7-27-2015

Schema Induction through Evaluation and Correction

Clayton Estey

SCHEMA INDUCTION THROUGH EVALUATION AND CORRECTION

by

Clayton Estey

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: Psychology

The University of Memphis

August 2015

Abstract

Estey, Clayton Joseph. M.S. The University of Memphis. August, 2015. Schema Induction through Evaluation and Correction. Major Professor: Philip Pavlik Jr.

Schema induction occurs when people form mental representations of how to organize and interpret information through learning generalities across multiple events. An instructional intervention/task purported to enhance this process for students is Schema Induction through Evaluation and Correction (SIEC, aka "Text Editing"), whereby students label problems based on their structural quality and edit their flaws away if necessary. We compared the relative learning efficiency of a SIEC condition with a conventional problem solving condition using word-problems one would find in basic probability theory. In a randomized control design with a large and diverse internet sample, we found SIEC to offer no advantage over the control in preparing participants for a problem solving posttest, both in terms of relative learning efficiency and raw pretest-posttest gains. The results and their implications are discussed.

Table of Contents

Schema Induction through Evaluation and Correction

As shown in the seminal paper by Gick and Holyoak (1983), "schema induction," whereby students develop mental representations of concepts/problems through transferring knowledge across learning events, may be a way to realize the learning of complex procedures. During schema induction, students progressively transfer their knowledge across similar and dissimilar learning events and begin to abstract which concepts/rules are relevant and which ones are not. The terms "progressive abstraction" (e.g., Cameron, 1993) and "concreteness fading" (e.g., McNeil & Fyfe, 2012) are additional labels, both capturing a different perspective on the same process, because increasing abstraction implies decreasing concreteness in the same learning context.

The basis behind schema induction is analogical reasoning. More specifically, schema induction is a consequence of matching features across stimuli/events, forming relations among these features (forming analogies), then forming inferences based on the regularities found across stimuli/events (e.g., Genter, Loewenstein, & Thompson, 2003). The mental conception of the regularity (i.e., the features and their relations learned across stimuli/events) is termed the "schema," and the inferences we make given such schemas are based on what is relevant or irrelevant to that schema's implementation. Interestingly enough, this makes the process of schema induction very similar to, if not the same as, modern thinking on category/concept induction[1]. In a review (Ross, Taylor, Middleton, & Nokes, 2008), the author notes more modern conceptions of concept/category learning incorporate both features *and* non-trivial relationships between

[1]Ross (2008) clarifies that concepts are mental representations people use to pick out a class of entities, while categories are the class of entities the concept targets. Despite this, we will use these terms interchangeably given there is no difference between a concept and an abstract "meta-category" under these definitions.

them, reflecting the learning of more complex, realistic categories/concepts. Ross also notes past views concentrated on categorical membership and inference only being based on features while ignoring relations among them. This conception of categories was likely due to relying on simplistic laboratory stimuli. Indeed, the modern view is identical to the process outlined above with schema induction—it is the features and their relationships learned over time that characterizes a schema and subsequent inference using that schema.

This means when someone speaks about learning schemas, it is safe to assume they are talking about the learning of categories whereby the relationships among the features are important in representing and using the categories. Given the above clarifications and ties with category/concept learning, the present proposal thus concerns getting people to learn the proper relations among key features in problems. This would occur through trial-by-trial analogical reasoning and knowledge transfer and properly applying the induced schema (i.e., concept).

The research on theoretical contributions to schema induction principles and their applications in education have offered notable extensions of the Gick and Holyoak (1983) study and the processes outlined in Genter and colleagues (2003). In a study investigating positive and negative transfer among items of varying concreteness (i.e., specific versus abstract schema conditions) and later impact on problem solving, Chen and Daehler (1989) showed both the learning of an abstract (problem general) problem solving schema *and* knowing when to apply that schema were necessary for positive transfer, because while abstract schemas were necessary for positive transfer, participants failed to transfer toward problems with solution principles different from training.

Before this study it was less clear what role schema induction played in positive transfer and what the limitations were for that transfer to occur. Ten years later in a study varying specific solution procedures while holding the solutions themselves constant, Chen (1999) found participants were more likely to develop general problem solving schemas and use them appropriately when procedures to be learned varied across learning trials, compared to the procedures staying the same. Those in the varying procedure condition also showed positive transfer toward novel problems with unfamiliar procedural content. In a later study refining Chen's (1999) prior findings with procedural variation but adding the component of different *dimensions* within word problems (e.g., story lines, formulas, etc. are each "dimensions") and if schema induction was dimension specific versus general, Chen and Mo (2004) found that exposure to varying procedures across learning trials led to slower initial learning, but more flexible schemas, whereas vice versa occurred when procedures varied less. This replicated the prior finding with procedural variation. Additionally, they also found schema induction was *dimension specific*. This refers to positive transfer only existing along the same dimension participants were trained in.

There has also been considerable interest in "schema-based instruction" and "schema training" with studies consistently showing learning gains compared to non-schema based practice in the target domain (e.g., Fuchs et al., 2004; Jitendra & Hoff, 1996; Jitendra et al., 2009; Robins & Mayer, 1993; Xin, 2008). In addition to refinements and applications, there was a change in terminology in at least one research group when referring to schema induction ("relational schema induction," "relational schema theory"—Halford, Bain, Maybery, & Andrews, 1998; Halford & Busby, 2007).

This change perhaps explicates the fact schema induction is analogical and relational in mechanism and better focuses new readers to the core aspects of schema induction, but the psychological processes outlined in their studies do not differ from prior thinking about schema induction and investigations thereof.

Given the reasons why schema induction occurs, and past success in applying these principles for learning how to solve problems, it should be possible to get students to understand what the "deeper structures" of problems are while disregarding erroneous, surface level details, as this is the trademark distinction between novices and experts in a given problem solving domain (e.g., Chi, Feltovich, & Glaser, 1981). In terms of schema induction, the "deeper structure" is the parts of a problem instantiating a target concept, and the desired schema to be learned would be the students' internalization/modeling of either the concept itself or a problem solving strategy given that concept. All other information in a problem would be seen as irrelevant by an expert, and would be superficial details at best and deceptive details at worst for a novice.

To implement the above concepts and investigate schema induction, we have chosen the task of "text editing," whereby students label problems according to their solvability and relevance of the content, then edit the problem to remove its flaws (e.g., Birney, Fogarty, & Plank, 2005; Low & Over, 1989, 1990, 1992, 1993; Low, Over, Doolan, & Michell, 1994; Ngu, Low, & Sweller, 2002; Ngu & Yeung, 2013). The goal of text editing is to increase knowledge of, and attention toward, a problem's deeper structure. Increasing this knowledge and attention allows the learner to better understand the components of the problem most relevant to problem solving rules. In text editing, students label a problem as *sufficient* when there is no irrelevant information and the

problem is solvable, *irrelevant* when it is solvable but there is superficial/deceptive information, and *missing* when there is not enough information to solve the problem. Should a student answer correctly in that the problem is either irrelevant or missing, they are then asked which aspects to edit (i.e., add or delete) to make the problem sufficient.

Through classifying word problems in text editing tasks, the literature above claims students are led to focus on which features of problems are irrelevant and which ones are crucial in order to apply a concept the problem instantiates (e.g., a math problem might instantiate a compound interest concept). If a student knows why an *irrelevant* information problem should be classified as such, then they can learn to ignore related features in subsequent problems, encouraging the student to not let superficial details influence their problem solving. A student might also eliminate the misconception that an irrelevant detail is actually part of the problem solving procedure. If a student knows why a *missing* information problem should be classified as such, then they can learn to focus on the features necessary for reaching a correct solution, and thus incorporate this information in their problem solving. This contrasts with conventional practice, whereby students learn procedures for solving problems without explicit feature reinforcement or addressing misconceptions about the "deeper structure" of problems. Indeed, conventional practice is typically cited in this literature as resulting in lower problem solving performance in posttest (compared to text editing) for these reasons (e.g., Birney et al., 2005).

