How One Microtask Affects Another

Edward Newell
School of Computer Science
McGill University
edward.newell@mail.mcgill.ca

Derek Ruths
School of Computer Science
McGill University
derek.ruths@mcgill.ca

ABSTRACT

Microtask platforms are becoming commonplace tools for performing human research, producing gold-standard data, and annotating large datasets. These platforms connect *requesters* (researchers or companies) with large populations (crowds) of workers, who perform small tasks, typically taking less than five minutes each. A topic of ongoing research concerns the design of tasks that elicit high quality annotations. Here we identify a seemingly banal feature of nearly all crowdsourcing workflows that profoundly impacts workers' responses. Microtask assignments typically consist of a sequence of tasks sharing a common format (e.g., circle galaxies in an image). Using image-labeling, a canonical microtask format, we show that earlier tasks can have a strong influence on responses to later tasks, shifting the distribution of future responses by 30-50% (total variational distance). Specifically, prior tasks influence the content that workers focus on, as well as the richness and specialization of responses. We call this phenomenon *intertask effects*. We compare intertask effects to framing, effected by stating the requester's research interest, and find that intertask effects are on par or stronger. If uncontrolled, intertask effects could be a source of systematic bias, but our results suggest that, with appropriate task design, they might be leveraged to hone worker focus and acuity, helping to elicit reproducible, expert-level judgments. Intertask effects are a crucial aspect of human computation that should be considered in the design of any crowdsourced study.

Author Keywords

Microtask; crowdsourcing; priming; framing; task-design; bias; classifier

ACM Classification Keywords

H.1.2. Information Systems: Models and Principles; Human information processing

INTRODUCTION

There are many tasks that are trivial for people, but difficult to solve programmatically. Such tasks include tagging and categorizing images, coding and transcribing media, and performing surveys for academic or market research purposes (see Table 1 for a listing of task types seen on the Amazon's Mechanical Turk microtask platform). Many tasks are ill-defined, in the sense that they do not have a clear "correct" response, and require high-level, qualitative judgment. Microtask platforms are marketplaces that help fill the gap in current computational capabilities, by matching requesters, who need to have such tasks completed, with human workers. Amazon's Mechanical Turk (MTurk)[1], and CrowdFlower[2] are two popular microtask platforms.

This form of crowdsourcing embeds workers in a controlled but flexible task infrastructure. To the requester, workers seem almost like input-output devices. This provides much of the flexibility and cost-savings of fully automating the work in a computer program: the workforce is available on demand over the Internet using automated scripts, without the need for interviews or contracts [37, 11]. Work can be performed at a fraction of the cost of traditional methods for recruiting temporary workers or experimental subjects [4]. Many researchers consider microtask platforms as a new kind of *human computing* architecture [11].

The flexibility and cost-effectiveness of microtask labor has led to a surge in demand from industry and academia [37, 4]. More recently microtask platforms have been assessed as a way to supplement expert human resources to increase capacity in critical applications, such as in medical diagnostic functions, with promising results [35].

Naturally, researchers have investigated the factors affecting the reliability of microtask work, including the design of the task interface [14], the design of workflows (how work is divided into tasks) [24, 17, 25], and the framing of tasks [23, 7, 33]. Here we draw attention to a ubiquitous yet overlooked feature of microtask work: the tendency for workers to perform many similar tasks in quick succession.

This tendency arises from a combination of worker preferences and the logistics of serving small tasks. Workers have a preference for performing sequences of similar microtasks [9], probably because it reduces cognitive load arising from task-switching [1]. Moreover, as workers complete task assignments, they must continually switch between working on an assignment and choosing their next assignment (weighing such factors as the wage paid, and effort required). Since microtasks are very short, it makes sense to bundle tasks together into larger assignments, to reduce the overhead of switching between working on assignments and choosing them. This may explain why the predefined assign-

[1] mturk.com

[2] crowdflower.com

CHI'16, May 07–12, 2016, San Jose, CA, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3362-7/16/05 $15.00
DOI http://dx.doi.org/10.1145/2858036.2858490

Task Type	Count	Fraction (%)
Image transcription	57	28.5
Information gathering	46	23.0
Image labeling and classification	36	18.0
Copy writing and editing	18	9.0
Text tagging and classification	15	7.5
Audio/Video transcription	12	6.0
Survey	9	4.5
Unknown	7	3.5

Table 1. Frequency of various broad task types seen among the 200 most recently posted tasks on Amazon's Mechanical Turk, accessed on 12 February, 2015.

ment templates available on microtask platforms generally bundle many tasks together by default[3].

