Single Template vs. Multiple Templates: Examining the Effects of Problem Format on Performance

Yang Jiang, Educational Testing Service, yjiang002@ets.org
Ma. Victoria Almeda, TERC, mia_almeda@terc.edu
Shimin Kai, Teachers College Columbia University, smk2184@tc.columbia.edu
Ryan S. Baker, University of Pennsylvania, rybaker@upenn.edu
Korinn Ostrow, Worcester Polytechnic Institute, korinn.ostrow@gmail.com
Paul Salvador Inventado, California State University Fullerton, pinventado@fullerton.edu
Peter Scupelli, Carnegie Mellon University, pgs@andrew.cmu.edu

Abstract: Classroom and lab-based research have shown the advantages of exposing students to a variety of problems with format differences between them, compared to giving students problem sets with a single problem format. In this paper, we investigate whether this approach can be effectively deployed in an intelligent tutoring system, which affords the opportunity to automatically generate and adapt problem content for practice and assessment purposes. We conducted a randomized controlled trial to compare students who practiced problems based on a single template to students who practiced problems based on multiple templates within the same intelligent tutoring system. No conclusive evidence was found for differences in the two conditions on students' post-test performance and hint request behavior. However, students who saw multiple templates spent more time answering practice items compared to students who solved problems of a single structure, making the same degree of progress but taking longer to do so.

Introduction

Mathematics content can be encountered in a number of formats, presentations, and situations. For instance, the same number can be represented as a fraction, a percentage, a mixed number, or a decimal, and can pertain to money, portions of an item, a sports scenario, or many other real-life situations. Yet often the practice problems given to students tend to be represented in a specific format or form of representation, which may hinder the degree to which students see how the mathematical concepts are general in scope (e.g., Chang, Koedinger, & Lovett, 2003).

There has thus been a growing interest in developing better methods to help young students see patterns in the application of mathematics and transfer what they have learned to other similar problems and contexts (Chi & VanLehn, 2012; Quilici & Mayer, 1996). Too much similarity between the problems a student studies could lead to challenges in developing this ability, as most students tend to focus on surface features rather than the deep structural features (Chi & VanLehn, 2012). In math problems, the term "surface feature" is typically used to refer to the attributes described in a problem that are irrelevant to the underlying mathematical procedures needed for solution of the problem, such as cover story themes, colors, the wording and sentence structures in a word problem, or even mathematical features that do not change the problem-solving process (Chi & VanLehn, 2012; Quilici & Mayer, 1996). For example, changing a surface feature of a word problem could involve replacing the number of apples being distributed with the number of pencils. Structural features, on the other hand, refer to the concepts, rules, or principles that could directly impact a student's understanding of the mathematical processes needed to solve the problem (Chi & VanLehn, 2012). Examples of structural features include mathematical principles needed to generate solutions to problems such as the combination principle and the permutation principle, and features such as the number and types of variables involved in the problem (e.g. Quilici & Mayer, 1996). In other words, surface features are derived from the elements of the cover story while structural features determine solution procedures to the problem. Research on expertise showed that experts tend to recognize the deep structural features during problem solving while the novices tend to rely on salient irrelevant surface features to solve problems (Chi, Feltovich, & Glaser, 1981).

However, Ben-Zeev and Star (2001) found that even when pre-tests showed that students had a conceptual understanding of math concepts they still relied on surface features for cues on how to solve the problems. For instance, when asked to compare fractions with the presence of a logarithm or radical in the denominator, undergraduate Yale students still executed procedures commonly associated with logarithms and radicals even when these features were irrelevant to the underlying procedures needed to solve the problem.

