Revising Learner Misconceptions Without Feedback: Prompting for Reflection on Anomalies

Joseph Jay Williams  
Harvard University  
 joseph_jay_williams@harvard.edu  

Tania Lombrozo  
UC Berkeley  
lombrozo@berkeley.edu  

Anne Hsu  
University of London  
anne.hsu@eecs.qmul.ac.uk  

Bernd Huber  
Harvard University  
berdbhuber@gmail.com  

Juho Kim  
Stanford University & KAIST  
juhokim@cs.kaist.ac.kr  

ABSTRACT
The Internet has enabled learning at scale, from Massive Open Online Courses (MOOCs) to Wikipedia. But online learners may become passive, instead of actively constructing knowledge and revising their beliefs in light of new facts. Instructors cannot directly diagnose thousands of learners’ misconceptions and provide remedial tutoring. This paper investigates how instructors can prompt learners to reflect on facts that are anomalies with respect to their existing misconceptions, and how to choose these anomalies and prompts to guide learners to revise incorrect beliefs without any feedback. We conducted two randomized experiments with online crowd workers learning statistics. Results show that prompts to explain why these anomalies are true drive revision towards correct beliefs. But prompts to simply articulate thoughts about anomalies have no effect on learning. Furthermore, we find that explaining multiple anomalies is more effective than explaining only one, but the anomalies should rule out multiple misconceptions simultaneously.

ACM Classification Keywords
K.3.1 Computing Milieux: Computer Uses in Education; J.4 Computer Applications: Social and Behavioral Sciences

Author Keywords
Online learning; Explanation; Prompts; MOOCs.

INTRODUCTION
The Internet provides tremendous opportunities for learning, from Massive Open Online Courses (MOOCs) to Wikipedia. However, much of this learning occurs without access to a teacher or the corrective feedback a teacher would typically provide. In the absence of catered instruction and feedback, learners may be less likely to engage in belief revision, especially when doing so requires overcoming existing misconceptions. For example, facts that contradict existing beliefs could be ignored or simply memorized [5], without learners truly grappling with the implications and moving away from misconceptions towards more accurate understanding. As the number of learners grows, it becomes impossible for instructors to diagnose individual students’ incorrect beliefs and dynamically design activities to correct them. In light of these challenges, how can we design online environments to ensure large-scale active learning and belief revision without instructors’ real-time involvement?

Existing approaches try to leverage peers for discussion [16] or assessment [10], and intelligent tutoring systems for feedback [13]. For instance, some present an automatic hinting interface [9], use Natural Language Processing to coach answering domain-specific questions [6], or provide feedback on the correctness of multiple choice explanations in an intelligent tutoring system [2]. However, it is challenging to scale the success of these intelligent tutoring technologies to the many new online lessons and problems that are rapidly emerging.

Another approach, which has received less attention, is to guide learners to engage in belief revision themselves. This approach faces an obvious challenge: because learners are by definition ignorant of what they are trying to learn, they can’t replace the role of an informed instructor providing accurate feedback. On the other hand, there’s evidence that engaging in reflective cognitive processes, such as explanation, can help learners identify gaps and inaccuracies in their current beliefs and guide them to better alternatives, even in the absence of feedback [4, 11, 12]. Appropriate reflective prompts or “Socratic questions” may thus achieve some of the benefits of interactive instruction, while being broadly applicable across domains, easy to implement, and easy to scale.

Designing successful reflective prompts requires an understanding of how people learn. How can the cognitive processes for belief revision be elicited by the right interface features – such as reflective questions – and the right content – like particular facts or examples presented for reflection [14]? Accordingly, this paper tries to help instructors by investigating the following questions: Which reflective prompts most ef-
effectively promote belief revision? And towards what content should reflective prompts be directed? Specifically, to what extent should learners be prompted to engage with anomalies – material that challenges prevalent misconceptions?

RELATED WORK
Online environments frequently provide explanations of concepts to learners through lecture videos. However, research in education emphasizes the pedagogical value of using technology to prompt learners to generate their own self-explanations of what concepts or facts mean in their own words [1, 4]. Of particular relevance to belief revision is the finding that explaining why a fact is true does not merely boost attention or motivation, but drives people to interpret what they are explaining as one instance of a broader pattern [18, 19, 20, 8].

In light of this finding, it may be especially effective for learners to generate explanations for anomalies – facts that are inconsistent with their prior beliefs. Explaining anomalies could potentially overturn prior misconceptions and direct learners to alternatives that render the anomaly intelligible. For example, explaining why 2 is a prime number could conflict with a learner’s misconception (that all primes are odd) and help guide the learner to a more accurate principle that accounts for the anomaly (that primes have exactly two unique divisors).

