southern United States. The fifth and sixth grade curricula are "core" curricula; they replace the traditional mathematics class and are generally used for the students' entire scheduled mathematics instruction time, usually 3-5 days per week for 45-90 minutes each day.

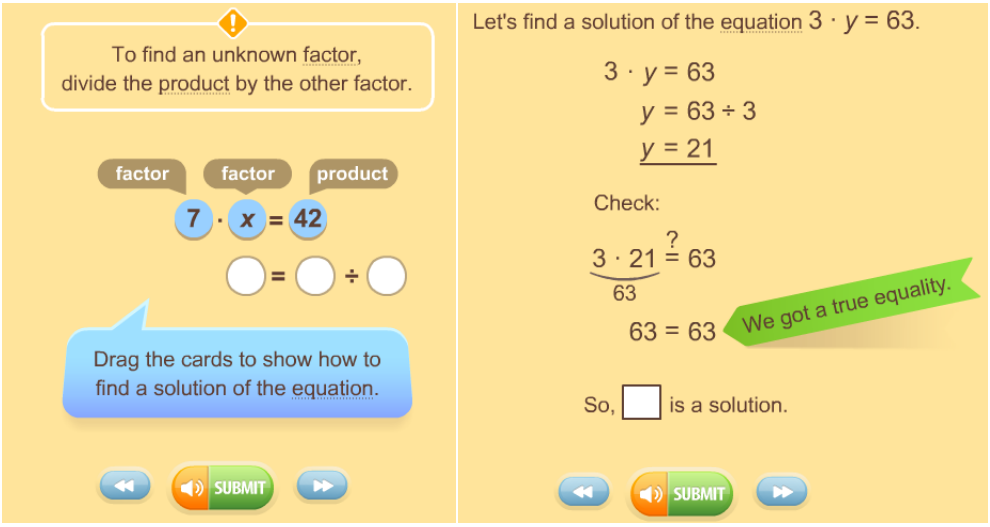

Figure 1. Examples of Theory items.

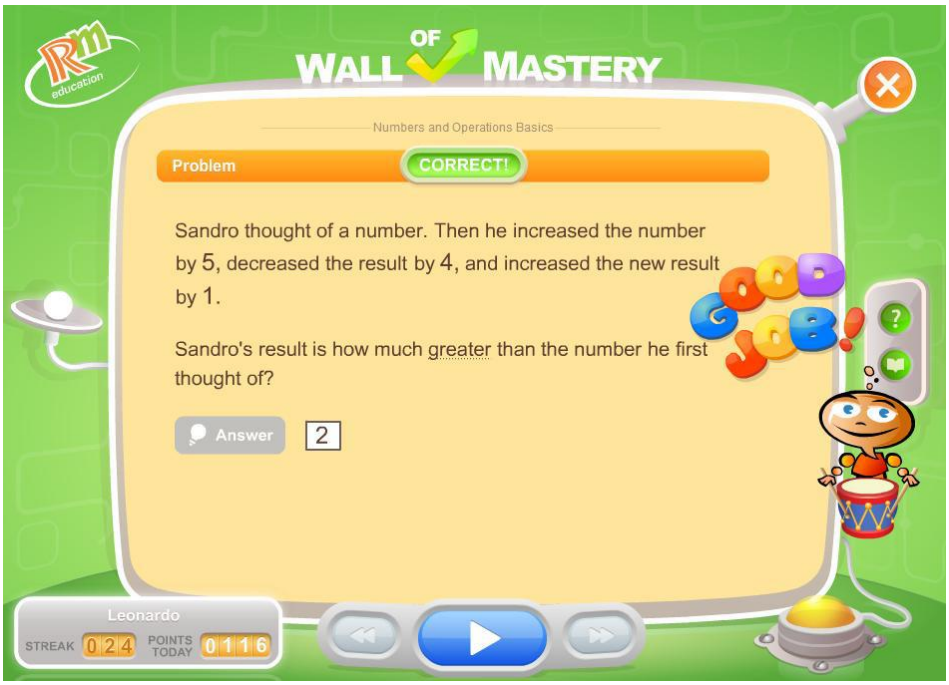

Figure 2. Example of a Problem.