In order to help students focus on the underlying similarities in the mathematical procedures across problems, researchers have suggested exposing students to a variety of problems, with different surface features

 © ISLS

but similar structural characteristics (Gick & Holyoak, 1983). In particular, Quilici and Mayer (1996) found that students were more likely to see structural features when they were presented with multiple examples of problems with different surface features (cover story used in the problems). Similarly, Chang, Koedinger, and Lovett (2003) found learning benefits associated with training students to solve varied boxplot problems with different cover stories and question wordings (varied condition), compared to solving boxplot problems with the same cover story and wording (spurious condition). When presented with problems with zero features matching to the materials seen in either study condition, students in the spurious group had significantly lower post-test scores than students in the varied group, suggesting that they were relying on spurious cues for solving problems.

Other forms of variety between problems have been found to facilitate student mathematical understanding. For example, McNeil and colleagues (2011) studied the impact of presenting practice arithmetic problems in a nontraditional format – such as $9 = 7 + 2$ instead of $7 + 2 = 9$ – on facilitating student understanding of mathematical equivalence, measured in a post-assessment.

With the increasing availability of intelligent tutoring systems (Ma, Adesope, Nesbit, & Liu, 2014), teachers now have the tools available to generate hundreds of practice problems using a single template with ease (Heffernan & Heffernan, 2014). This same functionality also makes it feasible to generate large numbers of problems comprising a variety of problem formats or representations. From the existing studies discussed above, it seems that exposing students to nontraditional problem formats or a variety of problem stories and formats could positively impact student learning outcomes. On the other hand, few studies thus far have examined whether problem-solving with multiple (or non-traditional) problem formats is sufficient if this problem-solving is not combined with additional instructional support or activities. For instance, McNeil and colleagues (2011) had human tutors conduct training sessions on both nontraditional and traditional problems. Additionally, studies by Quilici and Mayer (1996) and Ben-Zeev and Star (2001) asked students to sort math problems into meaningful groups, likely prompting students to review the features of each problem. This raises the question of whether simply providing problem-solving with multiple representations or cover stories – a relatively easy intervention – will be sufficient.

As such, the goal of this study is to explore how varying the presentation of problems influences students' math performance, both during practice and on post-test problems. Specifically, we hypothesize that being exposed to a variety of problem formats during practice within the ASSISTments platform will be associated with better student performance on a later post-test, as compared to when students only encounter one type of problem format during their practice. We study this issue in the ASSISTments learning platform by conducting a randomized controlled trial to compare differences in math performance during practice and on a post-test, between single template and multiple templates conditions.

ASSISTments

The ASSISTments platform (Heffernan & Heffernan, 2014) is a free online formative assessment and tutoring system for elementary, middle, and high school students. While ASSISTments can be used in a range of domains, it is primarily used for mathematics. Teachers use ASSISTments to assess students' knowledge of mathematical concepts and skills while facilitating their learning of these concepts. The system provides teachers with formative assessments of the students' learning progress in their acquisition of specific knowledge components. Currently, the ASSISTments platform has been adopted by 650 teachers across the United States, with an average of over 5,000 student users a day and over 50,000 student users a year. ASSISTments has a policy of allowing external researchers to conduct experimental studies within their platform, functionality that has been used by dozens of researchers in published papers.

We make use of data obtained from students' learning within Skill Builders (Heffernan & Heffernan, 2014), in which students complete several problems related to the same skill. In Skill Builders, students cannot proceed to the next problem until they submit the correct response. However, hints are available to assist them with problem solving. For each problem, students can attempt the problem multiple times and request multiple hints. After three or four hints, the last hint of each problem is a "bottom-out hint" that provides the correct answer to the problem.

Skill Builder problems assessing a skill are based on one or more problem templates. Each template represents a problem design that may be the basis for multiple problems by changing the values within each problem. Problem templates may differ from one another by having different cover stories or settings or providing different clues or structures to help students solve the problems. Some problem sets have a single template (16.71% of existing certified Skill Builders), with a single structure or word problem cover story for all problems, whereas other problem sets (the remaining 83.29%) have a variety of structures or several word problem cover stories. In the latter case, different problem templates within a single problem set may involve variants of a mathematical skill, or may require students to practice the skill on different forms of a problem. For example, a

problem set based on a single template could include problems of the same word problem cover story such as "Alan had 5 apples. He gave his brother 2 apples. How many apples does Alan have left?" The only differences between problems of the same template would be the numbers (Xiong, Adjei, & Heffernan, 2014) and in some cases the objects (e.g., replacing apples with pears). In contrast, a problem set based on multiple templates would include problems with a different cover story or structure, such as "Alan had 5 apples. He gave his brother some apples and now he only has 2 apples. How many apples did Alan give his brother?" Using a simple template, teachers can generate a large number of problems using the same cover story and wording but with different values in each problem.

