COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 40 (2016) 1036–1079
Copyright © 2015 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12275

Learning Problem-Solving Rules as Search Through a Hypothesis Space

Hee Seung Lee,[a] Shawn Betts,[b] John R. Anderson[b]

[a]*Department of Education, Yonsei University*
[b]*Department of Psychology, Carnegie Mellon University*

Received 4 January 2015; received in revised form 19 April 2015; accepted 22 April 2015

Abstract

Learning to solve a class of problems can be characterized as a search through a space of hypotheses about the rules for solving these problems. A series of four experiments studied how different learning conditions affected the search among hypotheses about the solution rule for a simple computational problem. Experiment 1 showed that a problem property such as computational difficulty of the rules biased the search process and so affected learning. Experiment 2 examined the impact of examples as instructional tools and found that their effectiveness was determined by whether they uniquely pointed to the correct rule. Experiment 3 compared verbal directions with examples and found that both could guide search. The final experiment tried to improve learning by using more explicit verbal directions or by adding scaffolding to the example. While both manipulations improved learning, learning still took the form of a search through a hypothesis space of possible rules. We describe a model that embodies two assumptions: (1) the instruction can bias the rules participants hypothesize rather than directly be encoded into a rule; (2) participants do not have memory for past wrong hypotheses and are likely to retry them. These assumptions are realized in a Markov model that fits all the data by estimating two sets of probabilities. First, the learning condition induced one set of Start probabilities of trying various rules. Second, should this first hypothesis prove wrong, the learning condition induced a second set of Choice probabilities of considering various rules. These findings broaden our understanding of effective instruction and provide implications for instructional design.

Keywords: Problem solving; Hypothesis testing; Search space; Examples; Verbal direction; Markov processes

Correspondence should be sent to Hee Seung Lee, Department of Education, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, Korea. E-mail: hslee00@yonsei.ac.kr

1. Introduction

Faced with a new class of problems, learners have to internalize a new set of rules for solving these problems. Naively, one might think the best way to learn such rules is by being directly told them via verbal instruction. Others might think learners learn better when they are given an example they can model from (e.g., Reed & Bolstad, 1991; Zhu & Simon, 1987). On the other hand, some studies showed that verbal instructions and examples were equally effective (e.g., Cheng, Holyoak, Nisbett, & Oliver, 1986; Fong, Krantz, & Nisbett, 1986). Given mixed results about relative efficacy of different types of instruction (for a review, see Lee & Anderson, 2013; Wittwer & Renkl, 2008), it seems that it is not really a matter of the student internalizing rules as instructed; it really is a matter of the instruction influencing the rules that the student constructs. What is important is how obvious the correct rules are given the whole instructional context.

1.1. Learning problem-solving rules as a hypothesis search process

Learning how to solve a class of problems can be characterized as a search through a space for a rule[1] that will solve the problems. Hypothesis testing in the domain of problem solving can be seen as a similar process to the hypothesis testing in concept learning (e.g., Bower & Trabasso, 1964; Levine, 1975; Nosofsky, Palmeri, & McKinley, 1994; Restle, 1962). For example, according to Restle's (1962) formulation of hypothesis-testing theory, a participant continuously generates hypotheses about the concept and tests those hypotheses. Whenever a hypothesis makes a correct classification, the participant retains it; whenever it leads to an incorrect classification, the participant rejects it. If the sample of hypotheses contains a correct hypothesis, the problem will be eventually solved when wrong hypotheses are eliminated by testing.

With regard to learning of problem solving, we can think of a similar process. Participants solve a series of problems while being told whether their solution is correct or not. Participants can come up with a hypothesis as to how to solve the problems based on whatever examples or explanations are present or just based on background knowledge if no instructional support is present. Based on the feedback provided about the correctness of the problem solution, participants can retain their initial hypothesis for further testing on successive problems or discard it and choose a new one. However, many problem-solving situations including the one we will study differ from classic concept learning experiments in that an error does not necessarily imply the hypothesized rule is wrong. Rather, the participant may have made a miscalculation while nonetheless applying the correct rule.

As we will see, the learning situation induces strong initial biases in the student as to what the correct rule is. If this first hypothesis is wrong, students have to engage in a search for some other rules. When there is a large search space of hypotheses, problem solvers might not get around to testing the correct rule, particularly when the learning situation leaves that rule obscure. A further challenge is remembering which hypotheses have already been tested. A student may wind up testing the same hypothesis over and over again. In fact, early theories of concept learning (Bower & Trabasso, 1964; Restle, 1962) assumed that participants did not use memory of information presented on the past trials. Bower and Trabasso supported this no-memory assumption via several experiments (e.g., Bower & Trabasso, 1963, 1964; Trabasso, 1963; Trabasso & Bower, 1964).

Learning a problem-solving rule can pose an additional challenge not true in many of the simple concept learning experiments. Problem-solving tasks can pose a heavy burden on working memory. Not only do problem solvers have to formulate and compare alternative hypotheses, they also have to perform the computations implied by these hypotheses. There are numerous demonstrations in the domain of problem solving that a heavy cognitive load during problem solving interferes with learning (e.g., Mawer & Sweller, 1982; Sweller, 1983, 1988; Sweller, Mawer, & Howe, 1982). One can imagine that this working memory load will make it even harder to remember what hypotheses they have already tested than in the classic concept learning experiments.

This study will present a model that embodies two hypotheses about the nature of the learning process, at least in our current experiments:

1. Participants learn by generating rules and seeing if the rules produce the correct results. They keep generating rules until they find a successful one. Instructions are not directly encoded into rules. The effect of instruction is to *bias* the hypothesis generation process that produces these candidate rules.
2. If participants come up with an incorrect hypothesis and reject it, the probability of their coming back to that incorrect hypothesis later is just the same as if they had never tried the hypothesis. That is, they have *no memory* for what has failed.

The model, incorporating just these two hypotheses without further detail, does a surprisingly good job in accounting for our data.

2. Current experiments

This series of experiments began with some earlier research (Anderson, Lee, & Fincham, 2014; Lee, Anderson, Betts, & Anderson, 2011; Lee, Betts, & Anderson, 2015; Lee, Fincham, Betts, & Anderson, 2014) where we studied students learning an extensive curricula of "data-flow" diagrams under different amount of instructional guidance. The task began with filling in the empty boxes in a diagram like Fig. 1 and most students showed considerable difficulty when the structure of the data-flow diagram involved multiple paths to combine. Our prior research mostly focused on how to help students overcome this learning difficulty by varying the amount of instruction.

Although students had considerable struggles with later parts of the curriculum, we observed that most students found beginning diagrams (the one like Fig. 1) remarkably easy. The majority of students could solve their first problem correctly without instruction. The correct hypothesis, that they were to evaluate the expression in a box and put the answer where the arrow pointed, seemed immediately obvious. In the current research

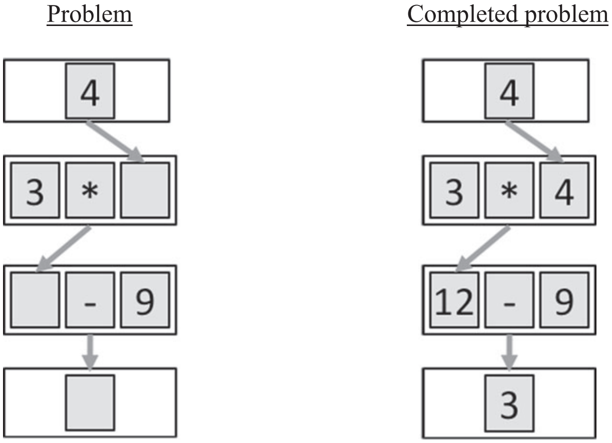
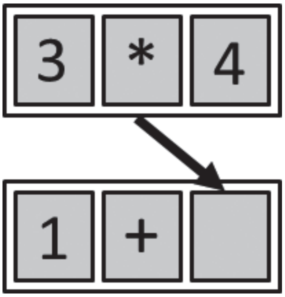


Fig. 1. An example of data-flow diagram. The task is to fill a number into the empty portions of the diagram.

we looked in more detail at the factors that would make such a hypothesis more or less obvious. We developed simplified tasks where the correct rule was either obvious or obscure. This allowed us to contrast learning an Easy rule vs. a Hard rule, while keeping the mathematical calculations comparable.

Fig. 2 shows an example of problems used in this study. In this example, participants have to determine a number to fill in the empty box given the three numbers and two operators already provided. Given just the problem alone there are a number of possible rules. On the basis of our prior work with problems like Fig. 1, we expected a strong tendency to perform the operation in the top rectangular box and fill the answer (in this case 12) into the empty box to which the arrow pointed. We called this the “Easy” rule. In our past research we had observed students assume this rule without any instruction, undoubtedly based on prior experience with mathematical expressions and conventions of such diagrams. However, there is a sense in which this is not such a good rule for the problem in Fig. 2 as it was for Fig. 1. It leaves the other number (i.e., 1) and operator (i.e., +) in the bottom box unexplained (Lewis, 1988). In some conditions of the experiments in this study, we observed participants use what we have called the “Equality” rule. This is that the top and bottom rectangular boxes should have the same net value and so the empty box should be filled with the number that achieves this equality (in this case 11). This approach explains all the numbers and operators but leaves the arrow unexplained. Sometimes, however, what students need to learn is not obvious and we wanted to study such a situation. Therefore, we had some of our participants try to learn what we called the “Hard” rule. Participants had to treat the bottom rectangular box (i.e., 1 + ?) as an expression and the number that is on the opposite side of the empty box (i.e., 3) as a result value. Because 1 + 2 = 3, the correct answer to this problem is 2. In this type of problem, the rule is not intuitive, on a number of grounds, including that it contradicts a natural interpretation of what the arrow might mean and that the relevant terms (i.e., 1, +, and 3) are not grouped in the same box. However, ignoring such perceptual features, applying it involves performing



Hypotheses	Solutions
Easy	3 * 4 = x → x = 12
Equality	3 * 4 = 1 + x → x =11
Hard	1 + x = 3 → x = 2

Fig. 2. An example of problems used in this study. Correct answer to this problem becomes different depending on the type of problem rules participants had to learn. In the Easy problem, the correct rule is $3 \times 4 = x \rightarrow x = 12$; thus, the correct answer is 12. In the Hard problem, the correct rule is $1 + x = 3 \rightarrow x = 2$; thus, the correct answer is 2.

a simple mathematical calculation ($3-1$ in this example) just like the Easy rule. In both cases, the rule uses two of the numbers and one operator and leaves the remaining number and operator unexplained. There are many other possible rules participants might adopt in attempting to solve these problems, but the Easy, Hard, and Equality rules were the ones we observed with some frequency.

Across four experiments, we had participants solve a series of problems to learn either Easy or Hard rules and examined how different properties of the learning situation influenced their success. Table 1 shows a list of experimental manipulations applied across the four experiments. In Experiment 1, we simply investigated properties of the problems when they were presented without instructional support. Specifically, we manipulated how easy it was to compute the answers for specific rules. In Experiment 2, we tested effects of example content on learning. By varying the example content, we manipulated whether the presented example supports one specific hypothesis (good example) that is consistent with correct solution rule or dual hypotheses (ambiguous example) that are consistent with both correct and incorrect solution rules. In Experiment 3, we contrasted examples with providing verbal instructions that directed the learner's attention to relevant elements of the problem. Both the examples and verbal instructions in this experiment stopped short of telling the participant what the rule was. In the final experiment, we tried to create verbal directions that specified the rule and provided scaffolding to the example to make the rule apparent. However, even with our best attempts we will see that learning was not always successful.

Table 1
A list of experimental manipulations applied across the four experiments

	Exp 1	Exp 2	Exp 3	Exp 4
Problem property				
Constrained (Non-integer answer for incorrect rules)	✓			
Unconstrained (Integer answer for incorrect rules)	✓	✓	✓	✓
Example				
No-example	✓	✓		
Good (Examples support one specific hypothesis)		✓	✓	
Ambiguous (Examples support dual hypotheses)		✓	✓	
Scaffolded (Equation plus good example)				✓
Verbal direction				
No-direction	✓	✓		
Good (Verbal direction supports one hypothesis)			✓	
Ambiguous (Verbal direction supports dual hypotheses)			✓	
Enhanced (Explicit rule is provided)				✓

Notes. The marked cells represent the characteristics of problem or instructions used in each experiment. The shaded cells indicate the major experimental manipulation.

To properly assess our theoretical model, we wanted to get large samples of participants and so ran the experiments on an online labor market, Amazon Mechanical Turk. Participants could quit anytime in the middle of the study and this resulted in a 12–17% dropout rate in each experiment (Table 2 provides a summary and breakdown by experiment). We kept sampling until we got at least 50 participants in each condition of each experiment. Also, we tested the first experiment with both Mechanical Turk participants (Experiment 1a) and undergraduate students (Experiment 1b) to confirm that our findings are not attributable to characteristics of specific population.

2.1. Markov model

To test our two major claims about the hypothesis-testing behavior of the participants (that the effect of instruction is to bias the hypothesis generation process and that there is no memory for past hypotheses) we developed a series of Markov models. Markov models were used to characterize learning in the early concept learning experiments (e.g., Richard, Cauzinille, & Mathieu, 1973; Trabasso & Bower, 1968) and have been applied to other domains in psychology (e.g., Brainerd, 1979). The development of dynamic programming techniques associated with Hidden Markov Models (Rabiner, 1989) has led to

Table 2
Number of participants who quit the study and those who completed the study in Experiments 1–4

	Drop-out	Completed
Exp 1a		
Easy		
Constrained	7	50
Unconstrained	6	50
Hard		
Constrained	14	50
Unconstrained	11	50
Exp 2		
Easy		
Good example	6	50
Ambiguous example	1	51
No example	3	50
Hard		
Good example	23	51
Ambiguous example	8	52
No example	19	50
Exp 3		
Easy		
Good verbal	1	50
Good example	5	50
Ambiguous verbal	9	50
Ambiguous example	5	50
Hard		
Good verbal	16	50
Good example	11	50
Ambiguous verbal	28	50
Ambiguous example	9	50
Exp 4		
Easy		
Enhanced verbal	5	54
Scaffolded example	4	54
Hard		
Enhanced verbal	15	51
Scaffolded example	5	51

their wide use in many domains outside of psychology, perhaps most notably in speech recognition. Our application of Hidden Markov Models will be close to work in educational data mining (e.g., Cohen & Beal, 2009; González-Brenes & Mostow, 2012).