Psychological experiments show that the exposure to stimuli immediately before performing a task influences performance, an effect known as *priming* [15, 21, 3, 36]. The effect of priming appears to be stronger when the modality of the prime is the same as the task [36]. Thus, for the typical microtask worker, who does many similar tasks in quick succession, there is a potential for earlier tasks to influence later ones, via priming or other mechanisms.

We seek to determine if any such effects in fact occur and if so, to measure and characterize them. Certain tasks admit a well-defined notion of "correct" and "incorrect" responses. But many tasks involving qualitative judgment do not. So, rather than simply characterizing effects on accuracy, we seek to provide a generalized measure of the extent to which prior tasks can alter the distribution of responses to later tasks. Measuring the *extent* of the shift in a distribution of responses is substantially harder than simply determining *whether* an effect has occurred. Our first contribution is a statistically grounded method for doing so.

To investigate the phenomenon in a highly generalizable setting, we sought a task that would involve qualitative judgment and enable relatively unconstrained responses. We also sought a task that would be a typical exemplar of the kinds of tasks actually seen on microtask platforms (see Table 1 for examples). For this purpose, we adopted image-labeling as a canonical microtask, a task in which workers provide descriptive labels to images using free-text inputs. Image labeling is a qualitative task without clearly "correct" or "incorrect" responses, and is among the most common kinds of tasks on MTurk (see Table 1). If effects from earlier tasks arise in our setup, similar effects could be expected in a large number of microtasks and crowdsourced workflows where workers are called upon to provide qualitative judgment.

In our experiments, workers label a series of images, one at a time. Depending on the treatment to which a worker is assigned, we vary the images in the first five tasks, while keeping those in the last five tasks the same. For example, in one experiment, the first five images shown to one group of workers contain food, while those shown to the other contain (non-food) objects. The last five images for both groups contain both food and objects.

Our results show that a worker's responses are strongly influenced by the content of tasks performed beforehand, leading to as much as 50% total variational distance (see Figure 1 for illustrative definition of total variational distance). Using the WordNet knowledge base, we analyze worker's word choices to characterize the nature of these effects in detail. We find that, when workers label a series of images that are more similar, their responses become more specialized and more diverse. Prior tasks can shift the topical focus of worker's labels, inducing them to focus on different aspects of the images.

As a point of comparison, we also conduct framing experiments, in which we alter the framing of microtask assignments, in terms of the stated purpose of the task, or by naming the funder. Remarkably, the effects of prior microtasks, which are virtually ubiquitous, are on par with, or stronger than, the effects of overtly framing an assignment.

We call the effects that earlier tasks exert on later ones *intertask effects*. If, as has been suggested [11], microtask platforms are to be considered as a new form of computing architecture, it will be necessary to reconcile the fact that the human computing elements exhibit *hysteresis*, meaning that microtask workers' outputs depend on the *history* of their inputs.

Unraveling the psychological mechanisms that give rise to such strong effects will require further investigation. We describe a possible mechanism based on positive and negative priming, consistent with the observations from our experiments. Irrespective of the underlying mechanism, this very strong effect might be exploited in task design to tune worker focus and acuity.

As our main contributions, we

- derive a method for measuring changes in response distributions;
- measure the strength of intertask effects;
- show that intertask effects are stronger than, or on par with framing;
- show that when completing a series of similar microtasks, worker's responses become more specialized and diverse;

The rest of the paper is organized as follows. In the next section, we review the relevant prior work. We then present our method for measuring changes to response distributions. Next we present our first experiment which measures intertask effects and compares them to framing. We then describe a second experiment which extends these results, addressing questions raised by the first experiment. We conclude by discussing interpretations and ramifications, suggesting a possible mechanism for intertask effects based on priming, and describing directions for future work.

[3] e.g. on mturk.com and crowdflower.com

PRIOR WORK

Effects of microtask design on response quality

Microtask platforms are increasingly used to clean and codify datasets, administer experiments with human participants, and distribute surveys. Accuracy and replicability are crucial in all of these applications.

There is ongoing research investigating how the design of microtasks, and their context, affect the quality of responses. One of the most basic issues with crowd work is that workers vary in the amount of attention they pay to tasks, and in their understanding of instructions. Workers who disregard or misunderstand instructions can be effectively screened out by including quiz-questions among the tasks, while instruction comprehension can be improved using a training phase before actual tasks begin [26, 19]. In addition, providing real-time feedback to workers encourages higher-quality responses, while prompting workers to review their work according to a set of criteria increases the worker's response quality over time [12].