In the next section, we discuss the research design for our randomized controlled trial, where we created multiple templates for a Skill Builder whose problems had previously been generated from a single template.

Methods

Research design

In order to investigate our research question, we selected a problem set, Solving System of Equations, from a pool of existing Skill Builder problem sets that had already been created with only one problem template. This problem set involved practice problems on one single mathematical concept, of solving linear algebraic equations by inspection. These problems correspond to the Common Core Standards 8.EE.C.8b EX (National Governors Association Center for Best Practices Council of Chief State School Officers, 2010) and aim to teach students to solve systems of two linear equations with two variables algebraically. We then expanded on this problem set by introducing seven additional templates to the same problem set. An example of a problem in the original template is shown below:

> Solve the following system of equations using linear combination.
>
> $2y + 7x = -2$
> $3y + x = 3$
>
> What is the value of x?
> (Enter as a fraction)

To create additional templates, we modified several parts of the linear equation in the original template. Because the practice problems are centered on solving algebraic equations, changes were made in the problem features that emphasize the principles of identifying equivalence between different algebraic expressions for dependent and independent variables. Specifically, we created additional problem templates by making the following changes:

1. Changing the variables from {x, y} to {a, b}
2. Switching the dependent variable from x to y
3. Changing the order of x and y.
   - e.g., {7x + 2y = −2; x + 3y = 3}
4. Changing the wording of the question. The following two templates were created with this change.
   a. Replace "What is the value of x? (Enter as a fraction)" in the original text with "x = ? Enter the value of x as a fraction."
   b. Change the wording to: "What is the value of x in the following system of equations? Use the linear combination to solve the equations and enter the answer as a fraction."
5. "Flipping" the equations such that x and y are on different sides of the equation (two templates were created with this change)
   a. {−2 − 7x = 2y; 3 − x = 3y}
   b. {−2 = 2y + 7x; 3 = 3y + x}

In total, seven additional problem templates were created. With these templates in addition to the original problem template, we created a problem set that generated randomized problems based on these templates, which would be attempted by students in the Multiple Templates (MT) condition. In other words, students in the Multiple Templates (MT) condition would attempt randomly generated problems from eight problem templates. Students in the Single Template (ST) control condition, on the other hand, would attempt randomly generated problems

based on one template — the original problem template. Figure 1 shows an example of the problems in the ST and MT conditions.

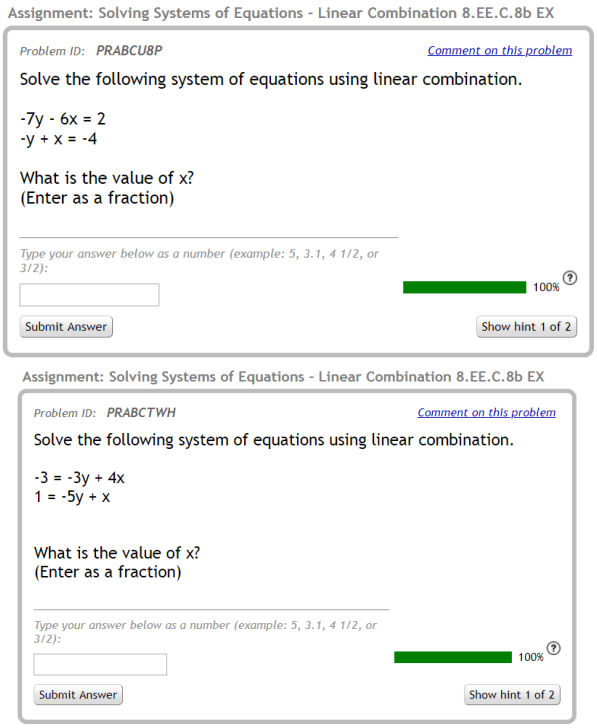

Figure 1. Screenshots of a problem in the ST condition based on the original template (see top), and a problem in the MT condition based on a modified template (template 5.b, see bottom) in ASSISTments.