Despite the potential for anomalies to revise beliefs, education research has revealed that merely presenting people with anomalies frequently fails to elicit cognitive processes that change entrenched beliefs [5]. People need to process the anomalies appropriately [7, 3], and prompts to explain the anomalies could be an effective way to elicit the requisite processing.

STUDIES: DESIGNING PROMPTS & ANOMALIES
Our goal was to guide online learners to revise prior misconceptions, even when feedback from an instructor was unavailable. The educational topic was learning how to compare samples from different populations [14, 8]. This task requires understanding statistical variability, which is central to many everyday decisions [15, 17]. Our design of prompts and anomalies was intended to be technologically easy for instructors to implement, yet psychologically potent in revising beliefs. Two experiments investigated which reflective prompts would be effective, and how to choose the number and distribution of anomalies targeted by the prompts.

Methods
Materials: Misconceptions & Anomalies in Statistics Problems
The experiments had participants learn how a university compared student grades that came from different courses. Participants reflected on observations like the ranking of the sample pair of students in Figure 1. Sarah’s and Tom’s scores were shown along with the respective class’s average, max, min, and deviation, and a statement about who was ranked higher by the university (Tom). Learning the correct ranking rule required integrating statistical knowledge with observations. Figure 2 shows how Sarah and Tom would be ranked by four different rules for comparing samples from populations.

The "More deviations above the average" rule was the true basis for the university’s ranking. In all five (six in Experiment 2) ranked pairs that participants saw, the higher-ranked student was whoever was a greater number of deviations above the average. This corresponds roughly to the higher standardized normal or z-score.

Figure 2 also shows how Sarah would be expected to be ranked higher, rather than Tom, according to three common misconceptions learners have about ranking [14]. For example, belief in the "Higher Score" rule would give a higher rank to the student with a higher score, without taking the course mean or variability into account. These "misconceptions" each neglect important statistical concepts captured by the "More deviations above average," or z-score.

We define a ranked pair as an anomaly with respect to a misconception about ranking, when the ranking contradicts what an incorrect belief predicts. The Sarah–Tom pair in Figure 1 is therefore an anomaly with respect to each of the three misconceptions.

\[\text{[18] is a non–archival and less extensive analysis and discussion of this data. [22] is a non-archival report of an earlier version of Experiment 1.}\]
We recruited 659 (Experiment 1) and 261 (Experiment 2) participants. In Experiment 1, learners were randomly assigned to receive an Explain vs. a Write Thoughts prompt. This manipulation allows us to assess whether explanation prompts are especially effective relative to generic prompts for explicit reflection.

Procedure

Before participants studied the ranked pairs, they completed a Pre-Test by making predictions of who would be higher ranked for four unranked pairs of students. Each student pair was similar to the Sarah-Tom pair in that the correct “More deviations above average” rule predicted the opposite ranking (e.g., Tom) to all three misconceptions (e.g., Sarah).

Participants then studied the five (Expt. 1) or six (Expt. 2) ranked pairs. Each ranked pair was onscreen for exactly two minutes, so participants could not take more or less time. Below the ranked pair was a Reflection Prompt and text box to type into, as shown in Figure 1. Participants’ beliefs were then measured by predicting rankings in a Post-Test. The Post-test used four unranked pairs that were statistically isomorphic to the Pre-Test, except for having different numbers and names.

We operationalize belief revision as the Pre- to Post-test increase in accuracy. This measures the degree to which learners are driven to believe in the correct rule over the misconceptions, after reflecting on the ranked pairs. Pre- to Post-test Accuracy Increase is the dependent variable in all graphs and statistical tests.

Participants

We recruited 659 (Experiment 1) and 261 (Experiment 2) participants on Amazon Mechanical Turk to do a 20-40 minute research study on learning. Compensation was around $3.00–$6.00 per hour. Our goal in using crowd workers was to obtain a more representative online sample than undergraduate laboratory participants, while enabling greater experimental control and measurement of learning than with students taking an online course.

Experimental Comparison of Prompts and Anomalies

Experiments 1 and 2 used between-subjects factorial designs that independently manipulated multiple variables. To investigate the design of reflection prompts to promote belief revision, both experiments varied the kind of Reflection Prompt, randomly assigning learners to receive an Explain vs. a Write Thoughts prompt. Prior research suggests that prompt to explain should be especially potent [20]. The Write Thoughts prompt was selected as a close comparison to explaining in terms of effort and engagement, and was also matched in requiring a verbal response. This manipulation allows us to assess whether explanation prompts are especially effective relative to generic prompts for explicit reflection.