A Hidden Markov Model can be characterized by three components describing state behavior, where the “states” are hypotheses in our case:

1. A start vector of probabilities of starting out in any state (i.e., with a specific hypothesis).
2. A transition matrix characterizing the probabilities of transitioning between states on each trial (i.e., sticking with the current hypothesis or trying a new one on the next problem).
3. An observation matrix giving the probability of different actions in a state (in our case, these are the answers the participant gives).

We will just consider four states: Easy (the participant is considering the Easy rule), Hard (the participant is considering the Hard rule), Equality (the participant is considering the Equality Rule), and Other (any other hypotheses). The bias interpretation of different instruction will be tested by seeing if we can reduce the transition matrix to the result of a bias for particular hypotheses. The no-memory assumption will be tested by whether we can use states that do not reflect any of the past hypotheses tested.

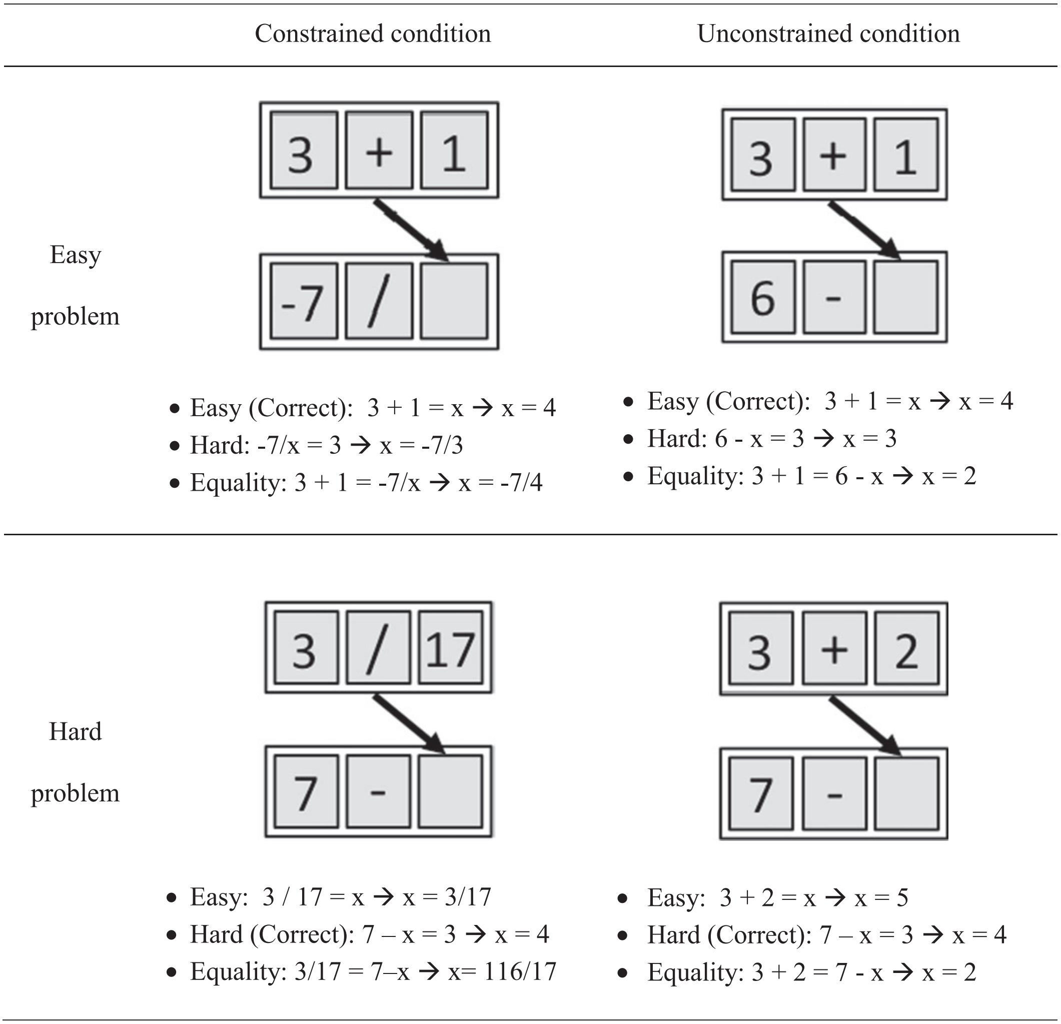
3. Experiment 1a: Effect of problems

In our early work with problems like those in Fig. 1, we found that participants could learn without any instructional support at all. Therefore, we decided to start with an experiment that had no instructional support. However, we did manipulate how difficult it was to apply the rules to the problems that participants saw. In the Constrained condition, we made it difficult to apply the Hard and Equality rules when the Easy rule was correct, while it was difficult to apply the Easy and Equality rules when the Hard rule was correct. In the Unconstrained condition, problems were constructed such that all three rules were easy to apply for both Easy and Hard problems. Fig. 3 shows examples of problems constructed for the Constrained and Unconstrained conditions. It also shows the required computations when each of the three rules is applied to the problem. In the Constrained condition, only the correct rule results in an integer number as an answer whereas incorrect rules result in non-integer answers (participants could enter the fractional answers). In the Unconstrained condition, application of all three rules results in integer numbers as answers. Thus, integer vs. non-integer answer manipulation served to constrain participants' behavior by creating computational ease or difficulty, also equally plausible, by using participants' general tendency of preferring integer numbers to non-integer numbers as answers in problem solving.

3.1. Method

3.1.1. Participants

A total of 238 participants took part in the study and they were recruited from Amazon Mechanical Turk. The participant pool (133 male and 105 female, M = 28.30 years, SD = 6.36) reported various levels of educational background (42% 4-year, 36% 2-year or some college, 9% high school, 11% masters, and 2% others). Of 238, 38 participants



Easy problem, Constrained condition:

- Easy (Correct): 3 + 1 = x → x = 4
- Hard: -7/x = 3 → x = -7/3
- Equality: 3 + 1 = -7/x → x = -7/4

Easy problem, Unconstrained condition:

- Easy (Correct): 3 + 1 = x → x = 4
- Hard: 6 - x = 3 → x = 3
- Equality: 3 + 1 = 6 - x → x = 2

Hard problem, Constrained condition:

- Easy: 3 / 17 = x → x = 3/17
- Hard (Correct): 7 – x = 3 → x = 4
- Equality: 3/17 = 7–x → x= 116/17

Hard problem, Unconstrained condition:

- Easy: 3 + 2 = x → x = 5
- Hard (Correct): 7 – x = 3 → x = 4
- Equality: 3 + 2 = 7 - x → x = 2

Fig. 3. Examples of Easy and Hard problems used in Constrained and Unconstrained conditions of Experiment 1. In the Constrained condition, application of correct rules results in an integer number as an answer and application of incorrect rules results in a non-integer number as an answer. In the Unconstrained condition, application of both correct and incorrect rules results in an integer number as an answer.

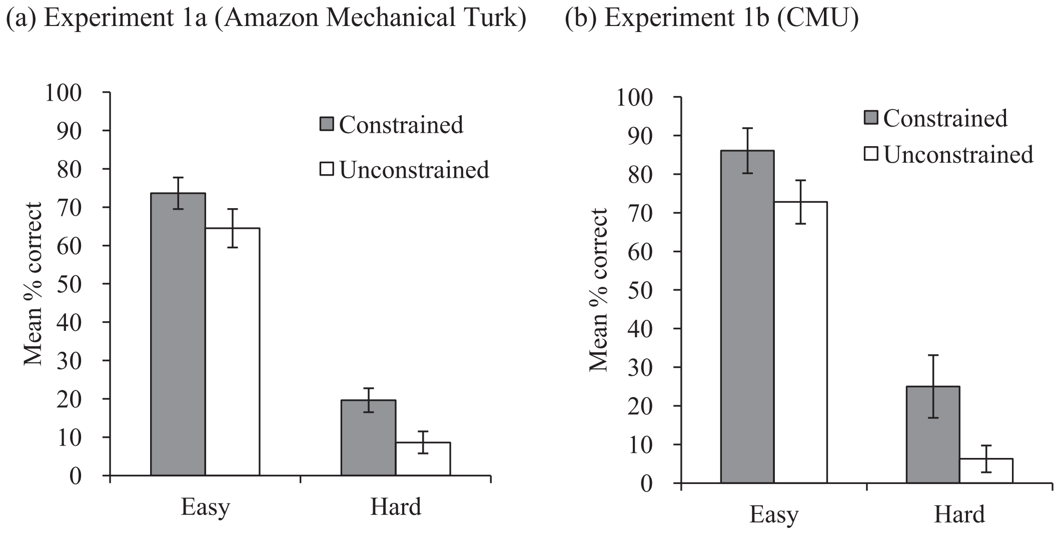
(16%) quit the study in various phases of the study. Table 2 shows the number of participants who quit the study and the number of those who completed the study. This left a total of 200 participants for data analysis. There were 50 participants in each of the Constrained and Unconstrained condition for each of Easy and Hard problems. Participants received a fixed amount of $1.5 plus a performance-based bonus (20 cents per correctly solved problem).

3.1.2. Design, materials, and procedure

A 2 × 2 between-subjects design was employed where we crossed the problem difficulty and properties of the problems. Participants learned to solve either Easy problems or Hard problems. We also manipulated problem property (Constrained vs. Unconstrained) by changing numbers and operators used in the presented problems. A total of 16 problems were constructed for each of the four conditions according to experimental manipulations. As in Fig. 3, the Easy problems in both conditions had the same arithmetic solutions as did the Hard problems in both conditions. The problems varied in terms of the location of empty box (left, right), direction of arrow (this always pointed to the empty box), and arithmetic operators (addition, subtraction, multiplication, division). Participants were given 1 minute for each problem. After entering a number they thought was the answer to the current problem, participants clicked a "done" button and the response was followed by immediate feedback presented for 2 seconds. The feedback page showed whether the submitted answer was correct or incorrect, but it did not show the correct answer or any kind of explanation. The entire experiment took about 15 minutes.

3.2. Results

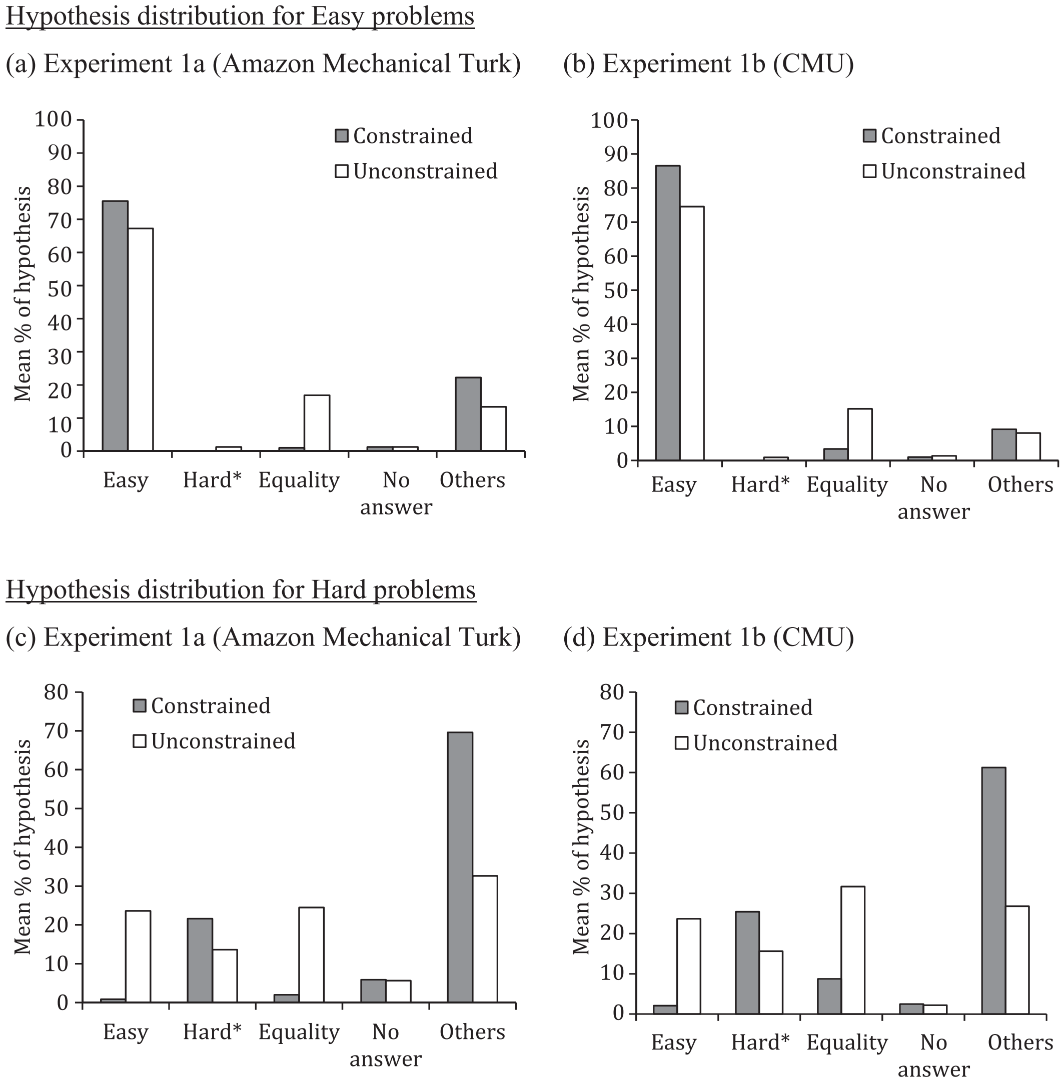
Fig. 4a shows mean percentage of correctly solved problems for the Easy and Hard problems in the Constrained and Unconstrained conditions. A 2 × 2 between-subjects analysis of variance (ANOVA) was performed on the percentage of correctly solved problems. Problem difficulty (Easy vs. Hard) and problem property (Constrained vs. Unconstrained) were included as between-subjects variables. There were significant main effects of both problem difficulty, $F(1, 196) = 200.78$, $p < .0001$, $\eta_p^2 = .506$, and problem


(a) Experiment 1a (Amazon Mechanical Turk)

(b) Experiment 1b (CMU)

Fig. 4. Mean percentage of correctly solved problems for Easy and Hard problems in Constrained and Unconstrained conditions of Experiments 1a and 1b. Error bars represent 1 standard error of mean.

property, $F(1, 196) = 6.74$, $p = .01$, $\eta_p^2 = .033$. Problem difficulty and problem property did not interact, $F < 1$. For both Easy and Hard problems, participants who were given Constrained problems solved significantly more problems correctly than those who were given Unconstrained problems. For Easy problems, Constrained participants (M = 74%, SD = 29%) solved about 9% more problems than the Unconstrained participants (M = 65%, SD = 36%). For Hard problems, Constrained participants (M = 20%, SD = 22%) solved about 11% more problems than the Unconstrained participants (M = 9%, SD = 20%).