Participants

294 eighth and tenth grade students from ten schools in the United States participated in this study during the 2017–2018 school year. They came from 17 math classes that were taught by 14 teachers.

Students were randomly assigned to the Single Template (ST) condition or the Multiple Templates (MT) condition when they began the problem set. In this study, 147 students were randomly assigned to the ST condition and 147 students were assigned to the MT condition. Only students who completed the Skill Builder were shown the post-test; students who did not complete the problem set were excluded from analysis. In total, 173 students completed the problem set and were included for analysis in this study. Specifically, 93 students who were assigned to the ST condition completed the problem set and 80 students in the MT condition completed the problem set. Seventy students were males, 73 were females, and 30 participants' gender was not reported. Despite the relatively low degree of completion, the completion rate was not statistically significantly different between the two conditions across all of the students who originally participated in our study, $\chi^2(1, N = 294) = 2.02$, $p = .155$.

Procedure

In this study, students in each condition (single template vs. multiple templates) were assigned to use ASSISTments to solve algebra problems, specifically systems of linear equations, as part of their regular math class or homework.

Student correctness on each item was evaluated by their first response attempt. They were considered to have mastered the skill involved in the problem set once they had correctly answered three consecutive questions using first attempts, the standard mastery criterion within the ASSISTments system (Heffernan & Heffernan, 2014). The students who completed the problem set needed an average of 6.42 ($SD = 4.21$) questions to achieve skill mastery. Upon achieving mastery and completing the problem set, students moved on to complete a two-item post-test.

Measures

This study compared the Single Template (ST) condition and the Multiple Templates (MT) condition on a set of measures related to students' use of ASSISTments and their performance in the platform. These measures were generated from the interaction log data produced by the 173 students who completed the assignment and included: 1) Mastery speed, 2) post-test performance, 3) hint usage during practice and post-test, and 4) time-based measures.

Mastery speed

As mentioned above, students in ASSISTments are considered to have mastered the problem set if they answer three questions correctly in a row on their first attempt at each question. A student's mastery speed is the number of problems attempted prior to the student achieving mastery (Xiong et al., 2014).

Post-test performance

Upon achieving mastery, students were asked to complete a post-test that was comprised of two items that were of different designs from templates used in the ST or MT conditions (see Figure 2). As was the case with the practice questions, hints and bottom-out hints were available to students during the post-test. Student performance on the post-test was evaluated by whether the student answered both post-test items correctly on their first attempt or not. Having only two post-test items, we adopted binary coding because correctly answering both items shows evidence of mastery of the skill, with less probability of correctness by chance. Students who accurately answered both questions were assigned a score of 1, while all other students were given a score of 0.

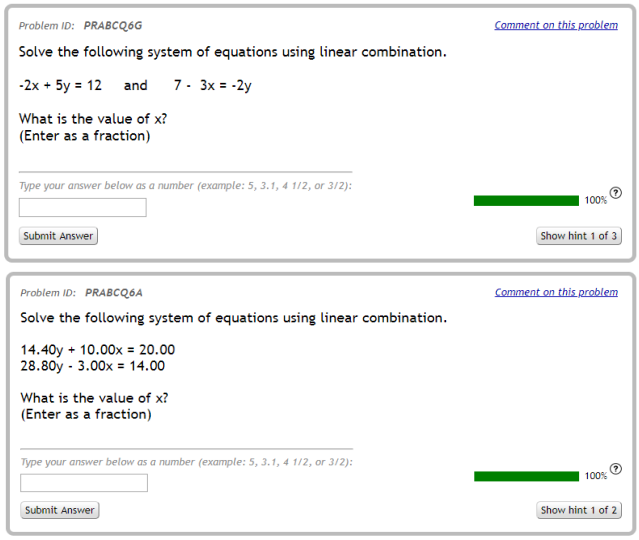

Figure 2. Screenshots of post-test problems.