Precisely how the editing process itself (removing irrelevant information and adding necessary information) interacts with the mechanisms of schema induction was never the focus of this more applied literature, and is therefore less understood. While an

elaborate hypothetical is possible relating the mechanisms of schema induction to the editing process, the editing may or may not add more to the students' problem solving performance than what is accomplished through the classification stage alone. After all, if a student classifies correctly, then they likely already have schematic knowledge of the deeper problem structure, with the editing stage being redundant. Delineating the contributions of these two stages of text editing toward problem solving performance on posttest would make for an interesting empirical investigation, but the focus of the present study was to test text editing as it had been used in the literature.

Although "text editing" is a name consistently used throughout the literature cited, this name does not clearly convey the process just described. For example, upon hearing about text editing, a researcher would likely think "Oh, I thought it was referring to something like editing a paper." To prevent this confusion, we will be referring to this "text editing" by a label we view as better reflecting the underlying learning phenomenon, Schema Induction through Evaluation and Correction (SIEC). SIEC is purported in the literature to be robust across domains, including algebra (Low & Over, 1990), geometry (Low & Over, 1992), and basic probability theory (Birney et al., 2005). Some researchers in the literature even claim this approach can increase problem solving ability when there is no practice solving problems during training (Low et al., 1994; Ngu et al., 2002; Ngu & Yeung, 2013). This might be a very counterintuitive claim for many and thus will be one of the claims tested in the present study. Although the domain of chemistry holds up to this pattern (Ngu et al., 2002; Ngu & Yeung, 2013), SIEC was shown to not work for stoichiometry problems (Ngu et al., 2002; Ngu & Yeung, 2013). One reason for the aforementioned result regarding stoichiometry problems is that SIEC

is only beneficial when it enables students to attend to a problem's underlying concepts--through fully representing the concepts in the word problem as textual information. This is not possible for stoichiometry problems which do not contain direct textual representations of concepts needed for problem solving.

However, Ngu and Yeung (2013) also demonstrated that seeing worked examples not only outperformed SIEC in the contexts it *can* be used in, but that it also increased domain knowledge in the contexts where SIEC is ineffective. The curious reader may wonder why anyone should study SIEC when there is a better alternative. We argue for the possibility that their study was an inadequate test of SIEC. In their discussion section, they speculated the performance of the SIEC group could better rival the worked examples group if SIEC were used in feedback during initial learning of problem solving for the SIEC group, and no feedback was provided during training in their study nor in Ngu et al. (2002) for the above type of problems. Therefore, their study may not invalidate SIEC as an intervention, as their findings regarding stoichiometry problems could represent exceptions to the norm for SIEC.

Despite our paper up to this point illustrating SIEC, there are issues in interpreting the literature. All studies were in school settings and varied on key factors possibly related to posttest performance (see Birney et al., 2005; Low & Over, 1989, 1990, 1992, 1993; Low et al., 1994; Ngu et al., 2002; Ngu & Yeung, 2013). These factors include school location, the amount of delay between pretest and intervention, whether a pretest or random assignment occurred, and other qualitative discrepancies making interpreting the findings difficult. Our study was the first to test SIEC under laboratory conditions.

In addition to laboratory replication, we addressed the literature's claim that SIEC can be effective despite lack of practice problem solving during training.

The final experiment targeted learning *efficiency*. In past investigations of SIEC, the researchers held time constant across groups so as to eliminate time as a factor. While this design is intuitive and usually appropriate, doing so limits investigations into the potential of one learning approach versus another. For example, if one approach yielded greater learning gains while students spent longer time learning, and students spent much less time learning under an alternative approach for slightly less gains, then such a time discrepancy could have been informative in determining the relative worth of these approaches. Thus, we also addressed whether or not SIEC training results in more efficient learning outcomes (pre-post gains) compared to conventional practice. We operationalized and justified a two-group learning efficiency measure in Appendix A. If greater schematic knowledge resulted in better learning even without problem solving practice as is claimed, such contributions might not have been as meaningful if the needed intervention took longer, as that could have been time spent with additional problem solving practice. The domain in which we investigated the above was basic probability theory with items similar to Birney et al.'s (2005). Specifically, the items were word problems pertaining to basic probability concepts with multiple choice answers.

Study 1

The purpose of the first study was to gather demographic information (see Appendix B) and conventional problem solving pretest scores in a sample separate from the main study (i.e., Study 2). One reason for having the pretest in a separate group from the intervention and posttest was for participants to have enough time during learning and posttest for the main study. However, the primary benefit was in eliminating carry-over effects between pretest and intervention, thereby allowing direct inferences about how performance was influenced only by the intervention and not due to interactions with prior problem solving. Resolving this issue required using the demographics as predictors in a model predicting pretest scores. Specifically, we input Study 2 demographic values into the model so each person in the Study 2 sample would have their own synthetic pretest scores, with the scores serving as a covariate in explaining posttest scores for Study 2.

Method

Participants

Eighty-one participants were recruited using the Mechanical Turk (MTurk) service from Amazon, restricted to people over the age of 18 who reside in the U.S. and speak fluent English. MTurk is an online labor market whereby people can be cheaply paid for online tasks. Studies have shown MTurk is at least as reliable as, and more generalizable than, undergraduate subject pool and conventional online samples for non-experimental studies (e.g., Buhrmester, Kwang, & Gosling, 2011; Mason & Suri, 2012; Paolacci, Chandler, & Ipeirotis, 2010) and recently, cognitive experiments (Crump, McDonnell, & Gureckis, 2013). Participants were asked to provide informed consent

before the study. This study was approved by the IRB and their compensation for participating is below.

Participants were paid $0.10 for participating in the demographic surveys even if they withdrew. In addition, they were paid $1 for attempting the pretest and even if they withdrew before answering the first problem, they were still paid $1.10. They were offered up to $1 in bonus based on their performance for correctly solving the problems (which scaled according to their percentage score). Thus, participants had the potential to earn $2.10.

Design, Materials, and Procedure

This study was a single group design with no conditions. We used the Fact and Concept Training (FaCT) system, a computerized tutoring system for enhancing student learning of concepts and facts (Pavlik Jr. et al., 2007). We also used the R statistical programming language (R Development Core Team, 2011) for the analyses.

To mitigate self-selection bias for/against the topic, and thus skewed demographics, the name of the task was content general ("Concept Learning") despite the task being mathematical. We hoped those with a prior bias against math were persuaded by a reasonable economic incentive, which would exist even with poor performance while increasing with better performance. Therefore, by the time they found out the task was math-based, they would have already had this incentive in mind. Likewise, we hoped to mitigate the influx of individuals with probability knowledge who would be attracted to the study strictly because of the title and contribute to ceiling effects.

After the participants gave their consent, they were exposed to the demographic survey, the items of which were randomized per participant to mitigate response bias via

impressions from prior questions (see Appendix B). Then they were shown instructions and expectations pertaining to the task. Appendix C contains the instructions and expectations, concrete examples of problems and relevant feedback, and precisely illustrated variations of such across conditions (latter is for Study 2). Then they were exposed to 12 pretest problems with different numbers across problems, randomized per participant. Each problem was multiple-choice with four answer options and only one correct answer each. They were all a possible combination of 2(number of probability concepts) X 3(sufficient, irrelevant, or missing information) X 2(each problem repeated once with different numbers) which totaled 12 problems. Two answer options reflected an answer applying a particular probability concept, one option was a distractor which could never be correct for this study, and the other option reflected acknowledging a problem could not be solved due to missing information.

