

University of Vermont
ScholarWorks @ UVM

UVM Honors College Senior Theses

Undergraduate Theses

2018

Worker Retention, Response Quality, and Diversity in Microtask Crowdsourcing: An Experimental Investigation of the Potential for Priming Effects to Promote Project Goals

Brian J. Colombini
University of Vermont

Follow this and additional works at: <https://scholarworks.uvm.edu/hcoltheses>

Recommended Citation

Colombini, Brian J., "Worker Retention, Response Quality, and Diversity in Microtask Crowdsourcing: An Experimental Investigation of the Potential for Priming Effects to Promote Project Goals" (2018). *UVM Honors College Senior Theses*. 227.
<https://scholarworks.uvm.edu/hcoltheses/227>

This Honors College Thesis is brought to you for free and open access by the Undergraduate Theses at ScholarWorks @ UVM. It has been accepted for inclusion in UVM Honors College Senior Theses by an authorized administrator of ScholarWorks @ UVM. For more information, please contact donna.omalley@uvm.edu.

WORKER RETENTION, RESPONSE QUALITY, AND
DIVERSITY IN MICROTASK CROWDSOURCING
AN EXPERIMENTAL INVESTIGATION OF THE POTENTIAL
FOR PRIMING EFFECTS TO PROMOTE PROJECT GOALS

An Undergraduate Honors Thesis Presented
by
Brian J. Colombini
to
The Faculty of the College of Engineering and Mathematical Sciences
of
The University of Vermont

May, 2018



The University of Vermont

Faculty Advisor: James P. Bagrow
Committee Member: Margaret J. Eppstein
Committee Member: Robert M. Erickson

Abstract

Online microtask crowdsourcing platforms act as efficient resources for delegating small units of work, gathering data, generating ideas, and more. Members of research and business communities have incorporated crowdsourcing into problem-solving processes. When human workers contribute to a crowdsourcing task, they are subject to various stimuli as a result of task design. Inter-task priming effects - through which work is nonconsciously, yet significantly, influenced by exposure to certain stimuli - have been shown to affect microtask crowdsourcing responses in a variety of ways. Instead of simply being wary of the potential for priming effects to skew results, task administrators can utilize proven priming procedures in order to promote project goals. In a series of three experiments conducted on Amazon's Mechanical Turk, we investigated the effects of proposed priming treatments on worker retention, response quality, and response diversity. In our first two experiments, we studied the effect of initial response freedom on sustained worker participation and response quality. We expected that workers who were granted greater levels of freedom in an initial response would be stimulated to complete more work and deliver higher quality work than workers originally constrained in their initial response possibilities. We found no significant relationship between the initial response freedom granted to workers and the amount of optional work they completed. The degree of initial response freedom also did not have a significant impact on subsequent response quality. However, the influence of inter-task effects were evident based on response tendencies for different question types. We found evidence that consistency in task structure may play a stronger role in promoting response quality than proposed priming procedures. In our final experiment, we studied the influence of a group-level priming treatment on response diversity. Instead of varying task structure for different workers, we varied the degree of overlap in question content distributed to different workers in a group. We expected groups of workers that were exposed to more diverse preliminary question sets to offer greater diversity in response to a subsequent question. Although differences in response diversity were revealed, no consistent trend between question content overlap and response diversity prevailed. Nevertheless, combining consistent task structure with crowd-level priming procedures - to encourage diversity in inter-task effects across the crowd - offers an exciting path for future study.

Table of Contents

1	Introduction	1
1.1	Crowdsourcing	1
1.2	Priming	1
1.3	Priming for Motivation and Quality in Crowd Work	2
1.4	Priming for Response Diversity in Crowd Work	2
1.5	Organization of Thesis	3
2	Related Work	3
2.1	Efficiency and Motivation	3
2.2	Diversity	5
3	General Experimental Design Considerations	6
3.1	External Questions	6
3.2	Domain of Crowd Work	6
4	Experiment 1(a): Sustained Participation	6
4.1	Experimental Design	7
4.1.1	Overview of HIT Structure and Slide Interface	7
4.1.2	Slide 1: Initial Subtask/Treatment	7
4.1.3	Slide 2: Mandatory Follow-Up Subtasks	8
4.1.4	Slides 3-6: Optional Bonus Subtasks	9
4.1.5	Sources of Constrained Input Term Options and Follow-Up Subtask Word Pairs	11
4.1.6	Data Collection	11
4.1.7	HIT Presentation and Payment Scheme	11
4.2	Results	12
4.2.1	Crowdsourcing Metrics	12
4.2.2	Time Spent	12
4.2.3	Number of Bonus Subtasks Completed	12
4.3	Discussion	13
4.3.1	Nuance of Priming Treatment	13
4.3.2	Individualized Reactions to a Common Priming Treatment	14
4.3.3	Future Work	14
5	Experiment 1(b): Response Quality	15
5.1	Experimental Design	15
5.1.1	Overview of HIT Structure and Slide Interface	15
5.1.2	Slide 1: Initial Subtask/Treatment	15
5.1.3	Slide 2: Mandatory Follow-Up Subtasks	16
5.1.4	Data Collection	17
5.1.5	HIT Presentation and Payment Scheme	17
5.1.6	Measuring Response Quality	17
5.2	Results	18
5.2.1	Crowdsourcing Metrics	18
5.2.2	Bonferroni Correction	18
5.2.3	Time Spent	18
5.2.4	Probability of "Best" Response Type	19
5.2.5	Response Type Frequency by Word Pair Type	19
5.2.6	Small Sample Size Simulations	21

5.3	Discussion	23
5.3.1	Free Group Response Bias	23
5.3.2	Future Work	23
6	Experiment 2: Response Diversity	24
6.1	Experimental Design	24
6.1.1	Overview of HIT Structure	24
6.1.2	Treatment: Content Overlap in Preliminary Question Sets	24
6.1.3	Test Question	25
6.1.4	Sources of Cause Terms	25
6.1.5	Data Collection	26
6.1.6	HIT Presentation and Payment Scheme	26
6.1.7	Measuring Response Diversity	26
6.2	Results	27
6.2.1	Crowdsourcing Metrics	27
6.2.2	Bonferroni Correction	27
6.2.3	Time Spent	27
6.2.4	Lexical Diversity: Response Frequency	27
6.2.5	Lexical Diversity: Mode Fraction	28
6.2.6	Semantic Diversity	29
6.3	Discussion	30
6.3.1	Nuance in Priming Treatment and Test Question	30
6.3.2	Future Work	31
7	General Discussion	32
8	Acknowledgments	33

List of Figures

1	Experiment 1(a): Screenshots of initial subtasks	7
2	Experiment 1(a): Screenshots of follow-up subtasks	10
3	Experiment 1(a): Complementary cumulative distribution functions of number of bonus subtasks completed	14
4	Experiment 1(b): Screenshot of initial Baseline subtask	16
5	Experiment 1(b): Response type probability heat maps	20
6	Experiment 1(b): Majority "best" response rate vs. bootstrapped sample size by word pair type	22
7	Experiment 1(b): Majority "best" response rate vs. bootstrapped sample size for all word pairs	22
8	Experiment 2: Screenshot of task	25
9	Experiment 2: Fraction of unique and fraction of singleton responses	28
10	Experiment 2: Boxplots of word2vec cosine similarity	30

List of Tables

1	Experiment 1(a): Experimental probabilities of number of bonus subtasks completed	13
2	Experiment 1(b): Response type frequency mean by word pair type	21
3	Experiment 2: Treatment groups	26
4	Experiment 2: Mode responses and mode fractions	28
5	Experiment 2: Pairwise comparisons of bootstrapped sample mode fraction distributions	29
6	Experiment 2: Pairwise comparisons of word2vec cosine similarity distributions	31

1 Introduction

1.1 Crowdsourcing

Over the course of the past decade, researchers, industry professionals, and entrepreneurs have turned to crowdsourcing as a means of harnessing human brainpower to complete tasks that escape the reach of what machines can reliably accomplish. Crowdsourcing is often utilized for the completion of monotonous tasks, such as image recognition and audio transcription, but it can also be targeted for human research or creative efforts, such as collaborative writing [1,2]. Online distribution platforms, like Amazon’s Mechanical Turk¹ (MTurk), enable large problems to be broken down into microtasks that can be readily accomplished by a crowd of human workers [3]. Microtask workers receive financial compensation for their work on what MTurk refers to as Human Intelligence Tasks (HITs).

1.2 Priming

Priming is a phenomenon of human memory by which exposure to a stimulus effects responses to subsequent stimuli. Priming occurs nonconsciously, producing automatic effects that influence many functions of memory and human life. Faculties of recognition [4], social behavior [5], and consumer choice [6] have all been shown to be vulnerable to priming effects. Although microtask crowdsourcing is sometimes thought of as “artificial artificial intelligence”, the human characteristics of workers - such as susceptibility to priming effects - can greatly influence their reactions to different task designs. The input of workers may therefore be directly affected by simple design and wording choices.

When overlooked, priming effects may produce undesirable crowdsourcing results, but priming can also be utilized as a design tool to promote project goals in crowd work. Such goals may include the maximization of creativity, accuracy, or diversity in responses. Affective priming, to induce positive or negative emotion, has been shown to significantly improve creativity in microtask responses. Moreover, priming for positive emotion (via a background photo of a laughing baby) was shown to lead to increases in response creativity, while selecting for workers with high levels of self-reported happiness had the opposite effect in the same study [7, 8]. The power of priming procedures is a result of their inconspicuous implementation.

The content of an initial set of questions can induce priming effects on subsequent questions. The presence of questions with verifiable answers has been shown to increase the quality of later responses to more subjective questions [9]. Furthermore, initial questions with greater levels of similarity in content have been shown to alter the focus of microtask workers such that subsequent responses exhibit greater specificity. Incredibly, this effect is at least as strong as explicitly framing the purpose of the task in a way that prompts for more specific responses [10]. The power of priming effects has been demonstrated in social science laboratories and microtask platforms alike. Those who publish crowdsourcing tasks (crowdsourcers) must acknowledge the effects of priming in order to either promote specific in-task biases, or mitigate the effects of in-task biases altogether.

¹<https://www.mturk.com/>

1.3 Priming for Motivation and Quality in Crowd Work

In order to optimize their time and the availability of workers, crowdsourcers must harness the capabilities of amassed workers to the maximum extent by executing thoughtful distribution methods, and both sparking and maintaining worker motivation. Previous research has illuminated ways to increase the efficiency of crowd workers [2, 3, 9, 11–17]. The focus of previous work has mainly been on decisions that affect the population of workers that a task draws, and the original motivation that workers have when choosing to undertake a task. Methods of increasing extrinsic motivation – motivation stemming from factors external to the task itself (i.e. financial compensation) [18] – have been shown to increase the quantity of work completed by individual crowd workers with respect to a given task [11, 12]. On the other hand, methods of increasing intrinsic motivation – motivation stemming from factors that can be attributed to the task itself (i.e. interest) [18] – have been shown to improve the quality of crowdsourced work [12].

On MTurk, workers are influenced by extrinsic and intrinsic motivations when initially choosing to work on a task. Priming mechanisms can potentially offer ways through which motivation can also be ignited or sustained *within* a task or series of microtasks. Explorations of these mechanisms could lead to the discoveries of novel methods for maintaining engagement, motivating workers to complete tasks once they have already begun them. Maintaining engagement can be crucial for microtask projects because retention positively affects response accuracy - workers who perform similar consecutive tasks gain contextual knowledge that helps improve their work [14, 19].

In our first experiments - Experiments 1(a) and 1(b) - we investigate the influence of the level of response freedom granted to workers on an initial question. Specifically, we study the effects of initial response freedom on sustained participation and response quality in subsequent work. We expect workers who are initially prompted to propose their own ideas, rather than select from a pre-generated list of ideas, will be motivated to complete a greater quantity of subsequent, more monotonous, tasks. We expect that workers who are enabled to respond more freely, and hence more creatively, will be stimulated by the initial work in a way that introduces priming effects that result in more sustained participation and better response quality throughout the task series.

Hypothesis_{1(a)}: Microtask crowd workers who are enabled to respond more freely in the initial work of a task will complete more optional work on subsequent parts of the task.

Hypothesis_{1(b)}: Microtask crowd workers who are enabled to respond more freely in the initial work of a task will produce higher quality responses in subsequent parts of the task.

1.4 Priming for Response Diversity in Crowd Work

For crowdsourcers collecting responses to questions with verifiable answers, correctness is often the most crucial measure of response quality. However, certain task types request responses that escape evaluation by purely objective judgment. In many cases, response *diversity* becomes a measure of the quality of a crowdsourced data set. For example, some crowdsourcing projects aim at the generation of new ideas. Ideas may be sought after for purposes in commerce, science, or another domain of invention. Two major avenues of idea generation are the identification of problems and the brainstorming of potential solutions. Response diversity is beneficial in both avenues because identifying an expansive list of areas

of concern or interest, and collecting a large number of alternative solutions, enables the realization and evaluation of previously non-obvious choices [20–22]. In this respect, response diversity enables crowdsourcing as a tool not only for data collection, but for supplementing the intelligence of crowdsourcers. Crowdsourcing can also enhance the creation of data sets for which diversity is an advantage. For example, large amounts of non-expert inputs can be utilized for the creation of robust machine learning training sets [23].