The design of the task interface also has important effects on response quality. A simpler user interface can improve accuracy [14], and certain input controls are inherently less effortful and error-prone than others [8].

Since microtask work is generally paid, wage is a basic parameter in the requester's control that might influence response quality. One might expect a higher wage to *buy* better performance, but results are ambiguous. One study found that increasing wage has no effect on response quality, but does increase the amount of responses [28]. Another study [19] showed that increasing wages does increase quality, but that this effect saturates, and eventually increases in wage attract a larger proportion of workers who disregard the task instructions.

Workers have other motivations for completing tasks aside from remuneration [19]. Researchers have attempted to appeal to worker's motivation to do work that has a meaningful purpose: when a task was framed as assisting with medical research, workers were more likely to participate and completed more tasks, than when they were not provided any context [7]. Providing a meaningful context did not, however, increase the *quality* of work.

Other studies have investigating different means of framing. When asked to rate various policy interventions, workers emphasized respectively more or less punitive approaches depending on whether the phrase "crime is a beast" or "crime is a virus" appeared in the problem description [33]. When workers perform a task within the context of larger workflow, explaining how the task fits into the workflow, which amounts to a kind a framing, increased both the quantity and quality of responses [23].

The breaking down of complex jobs into simpler tasks can increase efficiency by enabling greater parallelism. While this could disrupt the natural context provided by performing the whole job as a single task, it turns out that, for complex tasks such as writing an article, dividing the job into subtasks, including outlining, information-gathering, and paragraph-synthesis, actually improves the quality of the end-product [24].

Other work investigated the effects of interruptions on task performance [25]. There, workers completed a series of tasks that required them to answer questions about an illustrated map. When interruptions were introduced in the form of either a time delay, or a different task involving a different map, workers took longer to complete the task which followed the interruption. However this study did not test for effects on the quality of responses.

In a somewhat related vein, other researchers investigated the effects of task-switching [38]. During *switch-tasks* (tasks which differ in nature from the immediately preceding task), workers were more inclined to make mistakes, and took longer to complete the task. However, the researchers found that by offering a performance-based bonus, which paid if workers completed the task quickly and correctly, performance during switch-tasks was both faster and more accurate.

However, long periods of performing the same task can lead to fatigue, which can be relieved by well placed-interruptions. A study [10] found that introducing "micro-diversions" periodically between tasks caused workers to complete more tasks overall, and to complete them more quickly. This was more pronounced for more cognitively taxing tasks.

Priming

Priming is a psychological phenomenon whereby previous exposure to a stimulus leads to faster or more accurate responses to similar stimuli [15], or to lower response thresholds [21]. Priming occurs by means of pre-activation, which can influence stimulus encodings (during perception) [21], the top-down application of object-knowledge [15], and memory access [3]. Priming is more effective when the prime and the target task involve the same kind of activity. For example, exposing participants to an image helps them subsequently recognize that image when shown very briefly in a tachistoscope; however, being presented with a related *word*, and reading it aloud, does not [36].

As we have mentioned, workers prefer to perform a series of tasks that are similar to one another. Moreover, many microtasks involve tasks based on visual and auditory perception similar to the kinds used in priming studies. Thus, in a typical microtask assignment, there are many opportunities for priming to arise.

Researchers have begun to explore the potential applications of priming in microtask design. One study found that showing workers an image of a laughing baby (so as to stimulate positive affect), led to better performance in a task requiring creativity [27]. This shows that priming effects can play a role in microtask design. The question we seek to answer, however, is whether earlier tasks can affect responses to later tasks.

Taken together, the prior work suggests that, beyond the design of tasks themselves, the design of the context of tasks has a major effect on responses. There is reason to believe

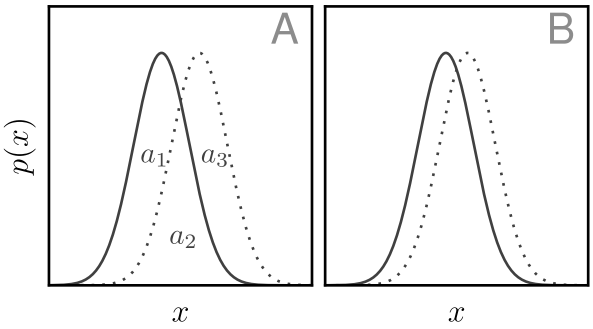

Figure 1. Normal distributions exhibiting a total variational distance of (A) 50% and (B) 30%. The total variational distance is the non-overlapping portion of the distributions, e.g. in (A) it is given by $\theta = a_1/(a_1 + a_2) = a_3/(a_2 + a_3)$**.**

that, even in a series of similar tasks, earlier microtasks may influence later ones, such as through a priming mechanism. However, no direct investigation of whether this phenomenon exists has yet been made, which is the goal of the present work.