Analyses

Iteratively re-weighted least squares (IWLS) multiple regression based on bootstrap resampling of observations was used to model pretest scores using demographic predictors. IWLS was chosen over the conventional, parametric ordinary least squares (OLS) regression because other than the observations being independent, the homoscedasticity of variances assumption for OLS was violated. The demographic predictors were gender, age, general education attainment, math education attainment, number of hours on average completing MTurk tasks (thus exposure to task practice in general)[2], and self-reported propensity to select math related tasks (desire for math related tasks). The same method was also used to model the participants' time-on-task

[2]Operationalized by multiplying results of two constructs: average number of hours per month completing MTurk tasks and number of months spent on MTurk. The result will be a predictor with values from 1 to 36.

using the same predictors, as well as incorporating time-on-task as an additional predictor in the former regression. In these latter two models, we wanted to determine if the time participants spent on the study could be described well using their demographics, if their times would add much more information about their performance than that provided by their demographics alone, and how much this addition could change our inferences about the other predictors. We thought about how behavioral economics issues with the payment incentive could have affected performance on the pretest. Therefore, we concluded it was best to control for this as best we could by incorporating time-on-task into our inferences about pretest scores relative to the other predictors.

Results

After removing three outliers $N = 78$. The three were determined to be outliers due to taking an abnormally long time on the task while also demonstrating low performance. The average number of items correct was 5.96 out of 12 with $SD = 2.36$, being slightly better than chance (3 out of 12). We also found that one of the initial predictors (number of hours on MTurk a month) was consistently at the maximum value (25+ hours) and therefore contributed no new information to the modeling. Therefore, total hours of practice on MTurk was replaced with number of months on MTurk. For below, the 95% confidence intervals are based on the lower and upper threshold percentiles of the bootstrapped sampling distributions for the coefficients, the standard error (*SE*) is based on the standard deviation of those distributions, and each β is the average of their respective distribution. Non-significant predictors were kept in the models because we observed that if they were to be excluded, the remaining coefficients would have had larger values than would occur in adding this extra information into the

model. Therefore, our inferences could have been unfairly biased had we removed non-significant predictors.

For the first model where pretest scores was the dependent variable being modeled by demographics and excluding time-on-task (adjusted R^2 = 0.36): gender (β = -0.07 ; *CI* = [-0.55 , 0.44] ; *SE* = 0.26), age (β = 0.08 ; *CI* = [-0.16 , 0.32] ; *SE* = 0.12), math education (β = 0.25 ; *CI* = [-0.09 , 0.54] ; *SE* = 0.16), general education (β = 0.23 ; *CI* = [-0.05 , 0.52] ; *SE* = 0.15), months on MTurk (β = 0.18 ; *CI* = [-0.07 , 0.42] ; *SE* = 0.12), and preference for math related tasks (β = 0.34 ; *CI* = [0.11 , 0.57] ; *SE* = 0.11). Thus in this model, preference for math related tasks had the only coefficient with an interval excluding zero.

For the second model where time-on-task was the dependent variable being modeled by demographics (adjusted R^2 = 0.13): gender (β = -0.15 ; *CI* = [-0.62 , 0.25] ; *SE* = 0.22), age (β = 0.1 ; *CI* = [-0.1 , 0.34] ; *SE* = 0.11), math education (β = 0.26 ; *CI* = [0.01 , 0.51] ; *SE* = 0.13), general education (β = -0.005 ; *CI* = [-0.22 , 0.23] ; *SE* = 0.11), months on MTurk (β = -0.03 ; *CI* = [-0.24 , 0.19] ; *SE* = 0.11), and preference for math related tasks (β = 0.22 ; *CI* = [0.07 , 0.51] ; *SE* = 0.11). Thus in this model, preference for math related tasks had an interval excluding zero just as before, as does math education's interval. In terms of the model fit, only the former significant predictor had a major sole contribution, as the model's fit with this predictor removed was only adjusted R^2 = 0.074, so about 57% of the model's fit, although low, is attributed to preference for math related tasks alone.

Finally, for the first model but with time-on-task added (adjusted R^2 = 0.40): gender (β = 0.009 ; *CI* = [-0.46 , 0.51] ; *SE* = 0.25), age (β = 0.06 ; *CI* = [-0.19 , 0.28] ;

SE = 0.12), math education (β = 0.18 ; *CI* = [-0.16 , 0.48] ; *SE* = 0.17), general education (β = 0.22 ; *CI* = [-0.06 , 0.52] ; *SE* = 0.15), months on MTurk (β = 0.19 ; *CI* = [-0.05 , 0.44] ; *SE* = 0.12), preference for math related tasks (β = 0.27 ; *CI* = [0.03 , 0.49] ; *SE* = 0.11), and time-on-task (β = 0.26 ; *CI* = [0.05 , 0.48] ; *SE* = 0.11). The inclusion of time-on-task did not change which predictors were significant versus not, although the coefficients for math education and preference for math related tasks did shrink substantially more than the other predictors through adding the extra predictor. Still, adding time-on-task only increased the adjusted R^2 by 0.04, suggesting this predictor describes pretest performance very poorly relative to the other predictors.

These results suggest that participants' use of time was insufficiently modeled. Additionally, their use of time had less influence on pretest performance than qualities reflected in their demographics (nor did it bias inferences about the other predictors to a meaningful degree). The coefficient being positive suggests participants spending longer time did not diminish performance on average. However, a positive coefficient could still allow the possibility of some participants "rushing through" the study and getting lower scores. Such a pattern was not consistently observed as there were participants who spent much less time than others while getting similar scores.

Overall, performance on pretest was too poorly explained by the offered payment incentive (if the time they spent on the task alone reflects such an incentive) and is more likely due to attributes not measured. At least one missing key predictor of pretest performance could be prior exposure to these particular types of probability questions (and not math education in general). Although the total model fit seems less than the ideal in the task of providing synthetic pretest scores for participants in Study 2, the fit

was sufficient to suggest this method is probably the better of two main alternatives[3] for the purpose of comparing the pretest score average with the posttest averages of Study 2. Additionally, as long as we can assume the coefficients for the model would be stable across an additional sample from the same population, the model should still be sufficient for the task of producing synthetic pretest scores. Through this method for eliminating carryover, it was also much easier to constrain our inferences about what exactly led to differences in posttest across groups in Study 2.

Study 2

Our goal for this study was to compare participants' learning of probability theory items with SIEC compared to a conventional problem solving control condition. Given design and sampling discrepancies across past studies in the SIEC literature, we had no adequate a priori hypothesis as to which condition would perform best on a problem solving posttest. This study compared the raw difference in pretest-posttest gain scores of the SIEC and control conditions before accounting for the potential differences in learning time across groups. Afterward, the difference in learning efficiency was compared across the groups.

[3]The first alternative is having the pretest, intervention, and posttest in the same sample but with the pretest and posttest items differing in their surface level details. This is the most common alternative in cognitive/educational experiments. While this mitigates prior recognition effects involving stimuli, some carryover still exists from prior practice before the intervention is shown as there will always be important similarities between pretest and posttest. The second alternative is much less common in research, but interactions between pretest, intervention, and posttest can be modelled and the carryover at least be "explained away." The issue here is precisely explaining potential multilevel interactions and how they would affect the posttest could be untenable, and the carryover never truly goes away. The present method used is subject to sufficient model fitting, but otherwise shares none of these weaknesses.

Method

Participants and Design

The 195 participants came from the same population as the first study. Participants were asked to provide informed consent before the study. This study was approved by the IRB and their compensation for participating follows. Participants were paid $0.05 for participating in the pre-experiment demographic survey even when they withdrew. In addition, they were paid $1.75 for attempting the experimental portion and even if they withdrew before answering the first problem, they were still payed $1.80. They were offered up to $1.90 in bonus based on their performance for correctly solving the problems during the testing phase (scaled by their percentage score). Thus, participants had the potential to earn $3.70.

The two between-subjects conditions were conventional problem solving (control) and SIEC. The conditions were divided into the learning phase block (first block; for the nominal factor) and the testing phase block (second block; for posttest performance), with an eight minute break for participants between blocks. Conditions were randomly assigned to participants as they entered the study.