An intuitive method for increasing the diversity of responses in a crowdsourcing task is to employ a diverse set of workers. However, screening the crowd for a diverse set of workers can stress budgetary constraints by costing crowdsourcers time and money. Furthermore, the nature of certain crowdsourcing tasks may require an appropriately qualified subset of crowd workers, thereby limiting the potential for diversity in the set of workers employed.

We propose a method of question distribution that exploits inter-task priming effects, with the goal of increasing response diversity within a given set of crowd workers. We do not aim to manipulate response diversity by priming for any particular scope of response from individual workers. Instead, we investigate the *magnitude of overlap* in the content of initial question sets distributed to workers, and its effect on responses to a subsequent question that is fixed for all workers. We expect that groups of workers that experience less overlap in initial question prompts will exhibit greater diversity in responses to a subsequent question.

Hypothesis₂: Groups of microtask crowd workers who are exposed to more diverse sets of preliminary question prompts will offer greater diversity in responses to subsequent work.

1.5 Organization of Thesis

Our work is presented as follows. First, in Section 2, we review previous crowdsourcing research related to worker efficiency, motivation, and response diversity. In Section 3, we describe experimental design considerations that apply to all three of our crowdsourcing experiments. In Sections 4, 5, and 6, we report on the design and results of the three experiments, and discuss the implications of each. Section 7 houses a general discussion informed by the results of all three experiments, and includes final thoughts on the work.

2 Related Work

2.1 Efficiency and Motivation

Identifying and understanding elements of task design that influence worker efficiency have been popular topics in crowdsourcing research. The presentation and mechanics of task interfaces have been found to play a role in both the distribution and execution of tasks. In a comprehensive review of quality control in online crowdsourcing, Allahbakhsh et al. advise that a simple and appealing user interface can draw more workers to a task, while avoidable interface complexities deter and delay reliable workers [13]. Crowd workers prefer tasks that are quick, easy, and unique [3, 14]. Supplemental input tools have been shown to affect the pacing of workers. For example, autocompletion interfaces for text input fields were found to produce slower response times [15]. Directing worker attention, in order to influence cognitive processing time,

presents another opportunity in task interface design. Tactical highlighting and input element placement have been shown to increase accuracy in crowdsourced data entry, but this work did not conclude any effects on temporal efficiency [16].

Less visual and mechanical methods of influencing work quantity and quality have also been investigated. One example of an inconspicuous approach concerns the reduction of interruptions and the optimization of microtask ordering, so that workers can build and draw from contextual knowledge. Interruption reduction and ordering strategies have been shown to increase temporal efficiency in crowdsourced work [14]. Kittur, Chi, and Suh found that initially prompting workers for verifiable information improved worker responses for the subjective task of evaluating a Wikipedia entry [9]. Little et al. demonstrated that showing workers the responses of previous workers – for the creative task of brainstorming a company name – produced a higher average rating for response quality. This iterative approach involves the recycling of worker responses to influence future work, a strategy that may help researchers sustain the use and value of individual worker responses [2, 17]. Evidently, a multitude of mechanical and conceptual approaches have been taken to examine ways of affecting the response time, reliability, and overall production of crowd workers.

Motivation is another intriguing facet of crowd work that strongly influences output. In crowdsourcing, there are several layers of motivation that must be present for maximum production. Workers must be motivated to: 1) contribute to crowdsourcing initiatives in general, 2) contribute to specific crowdsourcing initiatives, and 3) complete all of the available work requested by those initiatives. Quinn and Bederson’s classification framework for distributed human computation systems identifies motivation as a key factor in characterizing crowdsourcing tasks. The researchers describe motivational devices that are popularly targeted by crowdsourcing initiatives: pay, altruism, fun, and implicitness (work that may be completed as part of another task already being completed by an online user) [24].

Research on pay, an extrinsic motivator, has found that increased payment expedites initial task participation [25] and increases the amount of work completed by crowdsourced workers [11, 12]. However, increased payment does not increase the accuracy of crowd work [11, 25]. Furthermore, Mason and Watts find evidence that differences in workers’ intrinsic motivations may cloud the effects of variable payment [11]. In an examination of crowdsourcing contests, Zheng, Li, and Hou found that intrinsic motivation had a more significant influence on participation than pay or recognition [26]. Rogstadius et al. concluded that intrinsic motivation produces higher quality results in crowdsourced work. The researchers also found that intrinsic motivation does not increase the amount of work completed per worker. In this case, intrinsic motivation was harvested through the appeal of altruism – supporting a non-profit organization versus a private corporation [12]. Work concerning other forms of intrinsic motivation, such as those that may be prompted by intellectual or creative stimulation, remains to be carried out.

According to Amabile, task motivation may be thought of as the combination of: 1) a person’s general, usually unchanging feeling towards a task, and 2) the person’s situational reason for working on the task at any given time. The latter factor is affected by the circumstances surrounding each given instance of the task [18, 27]. It is reasonable to infer that priming effects play a role in the circumstances surrounding a unit of microtask work. Dietze and Gadiraju demonstrated that inconspicuously embedding inspirational quotes in microtask questions, such that the quotes were part of the questions themselves rather than explicitly posed in a disconnected manner, primed workers to complete significantly more questions in a crowdsourcing task [19]. Perhaps more discreet priming measures can produce the same effect.

Formulating task prompts to allow for freedom in worker responses is not a novel concept in crowdsourcing. The

initial adoption of crowdsourcing techniques by industrial entities and others was in large part undertaken for the purpose of innovation, fueled by the recognition of an untapped pool of knowledge and creativity [1, 28]. Crowd workers may experience not only senses of reward, but purpose and ambition as a result of contributing to crowdsourced initiatives in unique ways [28]. In Experiments 1(a) and 1(b), we investigate response freedom as a means to an end.

2.2 Diversity

Microtask crowdsourcing is often employed to leverage the “wisdom of crowds” that emerges from aggregating the responses of many independently working individuals, rather than a small selection of subjects, consultants, or experts. Surowiecki finds that crowd wisdom is dependent on the diversity of individuals in the crowd. The holding of private knowledge by individual participants plays a crucial role in the quality of crowd-based solutions [29]. However, online crowdsourcing places a few inherent limitations on crowd diversity. Brabham notes that the online environment of microtask crowdsourcing limits access to those who can afford, understand, and use the Internet [28]. Moreover, crowdsourcers may require that workers meet desired qualifications for the purposes of task communication and quality assurance. Limiting the geographic location of workers to predominantly English-speaking countries, for tasks that require English proficiency, decreases the chance of task misinterpretation; Only permitting workers with high approval ratings to participate in a task may reduce the risk of employing unreliable workers. Although groups of participants on microtask platforms may be more diverse than samples from other Internet sources or traditional research subject pools [25], the nature of crowdsourcing platforms and the requirements of certain tasks bound the diversity of crowdsourced workers.

Nevertheless, researchers have pursued efforts to maximize the diversity of workers who perform a given task. Wu et al. found algorithmic ways to maximize the diversity of the group of workers to whom a task is distributed. The proposed models appear successful in generating diverse sets of workers, but require surveying and profiling of the potential worker pool prior to task distribution [30]. These additional steps cost time (for the preparation and evaluation of surveys, and implementation of the assignment algorithms) and money (for the compensation of researchers spending the aforementioned time and workers who must spend time completing screening surveys).

Employing a diverse crowd may be the most intuitive and effective way to promote response diversity, but it is not the only method available to crowdsourcers. Given a particular crowd of workers - perhaps with limited diversity - there are possibilities to influence response diversity. The implementation of certain user interface tools, such as autocompletion interfaces, can increase the diversity of text-based responses among crowd workers [15]. Priming mechanisms offer other powerful ways for crowdsourcers to tune the diversity of responses. Chowdhury et al. found that priming procedures can be exploited to embed domain-specific knowledge in tasks, in order to train workers and reduce response diversity in favor of accuracy [31]. In the domain of image captioning, Newell and Ruths concluded that workers who label an initial set of similar images are primed to label subsequent images with greater specificity, leading to greater response diversity across the crowd [10]. In another method of question content manipulation, Jiang, Kummerfeld, and Lasecki found that prompting workers to paraphrase the proposed paraphrases of other workers led to greater response diversity than prompting all workers with the same original sentence to be paraphrased [23]. Evidently, the diversity of responses offered by a given set of workers may be influenced by task design.

In our experiment concerning response diversity - Experiment 2 - we investigate the distribution of prompt content, in terms of the overlap in the content of initial questions, between workers. Varying question overlap between workers is an easily implemented mechanism. With the discovery of emergent and significant effects, the mechanism could provide crowdsourcers with an efficient tool for promoting response diversity in a given set of workers.

3 General Experimental Design Considerations

3.1 External Questions

The tasks that we launched for our experimentation were designed to be compatible with MTurk’s external question system. External questions are tasks that are designed and hosted as web pages outside of the MTurk website. The use of external questions offers freedom in task design and data collection that is not reachable within the confines of developing with MTurk’s internal tools. The external web page is loaded and presented to MTurk workers in an iframe display. When workers view the titles and descriptions of tasks that they may want to contribute to, external question options appear no differently than other HITs.

3.2 Domain of Crowd Work

Much of microtask crowdsourcing work consists of monotonous identification and translation exercises. In contrast, other tasks prompt workers for non-obvious answers. A worker prompted to identify the content of an image likely has obvious answers on the screen in front of them, while a worker prompted for a paraphrase of a sentence must employ their lexical and grammatical knowledge to generate a response. The domain of work in our experiments is **causality**. This domain of microtask work – identifying causes, effects, and other relationship types – is uncommon in crowdsourcing settings, but it resembles other tasks administered for purposes of natural language processing (i.e. generating paraphrases, synonyms/antonyms, etc.). Humans maintain notions of cause and effect in order to logically inform systematic ways of thinking. This faculty can inform human and artificial intelligence researchers alike. In addition to potential applications that are similar to other crowdsourced microtasks, the causality domain invites a range of task designs that help us to investigate several experimental parameters.

4 Experiment 1(a): Sustained Participation

We restate the motivating hypothesis for this experiment:

Hypothesis_{1(a)}: Microtask crowd workers who are enabled to respond more freely in the initial work of a task will complete more optional work on subsequent parts of the task.

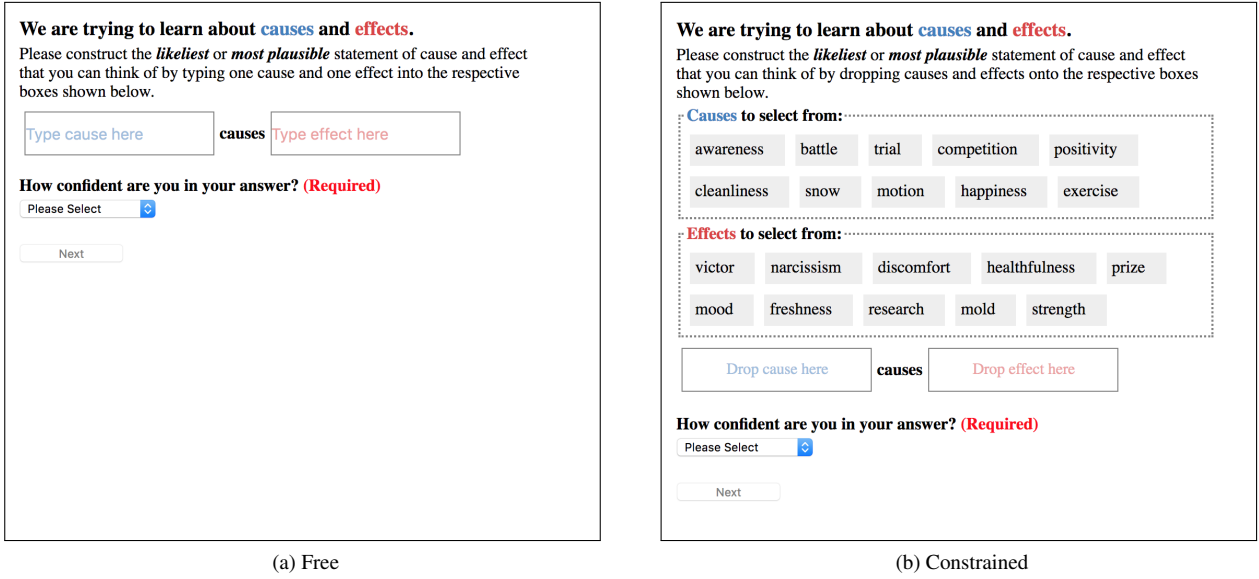


Figure 1: Screenshots of initial subtasks for the a) Free and b) Constrained treatment groups in Experiment 1(a).

4.1 Experimental Design

4.1.1 Overview of HIT Structure and Slide Interface

We first investigated the influence of initial response freedom on sustained participation. In order to measure sustained participation, we included bonus questions in the HIT. The HIT form was structured as a set of slides, with brief fade in/out animation (Figs. 1 and 2). The initial slide included the first subtask (one question). The following slide included a set of mandatory follow-up subtasks (10 questions). After completing the first two slides, workers could submit their work, or opt to complete bonus questions for additional compensation. Workers had access to up to four bonus question slides, featuring subtasks that were identical in form to the second slide’s mandatory subtasks (five questions on each bonus slide).