MEASURING CHANGES TO RESPONSE DISTRIBUTIONS

When a prior task influences a later one, this influence will manifest as a shift in the distribution of responses workers provide for the later task. A variety of tests exist to determine *whether* two samples, i.e. sets of responses, come from different distributions, most notably the χ^2 test. However, such tests do not tell us *how much* two distributions differ, and hence how strongly the initial task's influence is.

Determining the strength of intertask effects amounts to measuring a distance, or *divergence*, between response distributions. We will present a method for bounding a kind of divergence known as *total variational distance*[4], which we will henceforth denote by θ. The value of θ between two distributions P and Q is the fraction under the graph of P that does not overlap with that under Q (or vice-versa; see Figure 1). Formally:

$$\theta = \frac{1}{2} \sum_{x \in \mathcal{X}} |p(x) - q(x)|, \tag{1}$$

Where $p(x)$ and $q(x)$ are the respective probabilities that a worker submits response x, from a set of possible responses $\mathcal{X}$, under distributions P and Q. The divergence is constrained to $0 \leq \theta \leq 1$ (or from 0 to 100%). When $\theta = 100\%$, the distributions do not overlap at all. When $\theta = 0\%$, the distributions are identical. For reference, the distributions plotted in Figure 1A and B have $\theta = 50\%$ and $\theta = 30\%$ respectively.

Empirically measuring the divergence between distributions is a matter of ongoing research [34, 2, 6]. In the "naive" approach one first uses the samples to reconstruct empirical distributions, by calculating their maximum likelihood parameters. Then, one substitutes the empirical values of $p(x)$ and $q(x)$ into Equation 1 [2]. For example, one might directly estimate $\hat{p}(x) = N_x/N$, where N_x is the number of times response x is observed among N total responses.

However this approach can *drastically* overestimate θ [34]. As an illustrative example, suppose that we have sampled two sets of 1000 words from *possibly* different distributions, and that we wish to estimate the divergence between these distributions. It turns out that, if both sets of words were actually drawn from an identical Zipf distribution[5], the naive approach would typically lead one to report $\hat{\theta} \approx 65\%$, even though, in reality $\theta = 0\%$ (this can be shown by simulated sampling).

Nevertheless, we can establish a lower bound for θ by exploiting a fact about the theoretical limits on the accuracy of a classifier algorithm. The intuition is as follows: suppose workers are shown one of two alternative designs[6] for an otherwise similar task, and that we build a classifier which, based on a worker's response, infers which of the designs had been used. If worker's responses are not affected by the design alternative, then the classification problem will be hard, and the classifier's accuracy poor. Conversely, if the classifier accuracy is good, then design alternative must have had a strong effect on the response distribution.

Stated formally, any classifier algorithm, $\mathcal{A}$, that takes the response, x, of a worker, and guesses the design that elicited the response (from two possibilities: design$_1$ or design$_2$), will do so with accuracy $\eta_{\mathcal{A}}$, that is bounded according to:

$$\theta \geq 2\eta_{\mathcal{A}} - 1, \tag{3}$$

where θ is the total variational distance between the distributions of responses to design$_1$ and design$_2$.

We can establish Inequality 3 by considering an optimal classifier having accuracy η_*. Let us assume that workers are shown design$_1$ or design$_2$ with equal probability. If the worker gives the response x, the optimal classifier must guess that the worker was shown the design most likely to elicit x. In other words, if $p_i(x)$ is the probability that a worker shown design$_i$ responds with x, then it is optimal to guess that the worker saw design$_j$, where $j = \arg\max_j p_j(x)$.

Of course, neither $p_1(x)$ nor $p_2(x)$ are known. But, on seeing x, the probability that such a classifier would be correct is:

$$\Pr\{\text{correct}|x\} = \frac{1}{2} + \frac{|p_1(x) - p_2(x)|}{2(p_1(x) + p_2(x))} \tag{4}$$

Summing over all possible responses that a worker could provide, $x \in \mathcal{X}$, weighted by the probability of observing x, we

[4] Other formulations such as the Kullback-Leibler and Jensen-Shannon divergence can be related to the total variational distance.