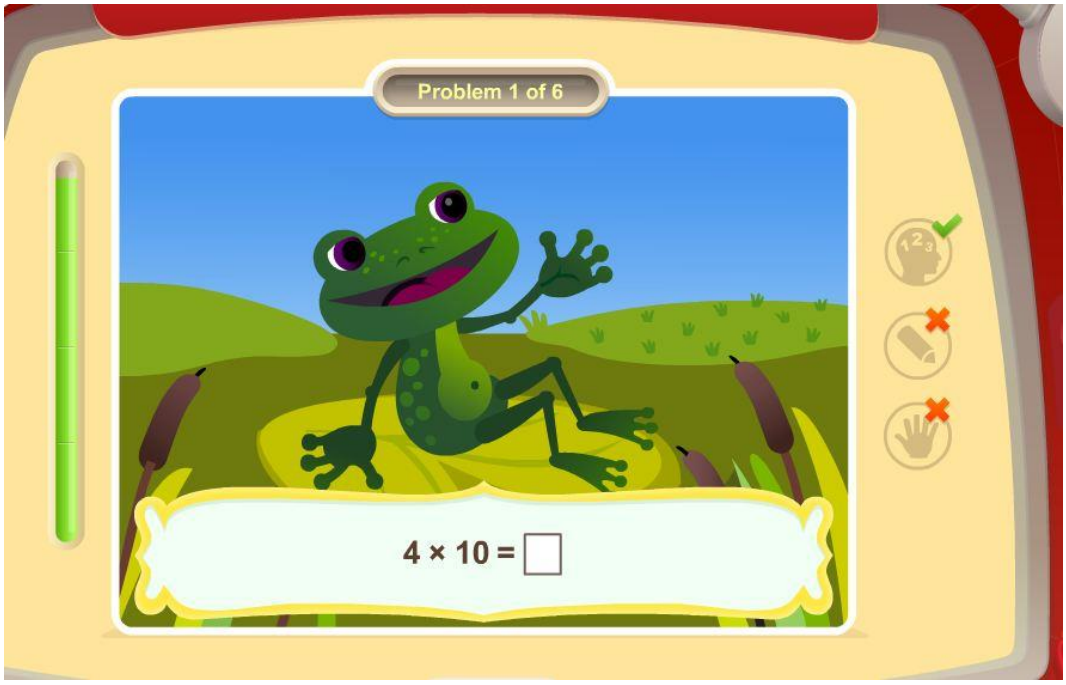

Figure 3. Example of a Speed Game item. The green bar on the left side of the image is a timer, and the icons on the right are reminders that Speed Game items are intended to be solved mentally.

Reasoning Mind has three types of activities where student knowledge of specific skills can be assessed: conceptual instruction ("Theory" items), and problem-solving ("Problems"), and tests of fluency presented in a game format ("Speed Games"). Theory items are exercises included during the instruction presented to the student, and are intended to help the student gauge his or her

understanding of the material. For example, the image on the left in *Figure 1* presents an instruction for finding an unknown factor and asks the student to show how that would be done using a concrete example; in the image on the right, a solution to an equation is found, and the student is asked to identify it. Problems are sequences of items given after the instructional block. They have varying levels of difficulty, but are intended to help the student practice and refine his or her knowledge; *Figure 2* shows a multi-step problem involving numbers and operations. Finally, Speed Games are timed, mental math problems which are intended to help the student develop fluency in mathematical operations with which they should already be proficient. An example is shown in *Figure 3*, which presents relatively simply multiplication problems to the student but requires them to answer in a very limited amount of time.

The path of a student working through a Reasoning Mind objective is as follows: first, the student is given a set of warm-up Problems, which are tests of knowledge that is prerequisite to the objective being covered. Next, the student is presented with one or more section of Theory, during which they may be asked questions which help them gauge their progress. Finally, they progress through up to three levels of problems: each student must solve a set of the easiest problems, with more challenging problem sets being presented depending on their accuracy and their progress through the curriculum. Speed Games occur primarily within the warm-up section, but may also appear within the Theory section (when they serve an instructional purpose) or within the lowest-level block of Problems after the Theory section.

The structure of Reasoning Mind lessons was developed primarily by expert Russian mathematics teachers. The Russian mathematics curriculum is among the strongest in the world (cf. Milgram 2005), with, in particular, very high content knowledge among middle school mathematics teachers (Tatto et al. 2009), and thus makes an appropriate choice on which to base a learning system such as Reasoning Mind. Reasoning Mind employs very experienced mathematics teachers and utilizes cognitive modeling approaches to model, as faithfully as possible, their classroom practices (Khachatryan et al., in press).

Russian teachers begin each day with a warm-up activity, which is usually a timed session of mental math. These activities are modeled in the Reasoning Minds system as Speed Games, which are simply timed sequences of mental math

(or other fluency practice) items in an appealing visual form. These warm-ups serve two purposes: they prepare students for the coming material in a process the Russian teachers explicitly refer to as "activation of prior knowledge," and they help the student to develop automaticity and fluency, which is seen as an integral part of concept mastery in Russian curricula.

The Theory section is conceptual instruction; it is the translation of teachers' classroom instruction practices to a computer interface. Because classroom instruction is very interactive, the Theory section contains frequent questions, with different student responses triggering different responses by the system: for example, an incorrect answer will trigger the presentation of a detailed solution to the student, and particular incorrect answers that can be interpreted as evidence of misconceptions result in targeted feedback. Russian teachers implement mastery learning (cf. Bloom, 1968): students are expected to master each concept before moving on to more advanced tasks. They check student mastery by requiring students to perform to at least a minimal standard on demonstration of learning tasks such as tests and quizzes. The Problem section is the online representation of such learning tasks; students must pass at least the most basic level of Problems in order to progress to the next objective.