Use of hints

As previously mentioned, hints and bottom-out hints are available to assist students with problem-solving in ASSISTments. Accordingly, we developed a series of measures representing hint usage, drawn from Inventado et al. (2016). These included measures of hint usage during the Skill Builder practice stage: 1) total number of hints requested during practice, 2) total number of bottom-out hints requested during practice, 3) average number of hints requested per problem during practice, and 4) average number of bottom-out hints requested per problem during practice. In addition, the total number of hints or bottom-out hints requested during the post-test was calculated and compared between groups.

Time

Time-based measures (computed in minutes) were computed for both the Skill Builder and post-test problems. The time-based measures computed were the total time spent on all of the Skill Builder problems attempted, the average time spent on each Skill Builder problem, and the total time spent on the post-test problems. Each of these three measures was computed separately for the first attempt response and the complete problem solution, which could include multiple response attempts and hint requests, resulting in six time-based measures (see the full list in Table 1). Similar to Inventado et al. (2016), the time measures on each problem were winsorized to a maximum of 15 minutes to exclude extreme cases possibly representing off-task behavior.

Data analysis

In this paper, a chi-square test was conducted to compare the post-test performance (a binary measure) between the students in the ST group who experienced one template and the MT group who experienced multiple templates. In addition, as data was not normally distributed, two-tailed Mann-Whitney U tests, a nonparametric alternative to the t-test, were conducted to compare the continuous measures such as mastery speed, the number of hint requests, and time between the two groups.

Results

Mastery speed

On average, students who were shown a single template showed a marginally significantly faster mastery speed ($M = 6.01$, $SD = 3.86$) than their counterparts who solved problems with multiple templates ($M = 6.90$, $SD = 4.57$), $U = 3184$, $p = .097$ (see Table 1).

Table 1: Comparison of the practice related measures (i.e., measures based on the Skill Builder problems) and the post-test related measures by condition.

Type	**Measure**	**ST**	**MT**	***p***
Practice	Mastery speed	6.01 (3.86)	6.90 (4.57)	.097 *
	Total hint use during practice	2.28 (3.75)	3.54 (6.62)	.152
	Total bottom-out hint use during practice	0.44 (1.34)	0.80 (2.22)	.203
	Avg hint use per item during practice	0.25 (0.34)	0.33 (0.41)	.163
	Avg bottom-out hint use per item during practice	0.05 (0.13)	0.07 (0.15)	.223
	Total first response time during practice	18.41 (16.81)	23.92 (19.63)	.041 **
	Avg first response time per practice item	3.06 (2.09)	3.59 (2.35)	.143
	Total time spent on problems during practice	24.53 (23.36)	33.60 (28.22)	.014 **
	Avg time spent on each practice item	3.80 (2.44)	4.66 (2.47)	.018 **
Post-Test	Post-test performance	34%	44%	.271
	Total hint use during post-test	1.41 (1.76)	1.45 (1.71)	.864
	Total bottom-out hint use during post-test	0.25 (0.52)	0.24 (0.53)	.812
	Total first response time during post-test	5.46 (4.42)	4.82 (4.14)	.337
	Total time spent on problems in post-test	7.87 (5.93)	6.61 (5.12)	.168

Note. Means with standard deviations in parentheses are reported for each group. Significant results ($p < .05$) are marked with **. Marginally significant results ($p < .10$) are marked with *.