4.1.2 Slide 1: Initial Subtask/Treatment

Upon accepting to work on the HIT, workers were randomly and blindly assigned to one of two groups. Group assignment only impacted the form of the HIT’s initial slide; the structures of the following slides were consistent for all workers, regardless of group assignment. The beginning of the instructions for the subtask on the initial slide were the same for both groups, and were worded as follows: “We are trying to learn about causes and effects. Please construct the *likeliest* or *most plausible* statement of cause and effect that you can think of. . .”. The instructions then diverged between groups to explain how input was to be performed.

Free Group Workers assigned to one group were presented with two textboxes for input (Figure 1a). The textboxes included placeholder text to clarify the desired input (“Type cause here”, “Type effect here”). Each box allowed textual input of up to four words. The two textboxes were separated horizontally by the word “causes”. Therefore, once input was entered into both textboxes, the resulting causal statement (“[cause input] causes [effect input]”) was presented clearly on

the page. The complete prompt for this group’s initial subtask read as follows: “We are trying to learn about causes and effects. Please construct the *likeliest* or *most plausible* statement of cause and effect that you can think of by typing one cause and one effect into the respective boxes shown below.” We will refer to this group as the **Free (F)** group.

Constrained Group Workers assigned to the other group were presented with a drag-and-drop interface for input (Figure 1b). Instead of having the freedom to input any terms into the cause and effect input boxes, workers in this group were presented with a set of 10 cause term options and a set of 10 effect term options. Workers were prompted to choose one word from each set to form one causal statement. The collection of cause term options was labeled as “Causes to select from”, while the collection of effect term options was labeled as “Effects to select from”. The same options were available for all workers assigned to the group. After being generated via randomized shuffling, the order of the term options in each set was also consistent for all workers in the group. As in the Free group’s initial subtask, the input boxes featured placeholder text to clarify the desired input (“Drop cause here”, “Drop effect here”). Additionally, the user interface only permitted cause and effect terms to be dropped into the appropriate input box. Dropped terms could also be replaced upon the dropping of another eligible term, or through an x-out element that appeared on dropped terms. In line with the presentation of the Free group’s initial subtask, the input boxes were separated horizontally by the word “causes” to result in the presentation of a complete causal statement. The prompt for this group’s initial subtask read as follows: “We are trying to learn about causes and effects. Please construct the *likeliest* or *most plausible* statement of cause and effect that you can think of by dropping causes and effects onto the respective boxes shown below.” We will refer to this group as the **Constrained (C)** group.

We strived to have the initial subtask take the same amount of time for all workers. We considered varying the amount of cause-and-effect relationships that were requested of workers in the Constrained group. However, results from a preliminary experiment informed us that the generation of one relationship took about the same amount of time in each group.

After inputting one cause and one effect, workers in both groups were required to select a numerical confidence level from a dropdown list box. The options in the list box ranged from 1 (“Least Confident”) to 5 (“Most Confident”). The confidence dropdown was included in order to slow down potentially inattentive workers. Since workers are financially compensated for completing MTurk microtasks, cheating can occur. Quality assurance measures, such as confidence checks, can be woven into the structure of tasks.

After the cause, effect, and confidence inputs were entered, a “Next” button was enabled beneath the confidence dropdown. The “Next” button provided access to the next slide of the HIT.

4.1.3 Slide 2: Mandatory Follow-Up Subtasks

The second slide of the HIT featured 10 follow-up questions. These questions all exhibited the same structure. For each, a pair of words was presented. The worker was prompted to identify the type of relationship between the two words in the pair. The questions were presented in a table display (Figure 2a) – each row of the table housed one question. The left column of the table was labeled “Word Pair”. Word pairs were presented in the form “[WORD1] and [WORD2]”. The right column of the table was labeled “Relationship”. In this column, each question included a dropdown field with five

input options. The relationship input options were listed in the dropdown as follows (in top-bottom order): “[WORD1] causes [WORD2]”, “[WORD2] causes [WORD1]”, “Something else causes both”, “[WORD1] and [WORD2] are related in another way”, “[WORD1] and [WORD2] are unrelated”. The set of word pairs displayed on this slide of the HIT was consistent for all workers, but the order of question presentation was randomly generated, as was the order of the two words within each question pair. After a short thank you message for the completion of the initial slide of the HIT, the prompt at the top of this slide of follow-up subtasks read: “The following pairs of terms often appear together in written documents, but we are unsure how they are related. For each pair, please select the best option.”

As on the initial slide, a numerical confidence level input was required for this set of mandatory follow-up subtasks. Upon completion of all 10 questions and the confidence dropdown, two buttons became enabled at the bottom of the page: a “Submit” button and a “Bonus?” button. At the same time, text located above the buttons changed from “*Not yet complete*” to “You can submit your work now OR complete additional bonus pairs if you so choose”. Clicking the “Submit” button allowed the worker to stop working on the HIT, having completed the initial subtask and 10 mandatory follow-up subtasks. Clicking the “Bonus?” button allowed the worker to proceed to a slide of bonus questions. Both actions launched an alert box on the screen for the worker to confirm the decision – either to end their work on the HIT at the conclusion of Slide 2, or to move on to bonus questions.

Another feature of Slide 2 was a dynamic question completion counter. Upon each input option selection for the 10 relationship input dropdowns, a “mandatory” pairs counter was incremented. A “bonus” pairs counter, which remained at zero on this slide was also visible. The counters were located alongside the “Submit” and “Bonus?” buttons in the bottom section of the slide.

4.1.4 Slides 3-6: Optional Bonus Subtasks

Upon electing to proceed to bonus questions, workers were presented with a new slide that was very similar in structure to Slide 2. On this slide, a table of five new word pairs was displayed (Figure 2b). The prompt on this slide included the statement: “You may submit your work to finish at any point. Below is a group of bonus questions.” A new “bonus pairs completed” counter could be seen above the table of subtasks, in addition to the original counter that remained next to the buttons in the bottom section of the slide. The bonus counters incremented with each input option selection for the five relationship dropdowns. A confidence level dropdown was presented below the question table, and was enabled upon completion of one question.











“Submit” and “More Bonus?” buttons were found in the bottom section of the slide. The “Submit” button was initially enabled, allowing the worker to submit their work for the HIT without completing any bonus questions. If the worker completed one or more questions on the slide, the “Submit” button was disabled until the confidence dropdown was completed. The “More Bonus?” button was only enabled if all five questions were completed, along with the confidence dropdown. In this case, the text located above the buttons changed from “*More bonus pairs available upon completion of all pairs above*” to “More bonus pairs available”. As in the previous slide, the clicking of either button launched an alert box on the screen for the worker to confirm their decision.

If workers elected to continue on to more bonus questions, the next slide that appeared was identical to the first bonus slide, except the questions consisted of different word pairs. The same was true for the third and fourth bonus slides. With


We are trying to learn about **causes and **effects**.**

Thanks! Please take a moment to complete these **follow-up questions**.

The following pairs of terms often appear together in written documents, but we are unsure how they are related. For each pair, please select the best option.

Word Pair	Relationship
LOSER and DEFEAT	Please Select 
DRILL and DRESSES	Please Select 
HOLDS and FLESH	Please Select 
YOGURT and FROZEN	Please Select 
SHIVERING and COLD	Please Select 
SNAPSHOT and PERIODS	Please Select 
OLDEST and WIVES	Please Select 
HOMELAND and PUSHING	Please Select 
STRONGLY and ILLNESS	Please Select 
PRIDE and ESTEEM	Please Select 

How confident are you in these answers? **(Required)**

Please Select 

Not yet complete

Submit **OR** Bonus? **0/10** mandatory + **0** bonus pairs completed






(a) Slide 2: Mandatory follow-up subtasks slide

We are trying to learn about **causes and **effects**.**


Thanks! You may **submit your work to finish at any point**. Below is a group of **bonus questions**.

The following pairs of terms often appear together in written documents, but we are unsure how they are related. For each pair, please select the best option.

0 bonus pairs completed

Word Pair	Relationship
AUTHOR and EDITOR	Please Select 
RITUAL and PRAYER	Please Select 
LAUGHTER and SMILES	Please Select 
DECAY and GROWTH	Please Select 
MUSCLES and POWER	Please Select 

How confident are you in these answers?

Please Select 

More bonus pairs available upon completion of all pairs above

Submit **OR** More Bonus? **10/10** mandatory + **0** bonus pairs completed






(b) Slide 3: First bonus subtasks slide

We are trying to learn about **causes and **effects**.**


Thanks! You may **submit your work to finish at any point**. Below is a group of **bonus questions**.

The following pairs of terms often appear together in written documents, but we are unsure how they are related. For each pair, please select the best option.

15 bonus pairs completed

Word Pair	Relationship
HEALTH and WELLNESS	Please Select 
MAJOR and GARBAGE	Please Select 
SWIMMER and WATER	Please Select 
VICTORY and WINNER	Please Select 
SUCCEED and ACCOMPLISH	Please Select 

How confident are you in these answers?

Please Select 

Submission

Submit **10/10** mandatory + **15** bonus pairs completed

(c) Slide 6: Final bonus subtasks slide

Figure 2: *Screenshots* of mandatory and bonus follow-up subtask slides in Experiment 1(a).

five questions for each of four bonus slides, workers were able to complete up to 20 bonus questions total. The final slide of bonus questions differed slightly from the first three. No more bonus questions were available after the final bonus slide, so only one button – the “Submit” button – appeared at the bottom of the slide (Figure 2c).

4.1.5 Sources of Constrained Input Term Options and Follow-Up Subtask Word Pairs

Constrained Group Initial Subtask The terms presented as input options in the Constrained group’s initial subtask were selected based on responses from previously executed crowdsourcing experiments in the domain of causality. In the experiments, workers identified the relationship between pairs of words (some of which exhibited a cause and effect relationship), and were given the opportunity to propose a new effect term for a given cause. We sampled five cause terms and five effect terms that were part of the pairs most often identified as having a causal relationship. In order to complete the sets of 10 cause options and 10 effect options in the Constrained group’s initial subtask, we also sampled five of the cause terms and effect terms that were part of the pairs more seldom identified as exhibiting a causal relationship.

Mandatory and Bonus Follow-Up Subtask Word Pairs Together, Slides 2-6 contain a total of 30 word pairs for the 10 mandatory and 20 bonus questions. 10 of the pairs were sampled from the responses of the previous causality experiments. This sample consisted of the pairs most often identified as exhibiting a causal relationship, excluding pairs that contained a word already used in the Constrained group’s initial subtask. Another 10 of the pairs were sampled from a database² of word pairs that was generated by psychology researchers, via experimentally administered free association exercises [32]. We consider the words in each of these pairs to be related in some way. The remaining 10 pairs were randomly constructed from a sampled set of common English words³, as determined by an analysis of Google’s Trillion Word Corpus. This sample was limited to words of 5-8 characters in length. The 30 word pairs were randomly assigned to Slides 2-6 of the HIT - 10 for Slide 2 and five for each of Slides 3-6. The slide assignment for each word pair was consistent for all workers in the experiment, although word pair ordering within a given slide was randomized for each worker.

4.1.6 Data Collection

For each worker, we recorded input responses for all subtasks completed. In addition to response input, we recorded: the number of bonus questions completed (our measurable quantity of sustained participation), the total time spent on the HIT, the time spent on each slide of the HIT, the order of the questions presented on Slides 2-6, the order of the words presented in each word pair, and the number of blurs and focuses on the web browser tab.

4.1.7 HIT Presentation and Payment Scheme

While browsing HITs on the MTurk marketplace, workers can view titles, descriptions, potential rewards, and previews. Our HIT was posted with the following title: “Identifying statements of cause and effect”. The description was listed as: “Create and identify short written statements relating causes and effects”. Initially, a potential reward of \$1.50 was listed for completing our HIT. \$1.50 represented the financial compensation offered for the completion of the mandatory work

²Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. <http://www.usf.edu/FreeAssociation>

³<https://github.com/first20hours/google-10000-english>

in the HIT (Slides 1-2: the initial subtask, and the 10 mandatory follow-up subtasks). However, the preview of the HIT offered a complete explanation of our payment scheme.

The compensation of \$1.50 was comprised of \$1.00 for the initial subtask and \$0.50 for the 10 mandatory follow-up subtasks. Each follow-up subtask was designated a value of \$0.05. As such, workers earned \$0.05 for each completed bonus subtask. Since workers could complete up to 20 bonus subtasks – four slides of five subtasks each – they could earn up to \$1.00 for completing all bonus subtasks. Workers who completed all of the work in the HIT earned a total of \$2.50 (\$1.00 for the initial subtask, \$0.50 for the mandatory follow-up subtasks, and \$1.00 for all bonus subtasks).

The preview of the HIT also reiterated the description of the work, and concluded with the following message: “Please note: Your submission may not be approved if you answer too quickly and/or other workers disagree with your responses.”

Workers could only accept to work on the HIT once. In order to qualify to work on the HIT, workers had to have: 1) had at least 90% of their previous MTurk responses approved by crowdsourcers, 2) been eligible to view adult content, and 3) been located in the United States.

4.2 Results

4.2.1 Crowdsourcing Metrics

The HIT was made available to MTurk workers on December 22 and December 23 of 2017. 503 workers submitted responses to the HIT. The Free group consisted of 250 workers, and the Constrained group was made up of the other 253 workers. Payment for completed bonus work was distributed on December 26, 2017.