3. Method

3.1 Participants

The data collected for this paper were from students using the Reasoning Mind fifth grade curriculum in the classroom as their regular curriculum, 5 days a week.

To increase the probability of model generalizability, data was collected from a diverse sample of students in six schools, broadly representative of the population currently using Reasoning Mind. Five of the six schools were in the Texas Gulf Coast region. Three of these Texas schools were in urban locations and served economically disadvantaged populations (defined as a high proportion of students receiving free or reduced lunch); of these three, two served predominantly African-American student populations, and one served a predominantly Hispanic student population. The other two schools in this region were in suburban locations, one serving mostly White students, and the other with

a mix of student ethnicities; both of these schools had a lower proportion of economically disadvantaged students. The sixth school was a rural school in West Virginia, with an economically disadvantaged, majority White population.

We collected these students' entire data across the entire year. The total data set involved 1,951,829 distinct actions across 462 students, across 172 knowledge components (KCs, defined below), within 45 content objectives, over the course of the entire fifth grade year.

3.2 Bayesian Knowledge Tracing

Corbett and Anderson's Bayesian Knowledge Tracing model (Corbett and Anderson 2008) computes the probability that a student knows a given skill at a given time, combining data on the student's performance up to that point with four model parameters. In the model's canonical form, each problem step in the tutor is associated with a single cognitive skill. The model assumes that at any given opportunity to demonstrate a skill, a student either knows the skill or does not know the skill, and may either give a correct or incorrect response (help requests are treated as incorrect by the model). A student who does not know a skill generally will give an incorrect response, but there is a certain probability (called P(G), the Guess parameter) that the student will give a correct response. Correspondingly, a student who does know a skill generally will give a correct response, but there is a certain probability (called P(S), the Slip parameter) that the student will give an incorrect response. At the beginning of using the tutor, each student has an initial probability ($P(L_0)$) of knowing each skill, and at each opportunity to practice a skill the student does not know, the student has a certain probability (P(T)) of learning the skill, regardless of whether their answer is correct.

The system's estimate that a student knows a skill is continually updated, every time the student gives a first response (correct, incorrect, or a help request) to a problem step. First, the system re-calculates the probability that the student knew the skill before the response, using the evidence from the response (help requests are treated as evidence that the student does not know the skill), using the first two equations of Figure 1. Then, the system accounts for the possibility that the student learned the skill during the problem step, using the third equation of Figure 1. Within the Cognitive Mastery algorithm used in most Cognitive Tutors

(Corbett and Anderson 2008), the student is assigned additional problems on skills that the system does not yet believe that the student has learned (e.g. skills that the student has less than 95% probability of knowing).

$$P(L_{n-1}|Correct_n) = \frac{P(L_{n-1}) * (1 - P(S))}{P(L_{n-1}) * \big(1 - P(S)\big) + \ (1 - P(L_{n-1})) * (P(G))}$$

$$P(L_{n-1}|Incorrect_n) = \frac{P(L_{n-1}) * P(S)}{P(L_{n-1}) * P(S) + \ (1 - P(L_{n-1})) * (1 - P(G))}$$

$$P(L_n|Action_n) = \ P(L_{n-1}|Action_n) + \big((1 - P(L_{n-1}|Action_n)\big) * P(T))$$

Figure 1. The equations used to predict student knowledge from behavior in Bayesian Knowledge Tracing.

The four parameters in Bayesian Knowledge Tracing are fit, for each skill, using data from students using that skill within an intelligent tutor. The goal during parameter fitting is to figure out which combination of parameters best predicts the pattern of correct and incorrect responses in the existing data, and then to use that model to make predictions about new students' knowledge as they use the tutor.

Within this paper, we study sixteen model variants of BKT, expressing a range of ways to model skill manifesting across three different activities. For each activity, we consider three different ways to model how student skill learning in this activity may influence learning and performance in the other activities. Each activity can be 1) fully included in the model (with student performance resulting in changing a unified estimate of student knowledge across activities, and each encounter with the activity serving as an opportunity for the student to learn), 2) only partially included (with student knowledge treated as separate in different activities, but each encounter still potentially causing student learning across activities), or completely separated from other activities. Table 2 below indicates how this set of possibilities across activities is mapped in 16 possible models. We hypothesize that uniform application of the full BKT model across the three activities will yield the best result, as this variant makes full use of all student data for each knowledge component.

3.3 Data Set

A collection of BKT models was constructed for Basic II, Reasoning Mind's fifth grade curriculum, which is made up of 45 distinct objectives, each taking 1-2 hours for the student to complete. Items in this curriculum were labeled with 172 knowledge components (KCs) by knowledge engineering. As is standard for Bayesian Knowledge Tracing, only first attempts at each item were included, regardless of what type of item it was. As discussed above, the full data set consisted of 1,951,829 distinct actions made by 462 students over the course of the year.