Materials and Procedure

All of the items in Study 1 were used in Study 2. The procedure for Study 2 up to but excluding the part where they were given instructions related to the task was identical to that of Study 1 (i.e., the name of the study and the demographic survey). For the control group, after the instructions they entered the learning phase and were presented isomorphs of all the items in Study 1 in a randomized order. In the event they answered incorrectly, they received feedback customized to the answer option they

selected then went to the next problem. The feedback contained an affirmation that they were incorrect and the solution procedure for the correct answer. If they answered correctly, they moved on to the next problem without feedback. This continued until the break. During the break, all participants viewed an eight minute mildly entertaining video playlist of cats such as is common on the internet. This was to clear working memory with an unrelated, entertaining stimulus while giving them rest between sessions. After the break they entered the testing phase (posttest) and were presented with all of the randomized items from Study 1 so changes in performance from pretest to posttest could only be attributed to the intervention and not to changes in testing items (i.e., identical; not isomorphic). The learning phase items being isomorphic to the testing phase items prevented simple memorization as a strategy for Study 2 participants. This allowed the Study 1 and Study 2 posttest items to be identical, and all of the learning phase items for the control condition to be isomorphs of every Study 1/posttest item. The total then was 12 pretest items, 12 learning phase items, and 12 posttest items.

For the SIEC group, the procedure leading up to the instructions was the same as that of the control condition. After the instructions they entered the learning phase and were presented a randomized order of SIEC items. Items in this condition either had one step to complete or two steps, but neither involved practice problem solving at any step.[4] The problems with "sufficient with no irrelevant information" as the correct answer were the one-step problems as they did not need an editing step. For two-step problems, the first step was in classifying the problem as a possible combination of *sufficient*,

[4]The most that could be done to mitigate participants mentally problem solving to help with the classification and editing tasks was to instruct them not to problem solve and only follow the task directions. Given the answering options were not numeric and problem solving wasn't necessary to be correct, they should have been less likely to problem solve.

irrelevant, or *missing* and there were four answer options. As before, one of the answer options was a distractor which could never be correct for this study. The distractors were answers which would result from applying an existing probability concept not in the study. If they answered incorrectly in this step, they were given feedback specific to their answer and moved on to the second step. The specific nature of the feedback on items in this study can be found in Appendix C. More briefly, the feedback contained an affirmation of correct/incorrect, followed by an elaboration on why the affirmation was so, and then concluding with an illustration of the solution procedure. If they answered correctly, they also went to the second step but without feedback. In the second step they classified what aspect of the problem needed to be changed so it could be re-classified as *sufficient with no irrelevant information*, which also had four answer options. If they answered the second step correctly or incorrectly they were given customized feedback and moved on to the next problem. For one-step problems only the first step occurred and they moved to the next problem after completion.

The one-step items in this condition had a feedback structure similar to the learning items in the control group (all of which were one-step problems), except problems in this condition also gave feedback illustrating a solution procedure even when they answered correctly. Illustrating the solution steps in a declarative manner allowed this type of feedback to be held constant across SIEC and control groups. This ensured it was not the absence of a solution procedure in the SIEC group's feedback that was responsible for different outcomes. This process continued until the break which was the same as that of the control condition. The posttest was the same as that of the control condition.

Results

After removing 19 observations with incomplete number of trials and some entering the study in one group then re-entering into the other group, $N = 81$ for the SIEC group and $N = 95$ for the control group for a total $N = 176$. The average number of posttest items correct for SIEC was 7.73 out of 12 with $SD = 3.26$. For the control group, the number was 7.56 out of 12 with $SD = 3.23$. The average number of synthetic pretest items correct for SIEC was 4.77 out of 12 and the mean of the prediction standard error was 2.47. For the control group, the average was 5.48 out of 12 and the mean of the prediction standard error was 2.46. The average learning time duration for SIEC was 17.67 minutes with $SD = 3.71$. For the control group, the duration was 9.35 minutes with $SD = 4.5$. For comparing the mean posttest items correct, a two-sided t-test based on unequal N and equal variances assumed showed a non-significant difference in means ($t = 0.35$; $df = 174$; $p = 0.729$). To determine if the non-significance was due to low N versus a trivial mean difference actually existing, we examined the 95% confidence interval for the above test ($CI = [-0.8, 1.14]$), with the precision of the interval suggesting the sample size was large enough to determine it was the effect difference that was "non-significant" as opposed to there being insufficient data.

To determine if the above non-significance extended to the difference in learning efficiency (DLE) statistic, an approximate permutation test[5] was conducted to compare the learning efficiency of the two groups. These results were also non-significant ($DLE =$

[5]We generated an empirical sampling distribution for the statistic under the null hypothesis, and then the observed value was compared to its null distribution to calculate a p value.

-0.74[6] ; $p = 0.19$) along with there being a significant difference in the learning phase time across groups ($t = 13.25$; $df = 174$; $p < 0.0001$). These findings suggest that the difference in learning efficiency between conditions being non-significant is due to the prior finding that the raw posttest differences themselves were trivial. Additionally, the triviality of the mean difference outweighed the significant difference in time spent in the learning phase in determining learning efficiency.

Based on the above findings, however, it was still possible that the non-significant raw difference was largely mediated by prior aptitude (i.e., synthetic pretest), with pretest scores predicting posttest performance differently across groups. Additionally, it was still unclear if minute demographic differences across studies interacting with the group factor would have explained the minute, non-significant difference in posttest. Therefore, a between subjects ANCOVA was conducted with synthetic pretest as the continuous covariate, the conditions as the nominal factor, and their interaction. If the interaction term was found to be significant, then this would be evidence of the intervention behaving differently for participants with differing initial performance. The interaction was found to be non-significant ($F = 1.89$; $df = 1$; $p = 0.17$). Just as expected from the prior t-test results, the group main effect was still found to be non-significant ($F = 2.32$; $df = 1$; $p = 0.13$). However, the synthetic pretest main effect was found to be significant ($F = 23.7$; $df = 1$; $p < 0.0001$). The interaction term being non-significant suggests that prior aptitude differences across groups minimally effected SIEC and the control. The group main effect being non-significant while the pretest main effect being significant

[6]The number would be interpreted as "the SIEC condition was 0.74 items *less* efficient in pre-post gains than the control condition given the discrepancy in learning time averages across groups relative to the grand learning time average across the groups."

suggests that, after accounting for synthetic pretest and the demographics producing them, one group did not show an advantage over the other.

Lastly, we wanted to determine if it was solely the posttest stimuli that were responsible for the non-significant difference or if the findings were due to effects of the intervention and control groups interacting with the stimuli as intended. If only the stimuli were responsible, then we would expect a non-significant difference between pretest and posttest for both groups since the pretest and posttest stimuli were identical. After running two one-tailed paired-sample t-tests, statistically significant results were found for SIEC synthetic pretest-posttest gains ($t = 9.2$; $df = 80$; $p < 0.0001$) and the same for the control group ($t = 6.4$; $df = 94$; $p < 0.0001$). These results suggest it was not issues in the stimuli alone that were responsible for the null finding. Instead, it is more likely interactions between stimuli and learning phase factors in Study 2 during and/or after the learning phases that caused the null difference.

Discussion

Our goal was to investigate the problem solving posttest performance produced from SIEC compared to the same for a conventional problem solving control group. In doing so we addressed both differences in raw posttest scores as well as the potential difference in learning efficiency. We obtained null results for both raw posttest difference and learning efficiency difference, and we performed follow-up analyses clarifying the most likely possibilities for these findings. Overall, the sample size was large enough to determine the null result from comparing raw posttest mean difference was due to a genuinely trivial difference in the posttest means between SIEC and control groups instead of there not being enough data to reject a true null hypothesis, and that this

finding resulted from circumstances during/after the Study 2 learning phases. This inference took into account whether or not initial aptitude during pretest interacted with the group factor as well as the possibility minor demographic differences across groups affected the minute difference in posttest means. Additionally, there was also a non-significant difference found in the learning efficiency across groups and that this non-difference was mostly mediated by the trivial difference in posttest means (since SIEC taking nearly twice as long in its learning phase as the control group still failed to yield a significant difference in learning efficiency). For this reason, we concentrate primarily on the raw mean differences instead of the learning time differences in our discussion.