4.2.2 Time Spent

Workers in the Free group spent more time on the initial subtask (Slide 1) than did workers in the Constrained group (Median_F: 76.56 seconds; Median_C: 63.01 seconds; Mann-Whitney: $U = 28477.5$, $p < 0.05$). Free group workers also spent more total time completing the HIT than did Constrained group workers (Median_F: 321.94 seconds; Median_C: 305.10 seconds; Mann-Whitney: $U = 28899.0$, $p < 0.05$). The timing distributions for each of Slides 2-6 were not significantly different between groups.

4.2.3 Number of Bonus Subtasks Completed

On average, workers completed 11.83 of the 20 bonus subtasks. Table 1 shows the experimental probabilities of workers completing 0, 5, 10, 15, or 20 bonus subtasks. 42.5% of workers completed all 20 bonus subtasks, a higher rate than for any other quantity of bonus work. The second most common behavior was for workers to submit the HIT having completed zero bonus subtasks. 23.1% of workers submitted the HIT after completing only the mandatory subtasks on Slides 1 and 2 of the HIT. Almost all workers who did not complete zero or 20 bonus subtasks completed either 5, 10, or 15. Only two workers, both randomly assigned to the Free group, completed a number of bonus subtasks that were not multiples of five (14, 19). The segmented slide structure of the bonus subtasks influenced workers to perform work in increments that corresponded to the number of questions presented on each slide. Even though workers could submit

Group	Number of Bonus Subtasks Completed				
	0	5	10	15	20
Free	.216	.096	.168	.096	.416
Constrained	.245	.126	.123	.071	.435
Combined	.231	.111	.145	.083	.425

Table 1: *Probability of Number of Bonus Subtasks Completed*. Experimental probabilities of workers completing 0, 5, 10, 15, or 20 bonus subtasks in Experiment 1(a). Bold indicates values for which the corresponding treatment group exhibited a higher probability than the other group.

their work at any point while working on the bonus slides, 99.6% of workers submitted the assignment upon completion of all questions on a slide.

Workers in the Free group completed slightly more bonus subtasks (Mean_F: 12.05; Median_F: 15) than did workers in the Constrained group (Mean_C: 11.62; Median_C: 15). Constrained group workers were 2.9% more likely than Free group workers to forego bonus subtasks altogether. Workers in the Constrained group were also more likely to complete 5 or, interestingly, all 20 bonus subtasks, by margins of 3.0% and 1.9%, respectively. Free group workers exhibited higher probabilities of completing 10 or 15 bonus subtasks, by margins of 4.5% and 2.5%, respectively. None of these differences in probability were dramatic, as none exceeded 4.5%. Moreover, a consistent trend concerning which treatment group completed greater increments of bonus subtasks was not revealed.

Figure 3 depicts the distributions of the number of bonus subtasks completed per worker in each group. Standard statistical tests reveal no significant difference between the distributions (Mean_F: 12.05; Mean_C: 11.62; T: $t = 0.589$, $p = 0.556$; Cohen's d : 0.053) (Median_F: 15; Median_C: 15; Mann-Whitney: $U = 31029.0$, $p = 0.351$). A test more appropriate for discrete distributions, a one-sided statistical test of Poisson rates - with 1) the hypothesis that Free group workers complete more bonus subtasks than Constrained group workers, and 2) the treatment of each bonus subtask as an individual trial - also does not yield statistical significance ($p = 0.080$). Although workers in the Free group completed slightly more bonus subtasks than did workers in the Constrained group, the effect size is negligible and the difference is not significant.

4.3 Discussion

No consistent relationship was revealed between the initial subtask treatment and resulting levels of sustained participation. Workers in the Free group completed slightly more optional bonus subtasks than workers in the Constrained group, but not by a significant margin. The results suggest that at most only a weak relationship exists between initial response freedom and subsequent motivation as a result of priming effects. However, this should not be taken as a general conclusion, but rather holds for the specific combination of priming treatments and task design studied here.

4.3.1 Nuance of Priming Treatment

There are many factors that could have contributed to the null result. It is possible that the experimental treatment was too nuanced to reveal any emergent behavior. Perhaps granting complete response freedom (textboxes) with respect to a single prompt, as was done in the Free group's initial subtask, does not provide enough volume to act as an effective priming stimulus. On the other hand, perhaps the number of response options offered in the Constrained group's initial subtask was too great to introduce a level of response freedom that sufficiently contrasted that of the Free group's initial

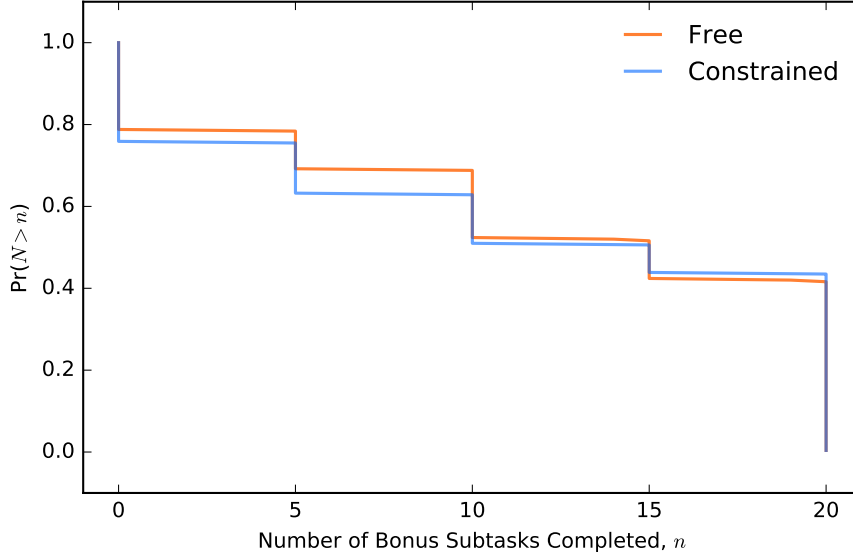


Figure 3: *CCDFs of Number of Bonus Subtasks Completed.* The complementary cumulative distribution functions (CCDFs) of the number of bonus subtasks completed per worker in Experiment 1(a) illustrate that there are no drastic differences between the treatment groups. Workers in the Free group completed slightly more bonus subtasks than did workers in the Constrained group, but the difference is not statistically significant.

subtask. A certain threshold of constraint may be necessary to hinder sustained participation in microtask work as a result of priming. A study involving many treatment groups that exhibit different levels of initial response constraint may reveal the existence of such a threshold.

4.3.2 Individualized Reactions to a Common Priming Treatment

Another potential explanation for the null result is that, within each treatment group, workers' personal reactions to the initial subtask prime may have opposed one another. The same prime - either the initial subtask for the Free group or that of the Constrained group - may have led to different priming effects for different workers within each treatment group. Based on underlying attitudes towards choice, some workers may consciously or subconsciously prefer - and be motivated by - greater levels of response freedom, while others may actually prefer lower levels of response freedom. This is a case of the same prime leading to different and opposing effects [33]. Greater degrees of choice are often considered advantageous in modern society, but freedom can also result in hesitation and increased pressure to make the best choice [34]. The latter reaction to freedom can hinder action and motivation. The overall result of both primes leading to opposite effects within each treatment group would have produced a cancelling-out effect. If the subsets of workers with contradicting reactions to response freedom were similarly represented in each treatment group, the null result would have been revealed despite any impact that the treatment had on the motivation of individual workers.

4.3.3 Future Work

The subtlety involved in implementing useful priming procedures requires the testing of many experimental parameters and the risk of insignificant results. However, each iteration of priming studies reveals new avenues for investigation that may lead to the discovery of high-impact priming procedures in real-world crowdsourcing. The degree of nuance in

the design of Experiment 1(a) offers many possibilities for future work. As discussed, a study involving many treatment groups that are assigned varying levels of response freedom may reveal the existence of a threshold that results in significant priming effects on sustained participation. The challenge lies in determining what amounts to a particular level of freedom or constraint. Similar priming treatments may be studied in other domains of crowd work as well.

Future studies may also seek to investigate the notion of the same prime leading to different - and, perhaps, counteracting - effects. Methods of pre-screening workers can be explored to assign priming treatment based on individual predispositions to the priming agent. Pre-screening should be done in a way that does not reveal the presence of a priming treatment. Also, in order for a combination of pre-screening and priming techniques to be deemed useful in microtask design, output returns must outweigh the costs of implementation.

5 Experiment 1(b): Response Quality

We restate the motivating hypothesis for this experiment:

Hypothesis_{1(b)}: Microtask crowd workers who are enabled to respond more freely in the initial work of a task will produce higher quality responses in subsequent parts of the task.

5.1 Experimental Design

5.1.1 Overview of HIT Structure and Slide Interface

Following the execution of Experiment 1(a), we investigated the influence of initial response freedom on subsequent response *quality*. The structure of the HIT for Experiment 1(b) resembled that of the HIT for Experiment 1(a), except for a few important modifications. No bonus work was made available. All work in the HIT was mandatory for results to be collected and the reward paid. Also, the follow-up mandatory subtask set consisted of 15 questions instead of 10. Finally, a third form of initial subtask – and hence, a third experimental group – was studied. This group will be referred to as the **Baseline (B)** group.








5.1.2 Slide 1: Initial Subtask/Treatment

Upon accepting to work on the HIT, workers were randomly and blindly assigned to one of the three groups: Free, Constrained, or Baseline. As in Experiment 1(a), group assignment only impacted the structure of the HIT's initial slide. Initial subtasks for the Free and Constrained groups were identical to the corresponding initial subtasks in Experiment 1(a). However, the order of cause terms and effect terms presented as input options in the Constrained group's initial subtask was randomized for each worker in the Constrained group.


Baseline Group Instead of initially being prompted to generate a cause and effect relationship, as were workers in the Free and Constrained groups, workers assigned to the Baseline group were presented with a table of seven relationship identification subtasks (Figure 4). The relationship identification questions comprising this initial subtask were constructed with the same structure as the follow-up subtasks in Experiment 1(a). Timing data from Experiment 1(a) informed the

We are trying to learn about causes and effects.

The following pairs of terms often appear together in written documents, but we are unsure how they are related. For each pair, please select the best option.

Word Pair	Relationship
WELLNESS and HEALTH	Please Select 
SHOWER and WETNESS	Please Select 
HOLDS and FLESH	Please Select 
SATISFACTION and FULFILLMENT	Please Select 
CHART and GRAPH	Please Select 
AUTHOR and EDITOR	Please Select 
TIGER and REFORMS	Please Select 

How confident are you in your answer? (Required)

Please Select 

Next

Figure 4: Screenshot of initial subtask for Baseline treatment group in Experiment 1(b).

decision to include exactly seven follow-up questions on the slide. We aimed to have the initial Baseline subtask require the same amount of time as the initial subtasks of the other two groups. The purpose of including the Baseline group in Experiment 1(b) was to consider the possibility that an initial set of questions – with a structure identical to that of a subsequent set of questions – could have a more favorable effect on subsequent response quality than the other treatments. If learning effects produced by task consistency lead to better results than either priming treatment, there would be no reason to implement the more invasive priming procedures. The prompt for the Baseline group’s initial subtask read as follows: “We are trying to learn about causes and effects. The following pairs of terms often appear together in written documents, but we are unsure how they are related. For each pair, please select the best option.” A confidence level dropdown and “Next” button, with the same functionality as those included in the initial subtasks of the other treatment groups, were included in the initial Baseline subtask.

5.1.3 Slide 2: Mandatory Follow-Up Subtasks

The second slide of the HIT was very similar to the second slide of the HIT in Experiment 1(a). However, 15 questions were included in the follow-up subtasks table instead of 10. As in Experiment 1(a), the set of word pairs displayed on Slide 2 of the HIT was consistent for all workers, but the order of question presentation was randomly generated, as was the order of the words within each question pair. The set of 15 word pairs was constructed as follows: five pairs were *Causal* pairs, sampled from the responses of the previous causality experiments; five pairs were *Related* pairs, sampled from the word association set generated by human psychology researchers; and five pairs were *Random* pairs, randomly constructed from the set of common English words. Since this slide was the final slide of the HIT, only a “Submit” button was displayed in the bottom section of the page. The “Submit” button became enabled upon completion of all 15 questions and the confidence dropdown. When the button was enabled, text displayed above the button changed from “*Not yet complete*” to “You can now submit your work”. An alert box prompted workers to confirm their submission.

5.1.4 Data Collection

For each worker, we recorded: the content of input responses, the total time spent on the HIT, the time spent on each slide of the HIT, the order of the questions presented on Slide 2, the order of the words presented within each word pair on Slide 2, and the number of blurs and focuses on the web browser tab. For Constrained group workers, we also recorded the order of input option terms presented on the initial subtask slide. For Baseline group workers, the order of the seven word pairs presented on the initial subtask slide, and the order of the two words within each pair, were recorded.

5.1.5 HIT Presentation and Payment Scheme

The HIT was posted with the following title: “Identifying statements of cause and effect”. The description was listed as: “Identify short written statements relating causes and effects”. The financial reward for completing the HIT was \$0.50, calculated based on timing results from Experiment 1(a) and a wage of \$10/hour. The preview available for the HIT reiterated the description of the HIT and noted the two-slide structure of the work. The preview concluded with the message: “Please note: Your submission may not be approved if you answer too quickly and/or other workers disagree with your responses.”

Workers could only accept to work on the HIT once. In order to qualify to work on the HIT, workers had to have: 1) never accepted to work on the HIT for Experiment 1(a), 2) had at least 90% of their previous MTurk responses approved by crowdsourcers, 3) been eligible to view adult content, and 4) been located in the United States.