Results

Experiment 1

Accuracy Increase in Experiment 1 is shown in Figure 3. We conducted a 2 (Reflection Prompt: Explain vs. Write Thoughts) x 2 (Number of anomalies: 1 vs. 4) ANOVA on Accuracy Increase from Pre- to Post-test. There was a significant overall effect of Reflection Prompt, with greater belief revision from Explain than Write Thoughts prompts ($F(1, 659) = 13.23$, $p < 0.01$). Belief revision was also increased by a greater Number of Anomalies ($F(1, 659) = 24.53$, $p < 0.01$).

However, explaining was only beneficial when sufficiently many anomalies were being explained (Figure 4). There was an interaction between Reflection Prompt and Number of anomalies ($F(1, 659) = 8.20$, $p < 0.01$). Prompts to Explain promoted belief revision when the targets included 4 anomalies ($t(260) = 4.07$, $p < 0.01$), but had no effect when the targets only included 1 anomalous fact ($t(367) = 0.62$, $p = 0.54$).
### Experiment 2

To analyze the data from Experiment 2, we conducted a 2 (Reflection Prompt: Explain vs. Write Thoughts) x 2 (Number of anomalies: 2 vs. 4) x 2 (Distribution of anomalies: Overlapping vs. Distributed) ANOVA on Accuracy Increase from Pre- to Post-test. This analysis revealed a significant interaction between Reflection Prompt and Distribution of anomalies \((F(1, 259) = 6.11, p < 0.05)\). To visually represent this interaction, Figure 6 presents Accuracy Increase as a function of these two factors. When the distribution of anomalies was Overlapping, as in Experiment 1, prompts to Explain promoted significantly greater belief revision than prompts to Write Thoughts \((t(127) = 2.20, p < 0.05)\). However, when anomalies were Distributed, the relative benefits of engaging in explanation disappeared \((t(129) = 1.32, p > 0.19)\).²

### Summary of Key Results

Experiments 1 and 2 asked three questions. First, which reflection prompts most effectively promote learning? The answer is that not all reflective prompts are equal: prompts to *explain* anomalies were significantly more effective than prompts to write thoughts, even though both tasks were similarly demanding and required a verbal response.

Second, how many anomalies should be explained? The findings suggest that explaining a single anomaly is insufficient. In Experiment 1, the benefits of explanation were only observed when 4 anomalies were explained, while in Experiment 2, the benefits of explanation were not significantly different across 2 vs. 4.

Third, how should anomalies be distributed? The findings suggest that anomalies are more effective in promoting learning when they are "overlapping" in the sense that they simultaneously rule out multiple misconceptions. In fact, when anomalies were distributed (in Experiment 2), the benefits of explanation vanished. This is likely because participants were able to explain the anomalies by appeal to alternative misconceptions [21], and therefore lacked unique guidance towards the correct ranking rule.

In sum, the experiments showed how to design reflective prompts in online lessons to guide learners to revise their misconceptions towards accurate beliefs, even in the absence of instructor feedback. However, the effectiveness of the prompts depends both on the prompt itself and on its target. Prompts to *explain* multiple anomalies were most effective, and it mattered that the anomalies simultaneously rule out multiple potential misconceptions.

### DISCUSSION & LIMITATIONS

There are limitations to the current work. It could be hard or time-consuming for instructors to identify learners’ misconceptions and generate corresponding anomalies; here we benefited from prior research on misconceptions within our content domain. Also, further studies are needed to see if our results generalize to learning tasks with other materials and other populations of online learners, who may differ from Mechanical Turk workers in their goals and motivation. For example, it would be interesting to investigate prompting users to explain anomalous facts in informal learning materials like Wikipedia pages.

Our work presents an example of how cognitive science theories can be used to experimentally identify scalable principles for instructional design. A sophisticated understanding of how people learn allowed us to formulate and rule out novel hypotheses about how to prompt learners so that they themselves would actively revise their beliefs, even without instructor feedback. With efficient technical implementation in a broad range of online lessons and problems, these prompts can leverage and empower learners’ active engagement to promote learning at scale.

### REFERENCES


---

²The omnibus ANOVA also revealed a main effect of Number of anomalies, with greater learning when more anomalies were present, \(F(1, 259) = 8.20, p < 0.01\). This factor did not interact with Reflection Prompt (as it did in Experiment 1), potentially due to the shift from 1 vs. 4 anomalies (in Experiment 1) to 2 vs. 4 (in Experiment 2).