Each action in the log was labeled with the item type, either Problem, Speed Game, or Theory. The data set contained 814,805 Problem items, 358,749 Speed Game items, and 778,275 Theory items. Not all Theory actions within the log corresponded to student correct or incorrect attempts; approximately 26% of the time, they just represented presentation of the information to the student. In these cases, for the purposes of BKT, each action could be treated as an opportunity for the student to learn (and in this case, each action would affect the student's knowledge estimate for the relevant KC via the transition probability P(T)), but because there was no student action that could be marked right or wrong, this 26% of Theory actions could not be used to update the probability of student knowledge based on student performance. Note that in the remaining 74% of cases, the student does provide a correct or incorrect response, and BKT can be affected in both fashions.

3.4 Fitting BKT Parameters

A variation on Baker et al.'s Brute Force BKT model fitting code (Baker et al. 2010) was used. In the original Brute Force method, a very large number of combinations of BKT parameters are tested (down to a resolution of 0.001). While this is feasible when fitting a traditional BKT model with four parameters per KC, its running time is exponential in the number of parameters per KC, and thus takes a very large amount of computational time for larger numbers. Because this study involved fitting up to ten separate parameters (see below), a simulated annealing procedure was used rather than brute force. In this algorithm, BKT algorithms are initially seeded with random values. An initial RMSE value is

calculated for using that set of parameters used to predict student performance via BKT.

For each step in the process, a random one of these values is incremented or decremented by a random amount between 0 and 0.05, and a new RMSE value is calculated. This change in value can then be accepted with probability

$$P_{\text{accept}}(\bar{\theta}_{\text{old}}, \bar{\theta}_{\text{new}}) = \min\left[1, \exp\left(\frac{\epsilon(\bar{\theta}_{\text{old}}) - \epsilon(\bar{\theta}_{\text{new}}}{\tau}\right)\right]$$

In this equation, $\bar{\theta}_{\text{old}}$ is the original set of parameters, $\bar{\theta}_{\text{new}}$ is the set of parameters after the random change, $\epsilon(\bar{\theta})$ is the RMSE of student predictions using the parameter set $\bar{\theta}$ in BKT, and τ is an adjustable parameter. This parameter was initiated to 0.005, and halved every 10,000 annealing steps. If the move is accepted, the process is repeated with the new parameter set; otherwise, a new random change on the original set is attempted.

The parameter set resulting in the lowest achieved RMSE is tracked. Every 10,000 annealing steps, the current lowest RMSE is compared to the best RMSE at the end of the previous 10,000 steps; if it is unchanged, calculation is assumed to have converged and annealing ends.

This method produces results as good as those produced by the original Brute Force method, but significantly faster (in about 2% of the total time). Because of this speed, despite the fact that the τ parameter above allows for escaping local minima, this method could be coupled with random restarts to ensure that a global minimum is found; in practice, this was found to be unnecessary for the data set and parameters analyzed here.

To evaluate model fit, model predictions were generated using the fit parameters, and A′ was calculated. A′ is the probability that, given one correct and one incorrect student attempt, the model can correctly identify which is which. It is mathematically equivalent to the area under the ROC curve (AUC) used in signal detection and to W, the Wilcoxon statistic (Hanley and McNeil 1982). A value of 0.5 for A′ indicates performance exactly at chance, and a value of 1 indicates perfect performance. In these analyses A′ was calculated at the level of students, rather than actions, and all reported values are computed using ten-fold cross-validation. A′ was calculated using Baker et al.'s "A′ per Student" code (Baker et al. 2008), available from http://www.columbia.edu/~rsb2162/edmtools.html. Throughout this paper, we

also use calculations of statistical significance computed using this package. For statistical significance, a separate statistical comparison was made for each student using a Z test, and then these comparisons were aggregated using Stouffer's Z. Using this approach addresses non-independence assumptions better than conducting a single Z test across all data points together -- see discussion of this issue in (Baker et al., 2008).

To ensure that this method produces good results, it was applied to known data sets and compared with the result from the Brute Force Algorithm. The "Algebra I 2006-2007" and "Algebra I 2008-2009" used in the 2010 KDD Cup were retrieved from http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp, and in each case the "training" data set was used. These data sets contained student logs from the Algebra I Cognitive Tutor for the indicated school years. The results are shown in Table 1. For each data set, the simulated annealing algorithm took approximately 2% of the total computational time of the brute force algorithm, and produced predictions that were at least as accurate, as determined by ten-fold cross-validated, student-level A′. In fact, while the difference in A′ values for the Algebra 2006-2007 data set was not statistically significant ($p > 0.1$), for the Algebra 2008-2009 data set, the result of the simulated annealing algorithm was better with $p < 0.001$. The better performance may be due to the fact that the simulated annealing algorithm is not constrained to a grid, and so is able to more precisely determine the optimal parameters (but is less prone to local minima that the expectation maximization algorithm).