5.1.6 Measuring Response Quality

For the word pairs - whose relationships were to be categorized by all workers in Slide 2 of the HIT - there are no ground truth measures of relationship type. There is no obvious way to measure the "correctness" of the relationship identification subtask responses because the task domain is subjective. Relationship recognition is informed by experience and prevailing associations. To administer a crowdsourcing task for which a simple measure of response correctness or quality does not exist, it is common for crowdsourcers to launch a follow-up crowdsourcing task for new workers to subjectively evaluate the quality of responses to the original task. We forewent this validation step because the three original sources of the word pairs comprising the questions on Slide 2 each somewhat inherently correspond to different relationship types. The word pairs that were sourced from previous experiments in the realm of causation are considered to exhibit a causal relationship. In our analysis, responses of “[WORD1] causes [WORD2]” and “[WORD2] causes [WORD1]” are considered to be responses of the “best” quality for these *Causal* pairs. These response options are defined as *Causal*-type responses. Word pairs that were sourced from the word association set are considered to exhibit a related relationship - the two words are related in some way other than causally. Responses of “Something else causes both” and “[WORD1] and [WORD2] are related in another way” are considered the “best” responses for these *Related* pairs. These response options are defined as *Related*-type responses. Finally, word pairs that were randomly constructed from the set of common English words are considered to exhibit a random relationship. The response of “[WORD1] and [WORD2] are unrelated” is considered “best” for these *Random* pairs, and is defined as a *Random*-type response. Defining the types of word pairs, and hence the quality of relationship identification responses, by way of word pair source allowed us to save

resources by not employing an additional set of workers to grade responses. Manual examination of assigned types did not reveal any glaring issues with the classification of word pairs. However, as noted, classification of the relationship between two words is a task that escapes the realm of objective judgment. Nevertheless, discussions of quality and performance regarding the crowd work of this experiment are based on the standards described in this section.

5.2 Results

5.2.1 Crowdsourcing Metrics

The HIT was made available to MTurk workers on January 24 and January 25 of 2018. 628 workers submitted responses to the HIT. 211 workers were assigned to the Baseline group. Another 211 workers were assigned to the Constrained group. The Free group consisted of 206 workers.

5.2.2 Bonferroni Correction

This experiment featured three treatment conditions. In order to account for multiple comparison effects when hypothesis testing, we apply a Bonferroni correction to the reported level of significance in pairwise comparisons. For three groups, three comparisons are necessary to account for all two-group combinations (Baseline, Free; Baseline, Constrained; Free, Constrained). In the reporting of hypothesis testing results in this section, the level of significance is $0.05/3 = 0.017$.

5.2.3 Time Spent

Baseline workers were slowest in completing their initial subtask. Workers in the Baseline group spent significantly more time on Slide 1 of the HIT than did workers in the Free group (Median_B: 79.64 seconds; Median_F: 53.98 seconds; Mann-Whitney: $U = 14111.0$, $p < 10^{-10}$) and Constrained group (Median_B: 79.64 seconds; Median_C: 55.59 seconds; Mann-Whitney: $U = 13969.0$, $p < 10^{-11}$). Unlike in Experiment 1(a), Free group and Constrained group workers did not exhibit a significant difference in time spent on the initial subtask (Median_F: 53.98 seconds; Median_C: 55.59 seconds; Mann-Whitney: $U = 21554.0$, $p = 0.442$).

Timing trends were the reverse on Slide 2 of the HIT. Workers in the Baseline group spent less time completing the follow-up subtasks than did workers in the Free group (Median_B: 104.67 seconds; Median_F: 116.94 seconds; Mann-Whitney: $U = 17813.0$, $p < 10^{-3}$) and Constrained group (Median_B: 104.67 seconds; Median_C: 116.72 seconds; Mann-Whitney: $U = 19073.5$, $p < 0.01$). As in Slide 1, Free group workers and Constrained group workers did not exhibit a significant difference in time spent (Median_F: 116.94 seconds; Median_C: 116.72 seconds; Mann-Whitney: $U = 20790.0$, $p = 0.222$).

Regarding total time spent on the HIT, the relative speed of Baseline group workers in completing the follow-up subtasks balanced out their slower performance on the initial subtask. No significant difference existed between groups for total time spent on the HIT (Median_B: 199.05 seconds; Median_F: 192.70; Median_C: 187.66; Kruskal-Wallis: $H = 2.770$, $p = 0.250$). The follow-up subtasks had exactly the same structure as the questions in the initial Baseline subtask. This consistency may have resulted in Baseline group workers bypassing Slide 2 instructions. Increased familiarity with the question structure may have also introduced learning effects that led to increased pace. While Baseline workers may

have needed more time than other workers on Slide 1 to process instructions and complete the work, learning effects likely allowed them to complete the follow-up questions more quickly.

5.2.4 Probability of "Best" Response Type

In analyzing the quality of responses, we limit our scope to responses to the follow-up subtasks presented on Slide 2 of the HIT, which were completed subsequent to the initial subtask. For the set of all word pairs (Random, Causal, and Related), Baseline group workers demonstrated the highest probability, 0.798, of choosing responses of the “best” type. Free group workers chose “best”-type responses with a probability of 0.771, and Constrained workers with a probability of 0.764. The tendency of Baseline workers to select responses of the "best" type was significantly different than that of both Free group workers ($\text{Mean}_B: 0.798; \text{Mean}_F: 0.771; T: t = 2.485, p = 0.013; \text{Cohen's } d: 0.244$) and Constrained group workers ($\text{Mean}_B: 0.798; \text{Mean}_C: 0.764; T: t = 2.694, p = 0.007; \text{Cohen's } d: 0.262$). The distributions of "best" response type probabilities for workers in the Free and Constrained groups were not significantly different ($\text{Mean}_F: 0.771; \text{Mean}_C: 0.764; T: t = 0.517, p = 0.605; \text{Cohen's } d: 0.051$). Figure 5 shows a segmented heat map display of response probabilities for each group, breaking down the results by word pair type (Random, Causal, Related) and response type (Random, Causal, Related).

5.2.5 Response Type Frequency by Word Pair Type

Table 2 displays the mean frequency of each response type, segmented by word pair type and treatment group. Baseline group workers were more likely than Free group workers ($\text{Mean}_B: 4.63; \text{Mean}_F: 4.24; T: t = 3.853, p < 10^{-3}; \text{Cohen's } d: 0.378$) and Constrained group workers ($\text{Mean}_B: 4.63; \text{Mean}_C: 4.16; T: t = 4.427, p < 10^{-5}; \text{Cohen's } d: 0.431$) to select “best” responses for the five Random word pairs. For the five Causal pairs, however, none of the treatment groups definitively outperformed the other two. Baseline workers did significantly better than Free group workers on Causal pairs ($\text{Mean}_B: 4.27; \text{Mean}_F: 3.95; T: t = 3.179, p < 0.01; \text{Cohen's } d: 0.312$), but Constrained group workers did not significantly differ from either. Free group workers more commonly selected “best” responses to the five Related pairs than Baseline group workers ($\text{Mean}_B: 3.07; \text{Mean}_F: 3.38; T: t = -2.548, p < 0.017; \text{Cohen's } d: -0.250$) by a significant margin. The Constrained group, again, did not demonstrate significant difference from either the Baseline group or Free group.

T-test results indicate significant difference in response type frequency between Free group workers and Constrained group workers in only one of the nine combinations of word pair type and response type. Free group workers were significantly less likely than Constrained group workers to select responses of the Causal type for Random word pairs ($\text{Mean}_F: 0.09; \text{Mean}_C: 0.28; T: t = -2.959, p < 0.01; \text{Cohen's } d: -0.289$). There was no significant difference in response type frequency between Free group workers and Constrained group workers for the other eight combinations of word pair type and response type, including the three combinations that corresponded to "best" quality responses. This serves as confirmation that differences in the initial subtask structures of the Free and Constrained groups did not have an impact on subsequent response quality.

For both Causal and Related word pairs, workers in the Free group were more likely than workers in the other two groups to select Related responses. For these pair categories, Free group workers were less likely to select Causal responses than were workers in the other two groups. Note that the differences between the Free and Constrained groups were not

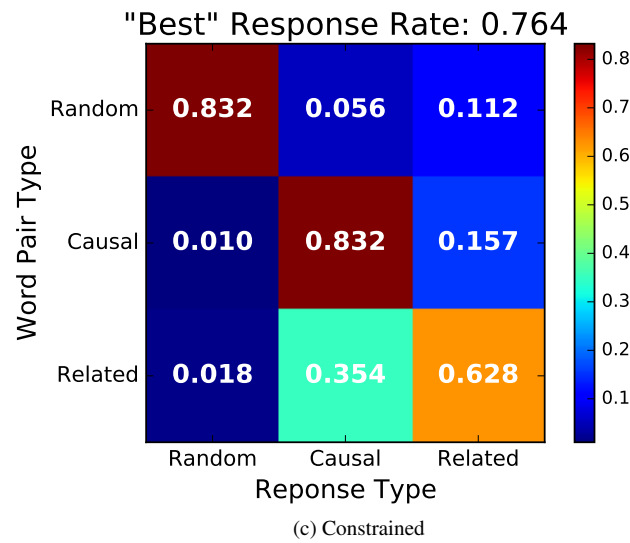
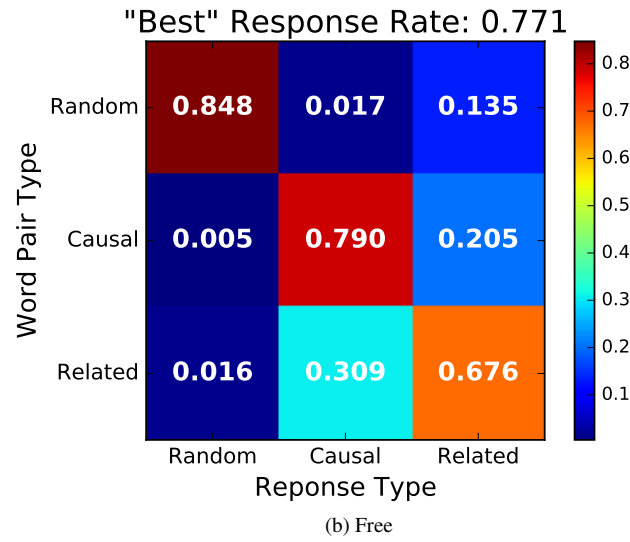
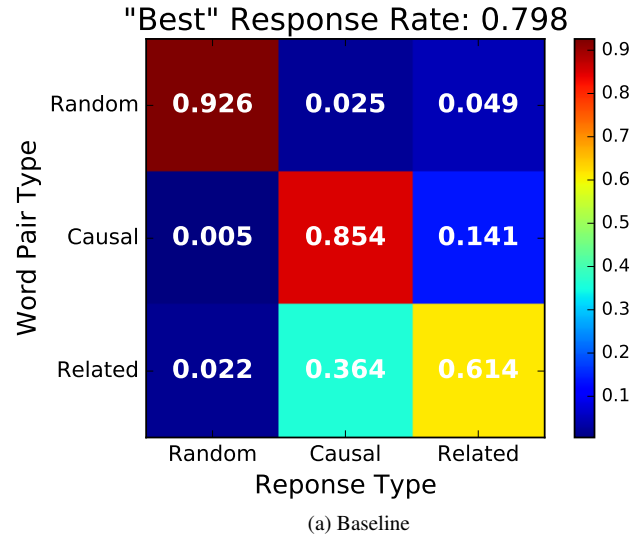


Figure 5: *Response Probability Heat Maps*. Experimental probability values of a worker selecting a Random, Causal, or Related response type, given word pair type (Random, Causal, or Related). A heat map matrix is presented for each treatment group. "Best" response type probabilities are located on the top-left to bottom-right diagonal of each matrix. The "*Best*" Response Rate reported above each matrix is the probability of a worker responding with the "best" response type for all word pairs.

Response Type	Word Pair Type								
	Random			Causal			Related		
	B	F	C	B	F	C	B	F	C
Random	4.63 ^a	4.24 ^b	4.16 ^b	0.02	0.02	0.05	0.11	0.08	0.09
Causal	0.12 ^{ab}	0.09 ^b	0.28 ^a	4.27 ^a	3.95 ^b	4.16 ^{ab}	1.82	1.54	1.77
Related	0.25 ^b	0.67 ^a	0.56 ^a	0.71 ^b	1.02 ^a	0.79 ^{ab}	3.07 ^b	3.38 ^a	3.14 ^{ab}

Table 2: *Mean Response Type Frequency By Word Pair Type*. Response type frequency represents the number of responses of one type (Random, Causal, or Related) that were input by a worker. For each combination of word pair type and response type, the table presents mean response type frequency values for each group of workers. The follow-up subtasks included five word pairs of each word pair type, so the theoretical frequency range is [0, 5] for each combination of word pair type and response type. Regions with a light gray background signify that the response type is considered the "best" option for the given word pair type (Sec. 5.1.6). Pairwise t-tests were conducted to investigate differences in the distributions of response type frequency between treatment groups. Within each combination of word pair type and response type, values with shared superscripts (*a* or *b*) indicate that there is no significant difference between the corresponding treatment groups. Values with opposing superscripts exhibit significant difference between treatment groups, for the given combination of word pair type and response type. For example, for Causal word pairs, the Baseline group (superscript: *a*) responded with significantly more Causal responses than did the Free group (superscript: *b*). However, neither of these groups exhibited significant difference with the Constrained group (superscript: *ab*) regarding Causal response frequency for Causal pairs.

significant in these cases. However, these tendencies suggest that the initial subtask for Free group workers led workers to be more strict than they otherwise would have been in labeling word pairs as Causal. For Random word pairs, both Free group and Constrained group workers were more likely than Baseline group workers to select Related responses. Workers in these groups may have been primed by their initial subtasks to stretch their notions of relatedness more often than workers in the Baseline group.