Hypothesis distribution for Easy problems

(a) Experiment 1a (Amazon Mechanical Turk)

(b) Experiment 1b (CMU)

Hypothesis distribution for Hard problems

(c) Experiment 1a (Amazon Mechanical Turk)

(d) Experiment 1b (CMU)

Fig. 5. Distribution of different hypotheses tested by participants in Constrained and Unconstrained condition for Easy problems (a–b) and Hard problems (c–d) in Experiments 1a and 1b. * indicates the correct rule.

Fig. 4a provides a measure of overall success but does not tell us about the various hypotheses that participants were trying. To address this, Fig. 5 categorizes participants' responses based on the Easy, Hard, and Equality rules. To accommodate sign errors, we included negative of each rule into the stated rule in this analysis.[2] There were also cases where participants did not provide an answer or they timed out, and such cases were included in the "no answer" category. Any other responses that did not belong to the above four categories were categorized as "others."

For Easy problems (Fig. 5a), most trials were solved via the correct Easy rule. However, in the Unconstrained condition a good number of responses were classified as arising from use of the Equality rule. This seemed to be because of tendency to include all numbers and operators in the computations so that nothing was left unexplained in their problem solving. Although the Easy rule was correct in this case, the Easy rule left one number and one operator unexplained and this might not be natural to some participants. For Hard problems (Fig. 5c), very different patterns of hypotheses distribution were obtained between the Constrained and Unconstrained conditions. In the Unconstrained Hard condition, the Easy and Equality rules were tried with approximately equal frequency ($t(49) = 0.18$) and more often than the correct Hard rule ($t(49) = 2.06$ for contrast with Easy and 2.13 for contrast with Equality, both p's $< .05$). The only rule tried with much frequency in the Constrained Hard condition was the correct Hard rule, although its frequency was much lower than for the correct Easy rule in Fig. 5a. Participants were quite reluctant to consider the Easy and Equality rules which required fractional answers, probably reflecting both their past experience with answers and the extra difficulty of entering fractions.

4. Experiment 1b: Effect of problems

We also conducted Experiment 1 with undergraduate students to see whether our findings can be generalized to other population and to confirm that the experimental results in Experiment 1a are not dependent on the characteristics of specific population recruited from the online labor market. Fifty-six undergraduate students (32 male and 24 female, $M = 19.41$ years, $SD = 1.23$) from Carnegie Mellon University (CMU) participated in the experiment for course credit. Each participant was randomly assigned to one of the four conditions (13 constrained easy, 15 constrained hard, 14 unconstrained easy, and 14 unconstrained hard). Experimental design, materials, and procedures were identical to those of Experiment 1a.

Fig. 4b shows mean percentage of correctly solved problems for the Easy and Hard problems in the Constrained and Unconstrained conditions. Although the CMU population, in general, performed better than the Mechanical Turk population, overall patterns of the results were the same between the two populations. There were significant main effects of both problem difficulty, $F(1, 52) = 109.06$, $p < .0001$, $\eta_p^2 = .677$, and problem property, $F(1, 52) = 6.88$, $p = .01$, $\eta_p^2 = .117$. Problem difficulty and problem property did not interact, $F < 1$. For both Easy and Hard problems, participants who were given

Constrained problems solved significantly more problems correctly than those who were given Unconstrained problems. For Easy problems, Constrained participants (*M* = 86%, *SD* = 21%) solved about 13% more problems than the Unconstrained participants (*M* = 73%, *SD* = 21%). For Hard problems, Constrained participants (*M* = 25%, *SD* = 31%) solved about 19% more problems than the Unconstrained participants (*M* = 6%, *SD* = 13%). The overall patterns of strategy distribution were also consistent with those of Experiment 1a. Fig. 5 categorizes participants' responses based on the Easy, Hard, and Equality rules for Easy problems (Fig. 5b) and Hard problems (Fig. 5d).

4.1. Applying the Markov model

While Fig. 5 provides a summary characterization of the overall sampling of rules, it does not tell us about the sequential characteristics of participants' choice behavior over the 16 problems. To get at this sequential structure, we fit the 4-state Markov model described in the introduction to this experiment. We used standard hidden Markov modeling software (Hidden Markov Model Toolbox for Matlab; Murphy, 1998) for obtaining maximum likelihood estimates of the probabilities in the start vectors, transition matrices, and observation matrices.

The meaning of a state is given by how it affects the probability of different actions, which is given by its observation matrix. To keep the meaning of the states the same across all conditions and experiments, we estimated the single observation matrix in Table 3 from all four experiments. The details of its estimation are given in the Appendix.[3] As described there, it was estimated under the constraints that the association of the three rules with designated answers for that be the same (first three .92 values down the main diagonal in Table 3), that the probability of producing a response associated with another rule be the same (the .002 values), and the probability of an Other rule producing an answer that matched one of the three focus rules be the same (the .105 values).

With the observation matrix fixed, we are left to estimate the start vector and transition matrix for each condition. The start vector represents the probability that various rules will be the first idea that participants try out to solve these problems. There are three degrees of freedom associated with the estimated probability of the three focus rules

Table 3
Observation matrix used for all conditions of all experiments: Probability of each response category in each state

	Response			
No Memory Model	Easy	Hard	Equality	Other
State				
Easy	0.920	0.002	0.002	0.076
Hard	0.002	0.920	0.002	0.076
Equality	0.002	0.002	0.920	0.076
Other	0.105	0.105	0.105	0.685

(Hard, Easy, Equality). The probability of an Other rule is one minus their sum. We will refer to these as the Start probabilities.

The transition matrix reflects the probability that a participant will change from one state (hypothesis) to another. Since there are four states and the transition probabilities out of each state must sum to 1, there are 4 × 3 = 12 degrees of freedom in estimating a transition matrix. We explored the question of whether we could simplify this parameter complexity down to four more meaningful parameters. Under this simplified model, if a participant found the correct rule he or she would stay in that state. This assumption already eliminates three degrees of freedom by eliminating transitions out of one state. The reduction from the remaining nine degrees of freedom to four is a test of our bias interpretation of the influence of instruction. If the rule participants were trying was not the correct one, there was a Stay probability that they would stay with that rule again (perhaps because they did not believe the feedback or thought they had made a calculation error). We constrained the Stay probability to be the same for all conditions in an experiment. If participants did choose to try another hypothesis, we assumed there were three Choice probabilities that they would consider each of the rules. These Choice probabilities reflect biases that depend on the instruc-

Table 4
Example parameters from the Unconstrained Easy condition of Experiment 1

(a) Probability of Starting with various rules and probability of later choosing the rules

	Start	Choice
Easy	0.496	0.086
Hard	0.000	0.002
Equality	0.165	0.407
Other	0.339	0.504

(b) Transition matrix constructed from the Choice probabilities in part (a) and a Stay probability of .19

		To			
		Easy	Hard	Equality	Other
From	Easy	1.000	0.000	0.000	0.000
	Hard	0.070	0.192	0.330	0.409
	Equality	0.070	0.002	0.520	0.409
	Other	0.070	0.002	0.330	0.599

(c) Transition matrix estimated with all 12 degrees of freedom

		To			
		Easy	Hard	Equality	Other
From	Easy	0.998	0.002	0.000	0.000
	Hard	0.510	0.490	0.000	0.000
	Equality	0.067	0.000	0.545	0.388
	Other	0.069	0.000	0.320	0.611

tional condition. Thus, according to the bias interpretation, the only thing that matters is the condition-induced attractiveness of the hypothesis that the participants are going to, not the hypothesis they are leaving.

Table 4 illustrates how the Markov model was applied to the data collected from Experiment 1. (In this experiment Stay probability is estimated to be .19.) Table 4a shows the estimated Start and Choice probabilities for the Unconstrained Easy condition. Table 4b shows the transition matrix constructed from these choice probabilities and the Stay probability of .19. For comparison, Table 4c shows the transition matrix that would be estimated if there were no constraints (the Start vector estimated in this case is virtually identical to that in Table 4a). There are similarities and differences between Table 4b and c, but the question is whether the differences in Table 4c reflect things not captured by the constrained estimation in Table 4b or whether they reflect overfitting. To address this question, we compared a Full model, which estimated all 12 degrees of freedom in a transition matrix per condition with the Reduced model with only 3, plus the Stay probability shared across conditions. The Appendix reviews how well the Reduced model does compared to the Full model by three statistical measures: a chi-square test, AIC, and BIC (Lewandowsky & Farrell, 2011). For this experiment and for all experiments combined, the Reduced model is preferred by all three measures (but the chi-square test does favor the Full model in Experiments 2 & 4). Given these tests, we feel that we can focus on the Choice probabilities as capturing participants' behavior after their first choice (the Start probabilities). The Appendix also reports a test supporting the underlying Markov assumption that choice behavior only depends on the current hypothesis the participant is entertaining and not prior history. This involves showing that the choice behavior is the same in the first and second half of the experiment. After reporting all the experiments, we will describe some aggregate analyses that illustrate how well this modeling is capturing the rule learning of the participants.

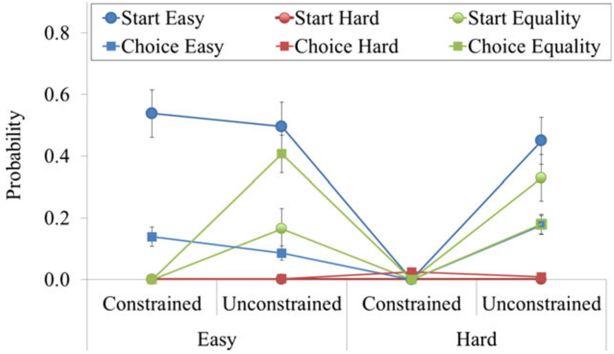


Fig. 6. The probabilities of starting with various rules and of later choosing those rules in the various conditions of Experiment 1. Standard errors of parameter estimates are calculated by bootstrapping (see Appendix). The Start probabilities for the Other category is one minus the sum of the Start probabilities of the three focus rules, and the Choice probability for the Other category is similarly calculated.

Fig. 6 shows the probabilities of starting with various rules (circles) and probability of later choosing different rules (squares). In all but the Hard Constrained condition participants choose to start with the Easy rule about 50% of the time. After this first choice, the Choice probabilities for the Easy rule drop to an average of little more than 10% in these conditions. The Equality rule is tried quite often in the Unconstrained conditions (with Start and Choice probabilities averaging a little over 25%). The Constrained Hard condition provides a striking contrast to the other conditions, with participants not choosing any of the three rules with high frequency. According to the model estimates, participants never start with the Hard rule in this condition and only chose it on later trials with a 2.4% probability. Still by the end of the 16 trials, the Markov model estimates that 25% of the participants in the Constrained Hard condition have identified the Hard rule as correct.

4.2. Summary

This experiment has shown that participants come in with strong biases as to what the correct rule might be and that these biases can vary even in the absence of instruction. Both the Easy and the Equality rule were frequent hypotheses, but the Easy rule had a preference as a first choice. However, participants were not so biased toward these rules that they would perform fractional arithmetic in the Constrained conditions to apply them. In the Hard Constrained condition, both the Easy and Equality rules were suppressed and some participants eventually hit upon the Hard rule based only on correctness feedback without any instructional support.

The model fit to the data assumes that participants have no memory for past hypotheses they have abandoned and are prone to repeatedly trying the same wrong rule. Table 5 shows the average number of times participants tried a wrong rule, defined as giving an answer that conformed to that rule either on the first trial or later after giving an answer that did not conform to the rule. It also shows how many of these were retries, defined as later giving the rule answer after having tried it and then producing an answer that did not correspond to the rule. In the first experiment retries were particularly prevalent in the Unconstrained Hard condition. Of the 50 Unconstrained Hard participants, 45 offered an Easy or Equality answer, tried some other answers, and then returned to trying the answer for the same rule.

The assumption that participants have no memory for past hypotheses is rather extreme. Therefore, we tried a model that assumed there was a certain probability that the participant would be able to remember a failed hypothesis and would not return to it. The Appendix includes a description of such a model. This Memory model requires 20 states to keep track of what hypotheses have been remembered. It requires one more parameter than the model we have fit, which is the probability of remembering a hypothesis. To find the best fit of this model to the data requires searching the two-dimensional space of possible Stay probabilities and Memory probabilities. Fig. 7 displays the search of this space for the four experiments. It can be seen that in all experiments the best fits involve choices of Stay and Memory probabilities that are rather low. The best Memory

Table 5
Mean number of times participants tried the wrong hypothesis (Hard or Equality when Easy was correct; Easy or Equality when Hard was correct) and number of times they retried the same hypothesis after trying something else

	Tried	Retried
Exp 1		
Easy		
Constrained	0.10	0.02
Unconstrained	1.34	0.68
Hard		
Constrained	0.36	0.12
Unconstrained	5.16	3.50
Exp 2		
Easy		
Good example	0.31	0.04
Ambiguous example	0.32	0.12
No example	0.66	0.28
Hard		
Good example	3.67	2.19
Ambiguous example	2.43	1.18
No example	5.54	3.64
Exp 3		
Easy		
Good verbal	1.32	0.74
Good example	0.28	0.04
Ambiguous verbal	0.54	0.18
Ambiguous example	0.22	0.06
Hard		
Good verbal	4.82	3.04
Good example	3.42	1.88
Ambiguous verbal	3.36	1.70
Ambiguous example	2.16	1.02
Exp 4		
Easy		
Enhanced verbal	0.07	0.00
Scaffolded example	0.26	0.06
Hard		
Enhanced verbal	1.02	0.33
Scaffolded example	1.96	0.94

probabilities were 0.00 for Experiment 1, 0.11 for Experiment 2, 0.05 for Experiment 3, and 0.00 for Experiment 4.[4] Thus, at best there is very little memory for past hypotheses. As the Appendix describes, the BIC measure indicates that the extra memory parameter is not justified.