Table 1

Comparison of Results of Simulated Annealing and Brute Force Fitting BKT to KDD Cup Data

Data Set	N	Time ratio	Simulated Annealing A′	Brute Force A′
Algebra I 2006-2007	1,852,340	0.019	0.6705	0.6711
Algebra I 2008-2009	8,648,976	0.018	0.7485	0.7475

Note. N is the total number of actions in the data set. The time ratio is total computational time for fitting BKT parameters using simulated annealing vs. brute force.

3.5 Different Item Types

There are many different ways that the separate item types described above could be treated in applying BKT to the Reasoning Mind data set. For example, the item type label could be ignored and all items could be treated in the same manner, or some but not all parameters could be fit differently for different types of items, or some subset of item types could be ignored altogether. In this paper, sixteen different scenarios for treating separate item types are considered. The scenarios are shown in *Note*. In the columns labeled "Learning Opportunity," an 'X' indicates whether Speed Games (SG) and/or Theory (T) items were included as opportunities to learn – that is, there was a probability that the student would transition from not knowing to knowing the skill each of those items. The "Performance Included" column indicates whether the estimate of student knowledge was updated based on performance on the indicated item types. The two rightmost columns indicate whether separate transition probabilities P(T), in the first, and/or guess probabilities P(G) and slip probabilities P(S), in the second, were fit for each item type (if not, a single value was applied across all item types).

Parameters were fit and evaluated under ten-fold cross-validation. A′ values were calculated by evaluating performance of the resulting BKT model on the Problem, Speed Game, and Theory items independently, and calculated at the student level.

Table 2

Summary of BKT Modeling Scenarios

	Learning Opportunity		Performance Included		Separate	Separate
	SG	**T**	**SG**	**T**	p(T)	p(G), p(S)
1						
2	X					
3	X				X	
4	X		X			
5	X		X		X	
6	X		X		X	X
7		X				
8		X			X	
9		X		X		
10		X		X	X	
11		X		X	X	X
12	X	X				
13	X	X			X	
14	X	X	X	X		
15	X	X	X	X	X	
16	X	X	X	X	X	X

Note. Problems were always included as learning opportunities and performance on problems was always included, and therefore were not included in this table. See the text for a more detailed explanation of this table.

4. Results

Error! Reference source not found. summarizes the resulting model goodnesses for each of the scenarios described above. Ten-fold cross-validated, student-level A′ values between 0.633 and 0.664 were obtained on Problem items. The highest A′ on Problems was attained for Scenario 2, which included Problems and Speed Games as learning items (using the same transition probability P(T) for both), but updated BKT estimates based on performance only for Problems. Scenario 2 was very slightly higher than Scenario 1. The A' value for Scenario 2 was statistically significantly higher than nine scenarios, specifically those that incorporated performance on non-problem items. It was not statistically significantly higher than the A' in the scenarios where non-problem item performance was not considered. Scenarios 1, 3, and 8 also performed statistically significantly better than the nine scenarios which included performance on Speed Games or Theory. This set of findings suggests that students can learn problem-solving skill from completing the Speed Games, but that actual performance on the Speed Games is not relevant to success in the problems. Completion of the

Theory problems does not seem to influence student success on the Problems in either fashion. These results broadly contradict our initial hypothesis, that modeling student skill within one activity and transferring that information to a second activity would lead to better prediction of student performance within the second activity.

Ten-fold cross-validated student-level A′ values between 0.567 and 0.679 were obtained on Speed Games items. Note that A′ was not calculated for Speed Games for all scenarios; in particular, if performance on Speed Games was not used to drive Bayesian updating, then A′ was not computed. The best-performing models for Speed Games was Scenario 4, which updated Bayesian estimates based on performance on both the Speed Games and Problems, as well as applying P(T) whenever either a Speed Game or Problem was completed. Scenario 5 was only very slightly lower than Scenario 4, suggesting that whether or not a different P(T) was used for Speed Games or Problems (scenario 4 versus 5) had little impact on A′. Performance decreased when data from Theory problems was incorporated, and when separate P(G) and P(S) were used for the Speed Games versus the Problems, probably due to over-fitting. When computing A′ for the Problems, scenarios 4 and 5 were only slightly (though statistically significantly) worse than scenario 2, the best scenario for Problems, at 0.655 and 0.656 compared to 0.664.

Ten-fold cross-validated student-level A′ values between 0.613 and 0.655 were obtained on Theory items. Note that A′ was not calculated for Theory for all scenarios; in particular, if performance on Theory was not used to drive Bayesian updating, then A′ was not computed. The best-performing scenario for Theory was scenario 11, which updated Bayesian estimates based on performance on both the Theory and Problems, as well as applying P(T) whenever either a Theory section or Problem was completed. In this model, there were separate P(T), P(G), and P(S) for Theory and Problems, suggesting that the Theory sections and Problems have very different properties. Incorporating data from Speed Games led to only minor degradation to performance on Theory (0.648 versus 0.655), but using the same P(G) and P(S) for all three types of data led to more substantial degradation (0.613). However, using the same P(G) and P(S) for both Theory and Problems led to less degradation (0.641) than using the same P(G) and P(S) for all

three. Also, using the same P(T) for both Theory and Problems led to only minor degradation (0.642).