[5] The Zipf distribution is a model for word frequencies [31, 40]:

$$p(x) = \frac{x^{-1}}{\sum_{n=1}^{\infty} x^{-1}}, \tag{2}$$

where $p(x)$ is the probability of the xth most common word.

[6] We use "design" in a general sense, to include both the design of the task and of its context. In particular, the "design" might include the selection of tasks that were shown beforehand, or the use of framing.

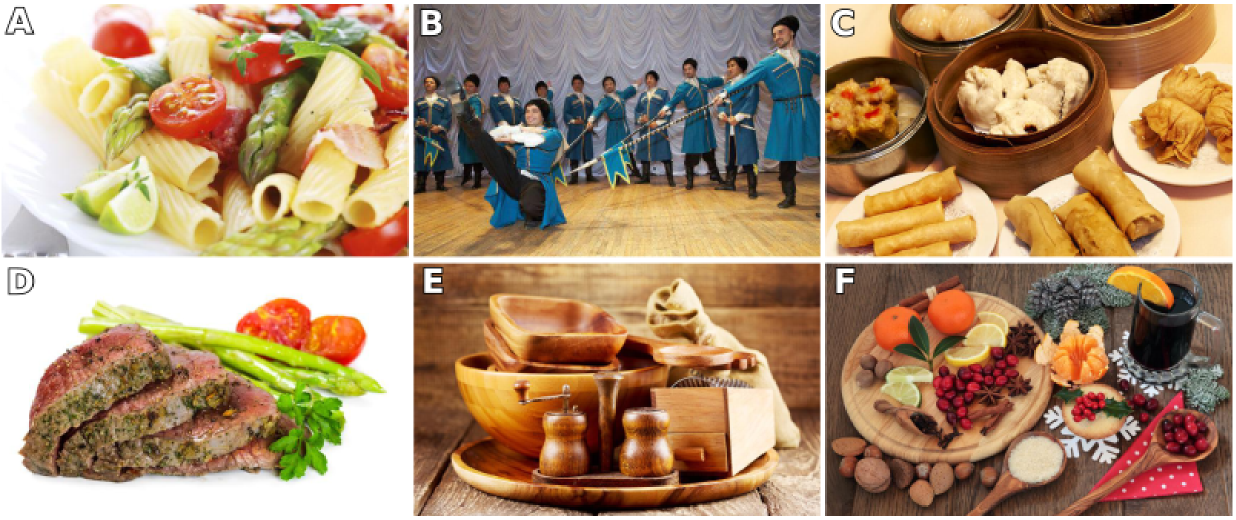

Figure 2. Examples of images used in initial tasks for the (A) *food* and (B) *objects* treatments of *intertask-food-objects*; (C) test tasks for *intertask-food-objects* and *frame-food-objects*; initial tasks for the (D) *food* and (E) *culture* treatments of *intertask-food-culture*; and (F) test tasks for *intertask-food-culture* and *frame-food-culture*.

obtain the accuracy of the optimal classifier:

$$\eta_* = \sum_{x \in X} \Pr\{\text{correct}|x\}\Pr\{x\} \tag{5}$$

$$= \sum_{x \in X} \left(\frac{1}{2} + \frac{|p_1(x) - p_2(x)|}{2(p_1(x) + p_2(x))} \right) \left(\frac{p_1(x) + p_2(x)}{2} \right) \tag{6}$$

$$= \frac{1 + \theta}{2} \tag{7}$$

Since no classifier can be more accurate than an optimal classifier, it follows that, for any *practical classifier* with accuracy $\eta_{\mathcal{A}}$, the bound $\theta \geq 2\eta_{\mathcal{A}} - 1$ holds.

Thus, we can establish a *lower bound* on θ by first building a classifier that infers the design shown to workers from their responses, and then measuring its accuracy.

EXPERIMENT 1

Setup

The first experiment we performed consisted of two sub-experiments which we will call “intertask” and “framing”, because they are respectively designed to measure intertask and framing effects. The experiment was performed on MTurk using 476 workers. Workers could only participate once in a given experiment, and could only participate in one of the experiments we present here. Workers were required to have had 90% of their previously submitted work on Mturk accepted by other requesters to participate. No other screening was applied to workers.