Hint usage during practice

Students in the ST condition and the MT condition did not differ significantly in the total number of hints that they requested during practice ($Ms = 2.28$ and 3.54, $U = 3279.5$, $p = .152$) or the total number of bottom-out hints requested during practice ($Ms = 0.44$ and 0.80, $U = 3408$, $p = .203$). Similarly, no significant differences were found in the average number of hints ($Ms = 0.25$ and 0.33, $U = 3290$, $p = .163$) or average number of bottom-out hints ($Ms = 0.05$ and 0.07, $U = 3421$, $p = .223$) requested for each practice item in the Skill Builder between conditions.

Post-test performance

Overall, 34% of the students in the ST condition correctly answered both post-test items and 44% of the students in the MT condition correctly answered both post-test items on their first attempt. The difference in post-test correctness was not statistically significant between the conditions, $\chi^2(1, N = 173) = 1.21$, $p = .271$. Further analysis indicated that students in the two conditions did not differ significantly on their correctness on the first post-test item ($\chi^2(1, N = 173) = 0.90$, $p = .343$) or the second post-test item ($\chi^2(1, N = 173) = 0.12$, $p = .733$).

Hint usage during post-test

On average, students who were exposed to one template and those who were exposed to multiple templates in the Skill Builder requested a similar number of hints (*Ms* = 1.41 and 1.45, $U = 3667$, $p = .864$) and bottom-out hints (*Ms* = 0.25 and 0.24, $U = 3774.5$, $p = .812$) while solving the post-test items.

Time-based measures

The ST group and the MT group did not differ significantly in the total post-test first attempt response time (*Ms* = 5.46 min. and 4.82 min., $U = 4036$, $p = .337$) or the total time spent on the post-test (*Ms* = 7.87 min. and 6.61 min., $U = 4173$, $p = .168$).

However, significant differences were found for the amount of time spent on practice. Specifically, it took students in the MT condition longer to provide a first attempt response to the problems based on multiple templates (*M* = 23.92 min., *SD* = 19.63 min.) compared to their counterparts in the ST condition (*M* = 18.41 min., *SD* = 16.81 min., $U = 3048$, $p = .041$). Similarly, the MT students spent more time overall completing the problems during learning (*M* = 33.60 min., *SD* = 28.22 min.) than ST students (*M* = 24.53 min., *SD* = 23.36 min., $U = 2916$, $p = .014$). The average time spent on each practice problem was also significantly longer for the MT condition (*M* = 4.66 min., *SD* = 2.47 min.) than the ST condition (*M* = 3.80 min., *SD* = 2.44 min., $U = 2943$, $p = .018$).

Discussion and conclusion

Results from our study show that multiple templates reduce students' mastery speed as they work on the practice problems, and increase the amount of time spent on each problem. However, these significant differences in student behavior did not translate into significant differences in student performance on the post-test.

The lack of significant differences in post-test performance may be explained by the design of each of the conditions. Switching between template types likely caused the increased time for each problem in the MT condition. Furthermore, it is possible that students may not have had enough opportunities to practice each of the different problem templates within the MT condition. This is in contrast to the ST condition where the students would have had multiple opportunities to practice problems of the same template. However, it is also possible that the study simply did not have sufficient statistical power to detect the effect size seen, although either way the effect size observed was quite small, suggesting that relatively little benefit was obtained from this intervention.

This result contrasts with many previous studies on the impact of varying problem formats, where varying problem formats or representations was associated with better post-test scores. In many of those previous studies, however, students were given substantial instruction or additional activities along with varied problem templates. As such, it is possible that the use of multiple templates in a problem set may have had a bigger impact if accompanied by explicit instruction and training before or after practice on the problem set. For instance, if time was allocated to teachers in each group to explain the mathematical concepts behind the presentation of linear equations in the problem set (e.g., with more emphasis on the structural features in the MT condition), this could have helped students belonging to the MT group understand the mathematical processes necessary to practice solving linear equations across a variety of problems. However, adding this would produce a more complex intervention that is harder to scale up than simply modifying the content available in an intelligent tutoring system. Future work could involve exploring the effects of implementing automated scaffolding and instruction within intelligent tutoring systems (e.g., in the form of highlighting the similarities and differences between problems of different templates) on student learning and performance.