Table 3

Summary of Results for Each Modeling Method

	Parameters per KC	Number of relevant items	A′ (Problems)	A′ (Speed Games)	A′ (Theory)
Problems only					
1	4	3527	0.664	n/a	n/a
Problems and Speed Games					
2	4	3527	0.664[a]	n/a	n/a
3	5	3527	0.660[a]	n/a	n/a
4	4	4304	0.655[b]	0.679[a]	n/a
5	5	4304	0.656[b]	0.679[a]	n/a
6	7	4304	0.648[b]	0.672[c]	n/a
Problems and Theory					
7	4	3527	0.662[a]	n/a	n/a
8	5	3527	0.661[a]	n/a	n/a
9	4	5212	0.642[b]	n/a	0.642[d]
10	5	5212	0.643[b]	n/a	0.641[d]
11	7	5212	0.641[b]	n/a	0.655[a]
Problems, Speed Games, and Theory					
12	4	3527	0.662[a]	n/a	n/a
13	6	3527	0.660[a]	n/a	n/a
14	4	5212	0.636[b]	0.669[d]	0.638[e]
15	6	5212	0.640[b]	0.567[b]	0.613[b]
16	10	5212	0.633[b]	0.675[c]	0.648[c]

Note. Number of relevant items refers to the average number of items per student which were used in calculating the estimates of student knowledge. Model goodness is measured by cross-validated, student-level A′ on each of Problems, Speed Games, and Theory items. "N/a" indicates that performance on the specified item type was not included in the model, making it inappropriate to compute performance for this item type on this model.

[a] Performed statistically significantly better than [b] and [d].

[b] Performed statistically significantly worse than [a], [c], [d], and [e].

[c] Performed statistically significantly better than [b] and [e].

[d] Performed statistically significantly worse than [a] but better than [b].

e Performed statistically significantly worse than [c] but better than [b].

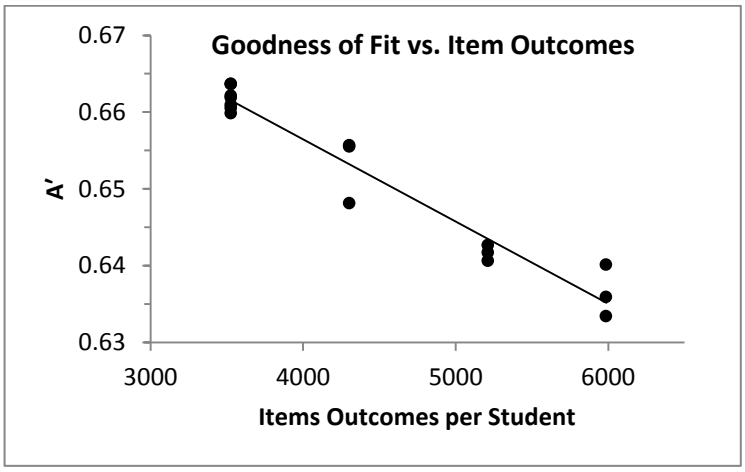

Figure 4. The ten-fold cross-validated, student-level A′ on Problems as a function of the total number of relevant items for each model. Each dot represents a scenario; the line is a linear fit with $R^2 = 0.95$.

To understand the impact of incorporating additional information on the system's ability to assess student knowledge on Problem items, A′ values on Problems were compared to the number of items on which the estimate of student knowledge was updated based on student performance (e.g. the number of items relevant to each model). The result is shown in Figure 4. A very strong negative correlation was found, with $R^2 = 0.95$. This indicates that using additional information is not beneficial for modeling student knowledge as relevant to the Problems, although the magnitude of the performance reduction is small. It is worth noting that this is not just a function of having more parameters. When A′ values on Problems were compared to the number of parameters per KC, a negative correlation was again found, but with a much smaller $R^2 = 0.29$.

5. Discussion and Conclusions

In this paper, we extend Bayesian Knowledge Tracing (BKT), a classic student modeling algorithm, to track student knowledge on tests of fluency ("Speed Games"), conceptual instruction ("Theory" items), and problem-solving ("Problems"). We found that for the purposes of predicting student performance

on Speed Games, they were most effectively modeled using the same BKT parameters as Problems; conversely, Theory items were best predicted using a different set of parameters. In specific, the use of separate P(G) and P(S) values for Speed Games (when not including Theory) led to a decrease in A′ on Speed Games, while the opposite was true for Theory. This seems to indicate that the skills being learned in Speed Games have similar properties to the skill being learned in Problems with regards to knowledge tracing, while Theory items have some very different properties.