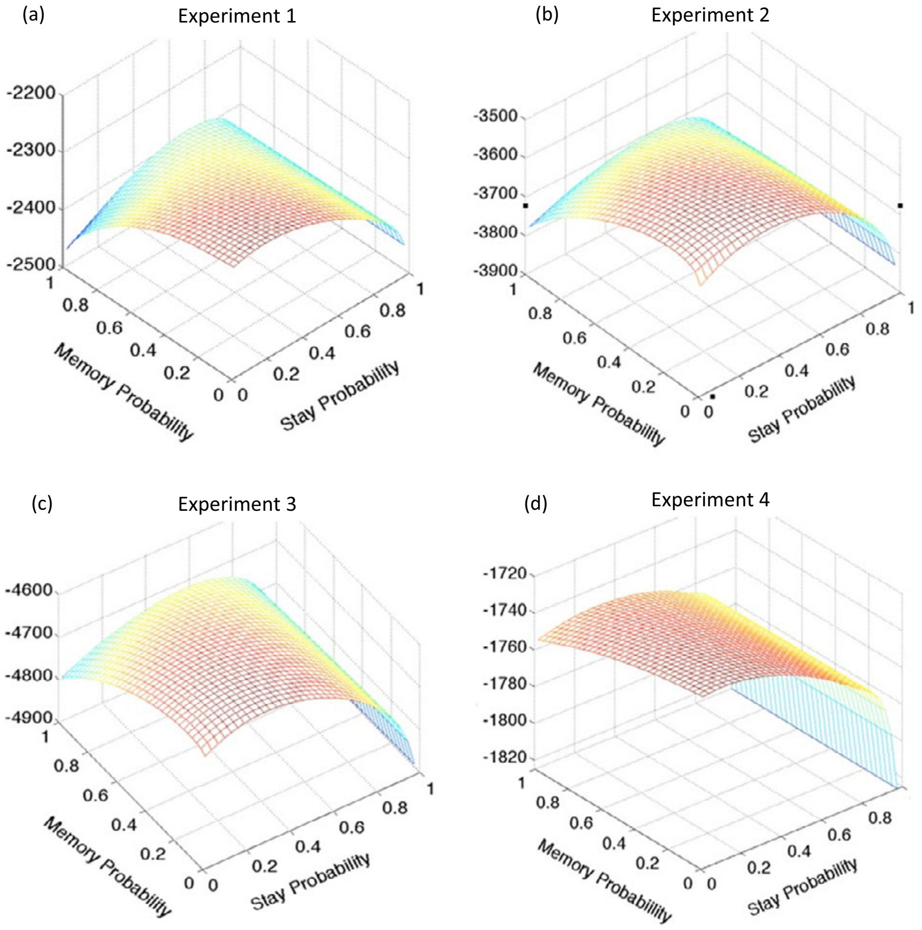


Fig. 7. The log likelihood of models that displayed various combinations of the Stay probability and Memory probability.

5. Experiment 2: Effect of examples

In the first experiment, participants' behavior looked very much like the classic hypothesis-testing behavior where they conducted a search through a space of rules. While there were strong biases in the probabilities of considering different rules, these probabilities describe a random search process that satisfied the Markov property of being independent of past history and only dependent on the current state. Indeed, we showed that it was simpler than even being dependent on the current state. Independent of state, there was a constant Stay probability of sticking with a rule for another problem. If they chose to consider another rule, the probability of choosing that rule depended only on the Choice probability for that rule and not on the current state.

However, this was in the absence of any instruction and participants might behave differently if there were instruction. Perhaps with instructional guidance learning would be less random, more controlled by the instruction, and less controlled by simple Start and Choice probabilities. As we noted in the introduction, studying examples can often be an effective source of instructional guidance. Therefore, in the second experiment we introduced examples and manipulated the content of the provided examples.

5.1. Method

5.1.1. Participants

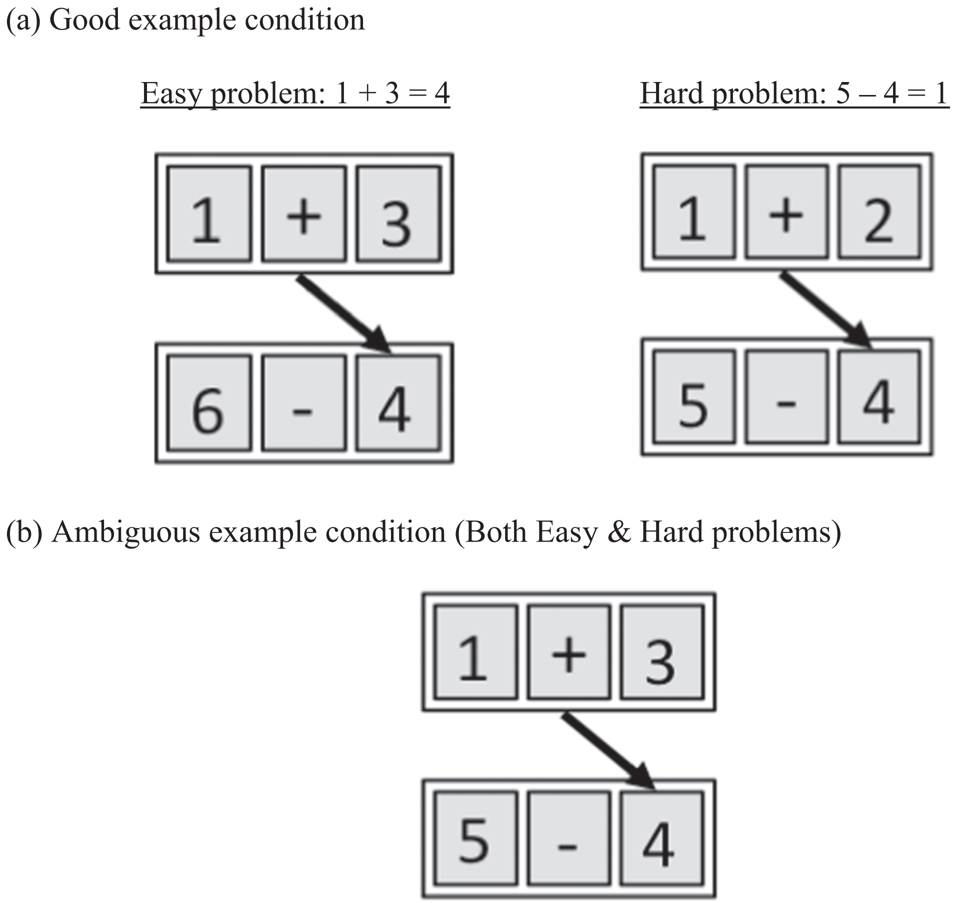
A total of 364 participants took part in this study and they were recruited from Amazon Mechanical Turk. The participant pool (193 male and 171 female, M = 28.31 years, SD = 6.92) reported various levels of educational background (39% 4-year, 36% 2-year or some college, 8% high school, 14% master, and 3% others). Of 364, 60 participants (16%) wanted to quit the study in various phases of the study and the data from those participants were excluded for data analysis. This left a total of 304 participants, at least 50 participants per condition. The payment was the same as Experiment 1a.

5.1.2. Design, materials, and procedure

A 2 × 3 between-subjects design was employed to test effects of example content on learning of two difficulty types of problems (Easy and Hard). The same sequence of 16 problems was used for both Easy and Hard conditions, but depending on the rule the correct answer varied. In terms of the previous experiment, all problems were unconstrained in that all three of the rules resulted in integer answers.

Participants were given a good example, an ambiguous example, or no-example. In the example conditions (both good and ambiguous), an example appeared with each problem simultaneously. The examples were already solved problems, but without any explanations. Fig. 8 illustrates good and ambiguous examples. In the good example condition, the example solution supported only the rule that participants had to find. In the Easy problems (Fig. 8a), the provided example supported only the Easy rule ($1 + 3 = 4$), and not the Hard rule ($6-4 \neq 1$). In the Hard problems, the provided example supported only the Hard rule ($5-4 = 1$), and not the Easy rule ($1 + 2 \neq 4$). In contrast, in the ambiguous example condition (Fig. 8b), the example solution always supported both Easy ($1 + 3 = 4$) and Hard rule ($5-4 = 1$). A total of 16 examples were constructed for each problem type. Therefore, participants were given a new example for every new problem.

In both example conditions, an example appeared on the left side of the screen while participants were solving a current problem presented on the right side of the screen. In the no-example condition, participants were not provided with any examples and they had to figure out problem rules in a trial-and-error manner. As in Experiment 1, participants were given 1 minute for each problem, and after their response they were shown feedback (2 seconds) that simply indicated whether or not their response had been correct.



(a) Good example condition

Easy problem: 1 + 3 = 4

Hard problem: 5 – 4 = 1

(b) Ambiguous example condition (Both Easy & Hard problems)

Fig. 8. Examples of good and ambiguous examples used in Experiments 2 and 3. In the good example condition (a), examples were constructed separately for the Easy and Hard problems. The Easy problem example supports only the Easy rule ($1 + 3 = 4$), and not the Hard rule ($6-4 \neq 1$). The Hard problem example supports only the Hard rule ($5-4 = 1$), and not the Easy rule ($1 + 2 \neq 4$). In the ambiguous example condition (b), the same example was used for both Easy and Hard problems. The example supports both the Easy ($1 + 3 = 4$) and the Hard rule ($5-4 = 1$).

5.2. *Results*

Fig. 9 shows mean percentage of correctly solved problems for Easy and Hard problems among the three example conditions. A 2 × 3 between-subjects ANOVA was performed on the percentage of correctly solved problems. Problem difficulty (Easy vs. Hard) and example content (good vs. ambiguous vs. no-example) were included as between-subjects variables. There were significant main effects of both problem difficulty, $F(1, 298) = 391.12$, $p < .0001$, $\eta_p^2 = .568$, and example content, $F(2, 298) = 9.05$, $p < .001$, $\eta_p^2 = .057$. More interestingly, there was a significant interaction between the problem difficulty and example content, $F(2, 298) = 4.54$, $p = .01$, $\eta_p^2 = .030$.

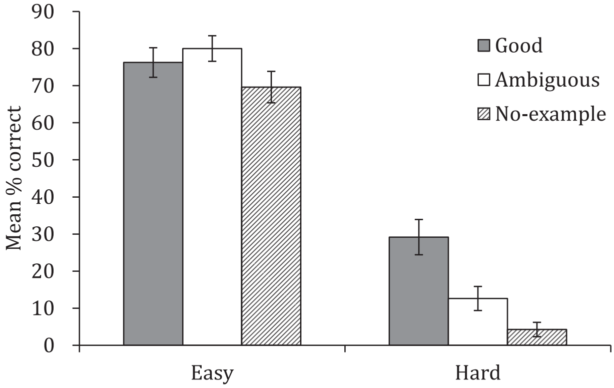
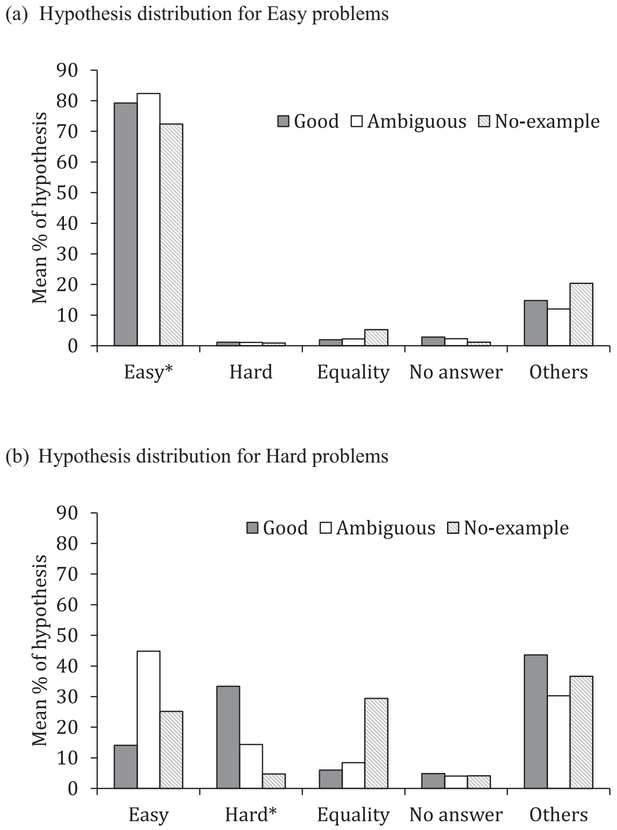


Fig. 9. Mean percentage of correctly solved problems for Easy and Hard problems in each example condition of Experiment 2. Error bars represent 1 standard error of mean.

Different patterns of results were observed for Easy and Hard problems. For Easy problems, most participants easily figured out the correct solution rule in all three example conditions (overall M = 75%, SD = 28%) and there were no mean differences among the three conditions, $F(2, 148) = 1.83$, $p = .165$, $\eta_p^2 = 024$. In contrast, for Hard problems, many participants never figured out the correct solution rule and thus overall performance was much lower than the Easy problems. Participants who were provided with good examples (M = 29%, SD = 34%) solved significantly more problems correctly than those provided with ambiguous examples (M = 13%, SD = 23%), $t(101) = 2.88$, $p = .005$, and than those provided with no-example (M = 4%, SD = 14%), $t(99) = 4.82$, $p < .001$. The mean difference between the latter two groups was also significantly different, $t(100) = 2.19$, $p = .031$.

Fig. 10 categorizes the participants' responses in the different conditions. As in Experiment 1, most of the responses for Easy problems (Fig. 10a) corresponded to the Easy rule hypothesis. The Hard and Equality hypotheses were tried less than 6% of the time in all conditions. For Hard problems (Fig. 10b), there were more cases of different types of hypotheses. Even though the Easy rule never resulted in the correct answer, participants in the ambiguous example condition repeatedly tried the Easy rule because the Easy rule was supported by the examples they were given. Again there was also an increased tendency to try the Equality rule in the no-example condition (29.4%).

As detailed in the Appendix, the Markov model once again provides a good characterization of the sequential structure of participants' behavior. Fig. 11 shows the probabilities of starting with various rules and later choosing them if the correct rule has not been identified. (The Stay probability of sticking with an incorrect hypothesis for another problem was estimated to be .35.) Once again, there are high probabilities of starting with a choice of the Easy rule. When there was an example that exemplifies that rule (Good Easy condition or Ambiguous Easy conditions) participants averaged over 75% probability of starting with the Easy rule. However, even in the remaining conditions that did not


(a) Hypothesis distribution for Easy problems

(b) Hypothesis distribution for Hard problems

Fig. 10. Distribution of different hypotheses tested by participants in each example condition for (a) Easy problems and (b) Hard problems in Experiment 2. *indicates the correct rule.

have such an example, participants still started with the Easy rule over 45% of the time (comparable to the Start probabilities in all but the Constrained Easy condition of Experiment 1). Again, the probability of later choosing the Easy rule was lower but still averaged over 20%. While participants seldom started with the Equality rule in this experiment, they averaged over 10% on later choices. In this experiment, participants do choose to start with the Hard rule on 10% of the trials given a good example and show 4% probability of later choosing that rule.