In conclusion, the results of this study do not appear to support our hypothesis that students in the MT condition would demonstrate better learning than students in the ST condition within the ASSISTments system. Analyses of students' behavior and performance during their work on the problem set and post-tests found no statistically significant difference in students' post-test performance or the number of hint requests between conditions. If anything, the results favored the ST condition. Students in the ST condition mastered the content marginally significantly faster than their MT counterparts. Students in the MT condition spent more time responding to the practice items, potentially indicating higher cognitive load and efforts required by the variety of problem templates. Differences in the amount of time spent on the problem attempts in each of these conditions did not translate into significant differences in post-test results. This finding suggests that it is warranted to measure time taken as well as success when comparing learning with varying surface and structural features and learning when features are not varied.

Limitations and future research

One of the limitations of this study is the relatively low completion rate. Students were not required to complete the problem set, and over half of the students who did not complete stopped during the first three problems of the problem set. Working with teachers to require completion (or at least completion of several problems) of the

problem sets studied could resolve this problem, but would represent a less authentic manner of use since many teachers using ASSISTments do not enforce completion requirements.

In this study, we decided not to include a pre-test to ensure the completion rate, which has already been low. However, the lack of a pre-test in the experimental design makes it difficult to measure and compare learning in the two conditions. In addition, participants in this study completed a post-test immediately following their practice in the Skill Builder. However, post-test performance may not necessarily be indicative of long-term learning, and it is possible that students exposed to the MT condition may have experienced desirable difficulty, leading to robust and long-term learning (Kapur, 2016). For future work, we plan to also study students' performance on transfer items over the long term and examine whether any different results are obtained.

Additionally, we created variations of an original template in an algebra problem set on systems of linear equations. Future research should test the generalizability of the findings to other topics and skills. For example, would similar results be found when more concrete problems such as word problems or other concepts such as geometry are used to create templates? How would the difficulty level of the problems influence the effects of templates?

Overall, we did not find significant differences in math post-test performance between ST and MT conditions, suggesting that more than just variations between problems may be required for students to learn general representations. Instead, additional activities or lecture discussing multiple representations or surface features may be needed.

References

Ben-Zeev, T., & Star, J. R. (2001). Spurious correlations in mathematical thinking. *Cognition and Instruction, 19*(3), 253-275.

Chang, N., Koedinger, K. R., & Lovett, M. C. (2003). Learning spurious correlations instead of deeper relations. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 25, pp. 228-233).

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Chi, M. T. H., & VanLehn, K. A. (2012). Seeing deep structure from the interactions of surface features. *Educational Psychologist, 47*(3), 177-188.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*(1), 1-38.

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education, 24*(4), 470-497.

Inventado, P. S., Scupelli, P., Van Inwegen, E. G., Ostrow, K. S., Heffernan, N., Ocumpaugh, J., . . . Almeda, V. M. (2016). Hint availability slows completion times in summer work. In *Proceedings of the 9th International Conference on Educational Data Mining* (pp. 388–393).

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist*, *51*(2), 289-299.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901-918.

McNeil, N. M., Fyfe, E. R., Petersen, L. A., Dunwiddie, A. E., & Brletic-Shipley, H. (2011). Benefits of practicing 4= 2+ 2: Nontraditional problem formats facilitate children's understanding of mathematical equivalence. *Child Development, 82*(5), 1620-1633.

National Governors Association Center for Best Practices Council of Chief State School Officers. (2010). Common Core State Standards for Mathematics. In *National Governors Association Center for Best Practices, Council of Chief State School Officers*. Washington D.C.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology, 88*(1), 144-161.

Xiong, X., Adjei, S., & Heffernan, N. T. (2014). Improving retention performance prediction with prerequisite skill features. In *Proceedings of the 9th International Conference on Educational Data Mining* (pp. 375–376).

Acknowledgments

This work was supported by NSF (DRL #1252297 and DRL #1535340). We would like to thank Anthony Botelho for his help in providing information about the number of templates in ASSISTments.