One possible explanation for these differences is the nature of the items. Theory items are given in the course of instruction on a new topic, and are given early in the student learning a topic; they are used primarily for the student to gauge his or her progress in learning the topic and secondarily to drive the direction of the instruction provided by the system. In contrast, both Problems and Speed Games occur later in the learning process: they are intended to help a student refine and solidify knowledge that the student has already begun to attain.

However, best performance was obtained if each type of item was treated as representing parallel performance, e.g. not adjusting student knowledge on Problems based on performance in Speed Games. In particular, the more items which are considered for skill assessment beyond Problems, the worse prediction of student performance on Problems becomes, with $R^2 = 0.95$. This combination of findings suggests that the corresponding skills (between Problems and Speed Games) are similar but that knowledge is not transferring between these situations.

On the whole, these results suggest that developing models which can generally model student skill and knowledge across a variety of contexts and problem types is a challenging problem. The approach here was successful at articulating the differences between these three types of activities, but was not successful at actually creating a model general across them, that leverages information from all three activities in predicting each of them. At this point, it is not clear whether another approach could be successful at this goal, or whether the seemingly similar skills represented by each activity type actually do represent fundamentally different skills. Understanding better which of these hypotheses is correct would be valuable for better understanding not just student modeling of mathematical knowledge, but also the fundamental nature of what students are

learning in Reasoning Mind. A clear limitation to this work is its application solely in one learning system and set of activities. Prior work has shown that some unification of skills is possible in tracking the same science inquiry skill across different science content domains but in the same activity (e.g. Sao Pedro et al., 2013). Figuring out how broadly skills can be unified within student models, and what factors inhibit or promote this process, will be an important area of future research, only feasible through studying these issues in a variety of contexts.

Within the specific context of Reasoning Mind, a system currently being used at scale in U.S. classrooms, these results can provide guidance for determining the best method by which to trace student knowledge depending on the precise goal of the knowledge tracing application; for example, if one's goal is the most accurate prediction of student performance on Problems, then knowledge tracing using only Problems may be the best algorithm. This is useful information, to the degree that it helps to optimize student practice within each activity, helping improve learning and learning efficiency (Cen, Koedinger, & Junker, 2007). Indeed, even small improvements in knowledge estimation can have large benefits for optimizing practice, indicating that it is important in Reasoning Mind to optimize practice based on estimation specific to a given activity.

However, there may be contexts in which more general methods are more appropriate, even if their internal predictive power is weaker. For example, if the goal of a learning system is to predict robust learning that transfers to new contexts, then it is likely that a model taking into account a variety of contexts would more readily generalize. For example, in long-term prediction of student success (e.g. Pardos et al., 2013; San Pedro et al., 2013), a more general model is likely to be more predictive in different activities. Similarly, a general estimate of student knowledge, independent of the specific activity, may be desired for the purpose of diagnosis and remediation of gaps and misconceptions in student knowledge. Teachers may be more interested in generally knowing what skills students have, rather than how likely they are to succeed in specific activities. As such, a broader model, or one focusing on a different set of activities, may be most relevant in this case.

In the long-term, the work presented here represents a step towards developing curricula that most effectively utilize and integrate multiple learning

activities, towards helping students develop robust and general mathematics knowledge and skill.

Acknowledgements

We thank support from the Bill and Melinda Gates Foundation, and also thank George Khachatryan for valuable suggestions and comments, and Belinda Yew for assistance in literature review.

References

Aleven, V., & Koedinger, K.R. (2002). An effective metacognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor. Cognitive Science, 26 (2), 147-179.

Archibald, T.N. & Poltrack, J. (2011) ADL: Preliminary Systems and Content Integration Research within the Next Generation Learning Environment. *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC).*

Baker, R.S.J.d., & Clarke-Midura, J. (2013) Predicting Successful Inquiry Learning in a Virtual Performance Assessment for Science. *Proceedings of the 21st International Conference on User Modeling, Adaptation, and Personalization*, 203-214.

Baker, R.S.J.d., Gowda, S.M., & Corbett, A.T. (2011) Automatically Detecting a Student's Preparation for Future Learning: Help Use is Key. *Proceedings of the 4th International Conference on Educational Data Mining*, 179-188.

Beck, J.E., & Chang, K.-m. (2007). Identifiability: A Fundamental Problem of Student Modeling. In Proceedings of the 11th International Conference on User Modeling (UM 2007).

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment,* 1(2), 1–12.

Cen, H., Koedinger, K., & Junker, B. (2007). Is over practice necessary? – Improving learning efficiency with the cognitive tutor through educational data mining. In Rose Luckin and Ken Koedinger (Eds.) Proceedings of the 13th International Conference on Artificial Intelligence in Education (pp. 511-518). Amsterdam: IOS Press.
Corbett, A.T., & Anderson, J.R. (1995). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. User Modeling and User-Adapted Interaction, 4, 253-278.