Although the treatment of each group may have influenced word pair categorization tendencies in some cases, none of the three groups significantly outperformed the other two groups in providing the "best" responses for each word pair type.

5.2.6 Small Sample Size Simulations

One reason why crowdsourcers utilize microtask platforms is to minimize the costs of data collection. Crowdsourcers often strive to minimize financial costs when structuring tasks and formulating task distribution procedures. One way to minimize cost is to employ only a small number of workers.

To simulate small numbers of workers in our HIT results, we conducted bootstrapping procedures for small numbers of responses. For each word pair type, Figure 6 shows the rates at which simulated small response samples (1 response - 10 responses) contain a majority response type that is considered the "best" response type for the word pair type. The trends align with the differences reported in response type frequency. Small samples of Random word pair responses indicate better performance by Baseline group workers. The performance of Free group workers, relative to both of the other two groups, was worse for Causal word pairs but better for Related word pairs. As is shown in Figure 7, when considering the set of all word pairs, small response samples do not reveal a definitive performance advantage for any treatment group.

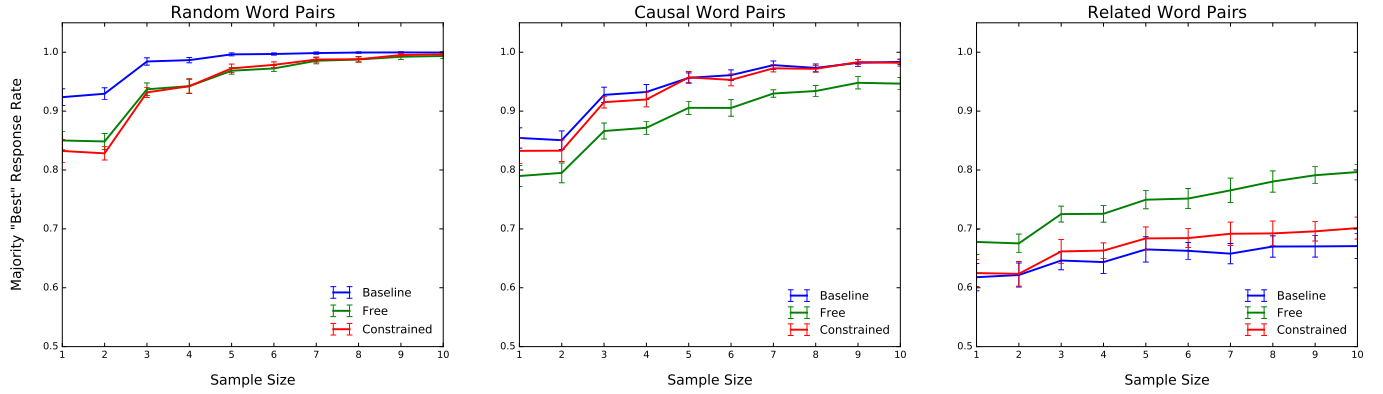


Figure 6: *Majority "Best" Response Rate vs. Bootstrapped Sample Size by Word Pair Type.* For simulated sample sizes ranging from 1 response to 10 responses, the set of responses for each word pair type was sampled (with replacement) 100 times. For each sampling, the mode response type was compared to the "best" response type for the given word pair type. We refer to the rate at which the mode response type matched the "best" response type as the *Majority "Best" Response Rate*. The plots report the mean Majority "Best" Response Rate for each treatment group, for 20 bootstrapping simulations. Error bars indicate one standard deviation.

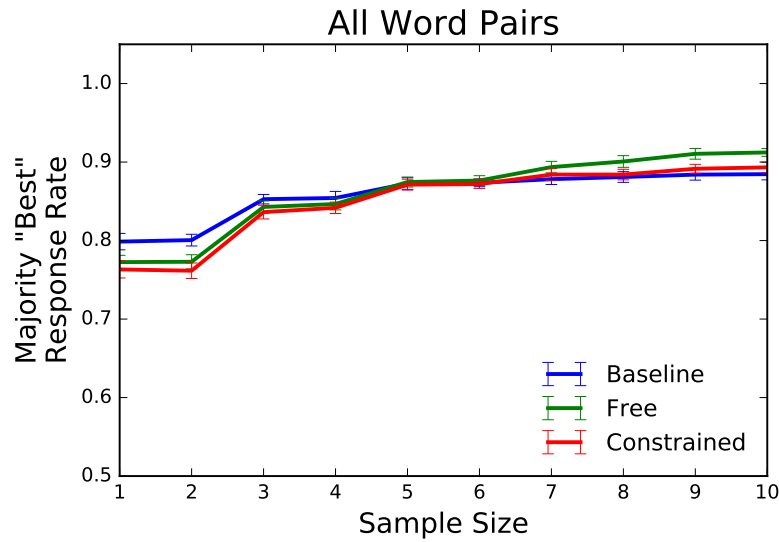


Figure 7: *Majority "Best" Response Rate vs. Bootstrapped Sample Size for All Word Pairs.* No consistent performance advantage among treatment groups is revealed when considering responses to all word pairs, in small sample bootstrapping simulations.

5.3 Discussion

The results of Experiment 1(b) suggest that response quality is not improved through priming for different levels of initial response freedom. Workers in the Free group demonstrated no significant performance gain over workers in the Constrained group. Our expectation that greater levels of initial response freedom would prime for better quality in subsequent responses was not supported by the results of the experiment. In fact, workers in the Baseline treatment group most frequently offered the best response quality. Baseline group workers also completed the follow-up subtasks faster than workers who received either of the other two treatments. The performance of the Baseline group may be explained by previous research conclusions on the benefits of task consistency regarding temporal efficiency and response quality [14, 19]. Learning effects, as a result of the consistency of task instruction and structure, likely played a role in fine-tuning the relationship identification performance of Baseline workers. Although the difference in overall “best” response rate for the Baseline group was statistically significant, effect sizes were small. Also, a finer-grained analysis of responses, segmented by word pair type, revealed that workers in the Baseline group did not perform significantly better than workers in the other two groups for all question types.

5.3.1 Free Group Response Bias

Free group workers performed better than Baseline group workers on Related word pairs. Furthermore, Free group workers were more likely than Baseline workers to select responses of the Related type for all word pair types. For Causal word pairs, Free group workers were less likely than Baseline workers to select responses of the Causal type. A potential explanation for this behavior is that the initial subtask for the Free group may have primed workers to maintain higher standards of causality. Prompting Free group workers to initially generate a cause-effect pair without any guidance or example terms may have stimulated the workers’ most immediate and prominent associations with causality. Subsequently, workers may have compared the relationships of presented word pairs to the relationships of their most recently active associations of cause and effect. On the other hand, workers in the Baseline and Constrained groups were initially presented with potential cause and effect terms, which may have prompted more lenient identification of cause-effect relationships.

5.3.2 Future Work

In the study of inter-task priming effects in crowdsourcing environments, it is important to keep in mind that consistency in structure, instruction, and design can have significant effects on the quality of responses. While an infinite amount of priming treatments can be studied and compared, priming procedures only offer significant benefits if they not only outperform others, but also outperform potentially overlooked and beneficial aspects of standard task design. Designing tasks to include repetitive structures, instead of targeting particular priming mechanisms through task structure, may be a strong force in optimizing the quantity and quality of crowd output. More subtle forms of priming treatment - embedded in repetitive task structures - may produce favorable effects without compromising the benefits of contextual consistency. One such form of treatment is addressed in Experiment 2.

6 Experiment 2: Response Diversity

We restate the motivating hypothesis for this experiment:

Hypothesis₂: Groups of microtask crowd workers who are exposed to more diverse sets of preliminary question prompts will offer greater diversity in responses to subsequent work.

6.1 Experimental Design

6.1.1 Overview of HIT Structure

The HIT for Experiment 2 consisted of a single page (Figure 8) - no fade-in/fade-out slide structure was involved. The questions posed on the HIT were causal statement completion exercises. For a given cause term, workers were asked to provide a corresponding effect term. For each question, an incomplete causal statement was presented in the following form: "[CAUSE TERM] is/are a cause of:". Horizontally adjacent to each incomplete statement was a textbox for worker input. The textbox allowed textual input of up to four words. A set of five preliminary questions was the subject of experimental treatment conditions. A sixth question - the *test* question - exhibited the same structure as those in the preliminary set, but the cause term was blocked from view until responses had been entered for all five preliminary questions.

The initial heading at the top of the HIT read: "We are trying to learn about effects of different causes". Following this initial text, two complete example statements were shown. Directly preceding the set of questions was the prompt: "Please answer the following as best as you can:".

When responses were entered for all six questions, a numerical confidence level dropdown was enabled. The "Submit" button became enabled once responses were entered for all causal statement completion exercises and the confidence level dropdown. Confirmation of submission was not enforced.

6.1.2 Treatment: Content Overlap in Preliminary Question Sets

Experimental treatment was applied via the distribution of cause terms included in the preliminary causal statement completion exercises. Specifically, we varied the degree in which common cause terms were presented to multiple workers in a group. Each worker was assigned to one of five experimental groups. It was predetermined that 100 workers would be recruited for each group. Each group was delegated a different level of overlap in the cause terms allocated to different workers. We refer to each group based on its corresponding *Cause Term Exposure Count (CTEC)*, the number of workers who were exposed to each cause term allocated to the group. In the group exhibiting the greatest level of overlap, each worker in the group completed causal statements given the same five cause terms in a randomized order. In other words, all 100 workers in the group were exposed to the five cause terms assigned to the group (CTEC = 100 workers). In the treatment group exhibiting the opposite extreme, each worker in the group completed causal statements for five cause terms that were all unique (within the scope of the group) to the worker (CTEC = 1 worker). In total, 500 cause terms were distributed to this group of 100 workers - five unique cause terms for each worker. Three additional treatments were implemented, exhibiting levels of overlap falling between the two extreme cases. The five groups exhibited the

We are trying to learn about **effects** of different **causes**. For example:

- "**coupon**" is a cause of "**savings**"
- "**dominant**" is a cause of "**undefeated**"

Please answer the following as best as you can:

1. "**achieved**" is/are a cause of:
2. "**merry**" is/are a cause of:
3. "**phone**" is/are a cause of:
4. "**simply**" is/are a cause of:
5. "**medic**" is/are a cause of:
6. "**[REDACTED]**" is/are a cause of:

How confident are you in your answers?

⌵

Figure 8: *Screenshot of Task for Experiment 2*. Questions 1-5 comprise the preliminary question set, while question 6 acts as the test question. The cause term for question 6 (the test question) is only revealed once input has been entered for all five preliminary questions.

following CTEC values: 100, 50, 25, 5, and 1. We refer to the groups as CTEC-100, CTEC-050, CTEC-025, CTEC-005, and CTEC-001, respectively. Lower CTEC values indicate lower levels of overlap in preliminary question content. Table 3 describes the differences between treatment groups. Variation in the overlap of preliminary question content was expected to produce differences in the variety of inter-task priming effects caused by exposure to cause terms. Differences in the diversity of inter-task priming effects, induced by the preliminary question set, were expected to produce differences in the diversity of responses to the test question.

6.1.3 Test Question

The cause term presented for the test question, the sixth question of the task, was fixed for all workers. The cause term for the test question was “benefit”. Responses to the test question are the subject of our analysis of the treatments’ influence on response diversity.

6.1.4 Sources of Cause Terms

A total of 500 words were used as cause terms in the HIT. As in Experiments 1(a) and 1(b), the terms were sampled from: 1) previous crowdsourcing responses in the domain of causality, 2) a word association database synthesized by psychology researchers, and 3) a public set of common English words. In order to increase the chance of sampled terms fitting into the grammatical structure of our incomplete causal statement questions, terms were only included if they were also listed as entities in the Microsoft Concept Graph [35]. Manual processing of an initially sampled word set was also performed. Names, foreign countries and cities, brand names, and potentially inaccessible technical terms (i.e. “metadata”) were removed. Other terms were manually removed due to a strong incompatibility with the grammatical structure of the questions (i.e. “which”). After manual cleansing, the set of potential cause terms included 558 words. The cause term for the test question, “benefit”, was randomly selected from this set. The potential cause term set was then randomly sampled

Group	Cause Term Exposure Count (Number of Workers to See Each Cause Term)	Total Unique Cause Terms Distributed to Group	Example Preliminary Cause Term List
CTEC-100	100	5	merry, phone, simply, achieved, medic
CTEC-50	50	10	broccoli, interview, roast, trauma, training
CTEC-025	25	20	joy, oxygen, zoning, frozen, swing
CTEC-005	5	100	challenge, capable, stone, alien, preview
CTEC-001	1	500	affair, tracking, brochure, debate, trade

Table 3: *Experiment 2 Treatment Groups*. We identify treatment groups based on their Cause Term Exposure Count (CTEC). This value represents the number of workers, within the group, that are allocated a given cause term in the preliminary question set. Hence, lower CTEC values indicate lower levels of overlap in preliminary question content. Notice that when the CTEC is multiplied by the number of total unique cause terms that are distributed to the group, the product is 500 for each treatment group. This is because a total of 500 preliminary questions (5 questions for each of 100 workers) are posed to each group.

down to 500 words. Of the 500 words, 36 originated from previous crowdsourcing responses, 167 originated from the word association database, and 297 originated from the set of commonly used English words.