The experiment was conducted using a single HIT[7]. Workers who accepted the HIT were randomly assigned to one of the two sub-experiments. Both sub-experiments consisted of two treatments, which we call “food” and “culture”, and workers were also randomly assigned to a treatment, resulting in 119 workers per treatment. In both treatments of both sub-experiments, workers were asked to perform ten image-labeling tasks, each of which required the worker to assign five descriptive labels to an image. Labels had to contain at least two letters to be accepted. Workers were paid USD$0.45 for their work. In all cases, the last five tasks were identical among all treatments (and across sub-experiments), and were presented in the same order. These last five tasks constituted the “test tasks”, whose distribution of responses could admit intertask or framing effects.

In the intertask sub-experiment, we varied the first five tasks given to workers based on the treatment. Workers assigned to the “food” treatment were shown images of food (see Figure 2A), while workers assigned to the “culture” treatment were shown other kinds of cultural depictions, such as dance, sport, and music[8] (see Figure 2B). However, within a given treatment, workers were shown the same images in the same order. As mentioned, workers from both treatments then performed the same five test tasks. The test tasks consisted of images depicting meals of diverse ethnic origin (see Figure 2C). From the perspective of the worker, there was no distinction between the initial and test tasks.

In the framing sub-experiment, we did not vary the initial tasks, but instead, introduced a *framing slide* before the tasks. Depending on the treatment, the framing slide read “This research is proudly funded by The National Foundation for Nutritional Awareness”, or “. . . by the Global Foundation for the Recognition of Cultures”. The intention was to frame the task by providing the name of a (fictitious) funder, while using names that invite the worker to focus on different aspects of the image content. Initial tasks were still included to ensure that learning or fatigue effects were equalized between the

[7] On MTurk, a HIT (or Human Intelligence Task) is the smallest assignable unit of work. The use of “Task” in HIT differs from that used here; here labeling a single image is considered to be one “task”, and so a HIT is actually composed of 10 tasks.

[8] We do not assert that cultural practices can be cleanly separated into food-related and non-food related ones, nor is a strict division necessary for our analysis.

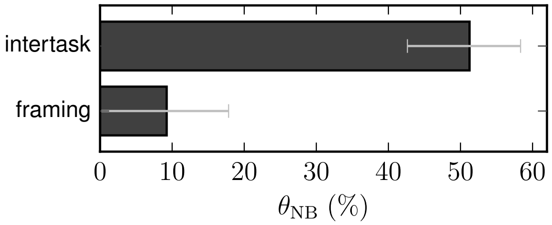

Figure 3. Lower bound total variational distance (θ_{NB}) between responses given by workers on test tasks in Experiment 1, when they have been shown different initial tasks (intertask), or been exposed to different framing (framing), as determined using naive Bayes classifiers. Standard error bars are shown.

sub-experiments. For the initial tasks, we used the same images as had been shown to the food treatment of the intertask sub-experiment. If intertask effects exist, any choice of images for the initial tasks will have an effect. This choice makes the food treatments of both sub-experiments very similar, so that the culture treatments of both sub-experiments can be viewed as perturbations from a similar baseline. The test tasks were the same as for the intertask sub-experiment.

Manual inspection showed that the vast majority of labels were of high-quality. Less than 1% of submissions were problematic, having redundant labels or entries that were not English words. We chose to include all submissions for analysis in the experiments presented here.

Results

Existence of intertask and framing effects.

Looking at the frequencies with which workers used words, we tested for heterogeneity between the two treatments of each sub-experiment. For the intertask sub-experiment, we found the responses for the two treatments differed significantly (by χ^2 test, $p < 0.001$), demonstrating that intertask effects do exist. However, we did not find significant effects due to framing ($p = 0.29$).

Testing for heterogeneity of word-frequencies was based on Pearson's χ^2 test with Yates' correction. In linguistics applications, the use of a χ^2 to test for heterogeneity between corpora has been called into question, because there is frequently heterogeneity *within* a given corpus [22]. This occurs because certain documents in a corpus will regard certain topics, leading to bursty word usage. While our setup is substantially remote from the applications for which this concern was raised, one could argue that the responses to a given treatment constitute a "corpus", and the responses of a given worker to a "document". To address this concern, we randomly split responses to given treatment into two sets, and tested for heterogeneity between the sets. In all cases, we failed to find heterogeneity *within* a given treatment (with $p > 0.3$ for all treatments and $p > 0.8$ for most). These results hold for Experiment 2, to be discussed later. Thus, the heterogeneity we observed *between* the intertask treatments indicates a meaningful effect.

Strength of intertask and framing effects.