Gobert, J.D., Sao Pedro, M., Raziuddin, J., & Baker, R. (2013) From Log Files to Assessment Metrics: Measuring Students' Science Inquiry Skills Using Educational Data Mining. *Journal of the Learning Sciences*, 22 (4), 521-563.

Greiff, S., & Funke, J. (2009). Measuring complex problem solving—The MicroDYN approach.
In F. Scheuermann (Ed.), The transition to computer-based assessment—Lessons learned from largescale surveys and implications for testing. Luxembourg, Luxembourg: Office for Official Publications of the European Communities.

Khachatryan, G., Romashov, A., Khachatryan, A., Gaudino, S., Khachatryan, J., Guarian, K., & Yufa, N. (in press). Reasoning Mind Genie 2: An Intelligent Learning System as a Vehicle for International Transfer of Instructional Methods in Mathematics. *International Journal of Artificial Intelligence in Education*.

Koedinger, K.R. (2002). Toward evidence for instructional design principles: Examples from Cognitive Tutor Math 6. In Proceedings of PME-NA XXXIII (the North American Chapter of the International Group for the Psychology of Mathematics Eduation).

Koedinger, K.R., Corbett, A.C., & Perfetti, C. (2012). The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36 (5),*757-798.

Lomas D., Ching D., Stampfer, E., Sandoval, M., & Koedinger, K. (2012). Battleship Numberline: A Digital Game for Improving Estimation Accuracy on Fraction Number Lines. Conference of the American Education Research Association (AERA).

Maass, J. K., Pavlik, P. I. (2013) Using Learner Modeling to Determine Effective Conditions of Learning for Optimal Transfer. Proceedings of the 16th International Conference on Artificial Intelligence in Education, Memphis, TN, 189-198.

McQuiggan, S. W., Rowe, J. P., Lee, S., & Lester, J. C. (2008). Story-based learning: The impact of narrative on learning experiences and outcomes. Proceedings of the 9th International Conference on Intelligent Tutoring Systems, 530-539.

Milgram, R. J. (2005). *The Mathematics Pre-Service Teachers Need to Know.* Retrieved from: ftp://math.stanford.edu/pub/papers/milgram/FIE-book.pdf.

Molnár, G., Greiff, S., & Csapó, B. (2013). Inductive reasoning, domain specific and complex problem solving: Relations and development. Thinking Skills and Creativity, 9, 35–45.

Pardos, Z.A., Baker, R.S.J.d., Gowda, S.M., & Heffernan, N.T. (2011). The Sum is Greater than the Parts: Ensembling Models of Student Knowledge in Educational Software. *SIGKDD Explorations, 13* (2), 37-44.

Pardos, Z.A., Baker, R.S.J.d., San Pedro, M.O.C.Z., Gowda, S.M., & Gowda, S.M. (2013) Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In Proceedings of the 3rd International Conference on Learning Analytics and Knowledge (pp. 117-124). Washington, DC: Association for Computing Machinery.

Pavlik, P. L. and Anderson, J. R. (2011) Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, 14 (2), 101-117

Quellmalz, E., Timms, M., & Schneider, S. (2009). Assessment Of Student Learning In Science Simulations And Games. National Research Council Report, Washington, D.C.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K. R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T.E., Upalekar. R, Walonoski, J.A., Macasek. M.A. & Rasmussen, K.P. (2005). The Assistment project: Blending assessment and assisting. In C.K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.) Proceedings of the 12th Artificial Intelligence in Education, Amsterdam: ISO Press. pp. 555-562.

Reye, J. (2004). Student Modeling based on Belief Networks. International Journal of Artificial Intelligence in Education, 14, 1-33.

San Pedro, M.O.Z., Baker, R.S.J.d., Bowers, A.J., & Heffernan, N.T. (2013) Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. In Proceedings of the 6th International Conference on Educational Data Mining (pp. 177-184). Worcester, MA: International Educational Data Mining Society.

Sao Pedro, M.A., Baker, R.S.J.d., Gobert, J., Montalvo, O. Nakama, A. (2013) Leveraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. *User Modeling and User-Adapted Interaction, 23* (1), 1-39.

Sao Pedro, M., Baker, R., & Gobert, J. (2013). Incorporating Scaffolding and Tutor Context into Bayesian Knowledge Tracing to Predict Inquiry Skill Acquisition. *Proceedings of the 6th International Conference on Educational Data Mining*, 185-192.

Schofield, J.W. (1995) Computers and Classroom Culture. Cambridge, UK: Cambridge University Press.

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., & Salden, R. (2009). The worked-example effect: Not an artefact of lousy control conditions. *Computers in Human Behavior, 25*, 258–266

Tatto, M. T., Schwille, J. S., Senk, S., Ingvarson, L. C., Peck, R., & Rowley, G. L. (2009). *Teacher education and development study in mathematics (TEDS-M): Conceptual framework*. Amsterdam, Netherlands: International Association for the Evaluation of Educational Achievement.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex Problem Solving. More than reasoning? *Intelligence, 40*, 1-14.