6.1.5 Data Collection

For each worker, we recorded: the content of input responses (six effect terms, confidence level), the timestamps of when the HIT was accepted and submitted, and an identifier of the ordered list of question content (cause terms) allocated to the worker.

6.1.6 HIT Presentation and Payment Scheme

The HIT was posted with the following title: “Identifying effects for given causes”. The description of the HIT was listed as: “Given a word or phrase, tell us what the word or phrase is a cause of.” The financial reward for completing the HIT was \$0.40, based on timing results from a similar previous experiment and a wage of \$10/hour. The preview of the HIT showed a frozen reconstruction of the HIT design, with placeholder text (i.e. “WORD_1”) displayed instead of sensible cause terms.

Workers could only accept to work on the HIT once. In order to qualify to work on the HIT, workers had to have 1) had at least 90% of their previous MTurk responses approved by crowdsourcers, 2) been eligible to view adult content, and 3) been located in the United States.

6.1.7 Measuring Response Diversity

In our analysis of responses to the test question, we study both lexical and semantic diversity. For our purposes, lexical diversity represents diversity in the actual textual content of input. Semantic diversity represents diversity in the meaning of words input by workers.

6.2 Results

6.2.1 Crowdsourcing Metrics

The HIT was made available to MTurk workers on February 25 and February 26 of 2018. 500 workers submitted responses to the HIT. In agreement with the intended distribution procedure, 100 workers were assigned to each treatment group. Preliminary cause term content was appropriately distributed as described in Section 6.1.2.

6.2.2 Bonferroni Correction

This experiment featured five treatment conditions. In order to account for multiple comparison effects when hypothesis testing, we apply a Bonferroni correction to the level of significance in pairwise comparisons. For five groups, 10 comparisons are necessary to account for all two-group combinations. In the reporting of pairwise hypothesis testing results in this section, the level of significance is $0.05/10 = 0.005$.

6.2.3 Time Spent

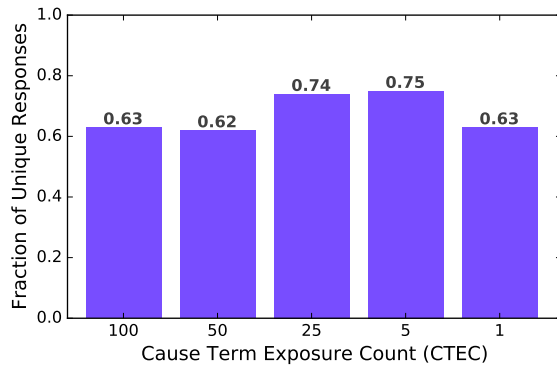
The time that workers spent on the HIT was consistent across all five treatment groups (Median_{CTEC-100}: 120.07 seconds; Median_{CTEC-050}: 121.78 seconds; Median_{CTEC-025}: 123.16 seconds; Median_{CTEC-005}: 109.31 seconds; Median_{CTEC-001}: 122.54 seconds; Kruskal-Wallis: $H = 1.044$, $p = 0.903$). The timing results agree with expectation, as the treatment did not substantially affect the size, structure, or interface of the task in any way; the treatment only affected the content of cause terms presented in the preliminary question set. The level of overlap in preliminary cause terms did not have an impact on the processing time of the task.

6.2.4 Lexical Diversity: Response Frequency

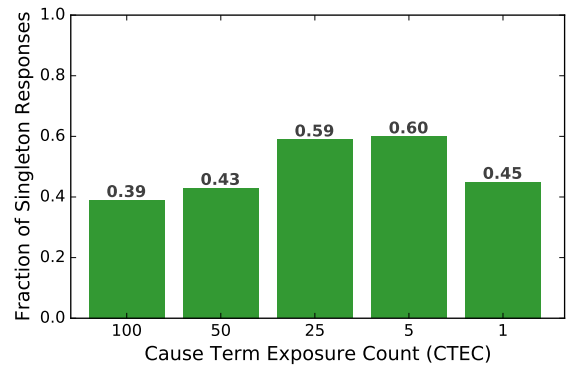
We analyze the lexical diversity of responses to the test question by investigating various measures of response frequency between treatment groups. A case-insensitive approach is taken in the analysis of responses.

Figure 9a shows the fraction of unique responses, f_{unique} , offered by each group. The fraction represents the length of the set of responses divided by the total number of responses for the test question. Figure 9b shows the fraction of responses that were only input by a single worker (singleton responses), $f_{singleton}$. Greater values of f_{unique} and $f_{singleton}$ imply greater lexical diversity in test question responses.

For both measures, the two groups treated with the highest level of preliminary cause term overlap, CTEC-100 and CTEC-050, exhibited relative consistency. Likewise, the two groups treated with the next two lowest CTEC values only differed by 0.01 in both f_{unique} and $f_{singleton}$. In line with expectation, the f_{unique} and $f_{singleton}$ values for the latter pair of groups, CTEC-025 and CTEC-005, were greater than the values of the former pair of groups by a margin of at least 0.11. Surprisingly, f_{unique} and $f_{singleton}$ for the CTEC-001 treatment group were more consistent with the values of the CTEC-100 and CTEC-050 groups than with those of the CTEC-025 and CTEC-005 groups. This result contradicts the notion that groups treated with lower levels of overlap in preliminary cause term content demonstrate greater lexical diversity in test question responses. No unequivocal trend is apparent.



(a) Fraction of Unique Responses



(b) Fraction of Singleton Responses

Figure 9: *Fraction of Unique and Singleton Responses*. The fraction of unique responses (the length of the set of responses divided by the total number of responses) and the fraction of singleton responses (the number of responses input by a single worker divided by the total number of responses) are both greatest for the CTEC-005 and CTEC-025 treatment groups. For both measures, greater values imply greater lexical response diversity. Surprisingly, the values for the CTEC-001 treatment group are similar to those for the CTEC-100 and CTEC-050 groups.

Group	Mode Response	Mode Fraction
CTEC-100	advantage	0.06
CTEC-050	advantage	0.06
CTEC-025	advantage	0.05
CTEC-005	happiness	0.05
CTEC-001	charity	0.06

Table 4: *Mode Responses*. The three groups treated with the greatest levels of overlap in preliminary question content produced the same mode response ("advantage") to the test question. The mode response is the effect term that was most commonly input to complete the test question statement ("benefit" is/are a cause of:).

Singleton responses represent test question responses that exhibited a frequency of one - each singleton response was only input by one worker in the given group. Responses that were input by two workers exhibited a response frequency of two, and so on. Regarding the distributions of such response frequency counts, no statistical difference between groups is evident (Kruskal-Wallis: $H = 7.009$, $p = 0.135$). This suggests that preliminary question content overlap did not have a significant effect on lexical response diversity.

6.2.5 Lexical Diversity: Mode Fraction

The mode response to the test question for each group is listed in Table 4. The mode fraction represents the mode response's frequency divided by the total number of responses in the group. Although mode frequency was relatively consistent across treatment groups, the mode response (most commonly input effect term for the cause term "benefit") was not the same for all treatment groups.

We conducted a bootstrapping procedure to compare the mode fraction distributions of simulated response samples for the different treatment groups. The distributions of the resulting mode fractions offer more robust comparisons of response frequency than stand-alone mode fraction values. The distributions exhibited differences between treatment groups (Kruskal-Wallis: $H = 32680.879$, $p = 0.0$). Results of the bootstrapping procedure are summarized in Table 5. In comparing the pairwise differences in mode fraction values and effect size, we observe a bootstrapped mode fraction

CTEC		Median		Mann-Whitney		Cohen's Effect Size
Group ₁	Group ₂	Group ₁	Group ₂	<i>U</i>	<i>p</i>	<i>d</i>
100	50	0.07	0.08	962195175.0	0.0*	-0.388
100	25	0.07	0.06	943002133.5	0.0*	0.439
100	5	0.07	0.06	864702505.5	0.0*	0.549
100	1	0.07	0.08	1081148140.0	0.0*	-0.222
50	25	0.08	0.06	674050955.5	0.0*	0.847
50	5	0.08	0.06	606545969.0	0.0*	0.959
50	1	0.08	0.08	1125004582.5	$<10^{-172}$ *	0.168
25	5	0.06	0.06	1165407509.0	$<10^{-81}$ *	0.114
25	1	0.06	0.08	779609170.0	0.0*	-0.676
5	1	0.06	0.08	706407287.5	0.0*	-0.788

Table 5: *Pairwise Comparisons of Bootstrapped Sample Mode Fraction Distributions*. For each treatment group, the set of responses to the test question was sampled (with replacement) 100 times. For each round of sampling, the mode fraction was calculated. 50,000 such bootstrapping simulations were performed for each treatment group to generate robust mode fraction distributions. Significant differences are revealed in pairwise comparisons. Asterisks represent the presence of statistical significance.

ranking of CTEC-050 > CTEC-001 > CTEC-100 > CTEC-025 > CTEC-005. Greater mode fraction values imply lower lexical diversity. This ordering appropriately complements the ordering of values for the fraction of unique responses, CTEC-005 > CTEC-025 > CTEC-100 > CTEC-001 > CTEC-050. Here, greater values imply greater lexical diversity. These rankings do not exhibit a consistent trend with respect to CTEC, nor do they align with the expected ordering of lexical response diversity (descending): CTEC-001, CTEC-005, CTEC-025, CTEC-050, CTEC-100.

6.2.6 Semantic Diversity

In order to analyze the semantic diversity of responses to the test question, we employed a 300-dimensional space of trained word vectors from the word2vec⁴ model. A word vector numerically represents a single word. Words that are found in similar contexts have been trained to be closer to each other in the vector space. The resulting space may be used as a tool for semantic operations and comparisons.

For each treatment group, we processed pairwise comparisons of the similarity between test question responses. First, within each group, for each pairing of test question inputs, we calculated the cosine similarity between the corresponding word vectors. Greater cosine similarity between word vectors implies greater semantic similarity between the words represented by those vectors. For reference, the word2vec cosine similarity between the vectors for “car” and “truck” is 0.674, while the cosine similarity between the vectors for “car” and “political” is 0.097. Greater levels of semantic similarity in a set of inputs imply lower levels of semantic diversity. In our semantic analysis, we exclude input pairings for which at least one of the inputs does not have a corresponding trained word2vec word vector.

Figure 10 shows a boxplot of the word2vec similarity distribution for each treatment group. The distributions exhibited statistical differences between treatment groups (Kruskal-Wallis: $H = 177.977$, $p < 10^{-37}$). Table 6 provides a summary of the pairwise comparison of treatment groups, regarding the distributions of word2vec cosine similarity. In comparing the

⁴<https://code.google.com/archive/p/word2vec>

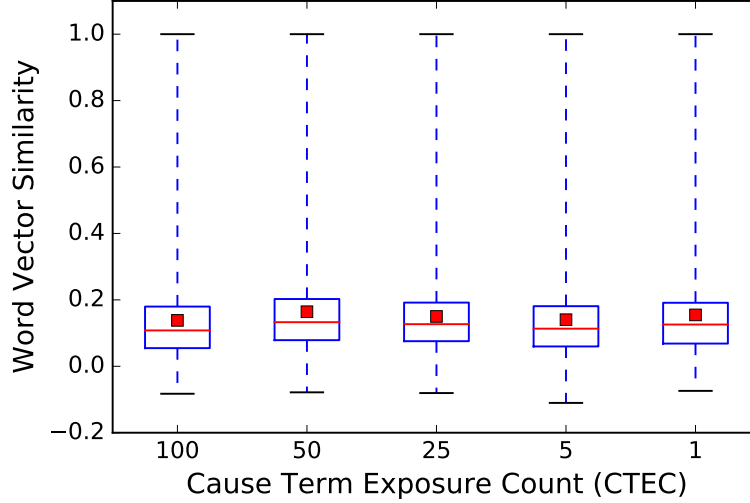


Figure 10: *Boxplots of Word2vec Cosine Similarity*. Cosine similarity is a measure of distance between word vectors in the trained word2vec space. Cosine similarity has a theoretical range of $[-1, 1]$. A word compared with itself has a cosine similarity of 1; completely unrelated words exhibit cosine similarities closer to 0; words with opposite meanings occupy the low end of the spectrum. We expected semantic similarity to gradually decrease with decreases in CTEC. Although a steady decrease in cosine similarity is exhibited by groups CTEC-050, CTEC-025, and CTEC-005, the cosine similarity distributions of the CTEC-100 and CTEC-001 groups defy the trend.

differences in word2vec similarity values and effect size, we observe a semantic similarity ranking of $\text{CTEC-050} > \text{CTEC-001} \approx \text{CTEC-025} > \text{CTEC-005} \approx \text{CTEC-100}$. These rankings do not exhibit a consistent trend with respect to CTEC, nor do they align with the expected ordering of semantic response diversity (ascending): CTEC-100, CTEC-050, CTEC-025, CTEC-005, CTEC-001. Moreover, the suggested trend in semantic similarity does not agree with the suggested trend in lexical similarity: $\text{CTEC-050} > \text{CTEC-001} > \text{CTEC-100} > \text{CTEC-025} > \text{CTEC-005}$. Our semantic analysis is limited by the non-existence of word2vec word vectors that correspond to certain inputs. As shown in Table 6, responses to the test question in the treatment groups with higher CTEC values were more likely to have corresponding word2vec word vectors.