Having affirmed the *existence* of intertask effects, we sought to measure their strength, i.e. θ. To measure θ we created a naive Bayes classifier based on a multinomial distribution over the word frequencies in worker's labels, and measured it's accuracy using leave-one-out cross-validation. We chose the naive Bayes classifier for three reasons. First, it performs well even when the number of features is large compared to the number of training examples [5, 16]. Second, there are no hyperparameters to optimize, which eliminates the need to partition the response data into dev and test sets. Third, the conditional independence assumption, normally undertaken for pragmatic reasons [39], is probably mild in relation to image labels, since they are likely to be less dependent on one another than in a coherent passage of text. We repeated our analysis using support vector machines and found very similar results.

To produce word frequency features, worker-submitted labels were split on white space and punctuation, and the resulting tokens were spell-corrected based on edit-distance to a dictionary of words, followed by stop-word removal and lemmatization. The spelling correction dictionary was compiled from a combination of WordNet [13] and words collected by crawling the World Food section of allrecipes.com.

Based on the performance of the classifier, intertask effects altered workers' responses, leading to (a lower bound of) 51.3% total variational distance in their response distributions (Figure 3). This represents a remarkably strong effect. Visually comparing two word frequency distributions is difficult, but Figure 1 provides an example of distributions having $\theta = 50\%$. In contrast, the effect due to framing, was not significantly different from zero (Figure 3). And so, intertask effects were stronger than the effects due to framing ($p < 0.001$)[9]. In other words, the *microtasks themselves* had a stronger effect than framing.

When observing the classifier's accuracy, in order to calculate θ, the number of correct inferences made by the classifier follows a binomial distribution. The fraction of successes yields the maximum likelihood estimate of the classifier's accuracy (used to generate θ), and we derived confidence intervals using the exact Clopper-Pearson method.

The results from the first experiment clearly show that intertask effects arise and are strong. On the other hand, we were scarcely able to observe any effect due to framing. This raises a few questions which we seek to answer using a second experiment. The first question is simply whether the result will replicate using different sets of images. If so, then as the next question, we may ask how intertask effects evolve as the worker performs tasks. Presumably, intertask effects wear off at some point, but is it after one task, two, or more?

Using the results from Experiment 1, we could look at the intertask effects for individual test questions, however, each

[9] To compare framing and intertask effects, we approximated the distribution of correct classifications as a normal distribution, and performed a (two-tailed) two-proportion z-test.

question uses a distinct image, whose own content may modulate the extent to which it can be influenced by intertask effects. We cannot disentangle the effects of task position from that of image content in Experiment 1.

In setting up Experiment 2, we also wish to increase the strength of framing. While it was remarkable that intertask effects were stronger than the framing effects elicited in the first experiment, perhaps the most meaningful comparison is one which shows just how extreme framing treatments need to be to produce effects on par with intertask effects. One factor that may have weakened the effects of framing in Experiment 1 is the inclusion of initial tasks between the framing slide and the test tasks. They were included to ensure that workers had always completed the same number of tasks before beginning the test tasks. However, their inclusion gives time for differential framing effects to subside, while also generating intertask effects that are *shared* between the treatments, potentially masking the framing effects that remain. In the next experiment, we measure framing effects immediately after the application of framing.

EXPERIMENT 2

Setup

In Experiment 2 we again had "intertask" and "framing" sub-experiments, but we included a third sub-experiment, "echo", which incorporated more extreme framing treatments. As before, each sub-experiment had two treatments, but this time they were "food" and "objects". Keeping a food treatment enabled us to perform deeper lexical analysis of both experiments (to be discussed in the results for Experiment 2). The experiment was again performed using a single HIT on MTurk, and involved 1666 workers. Workers were randomly assigned to a sub-experiment and treatment upon starting the HIT, resulting in 119 workers per treatment.

The intertask sub-experiment was similar to that in Experiment 1: workers performed a set of five initial tasks, whose images depended on the treatment to which they had been assigned, and then performed the test tasks, which were common to both intertask treatments (as well as to both treatments of the other sub-experiments in Experiment 2). In the food treatment, the initial tasks contained images of food (see Figure 2D). This time, care was taken to exclude any non-food items, such as utensils or place-settings, except, in some cases, for the plate supporting the food itself. In the objects treatment, initial tasks contained images depicting table settings and various non-food objects one might find in a kitchen, but no food (see Figure 2E). The images in the test tasks, depicted both food and non-food objects together (see Figure 2F).

In contrast to Experiment 1, we performed five replicates of the intertask sub-experiment, each time permuting the test tasks. The test tasks were "rotated" in such a way that, what was the first test task became the second, the second became the third, and so on up until the last test task, which became the first. In this way, each of the five positions was occupied by each test task, enabling us to disentangle the effect of test task position (relative to initial tasks) from test task content, and thereby see how intertask effects vary as a function of position in the task sequence.