From the results, we reason there is evidence of there being no difference between SIEC and conventional problem solving for this sample, and not that there is just no evidence of a difference, and that this non difference is due to either confounding or legitimate circumstances beginning in the Study 2 learning phases. Below we clarify five considerations and then offer a summary opinion as to why the posttest means converged to similar numbers given key similarities in the feedback and stimuli, yet with the learning task being so qualitatively different. These considerations are more broadly: Similarities in feedback across groups, the pretest coming from an independent sample, the nature of the stimuli, the conditions performing as expected without other issues, and the time duration of the learning phase for both groups.

First is the consideration that in the feedback for the SIEC condition, participants being told the procedure to solve the problems could have been what pushed the SIEC group's posttest scores toward equality with the control group's scores, with the SIEC group otherwise having lower posttest scores than the control. Given this, even though

participants were instructed not to practice problem solving, the statements alone could have been enough to achieve the observed pretest-posttest differences. This would make the posttest results be an artifact of them being told how to solve the problem. If this circumstance is what happened, it would still be evidence against SIEC being better than problem solving for these types of learning items. Seeing the solution procedures was likely itself a form of schematic reinforcement, making the text editing task of the SIEC condition redundant. Such schematic reinforcement would be due to the solution procedure illustrating the necessary and sufficient information for solving, along with the features of the problem pertaining to that information.

The second consideration is a consequence of giving the pretest to a separate sample from the intervention and posttest sample. Although this aspect of the research design is unorthodox, we explained why it would result in less confounded inference about the posttest scores. Namely, it prevents carry-over effects allowing more direct comparisons of group factors in producing posttest scores. By far the greatest initial concern with this design approach was the demographic variation across studies influencing the results more than the study itself. Specifically, the predictors could have interacted with learning phase performance differently across groups, and that minor differences in demographics could have somehow unfairly pushed posttest-differences toward equality. Such an objection would be unsubstantiated, both because the interaction term from our ANCOVA results show evidence to the contrary and the fact the posttest means were so close together despite the qualitative differences in the learning task across groups. A possible objection is that our synthetic pretest scores came from a model with lower than ideal fit for our data and, therefore, any analyses involving

such scores may not be trustworthy. The main effect of synthetic scores in our ANCOVA model being significant, however, instead illustrates the synthetic scores worked quite well for the task of explaining posttest scores while controlling for pretest performance.

The third consideration pertains to the structure of the stimuli themselves, instead of the feedback involving said stimuli. Out of the studies researching text editing, ours is the first to allow participants to acknowledge a problem's insolvability by answering "this problem cannot be solved" if that were true. This addition allows participants to reveal some baseline declarative schematic knowledge of the problem before being exposed to declarative schematic feedback in the SIEC condition, while allowing participants in the control condition to do the same despite not receiving SIEC feedback. As will be seen, however, the impact this design decision had for our study was likely negligible. The pretest and posttest stimuli were identical and thus both had these answering options. Yet, a highly statistically *significant* difference was found between pretest and posttest for both groups while no difference was found between posttests for both groups, despite good power of the test.

The fourth consideration is that the SIEC and control groups were behaving as planned with the task dynamics prescribed to the groups, except they happened to yield the same performance in posttest. Without further empirical study, we cannot know if this consideration is more plausible than the first—with the solution procedure being illustrated at the end of both SIEC and control group feedback. We assume the first consideration is at least more useful for the sake of discussion, as it appeals to a tangible,

directly observed aspect of the design as opposed to a more speculative, albeit possible, coincidental similarity in performance.

The final consideration pertains to intentionally limiting the learning time based on practicality in minimizing attrition rates for a population without institutional pressures to stay in the study. This contrasts with a classroom setting which would better reflect reality in the length of time a student has to learn the material, despite then having less control over potential confounds that could occur given the increased span of time before posttest. Even though the duration of the learning phase for the SIEC group was nearly double that of the control group, in an absolute sense 18 minutes on average for SIEC learning phase is insignificant compared to days, weeks, or months as could be seen in a classroom-oriented research design. It could be possible that SIEC's effects could be demonstrated as superior to the control only through longer time scales, despite the control condition being subject to that same time scale. The declarative illustration of the solution procedure across groups could have a more substantial impact in their schematic learning under lesser durations, while the classification and editing steps of SIEC become more important under greater durations. The influence of these factors cannot be known until a controlled classroom-oriented study is implemented.

Given the above considerations, what most likely caused the SIEC and control groups to have equal performance on posttest is that in the learning phases, the 12 word-problems (including their numbers) they were exposed to were the same (so only the task and not the content differed), as was how they were told the procedures to solve the problems (after the elaborative feedback in the case of SIEC). This equality between groups was found despite not needing to arrive at a numeric answer in the SIEC group's

learning phase. This cannot be attributed to the stimuli absent activity in the learning phase because there was still a difference found between pretest and posttest for both groups with the same items. Therefore, it seems that it was solely exposure to the structural features of the problems and the procedural feedback together (even without practice problem solving) that were needed by the participants, and that for our set of learning items and the time constraints provided, SIEC was redundant in preparing them for posttest. The procedural component of the feedback most likely provided all of the necessary information about solving the problem during posttest, through sufficiently displaying the relationships between key statements, key numbers, and the operations between them.

There are several contributions this study adds to the SIEC literature (e.g., Birney et al., 2005; Low & Over, 1989, 1990, 1992, 1993; Low et al., 1994; Ngu et al., 2002; Ngu & Yeung, 2013), both in the results and in its design. To start, the finding of a null difference between SIEC and conventional problem solving practice groups in posttest, while having a large sample size for this null difference, is a new contribution. Although it was unknown a priori which condition would outperform the other (given the aforementioned qualitative discrepancies across studies), we were not expecting there to be no difference in raw performance or in learning efficiency. The closest to this finding the literature achieved so far was in showing SIEC's ineffectiveness in two key applications, although with caveats our study at least partially took into consideration. These applications are for (1) posttest differences between SIEC and conventional practice with stoichiometry problems favoring conventional practice (Ngu et al., 2002;

Ngu & Yeung, 2013) and (2) for Ngu and Yeung's (2013) finding that a worked examples condition outperformed the SIEC condition for those same types of problems.

In the first application, SIEC's ineffectiveness was attributed to stoichiometry problems not containing direct textual representations of the needed problem solving concepts. In the second application, Ngu and Yeung (2013) speculated the performance of the SIEC group could better rival the worked examples group if SIEC were used in feedback during initial learning of problem solving for the SIEC group, and no feedback was provided during training in their study nor in Ngu and colleagues (2002) for the above type of problems. In our study, we did have stimuli containing direct textual representations of the concepts needed for problem solving, while also providing feedback during training in both conditions. Although using worked examples and/or stoichiometry problems were beyond the scope of our study, we did at least add their considerations into our design. Yet, we still found SIEC to be ineffective in our study compared to a conventional problem solving control group.

The second contribution our study provided was attempting to replicate the claim that SIEC is more effective than conventional problem solving practice (except for stoichiometry problems) even without problem solving training during learning in the SIEC conditions for those studies (see Low et al., 1994; Ngu et al., 2002; Ngu & Yeung, 2013). There was no training in problem solving in our SIEC condition and we were unable to replicate findings supporting that claim, due to our null difference. Although we could not directly observe the behaviors of our participants, we assume they did not problem solve because doing so was not required to answer the problems and they were given instructions to only classify and edit them, as would be done with SIEC. Even if,

hypothetically, people on average in the SIEC condition *did* problem solve without our knowledge, we would expect that to *increase* their performance in interacting with the SIEC condition. Yet, the result was still a null difference, so our findings contradict their claim. It is still unclear if their claim resulted from students practicing problem solving while outside the learning phase because the extra unmonitored time of a classroom-design allows such to take place, or if the claim resulted from true assumptions on their part and our demonstration of SIEC's ineffectiveness was due to a methodological reason unrelated to the lack of problem solving in our SIEC condition. Without a classroom-design allowing researchers to perhaps document student studying habits after the pretest and before posttest (coded to be a covariate controlling for discrepancies in study habits across groups), it is difficult to resolve these issues.