5.2.1. Summary

To summarize the results for the Easy problem condition, participants did not really need an example from which to learn and the effect of example content was little. The Markov model estimates that by the end of the experiment 92% of the participants have learned the Easy rule with good examples, 95% with ambiguous examples, and 88% with no examples. On the other hand, when problem structure was not clear as in Hard

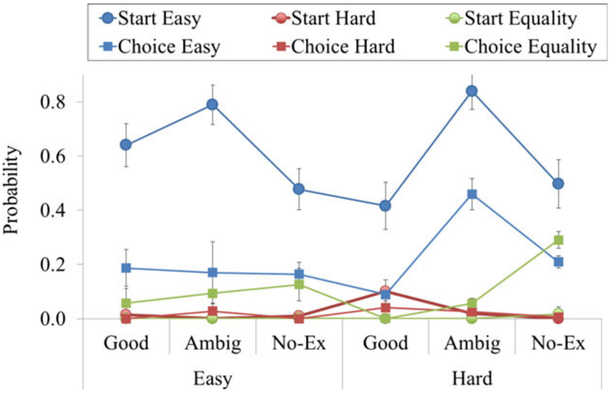


Fig. 11. The probabilities of starting with various rules and of later choosing those rules in the various conditions of Experiment 2. Standard errors of parameter estimates are calculated by bootstrapping (see Appendix). The Start probabilities for the Other category is one minus the sum of the Start probabilities of the three focus rules, and the Choice probability for the Other category is similarly calculated.

problems, learners were affected by example content. The Markov model estimates that by the end of the experiment 39% of the participants have learned the Hard rule with good examples, 22% with ambiguous examples, and 6% (just 3 of the 50 participants) with no examples. It seems that participants do use the examples as a source of possible rules, but this is only needed when the rule is not already obvious.

6. Experiment 3: Effect of instructions

From Experiment 2 we can conclude that the effect of examples is to make certain hypotheses more available as participants search through the space of possible rules. Just as much as in the first experiment, the effect of instructional condition seems to be captured by the probabilities that they will start with certain rules or later choose these rules. This characterization of learning (i.e., learning as a search through hypothesis space with a bias created by instructional conditions) might no longer hold when we provided verbal instruction. The goal of the third experiment was to investigate the effects of verbal directions and contrast these with the effects of examples. In the verbal direction condition participants were told what parts of the problem to attend to find the rule, but we did not tell them the explicit rule, whereas in the example condition participants were provided with an example to learn from as in Experiment 2. Crossed with this, we varied whether the instruction was ambiguous between alternative rules or pointed only to the correct rule. Because we were using a single verbal direction with all problems, we used a single example with all problems. This contrasts with Experiment 2 where each problem was accompanied by a distinct example.

6.1. Method

6.1.1. Participants

A total of 484 participants took part in this study and they were recruited from Amazon Mechanical Turk. The participant pool (246 male and 238 female, M = 28.97 years, SD = 7.83) reported various levels of educational background (35% 4-year, 39% 2-year or some college, 11% high school, 11% masters, and 4% others). Of 484, 84 participants (17%) wanted to quit the study in various phases of the study and the data from those participants were excluded for data analysis. This left 400 participants (50 per condition). The payment was the same as the previous experiments.

6.1.2. Design, materials, and procedure

A 2 × 2 × 2 between-subjects design was employed. As in Experiments 1–2, the first independent variable was problem difficulty type (Easy vs. Hard). Participants learned to solve either Easy or Hard problems. We also manipulated instruction type (verbal direction vs. example) and content of instruction (good vs. ambiguous) as between-subjects variables. In the verbal direction condition, participants were given verbal direction next to the problem instead of examples. Verbal directions used in this study are shown in Table 6. In the example conditions, the examples shown in Fig. 3 were used. The only difference from Experiment 2 was that the same example was used for all 16 problems. All procedures were identical with those of Experiments 1–2.

6.2. Results

Fig. 12 shows mean percentages of correctly solved problems in each experimental condition. A 2 × 2 × 2 between-subjects ANOVA was performed to test the effects of instruction type and instructional content on the percentage of correctly solved problems. Problem difficulty (Easy vs. Hard), instruction type (verbal direction vs. example), and instructional content (good vs. ambiguous) were included as between-subjects variables. As in previous experiments, there was a significant main effect of problem difficulty, $F(1, 392) = 380.54$, $p < .0001$, $\eta_p^2 = .493$. There were also significant main effects of

Table 6
Verbal directions used in Experiment 3

Good verbal direction	
Easy problems	To find the answer, use the two numbers and operator from the top rectangular box
Hard problems	To find the answer, use one number from the top rectangular box and use one number and the operator from the bottom rectangular box
Ambiguous verbal direction	
Both Easy & Hard problems	To find the answer, use two numbers and one operator from the rectangular boxes

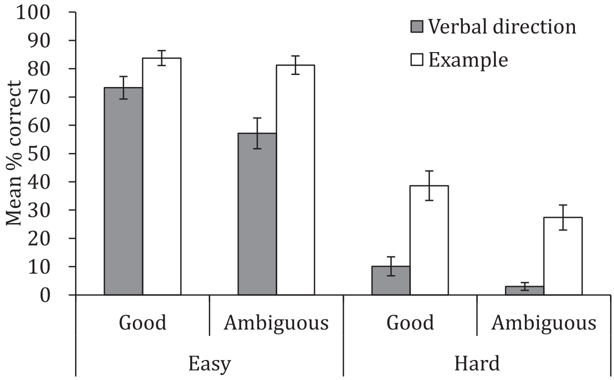
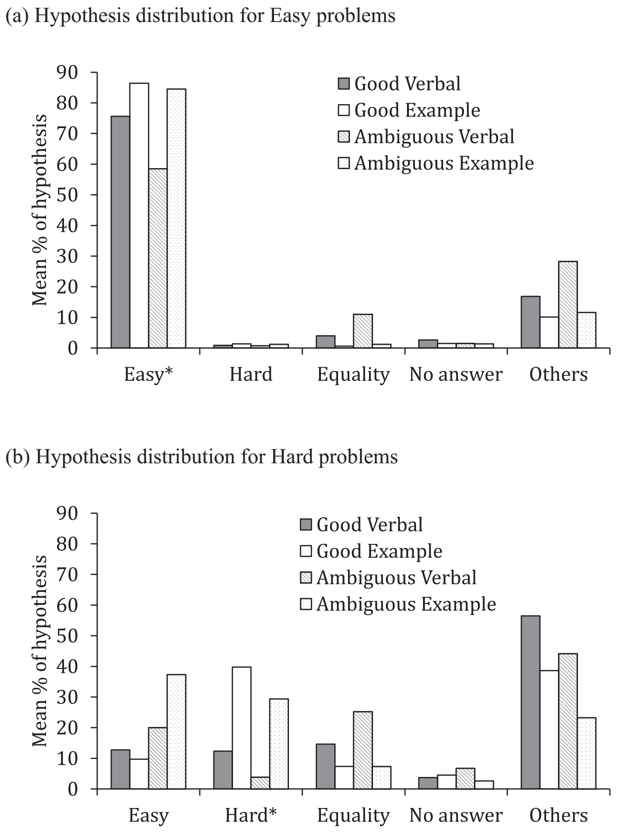


Fig. 12. Mean percentage of correctly solved problems for Easy and Hard problems in each condition of Experiment 3. Error bars represent 1 standard error of mean.

instruction type, $F(1,\ 392) = 62.30$, $p < .0001$, $\eta_p^2 = .137$, and of instructional content, $F(1,\ 392) = 11.14$, $p < .001$, $\eta_p^2 = .028$. These two instructional factors did not interact with each other, $F(1,\ 392) = 0.73$, $p = .392$, $\eta_p^2 = .002$. Also, problem difficulty did not interact with either instruction type, $F(1,\ 392) = 2.71$, $p = .101$, $\eta_p^2 = .007$, or instructional content, $F(1,\ 392) < 1$, $p = .982$, $\eta_p^2 < .001$. There was not a three-way interaction effect, $F(1,\ 392) = 2.56$, $p = .110$, $\eta_p^2 = .006$.

For Easy problems, regardless of instructional content (good vs. ambiguous), participants who were given examples solved more problems correctly than those who were given verbal directions, $F(1,\ 196) = 19.04$, $p < .0001$, $\eta_p^2 = .089$. In the good content condition, the example participants (*M* = 84%, *SD* = 19%) performed better than the verbal direction participants (*M* = 73%, *SD* = 28%). Likewise, in the ambiguous content condition, the example participants (*M* = 81%, *SD* = 23%) performed better than the verbal direction participants (*M* = 57%, *SD* = 39%). The same patterns of results were obtained from the Hard problems. Regardless of instructional content, example participants solved more Hard problems correctly than the verbal direction participants, $F(1,\ 196) = 46.67$, $p < .0001$, $\eta_p^2 = .192$. In the good content condition, the example participants (*M* = 39%, *SD* = 37%) performed better than the verbal direction participants (*M* = 10%, *SD* = 24%). Likewise, in the ambiguous content condition, the example participants (*M* = 27%, *SD* = 31%) performed better than the verbal direction participants (*M* = 3%, *SD* = 10%).

As in Experiment 2, good (vs. ambiguous) content had positive effects on learning on Hard problems, but this time the positive effects were found on Easy problems as well. For Easy problems, participants who were provided with good content solved significantly more problems than those who were given ambiguous content, $F(1,\ 196) = 5.51$, $p = .019$, $\eta_p^2 = .027$. Likewise, for Hard problems, the good content condition led to better performance than the ambiguous content condition, $F(1,\ 196) = 5.64$, $p = .019$, $\eta_p^2 = .028$. There were no significant interaction effects between instruction and content for Hard problems, $F < 1$.



(a) Hypothesis distribution for Easy problems

(b) Hypothesis distribution for Hard problems

Fig. 13. Distribution of different hypotheses tested by participants in each experimental condition for (a) Easy problems and (b) Hard problems in Experiment 3. *indicates the correct rule.

Fig. 13 shows how the different categories of responses are distributed. For Easy problems (Fig. 13a), regardless of experimental conditions, most responses fell into the Easy rule category. For Hard problems (Fig. 13b), various hypotheses were tested in all of the four conditions. Participants in the ambiguous example condition tried the Easy rule more often than in the other three conditions, implying that they were strongly affected by example content. This pattern was consistent with that of Experiment 2. Overall, verbal direction participants showed higher frequency of various incorrect hypotheses than the example participants. The verbal direction conditions were similar to the unconstrained conditions of Experiment 1 and the no-example condition of Experiment 2 in that participants tended to try the Equality hypothesis more often.

The Markov model once again provides a good characterization of the sequential structure of participants' behavior (see Appendix). Fig. 14 shows the probabilities of starting with various rules and later choosing them if the correct rule has not been identified. (The Stay probability of sticking with an incorrect hypothesis for another problem was estimated to be .30.) As in Experiment 2 (Fig. 11), when there was an

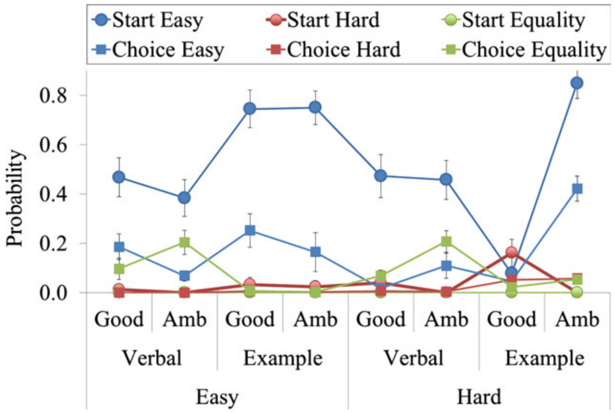


Fig. 14. The probabilities of starting with various rules and of later choosing those rules in the various conditions of Experiment 3. Standard errors of parameter estimates are calculated by bootstrapping (see Appendix). The Start probabilities for the Other category is one minus the sum of the Start probabilities of the three focus rules, and the Choice probability for the Other category is similarly calculated.

example that exemplifies the rule (Good Easy condition or Ambiguous Easy conditions), participants averaged over 75% Start probabilities. In all of the remaining conditions except for the Good example Hard condition, participants started with the Easy rule over 45% of the time. Again, the probability of later choosing the Easy rule was lower but still averaged about 16% across the conditions. While participants seldom started with the Equality rule in this experiment, they used it on average in 8% of later choices. The probabilities of choosing the Hard rule remained low but were highest in the Example conditions.

6.2.1. Summary

Participants learned better when they were given examples than verbal directions, but this was really because the biases induced by the examples better matched the rules we were asking participants to solve.[5] As Figs. 12 and 13 reveal, the biases induced by the examples better matched what participants were being asked to learn. Had the "correct" rule been Equality, participants would have done best with ambiguous verbal directions. However, given that this bias was inappropriate, the Ambiguous Verbal participants did worst. The Markov model estimates only 66% learned the Easy rule in the Ambiguous Verbal condition, which is much lower than the other Easy conditions (88% for Good Verbal, 98% for Good Example, and 94% for Ambiguous Example) in this experiment or any other experiments. In the Hard Ambiguous Verbal condition, the model only identifies one of the 50 participants as learning the rule in this condition (5 in the Good Verbal, 19 in both the Good Example and Ambiguous Example conditions).

Unlike Experiment 2, there was no difference in how many participants achieved mastery in the two Hard Example conditions. However, Fig. 14 reveals that mastery is achieved differently. Given a good example, participants have a .16 probability of starting right away with the Hard rule, whereas this never happens given an ambiguous example. However, Ambiguous Example participants actually have a slight edge in choosing the Hard hypothesis for subsequent problems—5.8% vs. 5.3%. Over the 16 trials they caught up with the Good Example participants. In contrast to Experiment 2, participants just saw the same example repeated for all problems. It seems that repeated exposure to the same ambiguous information increased the chance that participants would identify a rule that was not obvious on the first exposure.