6.3 Discussion

From the statistical analyses, we conclude that the treatment does not produce consistent crowd-level priming effects with respect to response diversity. It is likely that inter-task effects were at play in the HIT, given differences in mode response and lexical diversity measures between treatment groups. However, the experiment does not provide sufficient evidence that the studied mechanism (varying overlap in preliminary question content) primes the population of workers in a way that produces a consistent effect on response diversity. Overall, we observed no consistent trends in the lexical and semantic diversity of test question responses as a function of CTEC. The results do not provide support for our expectation that lower levels of overlap in preliminary question content lead to greater diversity in test question responses.

6.3.1 Nuance in Priming Treatment and Test Question

The priming treatment in this experiment exhibited an even greater degree of subtlety than the treatments prescribed in Experiments 1(a) and 1(b). Differences in task treatment were not applied to task structure, but instead were embedded

CTEC		Median		Mann-Whitney Test				Cohen's Effect Size
Group ₁	Group ₂	Group ₁	Group ₂	n_1	n_2	U	p	d
100	50	0.108	0.133	4278	4560	8321761.5	$<10^{-33*}$	-0.173
100	25	0.108	0.127	4278	4186	8054319.0	$<10^{-16*}$	-0.084
100	5	0.108	0.113	4278	3916	8124767.5	0.009	-0.015
100	1	0.108	0.126	4278	3916	7630674.5	$<10^{-12*}$	-0.109
50	25	0.133	0.127	4560	4186	9076695.0	$<10^{-5*}$	0.098
50	5	0.133	0.113	4560	3916	7881202.5	$<10^{-21*}$	0.164
50	1	0.133	0.126	4560	3916	8399350.5	$<10^{-6*}$	0.062
25	5	0.127	0.113	4186	3916	7622602.5	$<10^{-8*}$	0.071
25	1	0.127	0.126	4186	3916	8109461.0	0.205	-0.032
5	1	0.113	0.126	3916	3916	7221051.5	$<10^{-6*}$	-0.098

Table 6: *Pairwise Comparisons of Word2vec Cosine Similarity Distributions.* For each pairing of test question inputs within a treatment group, we calculated the cosine similarity between corresponding word2vec word vectors. Greater cosine similarity between word vectors implies greater semantic similarity between the words represented by those vectors. Greater levels of semantic similarity imply lower levels of semantic diversity. Asterisks represent the presence of statistical significance.

in the lexical content of question prompts. Moreover, the priming treatment (overlap in questions distributed to *different* workers) was applied at the group level, not within the scope of the task for individual workers. The potential for an emergent trend was dependent on the strength of variety in inter-question priming effects.

Our analysis only relied on responses to a single test question. It is possible that the broadness of the test question cause term, “benefit”, had a mitigating effect on the potential for evidence of expected group-level priming behaviors. The word “benefit” has several different semantic interpretations. Because of this, our test question offered a substantial degree of open-endedness. Many terms could have sensibly been input as an effect term for “benefit”. Inter-question priming effects introduced as a result of preliminary question content may have been overshadowed by the sheer number of appropriate responses to the test question.

6.3.2 Future Work

Previous priming studies in crowdsourcing have focused on varying stimuli at the scale of the individual worker. Despite the lack of a consistent result in our experiment, it is possible that different implementations of group-level, or crowd-level, priming can inform subtle task design choices (i.e. question content overlap) that significantly impact resulting crowd responses. Additional research is required to discover emergent trends as a result of crowd-level priming.

We recognize several opportunities for future study that may provide further clarity regarding the impacts of preliminary content overlap treatments on response diversity. We are interested in further investigating the influence of the specificity of the test question cause term. As discussed, “benefit” is a broad cause term. A test question cause term with a lower number of definitions - and a more limited set of sensible effect inputs - may reveal a more prominent and consistent effect of the group-level priming treatment. Task instructions provide another domain for the investigation of crowd-level priming treatments. In order to ensure that their work will be approved by crowdsourcers, reliable microtask workers often strictly abide by explicit HIT instructions. Varying the degree of content overlap in task instruction, instead of preliminary

questioning, may produce a stronger and more consistent emergent effect on response diversity. An experimental design very similar to that of Experiment 2 could be implemented to study the effect of treatments on task instruction overlap.

7 General Discussion

The results of Experiments 1(a) and 1(b) show that the granting of initial response freedom does not introduce outright priming effects that increase worker retention or improve response quality. In fact, the overall performance advantage of the Baseline group in Experiment 1(b) suggests that designing tasks in a way that promotes consistency in question structure, rather than imposing priming treatments through modified task structures, can lead to more efficiency and reliability in responses. The strength of the influence of subtly imposed inter-task priming effects was investigated in Experiment 2, through the implementation of consistent question structure and group-level treatment. The results of Experiment 2 did not reveal a consistent relationship between the level of overlap in preliminary question content and subsequent response diversity.

Although our experiments have not provided sufficient cause for the recommendation of any new guiding principles of microtask design, our confidence - in the potential for priming procedures to benefit crowdsourcing projects by improving worker retention, response quality, and response diversity - is not lost. The evidence of some influence of inter-task effects on responses in Experiments 1(b) and 2, and the avenues for future research encouraged by the nuance of our experimental designs, should motivate further work regarding crowd-level priming agents. The potential benefits of effective priming procedures are numerous for crowdsourcers. The power of the human subconscious may be harnessed in online microtask work environments to promote the goals of task administrators.

The utilization of priming procedures to influence worker output must always be viewed through an ethical lens. Priming effects act in the nonconscious dimension of human intelligence. Influencing work output with such mechanisms may be considered an infringement of free will. Priming mechanisms are constantly firing in our brains whether we like it or not. In order to uphold the integrity of microtask crowdsourcing environments and the research community, the power of priming treatments should never be applied for purposes of exploitation that adversely affect workers. Workers should always be justly compensated for their time and effort.

The power of priming reminds us to refrain from the oversimplification of viewing microtask workers as merely independent computational units. All humans are subject to different stimuli and exhibit behavior that is influenced by different biases. Even when we do not consider human microtask workers to be simple input-output machines, we must overcome any assumptions that all crowdsourced workers can be judged by similar standards. Although all workers must have some motivation to participate in crowd work, the causes and effects of different sources of motivation and bias complicate the realm of microtask crowdsourcing. The complexities of human cognition have facilitated the development of crowdsourcing into an exciting, challenging, and multidisciplinary world for researchers.

8 Acknowledgments

This work would not have been possible without the unwavering guidance of Dr. James Bagrow. We would also like to acknowledge the members of the thesis committee, Dr. Margaret Eppstein and Mr. Robert Erickson, for their contributions of time and useful feedback. Xipei Liu conducted preliminary work that informed the design of the experiments presented in this paper. A collective debt of gratitude is owed to the anonymous individuals who worked on our experimental crowdsourcing tasks. Without their willingness to work, crowdsourcing environments would not be worthy of such explorations. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1447634.

References

- [1] J. Howe, “The rise of crowdsourcing,” *Wired magazine*, vol. 14, no. 6, pp. 1–4, 2006.
- [2] G. Little, L. B. Chilton, M. Goldman, and R. C. Miller, “Exploring iterative and parallel human computation processes,” in *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pp. 68–76, ACM, 2010.
- [3] J. Cheng, J. Teevan, S. T. Iqbal, and M. S. Bernstein, “Break it down: A comparison of macro- and microtasks,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 4061–4064, ACM, 2015.
- [4] E. Tulving and D. L. Schacter, “Priming and human memory systems,” *Science*, vol. 247, no. 4940, pp. 301–306, 1990.
- [5] H. Aarts and A. Dijksterhuis, “The silence of the library: environment, situational norm, and social behavior,” *Journal of personality and social psychology*, vol. 84, no. 1, p. 18, 2003.
- [6] T. L. Chartrand, “The role of conscious awareness in consumer behavior,” *Journal of Consumer Psychology*, vol. 15, no. 3, pp. 203–210, 2005.
- [7] R. R. Morris, M. Dontcheva, and E. M. Gerber, “Priming for better performance in microtask crowdsourcing environments,” *IEEE Internet Computing*, vol. 16, no. 5, pp. 13–19, 2012.
- [8] R. R. Morris, M. Dontcheva, A. Finkelstein, and E. Gerber, “Affect and creative performance on crowdsourcing platforms,” in *Affective computing and intelligent interaction (ACII), 2013 humane association conference on*, pp. 67–72, IEEE, 2013.
- [9] A. Kittur, E. H. Chi, and B. Suh, “Crowdsourcing user studies with mechanical turk,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 453–456, ACM, 2008.
- [10] E. Newell and D. Ruths, “How one microtask affects another,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 3155–3166, ACM, 2016.
- [11] W. Mason and D. J. Watts, “Financial incentives and the performance of crowds,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 2, pp. 100–108, 2010.
- [12] J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic, “An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets,” in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, vol. 11, pp. 17–21, AAAI, 2011.
- [13] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, “Quality control in crowdsourcing systems: Issues and directions,” *IEEE Internet Computing*, vol. 17, no. 2, pp. 76–81, 2013.
- [14] W. S. Lasecki, J. M. Rzeszutarski, A. Marcus, and J. P. Bigham, “The effects of sequence and delay on crowd work,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1375–1378, ACM, 2015.
- [15] X. Liu and J. P. Bagrow, “Autocompletion interfaces make crowd workers slower, but their use promotes response diversity,” 2017. Preprint arXiv 1707.06939.
- [16] H. Alagarai Sampath, R. Rajeshuni, and B. Indurkha, “Cognitively inspired task design to improve user performance on crowdsourcing platforms,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3665–3674, ACM, 2014.
- [17] T. C. McAndrew, E. A. Guseva, and J. P. Bagrow, “Reply & supply: Efficient crowdsourcing when workers do more than answer questions,” *PLOS ONE*, vol. 12, no. 8, 2017. e0182662.
- [18] T. M. Amabile, *Creativity in context: Update to the social psychology of creativity*. Westview Press, 1996.
- [19] U. Gadiraju and S. Dietze, “Improving learning through achievement priming in crowdsourced information finding microtasks,” in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pp. 105–114, ACM, 2017.
- [20] C.-M. Chiu, T.-P. Liang, and E. Turban, “What can crowdsourcing do for decision support?,” *Decision Support Systems*, vol. 65, pp. 40–49, 2014.
- [21] T. Buecheler, J. H. Sieg, R. M. Füchslin, and R. Pfeifer, “Crowdsourcing, open innovation and collective intelligence in the scientific method—a research agenda and operational framework,” in *ALIFE*, pp. 679–686, 2010.

- [22] K. E. Bevelander, K. Kaipainen, R. Swain, S. Dohle, J. C. Bongard, P. D. Hines, and B. Wansink, "Crowdsourcing novel childhood predictors of adult obesity," *PloS one*, vol. 9, no. 2, p. e87756, 2014.
- [23] Y. Jiang, J. K. Kummerfeld, and W. S. Laseck, "Understanding task design trade-offs in crowdsourced paraphrase collection," *arXiv preprint arXiv:1704.05753*, 2017.
- [24] A. J. Quinn and B. B. Bederson, "A taxonomy of distributed human computation," *Human-Computer Interaction Lab Tech Report, University of Maryland*, 2009.
- [25] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?," *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011.
- [26] H. Zheng, D. Li, and W. Hou, "Task design, motivation, and participation in crowdsourcing contests," *International Journal of Electronic Commerce*, vol. 15, no. 4, pp. 57–88, 2011.
- [27] T. M. Amabile, "The social psychology of creativity: A componential conceptualization," *Journal of Personality and Social Psychology*, vol. 45, no. 2, pp. 357–376, 1983.
- [28] D. C. Brabham, "Crowdsourcing as a model for problem solving: An introduction and cases," *Convergence*, vol. 14, no. 1, pp. 75–90, 2008.
- [29] J. Surowiecki, *The wisdom of crowds*. Anchor, 2005.
- [30] T. Wu, L. Chen, P. Hui, C. J. Zhang, and W. Li, "Hear the whole story: Towards the diversity of opinion in crowdsourcing markets," *Proceedings of the VLDB Endowment*, vol. 8, no. 5, pp. 485–496, 2015.
- [31] S. A. Chowdhury, M. Calvo, A. Ghosh, E. A. Stepanov, A. O. Bayer, G. Riccardi, F. García, and E. Sanchis, "Selection and aggregation techniques for crowdsourced semantic annotation task," in *Sixteenth annual conference of the international speech communication association*, 2015.
- [32] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber, "The university of south florida free association, rhyme, and word fragment norms," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 3, pp. 402–407, 2004.
- [33] S. C. Wheeler and J. Berger, "When the same prime leads to different effects," *Journal of Consumer Research*, vol. 34, no. 3, pp. 357–368, 2007.
- [34] B. Schwartz, "Self-determination: The tyranny of freedom.," *American psychologist*, vol. 55, no. 1, p. 79, 2000.
- [35] Z. Wang, H. Wang, J.-R. Wen, and Y. Xiao, "An inference approach to basic level of categorization," in *Proceedings of the 24th acm international on conference on information and knowledge management*, pp. 653–662, ACM, 2015.