Our study also attempted to not only replicate findings supporting the prior discussed claim, but our study can also be seen as an attempt at extending the findings of Birney and colleagues (2005) for the same subject domain using nearly identical stimuli (basic probability theory). Our study's emphasis on how to test participants' schematic knowledge in this domain, however, was different. In their approach to assessing schematic knowledge, they provided a workbook to students with three sections. Section 1 was the SIEC section, section 2 pertained to classifying pairs of problems as similar vs. different, and section 3 was the problem solving section where students solved for numerical answers. Then they correlated accuracy in the first two sections with accuracy in problem solving for numeric answers (section 3), with high vs. low correlations suggesting high vs. low schematic knowledge. Sections 1 and 3 showed a strong

correlation ($r = 0.62$; $p < 0.001$) while sections 2 and 3 showed a weak one ($r = 0.25$; $p < 0.05$). This was taken as evidence of the SIEC section performing best.

The fact that the above element of their study was within-subjects, however, means their results may be sensitive to the order in which items or conditions are presented, as well as the role interactions between the sections could have played in altering what would have been the true relationships between the conditions and the solution test. In testing SIEC in this domain, we controlled for ordering effects and separated conditions allowing for a less confounded inference about SIEC's role in producing schematic knowledge compared to a control group. We found conventional practice, with feedback illustrating the solution steps, was sufficient to produce schematic knowledge beyond baselines and SIEC offered no additional advantage.

Finally, our attempt to measure the difference in learning efficiency across groups for SIEC vs. conventional practice offers a new methodological tool for future studies reviewing this literature to investigate instructional differences while allowing time to vary across conditions. Allowing time to vary allows task dynamics of differing complexity (such as what occurs for SIEC vs. conventional practice) to be compared in full without placing artificial constraints on those dynamics for the sake of keeping time equal. Although we failed to find a significant difference in learning efficiency due to the near equality in posttest means, past conclusions about SIEC's efficacy might have been altered had past studies used such a measure and allowed learning time to vary.

Conclusion

This paper offered a randomized-control design without order effects nor erroneous carry-over effects, a diverse and large sample of participants spanning the U.S.

who vary in their mathematical ability and feedback during learning which appropriately illustrates the differing task dynamics of the conditions while still allowing for a fair comparison of conditions. Such "fairness" comes into play when the conventional problem solving group is illustrated the solution procedure, as we would expect to happen in a classroom despite our study not being a classroom research-design. Illustrating the solution procedure at the end of the feedback for the SIEC condition allowed us to control for that element of feedback in light of participants being exposed to the full SIEC task dynamics beforehand. At least for the subject domain of basic probability theory with multiple choice answers, we found SIEC to offer no advantage over conventional practice in enhancing problem solving performance, which stands in contrast to the literature's findings in domains where problems can contain direct textual representations of problem solving concepts as is required for SIEC to be effective (e.g., such as in basic probability theory (Birney et al., 2005) and geometry (Low & Over, 1992).

We conclude with some comments about a couple of our study's weaknesses. The most obvious is one of generalizability to school settings where the conditions we were investigating would be implemented. Despite our conditions' feedback being realistic, the time constraints imposed on participants' learning may lead to alternative explanations due to those constraints, such as SIEC possibly yielding better outcomes over greater time intervals while falling short of its potential merits in the short term our study provides. We find this to be a viable possibility and one to be remedied in future research with a similar design except within a classroom. The second weakness is the multiple choice nature of the stimuli. Although we showed the stimuli themselves were

not solely responsible for our findings, past studies did allow open ended problem solving to arrive at a correct solution and to show their attempted solution steps to the researchers. Such a consideration would have allowed us to determine if SIEC influenced what participants reported in their solution procedures. Accounting for this in the future could yield a finer-grained investigation into SIEC's effects on how students reason with key statements in problems to arrive at a solution.

References

Baguley, T. (2009). Standardized or simple effect size: What should be reported?. *British Journal of Psychology, 100*(3), 603-617.

Birney, D. P., Fogarty, G. J., & Plank, A. (2005). Assessing schematic knowledge of introductory probability theory. *Instructional Science, 33*(4), 341-366.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi:10.1177/1745691610393980

Cameron, L. (1993). Degrees of knowing: An exploration of progressive abstraction in language awareness work. *Language Awareness, 2*(1), 3-13.

Chen, Z. (1999). Schema induction in children's analogical problem solving. *Journal of Educational Psychology*, *91*(4), 703.

Chen, Z., & Daehler, M. W. (1989). Positive and negative transfer in analogical problem solving by 6-year-old children. *Cognitive Development*, *4*(4), 327-344.

Chen, Z., & Mo, L. (2004). Schema induction in problem solving: a multidimensional analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 583.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive science*, *5*(2), 121-152.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*(3), e57410.

Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research*, 89-105.

Fuchs, L. S., Fuchs, D., Prentice, K., Hamlett, C. L., Finelli, R., & Courey, S. J. (2004). Enhancing Mathematical Problem Solving Among Third-Grade Students With Schema-Based Instruction. *Journal of Educational Psychology*, *96*(4), 635.

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, *95*(2), 393.

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

Halford, G. S., Bain, J. D., Maybery, M. T., & Andrews, G. (1998). Induction of relational schemas: Common processes in reasoning and complex learning. *Cognitive Psychology*, *35*(3), 201-245.

Halford, G. S., & Busby, J. (2007). Acquisition of structured knowledge without instruction: The relational schema induction paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 586-603.

Jitendra, A. K., & Hoff, K. (1996). The effects of schema-based instruction on the mathematical word-problem-solving performance of students with learning disabilities. *Journal of Learning Disabilities*, *29*(4), 422-431.

Jitendra, A. K., Star, J. R., Starosta, K., Leh, J. M., Sood, S., Caskie, G., ... Mack, T. R. (2009). Improving seventh grade students' learning of ratio and proportion: The role of schema-based instruction. *Contemporary Educational Psychology*, *34*(3), 250-264.

Low, R., & Over, R. (1989). Detection of missing and irrelevant information within algebraic story problems. *British Journal of Educational Psychology*, *59*(3), 296-305.

Low, R., & Over, R. (1990). Text editing of algebraic word problems. *Australian Journal of Psychology*, *42*(1), 63–73.

Low, R., & Over, R. (1992). Hierarchical ordering of schematic knowledge relating to area-of-rectangle problems. *Journal of Educational Psychology*, *84*(1), 62–69.

Low, R., & Over, R. (1993). Gender differences in solution of algebraic word problems containing irrelevant information. *Journal of Educational Psychology*, *85*(2), 331–339.

Low, R., Over, R., Doolan, L. & Michell, S. (1994). Solution of algebraic word problems following training in identifying necessary and sufficient information within problems. *American Journal of Psychology*, *107*(3), 423–439.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, *44*(1), 1–23. doi:10.3758/s13428-011-0124-6

McNeil, N. M., & Fyfe, E. R. (2012). "Concreteness fading" promotes transfer of mathematical knowledge. *Learning and Instruction*, *22*(6), 440-448

Ngu, B. H., Low, R., & Sweller, J. (2002). Text editing in chemistry instruction. *Instructional Science*, *30*(5), 379-402.

Ngu, B. H., & Yeung, A. S. (2013). Algebra word problem solving approaches in a chemistry context: Equation worked examples versus text editing. *The Journal of Mathematical Behavior*, *32*(2), 197-208.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making, 5*(5), 411-419.

Pavlik Jr., Philip I., et al. (2007). "The FaCT (Fact and Concept Training) System: A new tool linking cognitive science with educators." *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society*.

R Development Core Team. (2011). R: A language and environment for statistical computing. Retrieved from http://www.R-project.org

Robins, S., & Mayer, R. E. (1993). Schema training in analogical reasoning. *Journal of educational Psychology*, *85*(3), 529-538.

Ross, B. H., Taylor, E. G., Middleton, E. L., & Nokes, T. J. (2008). Concept and category learning in humans. *Learning and Memory: A Comprehensive Reference*, *2*(3), 535-556.

Valentine, J. C., & Cooper, H. (2003). Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes. Washington, DC: What Works Clearinghouse.

Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist*, *43*(1), 16-26.

Viskontas, I. V., Morrison, R. G., Holyoak, K. J., Hummel, J. E., & Knowlton, B. J. (2004). Relational integration, inhibition, and analogical reasoning in older adults. *Psychology and Aging*, *19*(4), 581-591.

Wylie, R., Koedinger, K. R., & Mitamura, T. (2009). Is self-explanation always better? The effects of adding self-explanation prompts to an English grammar tutor.

In *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 1300-1305).

Xin, Y. P. (2008). The effect of schema-based instruction in solving mathematics word problems: An emphasis on prealgebraic conceptualization of multiplicative relations. *Journal for Research in Mathematics Education*, 526-551.

Appendix A

There already exist metrics for "instructional efficiency" (e.g., Van Gos & Paas, 2008) and learning efficiency for pretest-posttest differences based on Z-scores of both normalized gain and time spent learning (e.g., Wylie, Koedinger, & Mitamura, 2009). The latter is only appropriate for within subjects studies and is standardized by *SD* (see Baguley, 2009 for the issues with *SD* standardizing), and therefore inappropriate for our goals. The former, although allowing comparisons across groups and therefore similar to what we attempted, standardizes by *SD* and incorporates constructs requiring additional theoretical assumptions and complicated measurements (e.g., incorporating "mental effort"). Our measure will be based solely on theory-free, easily obtained information (i.e., time) applicable even to those outside human factors/cognitive load research.

To reconcile these issues we propose the following metric. For a between groups, pretest-learning-posttest design, let I and A be the intervention and the alternative conditions, respectively. Then, let L_Time be the average time spent in the learning phase. Finally, let $L\_Average$ be the average of the two averages to standardize average learning times with respect to the grand average across groups. The metric is defined as:

$$DLE(I,A) = L\_Average\left(\frac{(I_{Post} - I_{Pre})}{I_{L\_Time}} - \frac{(A_{Post} - A_{Pre})}{A_{L\_Time}}\right)$$

where DLE is "Difference in Learning Efficiency" and the absolute[7] gain scores are the average gains across participants in a group.

[7]Normalized gains were not included because when posttest is 100% and gain is positive, gain = (100 - pre)/(100 - pre) = 1. Therefore, no matter the pretest score, improvement is impossible to ascertain. Absolute gain scores offer interpretations for all posttest scores.

Appendix B

Which gender do you most closely identify as? If another gender identity better describes you, select "other."

a. Male
b. Female
c. Other

What is your age (years)?

a. 18-24
b. 25-34
c. 35-44
d. 45-54
e. 55-64
f. 65+

What is the highest degree or level of school you have completed? If you are within 6 months of completing a particular level, choose that one instead.

a. No schooling completed
b. 8^{th} grade
c. High School or equivalent (for example: GED)
d. Trade/technical/vocational training
e. Associates degree
f. Bachelor's degree
g. Master's degree
h. Doctorate degree

How long have you been on Mechanical Turk (months)?

a. Less than a month
b. 1-6
c. 6-12
d. 12-18
e. 18-24
f. 25+

On average, about how many hours per month have you spent completing HITs on Mechanical Turk? If you have been on Mechanical Turk for less than a month, just answer how many hours.

a. Less than an hour
b. 1-6
c. 6-12
d. 12-18
e. 18-24
f. 25+

What is the highest level of math you have completed? If you are within 2 months of completing a particular course, choose that one instead.

a. No math courses taken
b. No math beyond Middle School Algebra
c. High School level Algebra
d. College level Algebra
e. College level Pre-Calculus 1
f. College level Pre-Calculus 2
g. College level Calculus 1
h. I have taken math beyond Calculus 1

Which is the closest to describing how you spend your time on Mechanical Turk when it comes to HITs involving math? If you have limited exposure to math HITs, answer based on how you *would* spend your time.

a. I always avoid them
b. I typically avoid them unless given another incentive
c. I participate in them like I would any other HIT
d. If it's a decision between a math and non-math HIT, assuming they would pay about the same, I usually choose the math one
e. I only seek out HITs involving math

Appendix C

Instructions for Study 1 as it appeared (including indents and spacing):

Before Demographics

"The study you're participating should last around 30 minutes. First, you'll be asked several demographic related questions. In answering these, you'd type the letter matching the answer best describing you. For example, if the question asks about your age and "b" reflects the age interval describing you, you'd press "b." After this you'll be given the final set of instructions on what to expect from the problems you'll solve."

After Demographics and Before Pretest

"Each problem you'll be given has 4 multiple choice answer options with only one correct answer, and you'll type in a letter for the correct answer just like before. Carefully read the problems as they may contain irrelevant information. One of the answer options will always be "This problem can't be solved." This means there's either not enough information in the problem for solving, or there is enough information for solving, but multiple answers are possible due to key information being missing. You're encouraged to grab a sheet of scratch paper and not to simplify your answers. As you won't be given feedback after each answer, you should do the best you can. Thank you for your participation!"

Instructions for Study 2 Control group:

Before Demographics

[Same as above except they're told the study should last around 60 minutes]

After Demographics and Before Learning Phase

[Same as above except "If you get a problem wrong you'll receive feedback customized to your answer. Once this task is completed, you'll see a link to a mildly entertaining 8-minute video. Watch the video then immediately return to this window and follow the final set of instructions."]

After Learning Phase and Before Testing Phase

"Now you'll be tested on what you learned from prior practice. This final set of problems will be similar to before except there won't be feedback this time, so do the best you can. Thank you for your participation!"

Instructions for Study 2 SIEC group:

Before Demographics

[Same as above]

After Demographics and Before Learning Phase

"Each problem you'll be given has 4 multiple choice answer options with only one correct answer, and you'll type in a letter for the correct answer just like before. Carefully read the problems as they may contain irrelevant information. In this task you'll classify each problem based on its statements and you won't use procedures to solve for an answer like you'd expect in a math course. If there's either not enough information in the problem to solve for *any* answer, or there *is* enough information for solving but multiple answers are possible due to key information being missing, then the problem has *missing* information. If there's deceptive or redundant information in the problem then the problem has *irrelevant* formation. If the problem is solvable, whether or not it has irrelevant information, then the problem is *sufficient*. You'll receive feedback after you answer. If the problem contains missing and/or irrelevant information and you're correct about that, you'll then be given a follow-up problem asking how the problem should be changed. This follow-up problem also gives feedback for answers. Once the task is completed, you'll see a link to a mildly entertaining 8-minute YouTube video. Watch the video then immediately return to this window and follow the final set of instructions."

After Learning Phase and Before Testing Phase

"Now you'll be tested on what you learned from prior practice. This final set of problems will be similar to before except you'll need to solve for the correct answer (a number) instead of labeling the problem. You're encouraged to use scratch paper. Don't simplify your answers. Also, there won't be feedback this time, so do the best you can. Thank you for your participation!"

Example problems exactly as listed in study (including indents and spacing):

Correct answers marked highlighted in grey. A-NC = addition rule with no co-occurring events; S = sufficient information; I = contains irrelevant information; M = contains missing information.

A-NC /S

A bag contains 6 balls in total: 2 red, 3 yellow, and 1 blue. If you randomly take one ball out, what is the probability that you'll get either 1 blue or 1 red?

a. Sufficient but with irrelevant information

b. Missing but no irrelevant information

c. Sufficient with no irrelevant information

d. Missing and with irrelevant information

IF CORRECT: Exactly. Since there are no flaws in this problem's content, there's no need to fix it before problem solving. Given this is the case, the following is the procedure for solving this particular problem.

There are 6 balls in total. Because there's 1 blue and 2 red, the probability of picking a blue is 1/6 and the probability of picking a red is 2/6. Lastly, because either picking one color ball or another can take place, but not both, we add the two probabilities. [Then they move on to the next problem.]