7. Experiment 4: Effect of enhanced instructions

Across Experiments 1–3, we intentionally made instructions (both verbal direction and examples) somewhat ambiguous so that we could examine what kinds of solution hypotheses participants generated during their learning of problem solving. However, this ambiguity may have been what pushed participants into the classic hypothesis-testing behavior that we saw in each experiment. In Experiment 4, we created a learning condition that would be similar to an educational setting where a teacher provides students with much more directive instructions. Experiment 4 tested two instructional conditions by providing either an enhanced version of verbal directions or a scaffolded example (an equation was added to an example). We expected that participants would show better learning performance in both conditions than in Experiments 1–3.

7.1. Method

7.1.1. Participants

A total of 239 participants participated and they were recruited from Amazon Mechanical Turk. The participant pool (131 male and 108 female, *M* = 29.76 years, *SD* = 7.84) reported various levels of educational background (40% 4-year, 38% 2-year or some college, 14% masters, 5% high school, and 3% others). Of 239, 29 participants (12%) quit the study in various phases of the study. This left a total of 210 participants for data analysis. The payment was the same as in Experiments 1–3.

7.1.2. Design, materials, and procedure

A 2 × 2 between-subjects design was employed. As in Experiments 1–3, participants learned to solve either Easy or Hard problems. We also manipulated instruction type by providing either verbal direction or an example. Experimental materials and procedures were identical to each of the verbal direction and example conditions of Experiment 3, with one exception. Experiment 4 used an enhanced version of verbal direction and a scaffolded example to make the problem solution more apparent to learners. In the verbal direction condition, Easy instructions were “Applying the top operator to the top two numbers results in the number that goes into the box where the arrow is pointing,” and the Hard instructions were “Applying the bottom operator to the bottom two numbers results in the top number that is on the opposite side of the empty box.” In the scaffolded example condition, we provided the same example that was used in the good example condition of Experiment 3 (see Fig. 8a) and added an equation below the example to indicate how to compute the answer. In Fig. 8a, the added equations were “1 + 3 = 4” and “5−4 = 1” for Easy and Hard problems, respectively. As in Experiment 3, the same verbal direction or the same example repeatedly appeared across the entire problems.

7.2. *Results*

Fig. 15 shows the mean percentage of correctly solved problems for Easy and Hard problems in the enhanced verbal direction and scaffolded example condition. A 2 × 2 between-subjects ANOVA was performed on the percentage of correctly solved problems to see the effect of problem type (Easy vs. Hard) and instruction type (enhanced verbal direction vs. scaffolded example) on learning. There were significant main effects of both problem difficulty, $F(1, 206) = 48.17$, $p < .0001$, $\eta_p^2 = .190$, and instruction type, $F(1, 206) = 6.77$, $p = .01$, $\eta_p^2 = .032$. The problem difficulty and instruction type did not interact, $F(1, 206) = 1.86$, $p = .17$, $\eta_p^2 = .009$. As in prior experiments, participants solved more Easy problems (M = 85%, SD = 20%) correctly than Hard problems (M = 61%, SD = 25%). Also, participants who were given a scaffolded example solved more problems correctly than those who were given enhanced verbal direction. Although the difficulty by instruction interaction effect was not reliable, the magnitude of the mean difference was greater for Hard problems. For Easy problems, example participants (M = 88%, SD = 15%) solved about only 4% more problems than the verbal direction participants (M = 84%, SD = 24%), whereas for Hard problems, example

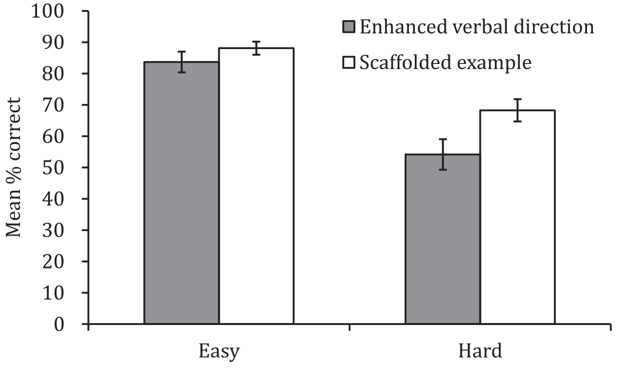
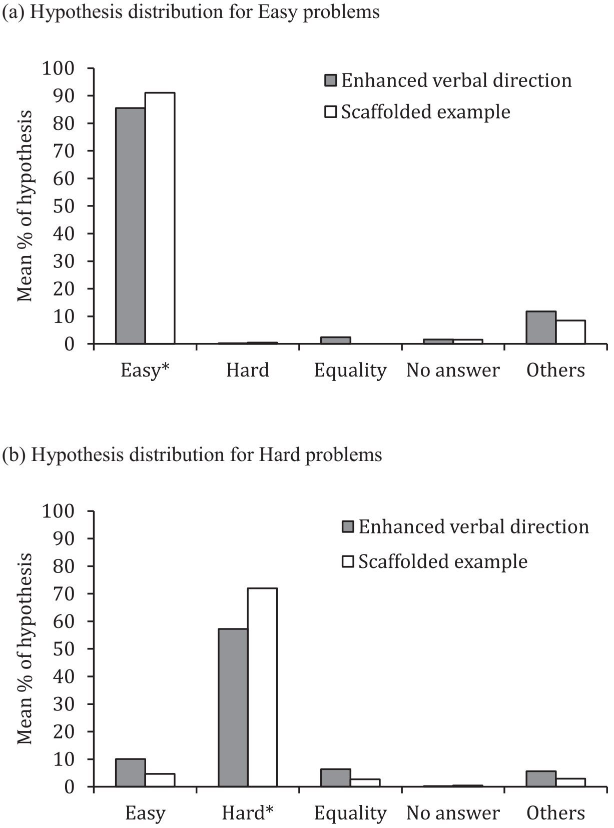


Fig. 15. Mean percentage of correctly solved problems for Easy and Hard problems in enhanced verbal direction and scaffolded example conditions of Experiment 4. Error bars represent 1 standard error of mean.


(a) Hypothesis distribution for Easy problems

(b) Hypothesis distribution for Hard problems

Fig. 16. Distribution of different hypotheses tested by participants in enhanced verbal direction and scaffolded example conditions for (a) Easy problems and (b) Hard problems in Experiment 4. *indicates the correct rule.

participants (M = 68%, SD = 25%) solved about 14% more problems than the verbal direction participants (M = 54%, SD = 35%). Overall, participants performed much better than the participants from any experimental conditions of Experiments 1–3. In none of the previous experiments was the average performance of Hard problems greater than 50%.

Fig. 16 shows how the different categories of responses were distributed. For Easy problems, as in Experiments 1–3, most cases were observed in the Easy rule category. For Hard problems, however, different from Experiments 1–3, most cases were now observed in the Hard rule category. Although we were able to greatly reduce the amount of incorrect hypothesis generation, there were still some responses that conformed to the Easy and Equality rules. This tendency was slightly greater in the verbal direction condition than in the example condition. This pattern is consistent with the findings from prior experiments in that the Equality hypothesis was tested more often when participants were not provided with examples that disprove the Equality hypothesis.

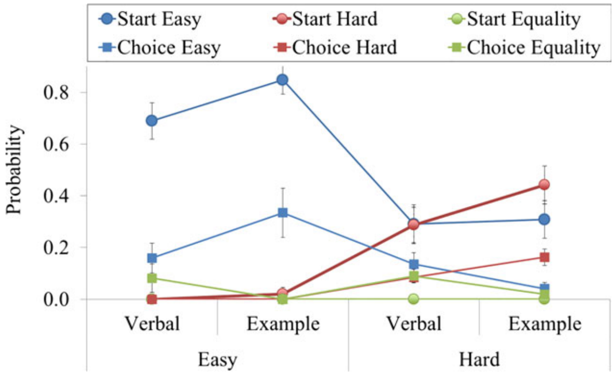


Fig. 17. The probabilities of starting with various rules and of later choosing those rules in the various conditions of Experiment 4. Standard errors of parameter estimates are calculated by bootstrapping (see Appendix). The Start probabilities for the Other category is one minus the sum of the Start probabilities of the three focus rules, and the Choice probability for the Other category is similarly calculated.

The Markov model once again provides a good characterization of the sequential structure of participants' behavior (see Appendix). Fig. 17 shows the probabilities of starting with various rules and later choosing them if the correct rule has not been identified (The Stay probability of sticking with an incorrect hypothesis for another problem was estimated to be .12). The probabilities of correct choices are generally higher than in previous experiments. The model identifies all 54 of the participants in the Easy Example condition as learning the rule, 51 of the 54 participants in the Easy Verbal condition, 48 of the 51 participants in the Hard Example condition, and 39 of the 51 participants in the Hard Verbal condition. However, it is worth noting that none of the Hard probabilities in Fig. 17 are greater than .5. The high levels of mastery are only achieved through repeated opportunities to learn the concept.

7.2.1. Summary

To summarize, Experiment 4 showed when we tried as best we could to make the problem structure apparent by either example or verbal instruction, participants enjoyed greater success in learning. The sequential behavior of participants still satisfies the test for the Markov property, but their probability of trying correct rules increased.

8. General discussion

In this study, we viewed learning of problem solving as hypothesis testing and conducted four experiments to see how instructional conditions can affect learning by changing the probabilities of trying various rules. Experiment 1 tested this with two different types of populations, and overall patterns of the results were consistent between the two populations. We also tested a Markov model that described how participants consider various hypotheses and transition among such hypotheses. Across the four experimental studies, we tested how instructional properties affected learning by constraining or relaxing the hypothesis search space. We created a task that could have various correct answers depending on the experimenter's rule and tested how well participants could find the correct rule in various experimental conditions. When the correct rule was the Easy rule, participants enjoyed high probabilities of success independent of instructional condition or indeed whether there was any instruction. In contrast, when it was the Hard rule, the majority of participants had difficulty learning. It was only with strong instructional intervention (Experiment 4) that most participants were successful at learning the Hard rule and even then it required multiple learning opportunities to assure that most participants achieved mastery. Over the four experiments we showed that success in finding the Hard rule was affected by computational difficulty (Experiment 1), quality of examples (Experiments 2–4), and quality of verbal directions (Experiment 3–4). In each case we showed that these instructional effects could be explained in terms of the effect these instructions had on the probability of starting with various rules and later choosing a rule should the initial rule hypothesis prove wrong.

8.1. Aggregate data analyses

Fig. 18 shows summary analyses aggregating the data from all four experiments. Part (a) shows the proportion of correct answers for the 559 Easy participants and 555 Hard participants. It displays what appears to be rather slowly improving learning curves with Easy problems much better than Hard problems. The smooth lines reflect the predicted average performance according to the Markov models that were fit to the experiments. However, as is the case in hypothesis-testing situations (Bower & Trabasso, 1964; Levine, 1975; Nosofsky et al., 1994; Restle, 1962), such continuous learning curves can hide an all-or-none discovery process where the correct rule is mastered on a single trial but different trials for different participants. The Markov models provide an estimate of the probability that a rule is mastered on each trial. We can identify the trial of mastery as the trial when estimated probability of the correct rule exceeds 50%. Given that there is a chance that the response does not reflect the rule being considered (i.e., solvers may make a computational error implementing their rule), one cannot be certain in identifying the trial of mastery. Still confidence is high: The inferred probability that the rule is mastered on the trial of mastery is 91% while the probability that it was mastered on the prior trial is 14%. Fig. 18c shows the probability of being correct as a function of the trial relative to the trial of mastery. It reveals a rather dramatic jump with the difference between Easy and Hard problems largely eliminated. Thus, the difference between Easy and Hard rules depends on differences in when their rules are identified, not accuracy in applying the rules.

One could criticize the analysis in Fig. 18c because the inference of trial of mastery is based on the accuracy data that we are plotting. Parts (b) and (d) of Fig. 18 look at latency, which we have not used in the modeling. Part (b) looks at average response time as a function of problem position. That figure reveals some tendency toward speed-up for both types of problems, but Hard problems average more than 7 seconds longer than

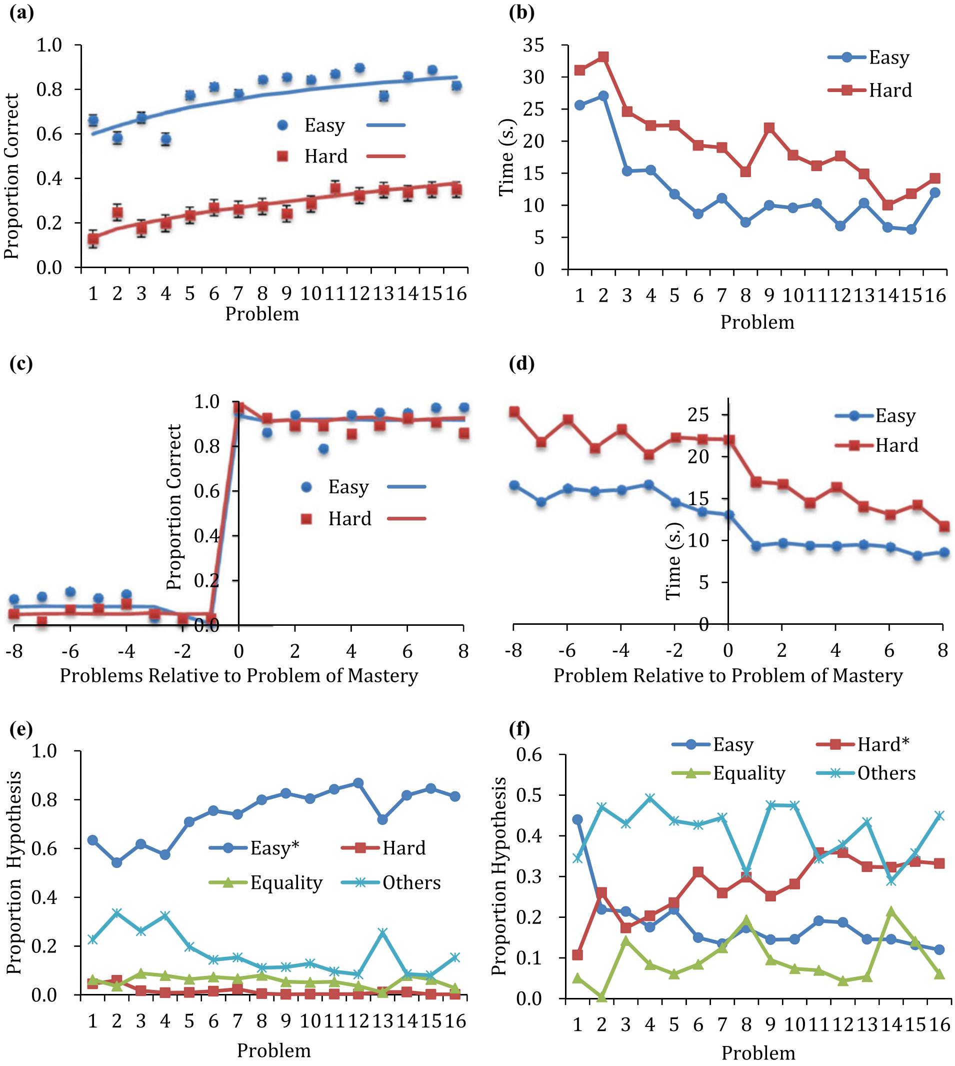


Fig. 18. Analyses average over the four experiments. (a) Probability of a correct answer as a function of problem position; (b) time to answer as a function of problem position; (c) probability of a correct answer relative to the trial of mastery; (d) time to answer relative to the trial of mastery; (e) proportion of a hypothesis as a function of Easy problem position; (f) proportion of a hypothesis as a function of Hard problem position.