IF INCORRECT: Initiate review period:

IF A, B, or D:

That's not quite right. In the problem there's information about the total number of balls in the bag, the number of blue and red balls within this total, and the fact only one ball could be taken out, either blue or red. This is the only information needed for solving this problem and there's no additional redundant or misleading information. Here's how you would solve this problem since it already had sufficient information:

There are 6 balls in total. Because there's 1 blue and 2 red, the probability of picking a blue is 1/6 and the probability of picking a red is 2/6. Lastly, because either picking one color ball or another can take place, but not both, we add the two probabilities. [Then they move on to the next problem.]

A-NC /I

A bag contains 4 balls in total: 2 red, 1 yellow, and 1 blue. If you randomly take one ball out, what is the probability that you'll get either 1 blue or 1 red if the blue and red are much lighter than the yellow?

a. Missing and with irrelevant information

b. Missing but no irrelevant information

c. Sufficient but with irrelevant information

d. Sufficient with no irrelevant information

IF CORRECT:

Which statement(s) should be deleted to make this problem "Sufficient with no irrelevant information?"

A bag contains 4 balls in total: 2 red, 1 yellow, and 1 blue. If you randomly take one ball out, what is the probability that you'll get either 1 blue or 1 red if the blue and red are much lighter than the yellow?

a. Both b and c

b. "…if the blue and red balls are much lighter than the yellow"

c. "…in total…"

d. Another statement

IF CORRECT:

Exactly. Once the statement you selected is removed there will be no flaws in this problem's content. The following is the procedure for solving this particular problem now that it's ready to be solved.

There are 4 balls in total. Because there's 1 blue and 2 red, the probability of picking a blue is 1/4 and the probability of picking a red is 2/4. Lastly, because either picking one color ball or another can take place, but not both, we add the two probabilities. [Then they move on to the next problem.]

IF INCORRECT: Initiate review period.

IF A:

Although b is correct, the problem requires there be information about the total number of balls. Without this information, the problem would be unsolvable. Once the correct statement has

been removed, the following is the procedure for solving this particular problem.

There are 4 balls in total. Because there's 1 blue and 2 red, the probability of picking a blue is 1/4 and the probability of picking a red is 2/4. Lastly, because either picking one color ball or another can take place, but not both, we add the two probabilities. [Then they move on to the next problem.]

IF C:

That's not quite right. The problem requires there be information about the total number of balls. Without this information, the problem would be unsolvable. Answer b is correct because how much the balls weigh is irrelevant in calculating the probability. Once the statement in b has been removed, the following is the procedure for solving this particular problem.

There are 4 balls in total. Because there's 1 blue and 2 red, the probability of picking a blue is 1/4 and the probability of picking a red is 2/4. Lastly, because either picking one color ball or another can take place, but not both, we add the two probabilities. [Then they move on to the next problem.]

IF D:

That's not quite right. There aren't any other statements in this problem warranting concern than what's reflected in the other answering options. Answer b is correct because how much the balls weigh is irrelevant in calculating the probability. Once the statement in b has been removed, the following is the procedure for solving this particular problem

There are 4 balls in total. Because there's 1 blue and 2 red, the probability of picking a blue is 1/4 and the probability of picking a red is 2/4. Lastly, because either picking one color ball or another can take place, but not both, we add the two probabilities. [Then they move on to the next problem.]

IF INCORRECT: Initiate review period.

IF A:

That's not quite right. Although this problem contains irrelevant information, there's still enough for problem solving. [They then go to the second step of this problem.]

IF B:

That's not quite right. Not only is there enough information to solve this problem, but the problem also contains irrelevant information. [They then go to the second step of this problem.]

IF D:

That's not quite right. Although this problem contains enough information for problem solving, there's also irrelevant information. [They then go to the second step of this problem.]

A-NC /M

A bag contains at least the following balls: 6 red, 1 yellow, and 1 blue. What is the probability that, after randomly selecting one ball, you'll get either 1 blue or 1 red?

a. Sufficient but with irrelevant information

b. Sufficient with no irrelevant information

c. Missing and with irrelevant information

d. Missing but no irrelevant information

IF CORRECT:

What content should be added to make this problem "Sufficient with no irrelevant information?"

A bag contains at least the following balls: 6 red, 1 yellow, and 1 blue. What is the probability that, after randomly selecting one ball, you'll get either 1 blue or 1 red?

a. That an additional ball needs to be withdrawn

b. Information about how many balls there are in total.

c. Either a or b but not both

d. Both a and b

IF CORRECT:

Exactly. Once the statement you selected is added there will be no flaws in this problem's content. [Then they move on to the next problem.]

IF INCORRECT:

IF A:

That's not quite right. For this problem there's no need to withdraw an extra ball. However, the statement about the total number of balls does need to be added. Currently the problem only clarifies balls we know exist in the bag. It doesn't tell us that's all there is in total. [Then they move on to the next problem.]

IF C:

Although both statements shouldn't be added, the choice between the two isn't arbitrary. For this problem, there's no need to withdraw an extra ball. However, the statement about the total number of balls does need to be added. Currently the problem only clarifies balls we know exist in the bag. It doesn't tell us that's all there is in total. [Then they move on to the next problem.]

IF D:

Although b is correct, there's no need to withdraw an extra ball in this problem. [Then they move on to the next problem.]

IF INCORRECT:

IF A:

That's not quite right. Although this problem lacks important information, there's still no irrelevant information in the problem. [They then go to the second step of this problem.]

IF B:

That's not quite right. There is there no irrelevant information in this problem, and the problem lacks the necessary information. [They then go to the second step of this problem.]

IF C:

That's not quite right. Although there's no irrelevant information in this problem, the problem still lacks the necessary information. [They then go to the second step of this problem.]

CONTROL

A-NC / S

A bag contains 6 balls in total: 2 red, 3 yellow, and 1 blue. If you randomly take one ball out, what is the probability that you'll get either 1 blue or 1 red?

a. 3/6

b. This problem can't be solved

c. 2/30

d. 16/36

IF CORRECT: [They move on to next problem]

IF INCORRECT: Initiate review period

IF B:

That's not quite right. There is enough information to solve the problem. Here's the procedure for solving the problem:

There are 6 balls in total. Because there's 1 blue and 2 red, the probability of picking a blue is 1/6 and the probability of picking a red is 2/6. Lastly, because either picking one color ball or another can take place, but not both, we add the two probabilities. [Then they move on to the next problem.]

IF C OR D:

That's not quite right. Here's the procedure for solving the problem:

There are 6 balls in total. Because there's 1 blue and 2 red, the probability of picking a blue is 1/6 and the probability of picking a red is 2/6. Lastly, because either picking one color ball or another can take place, but not both, we add the two probabilities. [Then they move on to the next problem.]

A-NC / I

A bag contains 4 balls in total: 2 red, 1 yellow, and 1 blue. If you randomly take one ball out, what is the probability that you'll get either 1 blue or 1 red if the blue and red are much lighter than the yellow?

a. 10/16

b. 3/4

c. 2/12

d. This problem can't be solved

IF CORRECT: [They move on to next problem]

IF INCORRECT: Initiate review period

IF A OR C:

That's not quite right. Here's the procedure for solving the problem:

There are 4 balls in total. Because there's 1 blue and 2 red, the probability of picking a blue is 1/4 and the probability of picking a red is 2/4. Lastly, because either picking one color ball or another can take place, but not both, we add the two probabilities. [Then they move on to the next problem.]

IF D:

That's not quite right. There *is* enough information to solve the problem. Here's the procedure for solving the problem:

There are 4 balls in total. Because there's 1 blue and 2 red, the probability of picking a blue is 1/4 and the probability of picking a red is 2/4. Lastly, because either picking one color ball or another can take place, but not both, we add the two probabilities. [Then they move on to the next problem.]

A-NC /M

A bag contains at least the following balls: 2 red, 1 yellow, and 3 blue. What is the probability that, after randomly selecting one ball, you'll get either 1 blue or 1 red?

a. 5/6

b. This problem can't be solved

c. 6/30

d. 24/36

IF CORRECT: [They move on to next problem]

IF INCORRECT: Initiate review period

IF NOT B:

That's not quite right. This problem can't be solved. [Then they move on to the next problem.]