Easy problems. It shows a complicating factor as well. The times for the first two problems average more than twice as long as the average for later problems (29.2 seconds vs. 13.9 seconds) and the time actually increases for the second problem (an increase for problem 2 is significant for both Easy and Hard problems, $t(558) = 2.00$ and $t(554) = 2.43$, both p's < .05). We are fairly confident that this delay reflects solvers' need to familiarize themselves with the input interface (the second problem is the first instance of an answer below zero that required entering a negative sign).

Fig. 18d replots latency as a function of position relative to trial of mastery, excluding the first two trials (because of the problem noted above) and problems that were mastered on these first two trials or never mastered. It reveals a major speed-up after the trial of mastery (from 13.1 to 9.4 seconds for Easy problems, and from 22.1 to 17.0 seconds for Hard problems). This Mastery Drop contrasts with the rather gradual speed-up before (Pre Slopes: average of .3 seconds per problem for both types of problems) and afterward (Post Slopes: .15 sec for Easy and .65 seconds for Hard). Looking at just those participants for whom we could compute both Pre and Post slopes excluding the first two trials, these differences are all significant (Mastery Drop > Pre Slope: $t(106) = 2.00$ for Easy and $t(100) = 2.56$ for Hard, both p's $< .05$; Mastery Drop > Post Slope: $t(106) = 4.39$ for Easy, $t(100) = 3.59$ for Hard, both p's $< .01$). Thus, there seems a qualitative shift in processing once the participants have determined the rule. The search process has been eliminated and participants are now just practicing the rule. Fig. 18d also reveals that the difference in times between Hard and Easy problems is not just a consequence of different points of mastery. Even after mastery, Hard problems average about 5 seconds longer than Easy problems. Although the arithmetic computations are not more difficult, participants find it harder to apply the Hard rule. In part, this is probably because the numbers they need to combine are not adjacent and they need to invert the operator.

Fig. 18e–f also shows the proportion of hypothesis used by participants as a function of problem position for the Easy problems and Hard problems. For the Easy problems (e), Easy rule was used most often and the proportion of the Hard rule was low across the trials. In contrast, for the Hard problems (f), the relative proportion of the Easy and Hard rule was reversed between the initial and later trials. The proportion of the Hard rule gradually increased, whereas the pattern was the opposite for the Easy rule.

Fig. 19 shows the relationship between Start probabilities and Choice probabilities, separately for Easy and Hard problems, plotting the points for the 22 conditions across the four experiments. Both sets of points are basically on the same function. The best-fitting function through the origin for Easy problems is

$$\text{Choice-Probability} = .329 \times \text{Start-Probability}$$

which accounts for 59.4% of the variance while the best-fitting function for Hard problems is

$$\text{Choice-Probability} = .345 \times \text{Start-Probability}$$

which accounts for 81.8% of the variance. Thus, each condition seems to induce a certain initial preference for a rule, which is reduced to about a third of its original value after the first problem, but the same experimental factors determine both Start and Choice probabilities.

Besides showing the correspondence between Start and Choice probabilities, this graph raises the question of why the Choice probabilities are so reduced. The reduction in the Easy and Hard probabilities is accompanied by an increase mainly in the probability of

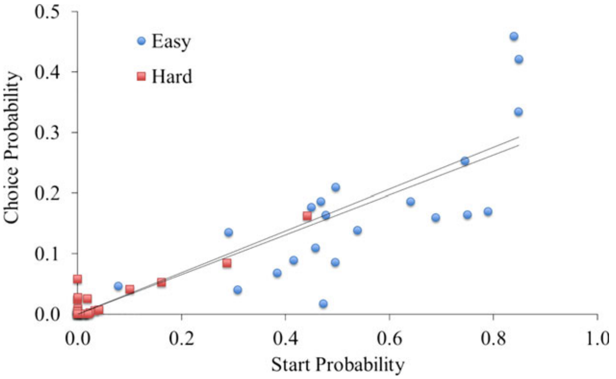


Fig. 19. Choice probabilities as a function of Start probabilities. The solid lines reflect the best-fitting linear functions through the origin.

the “Other” state (which is one minus the probabilities displayed in Figs. 6, 11, 14, and 17). It is hard to characterize generally what the rules are that participants are following in this state, but many seem specific to particular problems and reflect features that do not involve arithmetic computation. There is considerable copying of numbers of one of the numbers in the problems but many other sorts of answers. As instance of another sort of answer, in Experiments 2–4 one of the Hard problems had “30” appearing in two of the boxes and “−3” in the third. For this problem only, there were frequent answers of “3” (35 of 270 of all non-Hard answers). It seems that the participants’ hypothesis generation for this problem was influenced by the fact that one of the digits in all of the numbers was a “3.” The dramatic rise in these kinds of hypotheses after the first problem suggests that if the subjects are wrong with their first guess (typically some arithmetic computation) they open their hypothesis space to the possibility that the answer may not involve an arithmetic computation.

While the Markov model seems to capture the sequential structure of participants’ behavior across these experiments based on the probabilities in Fig. 19, it leaves open what determines these probabilities. Abstractly, we have described students as being in a space of hypotheses with biases determining the probabilities of trying various hypotheses. However, in more information-processing detail this search of the hypothesis space must involve constructing rules to try (perhaps revising former rules they have tried to produce new ones). There are numerous models of how rules are constructed and changed on the bases of examples (Gulwani, 2014; Matsuda, Cohen, Sewall, Lacerda, & Koedinger, 2007; Nosofsky et al., 1994). While these models exemplify a variety of approaches they can all be abstractly characterized as searching for a procedure, synthesized out of primitive components, that will produce the output given the input. Although this approach has not been applied to our task, such an approach could in principle be developed to learn from examples in our experiments. As multiple procedures can match any example, the procedure that these models produce is determined by various biases that guide their search. While simplicity is an inevitable part of these biases, there can be many other biases determined by how the models construct and search the space of possible procedures. If there is a random component in such models, some of these biases will be expressed probabilistically and enable the system to produce a range of programs given the same input. This random variation certainly was a characteristic of participants in our experiments. Such an approach would directly apply to the example conditions of our experiments. However, it could be further extended to explaining learning without an example by just generating a procedure according to the internal biases, without the constraint that it matches an example, and applying it to the problem. These approaches would explain the differences among conditions by essentially moving the probabilities we identified in our Markov model down into the biases in program generation. So, in a sense, they would still leave open what determines the probabilities.

8.2. Learning as a hypothesis testing

This study looked at learning to solve a class of problems as a hypothesis testing and demonstrated that constraining the search space could help learning by reducing trial-and-error search. Although the way of thinking of learning as a hypothesis testing is not new, this research provides an important clue on why some instructions are effective and others are not. Simply put, good instruction creates good biases, whereas poor instruction creates poor biases in hypothesis testing. The most surprising finding in our research was that participants showed no or little memory of failed hypothesis. This strongly suggests provision of instruction that creates good biases is critical to successful learning. Some may argue that no-memory assumption may not hold with extended amount of practice. Future studies will be needed to address how our no-memory assumption changes depending on the amount of instruction.

The current study suggests that focusing the search can improve learning of problem solving. One possible piece of supporting evidence comes from Zhu and Simon's (1987) work on example-based learning. They had students learn to factor quadratic expressions and demonstrated that students can learn problem solving by studying examples and/or by working problems without lectures if examples/problems are arranged in an appropriate way. They attributed students' learning success to the fact that students had to attend to only certain aspects of learning while solving a series of problems. This could have possibly reduced inefficient trial-and-error search and in turn working memory load. This is related to VanLehn's (1987) "one-disjunct-per-lesson" hypothesis that learning one subprocedure at a time is more effective than teaching two or more disjunctively related procedures in the same lesson.

Also, in the domain of science learning, several attempts have been made to help learning by supporting hypothesis formation in designing computer-supported learning environments. For example, Van Joolingen and De Jong (1993; De Jong & Van Joolingen, 1995; Van Joolingen & De Jong, 1991, 1996) introduced a cognitive tool called the hypothesis scratchpad. The scratchpad constrains the hypothesis that a learner could state and shows the contents of the hypothesis space. This can serve as a memory of the trace through hypothesis space. Kim and Pederson (2011) also introduced metacognitive scaffolds to strengthen the hypothesis-development process among sixth-grade participants.

Participants who used metacognitive scaffolds developed better hypotheses than the control group of participants and that hypothesis-development performance was predictive of solution-development performance.

One might think that the simplest way of constraining the search space is to provide explicit instruction about a correct rule. This is what we tried in the fourth experiment. While it had some success it was apparent that it did not constrain the search space down to just the correct rule. In fact, providing a scaffolded example was more successful, even though there was still search. As Fig. 19 reveals, each of our 22 different conditions can be characterized as inducing different biases to try the Hard and Easy rules. Learning was more successful when the induced biases matched what was to be learned. This is consistent with the general claim that good instruction is instruction that induces in the learner biases that hypothesize the correct rules and principles.

Although none of our conditions produced high rates of success in learning the Hard rule, we have observed high rates of success in a neuroimaging study conducted in our laboratory (H. S. Lee et al., 2015). Students first practiced learning from examples or verbal instruction for a set of 24 problems. Then, when we provided a novel verbal direction like “apply the top operator to the bottom two numbers” or an example illustrating the rule, participants were able to solve their first problem with over 90% accuracy. What subjects learned from the prior problems was appropriate biases for learning rules like our Hard rule.

8.3. Implications for instructional design

The findings from this research provide several implications for instructional design. First, when examples can be interpreted in more than one way, it will increase induction errors and negatively affect learning. Therefore, examples should be carefully selected so that learners can learn the correct rule, while not being misled by incorrect rules. Instructors should examine whether an example can be interpreted in multiple ways other than the correct way before they provide students with the instructional example. Although there is no doubt that ambiguity is problematic, recognizing and avoiding it might not be easy. Some ambiguity may not be obvious to instructors, especially when they already know problem-solving rules and they may experience “expert blind spot” (Koedinger & Nathan, 2004; Nathan & Petrosino, 2003).

Second, simply changing problem properties can help the discovery learning process. When students have to learn in a minimally guided instructional environment such as when doing homework alone, teachers can assign problems that are designed to help students focus on certain features of the problem. These could be formal features of problems such as numbers or operators as in our current study or informal perceptual features of problems such as spacing or visual salience (e.g., Landy & Goldstone, 2007, 2009). The way a problem is displayed can have a substantial effect on learning performance (Larkin & Simon, 1987; Zhang & Norman, 1994). For example, Landy and Goldstone (2007) demonstrated that accuracy was greater when visual layout of algebraic equations supported the mathematical convention via a non-mathematical grouping pressure (e.g., proximity, closure). Landy and Goldstone (2009) also showed that linear equation solving was facilitated when the background motion was consistent with the direction of the numeric transposition people would do to solve for the unknown variable. Using a graphic inference task, Hegarty, Canham, and Fabrikant (2010) also demonstrated the effect of visual salience on learning of inference task. Participants' performance increased when they were given visual materials that made task-relevant information more visually salient.

Third, provision of yes/no feedback appears insufficient to stop students from repeating their wrong strategies. The Markov modeling revealed that students had a tendency to stick with a wrong hypothesis even after they were provided with disconfirming evidence. We speculate that this is because they were not sure of why their answers were wrong and considered the reasonable hypothesis that their rule was correct but they had made a computational error. Effective feedback needs to make clear why an answer was wrong.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305A100109 to Carnegie Mellon University. We thank Aryn Pyke, Noboru Matsuda, and Christopher MacLellan for valuable comments on the study.

Notes

1. The word "rule" may have two senses. In a context of category learning, it may mean a criterion that divides class into different categories, whereas in a context of problem solving, it means a solution used for problem solving. In this study, we used the word to refer to the second meaning.
2. But participants were told that these answers were incorrect.
3. Estimation of all quantities is provided by the scripts associated with website https://www.dropbox.com/sh/nl4mrlocu46mnie/CekvOkmNJ5.
4. The corresponding Stay probabilities were .19, .25, .27, and .11. These are similar but not identical to the ones estimated with the four-state model used in the study. It will be noted that the curve tends to be rather flat for low Memory probabilities and Stay probabilities below .5, indicating a lack of sensitivity to the exact Stay probability.
5. Some may argue that this might be due to the language ability of the participants. Understanding of verbal instructions requires comprehension of language. Although we could not directly examine participants' native language, geo-locations of the Mechanical Turk population based on their ip address were identified as US territory in 98% of population. Also, results were not really different with our small CMU sample in Experiment 1.

6. Estimation of all quantities is provided by the scripts associated with website https://www.dropbox.com/sh/nl4mrlocu46mnie/CekvOkmNJ5.
7. The problems were only designed to guarantee that the Easy, Hard, and Equality answers would be different for each problem.
8. In this experiment, it was never the case that both the Hard and Easy rule would fail and so these states were really hypothetical and never reached.
9. The Stay probabilities for the no-memory models reported here are very slightly different than those in the main text because the estimated observation matrix was very slightly different.

References

Anderson, J. R., Lee, H. S., & Fincham, J. (2014). Discovering the structure of mathematical problem solving. *NeuroImage*, *97*, 163–177.

Bickenbach, F., & Bode, E. (2001). *Markov or not Markov-This should be a question* (No. 1086). Kieler Arbeitspapiere.

Bower, G., & Trabasso, T. (1963). Reversals prior to solution in concept identification. *Journal of Experimental Psychology*, *66*(4), 409–418.

Bower, G., & Trabasso, T. (1964). Concept identification. In R. C. Atkinson (Ed.), *Studies in mathematical psychology* (pp. 32–96). Stanford, CA: Stanford University Press.

Brainerd, C. J. (1979). Markovian interpretations of conservation learning. *Psychological Review*, *86*, 181–213.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology*, *18*, 293–328.

Cohen, P. R., & Beal, C. R. (2009). Temporal data mining for educational applications. *International Journal of Software and Informatics*, *3*, 31–46.

De Jong, T., & Van Joolingen, W. R. (1995). The SMISLE environment: Learning with and design of integrated simulation learning environments. In P. Held, & W. F. Kugemann (Eds.), *Telematics for Education and Training* (pp. 173–186). Amsterdam: IOS.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, *18*, 253–292.

González-Brenes, J. P., & Mostow, J. (2012). Dynamic cognitive tracing: Towards unified discovery of student and cognitive models. In K. Yacef, O. Za¨iane, H. Hershkovitz, M. Yudelson, & J. Stamper (Eds.), *Proceedings of the 5th International Educational Data Mining Society* (pp. 49–56).

Gulwani, S. (2014). Example-based learning in computer-aided STEM education. *Communications of the ACM*, *57*(8), 70–80.

Hegarty, M., Canham, M. S., & Fabrikant, S. I. (2010). Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 37–53.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kim, H. J., & Pederson, S. (2011). Advancing young adolescents' hypothesis-development performance in a computer-supported and problem-based learning environment. *Computers & Education*, *57*, 1780–1789.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *The Journal of the Learning Sciences*, *13*, 129–164.

Landy, D., & Goldstone, R. L. (2007). How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *33*, 720–733.

Landy, D. H., & Goldstone, R. L. (2009). How much of symbolic manipulation is just symbol pushing? In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society* (pp. 1072–1077). Austin, TX: Cognitive Science Society.

Larkin, J., & Simon, H. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*, 65–99.

Lee, H. S., & Anderson, J. R. (2013). Student learning: What has instruction got to do with it? *Annual Review of Psychology*, *64*, 445–469.

Lee, H. S., Anderson, A., Betts, S., & Anderson, J. R. (2011). When does provision of instruction promote learning? In L. Carlson, C. Hoelscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3518–3523). Austin, TX: Cognitive Science Society.

Lee, H. S., Betts, S., & Anderson, J. R. (2015). Not taking the easy road: When similarity hurts learning. *Memory & Cognition*, *43*, 939–952.

Lee, H. S., Fincham, J., & Anderson, J. R. (2015). Learning from examples versus verbal directions in mathematical problem solving. *Mind, Brain, and Education*, *9*, 232–245.

Lee, H. S., Fincham, J., Betts, S., & Anderson, J. R. (2014). An fMRI investigation of instructional guidance in mathematical problem solving. *Trends in Neuroscience and Education*, *3*, 50–62.

Levine, M. (1975). *A cognitive theory of learning*. Hillsdale, NJ: Lawrence.

Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. Thousand Oaks, CA: Sage.

Lewis, C. (1988). Why and how to learn why: Analysis-based generalization of procedures. *Cognitive Science*, *12*(2), 211–256.

Matsuda, N., Cohen, W. W., Sewall, J., Lacerda, G., & Koedinger, K. R. (2007). Predicting students' performance with simstudent: Learning cognitive skills from observation. *Frontiers in Artificial Intelligence and Applications*, *158*, 467.

Mawer, R. F., & Sweller, J. (1982). Effects of subgoal density and location on learning during problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*, 252–259.

Murphy, K. (1998). Hidden Markov Model (HMM) Toolbox for Matlab. Available at: http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html. Accessed August 13, 2014.

Nathan, M. J., & Petrosino, A. (2003). Expert blind spot among preservice teachers. *American Educational Research Journal*, *40*(4), 905–928.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological review*, *101*(1), 53–79.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Reed, S. K., & Bolstad, C. (1991). Use of examples and procedures in problem solving. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *17*, 753–766.

Restle, F. (1962). The selection of strategies in cue learning. *Psychological Review*, *69*, 329–434.

Richard, J. F., Cauzinille, E., & Mathieu, J. (1973). Logical and memory processes in a unidimensional concept identification task by children and adults. *Acta Psychologica*, *37*(5), 315–331.

Sweller, J. (1983). Control mechanisms in problem solving. *Memory & Cognition*, *11*, 32–40.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*, 257–285.

Sweller, J., Mawer, R. F., & Howe, W. (1982). Consequences of history-cued and means-end strategies in problem solving. *The American Journal of Psychology*, *95*, 455–483.

Trabasso, T. R. (1963). Stimulus emphasis and all-or-none learning in concept identification. *Journal of Experimental Psychology*, *65*, 398–406.

Trabasso, T. R., & Bower, G. H. (1964). Presolution dimensional shifts in concept identification: A test of the sampling with replacement axiom in all or none models. *Journal of Mathematical Psychology*, *3*, 163–173.

Trabasso, T., & Bower, G. H. (1968). *Attention in learning: Theory and research*. New York: Wiley.

Van Joolingen, W. R., & De Jong, T. (1991). Supporting hypothesis generation by learners exploring an interactive computer simulation. *Instructional Science*, *20*, 389–404.

Van Joolingen, W. R., & De Jong, T. (1993). Exploring a domain through a computer simulation: Traversing variable and relation space with the help of a hypothesis scratchpad. In D. Towne, T. de Jong, & H. Spada (Eds.), *Simulation-based experiential learning* (pp. 191–206). (NATO ASI series). Berlin: Springer.

Van Joolingen, W. R., & De Jong, T. (1996). Design and implementation of simulation-based discovery environments: The SMISLE solution. *Journal of Artificial Intelligence and Education*, *7*, 253–277.

VanLehn, K. (1987). Learning one subprocedure per lesson. *Artificial Intelligence*, *31*, 1–40.

Wittwer, J., & Renkl, A. (2008). Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations. *Educational Psychologist*, *43*(1), 49–64.

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

Zhu, X., & Simon, H. A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction*, *4*, 137–166.

Appendix[6]: Observation matrix

We estimated a single best-fitting model to the pooled data from all the experiments. In addition, we imposed the following constraints on the observation matrix to make the Easy, Hard, and Equality states equivalent in their association with their intended observations:

1. The probabilities of observing the appropriate response in each of these three states were the same. The best estimate of this probability was .92. The other .08 are interpreted as cases where participants made an error in calculating the answer appropriate to the state.
2. The .08 above includes responses that match either of the other two focus rules.[7] We constrained these probabilities to be the same. The best estimate of this probability was .002.
3. When in the Other state, the probability of observing a response that corresponded to one of the focus states was the same. The best estimate of this probability was .105. Most of these are probably cases where the participant's Other rule produced the same answer for that problem as one of the focus rules.

The estimation of these three probabilities completely determines the probability of observing a response that corresponds to Easy, Hard, and Equality in the four states. The probabilities of observing something else can then be calculated as $1-.9200-2 \times .002 = .0756$ for the Easy, Hard, and Equality states and $1-3 \times .105 = .685$ for the other state.

The resulting parameters are quite reasonable—a very high probability of giving a response that matches the state, a very low probability of giving a response in one of the focus state that matches another focus state, and a modest probability of giving something different.

Degrees of freedom for the transition matrices

The transition matrices for a condition have 12 degrees of freedom, but to increase interpretability we tested reducing this to three choice probabilities plus a stay probability that was shared by all the conditions in the matrix. Table A1a reports three tests ordered by how much they weight differences in significance vs. simplicity in choosing between the Full model and the Reduced model. That table gives the ingredients for calculating these various statistics:

1. Number of observations (N: 16 times the number of participants).
2. Number of parameters in estimating these models (K: 15 (3 start parameters, 12 transition parameters) x number of conditions for the Full model; 6 (3 start parameters, 3 choice parameters) x number of conditions plus 1 (Stay probability) for the Reduced model).
3. Log Likelihood (LL) of the data given those parameters.
4. BIC (Bayesian Information Criterion). This measure is calculated as

   $$\mathrm{BIC} = -2 \times \mathrm{LL} + K \times \ln(N)$$

   Smaller values are better. This measure most favors simplicity. The Bayes factor can be computed from the difference in BIC scores as exp($-\triangle$BIC/2) and can be interpreted as the relative probability of the two models (Kass & Raftery, 1995). It can be seen that in all cases BIC decisively favors the Reduced model.
5. AIC (Akaike's Information Criterion). This is measure is calculated as

   $$\mathrm{AIC} = -2 \times \mathrm{LL} + 2 \times K$$

   This measure penalizes number of parameters less. It is a measure of distance to the "true" model. Again in every case this measure favors the Reduced model.
6. Chi-square test. This test can be used when one model is nested within another, as in our case. Under the null hypothesis of no difference between the two models, twice the log-likelihood difference is distributed as a chi-square test with degrees of freedom equal to the difference in number of parameters. The null hypothesis can be rejected in the cases of Experiments 2 and 4, which is to be interpreted as there being some significant variance that is being captured by the additional parameters. Combined across the four experiments, the chi-square test is marginal just failing the .05 criterion value.

Tests of Markov property

A standard method (Bickenbach & Bode, 2001) of testing the Markov property is to determine whether different segments of the sequential data can be fit with the same parameters. To do this, we broke the data into the first eight problems and the second eight problems. The Single Model estimated a single set of Start probabilities and Choice probabilities to fit both segments. The Double Model estimated two sets for each half. In the Single Model a Start vector for the second half was calculated by using the start vector for the beginning of the experiment and applying the transition matrix eight times to it. Table A1b reports tests of the Markov property using the same structure as Table A1a. The Markov property is satisfied by all tests.

Bootstrapped estimates of errors of parameter estimates

To obtain bootstrapped estimates of the standard errors (Figs. 6, 10, 13, and 16) we randomly choose with replacement participants from the original experiment to create samples of the same size as in the original experiment. Thus, any participant could occur 0, 1, or more times in these random samples. The standard errors of the parameters are calculated by estimating parameters for 1,200 samples so generated.

The memory model

The memory model required creating a different state for considering a tracked rule (Easy, Hard, or Equality) for each possible set of past history of rules remembered as failed. This required creating four states for each of the Easy, Hard, and Equality rules. For instance, the Hard hypothesis could be considered without memory of any rules failed, after remembering Easy as failed, after remembering Equality as failed, or after remembering both rules as failed. As in the no-memory model, once it entered a state where it was considering the correct rule, it stayed in that state (but now there are four such absorbing states for the correct rule). There were eight Other states for each of the 2^3 combinations memory of Easy, Hard, and Equality rules as failed.[8] If participants did not stay in a state after an error, they could either transition to a state that included memory of the rule as failed or to a state that did not include such a memory. The probability of the first option was the choice probability for that state times the global memory probability for the experiment. The probability of the second option was the choice probability for that state times one minus the global memory probability for the experiment.

Just as we had for the no-memory model, we estimated the probabilities of giving each of the four categories of answers in any state (the observation matrix). These values were the same for all states that involved the same rule with different past histories. Initially the model would start in one of four states according to the Start probability, just like the no-memory model. We could derive the probabilities of subsequent histories from the Stay probability, the Memory probability, and three Choice probabilities for Easy, Hard, and Equality. If there were less than four states it could go to (because it remembered a rule as failed), it scaled its probabilities of going to these remaining states according to the choice probabilities associated with those states. For instance, if the choice probabilities for the two states were .1 and .3, the rescaled choice probabilities would be .25 and .75. The parameters estimated for different experiments and the software that implemented the memory model are available at the website that contains the data and code for this paper.

The best-fitting combination of memory and stay probabilities were 0.00 and 0.19 for Experiment 1, 0.11 and 0.25 for Experiment 2, 0.05 and 0.27 for Experiment 3, and 0.00 and 0.11 for Experiment 4. The best-fitting no-memory model for Experiment 2 had parameters of 0.00 and 0.35. It had a log-likelihood 5.40 worse than the best-fitting model, but considering the extra parameter the memory model's BIC measure was 2.32 worse.

The best-fitting no-memory model for Experiment 3 had parameters of 0.00 and 0.29. It had a log-likelihood 1.11 worse, but its BIC measure was 6.55 better.[9]

Table A1
Tests of the assumptions of the model

	Experiment				
(a) Full Model vs. Reduced Model	1	2	3	4	Combined
1. Numbers					
Participants	200	304	400	210	1,114
Observations	3,200	4,864	6,400	3,360	17,824
2. Parameters					
Full	60	90	120	60	330
Reduced	25	37	49	25	136
3. Log likelihood					
Full	−2,305.2	−3,509.2	−4,554.6	−1,683.3	−12,052.3
Reduced	−2,310.8	−3,561.4	−4,582.7	−1,710.4	−12,165.3
4. BIC					
Full	5,094.7	7,782.5	10,160.9	3,853.8	27,334.7
Reduced	4,823.4	7,436.9	9,594.8	3,623.8	25,661.8
5. AIC					
Full	4,730.4	7,198.4	9,349.2	3,486.6	24,764.6
Reduced	4,671.6	7,196.8	9,263.4	3,470.8	24,602.6
6. Chi-square test					
Chi-square	11.2	104.4	56.2	54.2	226.0
df	35	53	71	35	194
Probability	1.000	0.000	0.900	0.020	0.057

	Experiment				
(b) Markov Property	1	2	3	4	Combined
1. Numbers					
Participants	200	304	400	210	1,114
Observations	3,200	4,864	6,400	3,360	17,824
2. Parameters					
Double	49	73	97	49	268
Single	25	37	49	25	136
3. Log likelihood					
Double	−2,354.6	−3,622.8	−4,673.9	−1,723.2	−12,374.5
Single	−2,366.5	−3,642.5	−4,698.6	−1,733.4	−12,440.9
4. BIC					
Estimate	5,104.6	7,865.3	10,198.0	3,844.3	27,372.2
Re-Use	4,934.7	7,599.1	9,826.6	3,669.8	26,213.0
5. AIC					
Double	4,807.1	7,391.5	9,541.8	3,544.4	25,284.9
Single	4,782.9	7,358.9	9,495.1	3,516.8	25,153.8
6. Chi-square test					
Chi-square	23.8	39.4	49.3	20.4	132.9
df	24	36	48	24	132
Probability	0.474	0.320	0.421	0.675	0.463