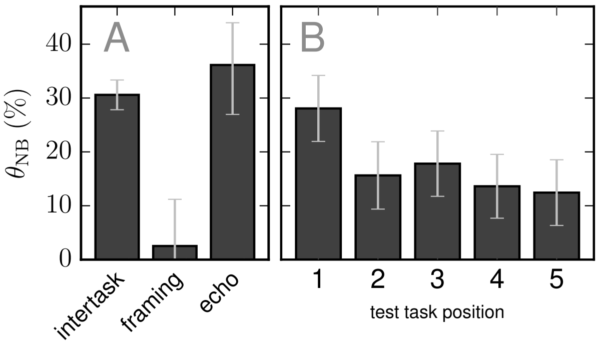

Figure 4. (A) Lower bound total variational distance (θ_{NB}), between responses given to test tasks in experiment 2 when workers were shown different initial tasks (intertask), or were exposed to different framing treatments (framing), or echoed framing treatments (echo). (B) Detail of the effects seen from showing workers different initial tasks (corresponding to intertask in panel A), now broken out for each of the five test task positions. As workers proceeded through test tasks, intertask effects waned but remained significant ($p < 0.05$) even for the fifth test task. Standard error bars are shown.

The framing sub-experiment was similar to that from Experiment 1, except that we omitted initial tasks altogether, so that the test tasks followed immediately after framing. This meant that workers would not have performed any tasks by the time they began the test tasks. While that will hold constant for both framing treatments, and thus not produce a differential effect, the workers may be more (or less) susceptible to influences during their first few tasks. One can expect framing effects to be stronger than in Experiment 1, because of the immediacy with which the test tasks follow the framing. The framing slide for the food treatment read "Funded by the laboratory for the visual perception of Food and Ingredients", and for the objects treatment it read "... perception of Objects and Tools".

We introduced the echo sub-experiment to produce the strongest framing effects. This sub-experiment was similar to the framing sub-experiment, but differed in two ways. First, the framing slide stated an explicit purpose for the study: "The purpose of this study is to understand the visual perception of Food and Ingredients" and "... of Objects and Tools". Second, before moving past the framing slide, workers were asked what the purpose of the study was, and had to respond by selecting the framing statement from among a short list in a combo-box. It is because workers had to echo the framing statement using a combo-box that we call this sub-experiment "echo".

Results

Existence and strength of effects.

As in the previous experiment, intertask effects led to significant changes in the words workers used to label images in the intertask sub-experiment ($p < 0.001$). Based on the performance of a naive Bayes classifier, the strength of the effect was (as a lower bound) $\theta = 30.6\%$ total variational distance (see Figure 4A). Again, it is difficult to visualize the distributions of word frequencies, but for reference, the distributions shown in Figure 1B have $\theta = 30\%$.

This time, the framing sub-experiment also showed a significant effect. Framing induced changes in the frequencies of word usage at significance (as determined by χ^2 test; $p = 0.0012$). However, the *extent* of the effect could not be distinguished from zero ($p = 0.37$) based on the performance of a naive Bayes classifier (Figure 4A). As in Experiment 1, we found that intertask effects were stronger than framing effects ($p < 0.001$).

It was only in the echo sub-experiment that framing effects were on par with intertask effects. In this sub-experiment, framing influenced workers responses ($p < 0.001$), and the performance of a naive Bayes classifier showed that it produced (lower bound) 36.1% total variational distance in worker responses (Figure 4A). The strength of effects due to echoed framing could not be distinguished from that of intertask effects ($p = 0.79$).

It is remarkable that intertask effects were on par with an explicit, actively reinforced statement of the tasks' purpose. Requiring the worker to reiterate the purpose signals our intent, as the requester, to ensure that the worker has taken note of it, possibly leading the worker to interpret the exchange as an instruction.

Persistence of intertask effects.

Taking advantage of the replicated intertask sub-experiment, we created a naive Bayes classifier for each test task, when it occurs at each test task position. Thus, for each of the five test task positions, we obtained five separate measures of the intertask effect strength (one for each of the test tasks, when occupying that position). We averaged these five measurements to produce the intertask effect strength for the given task position, and these values are shown in Figure 4B.

Not surprisingly, the first test task shows the strongest intertask effects, at 28.1% total variational distance. The effect appears to drop suddenly after the first test task, but remain significant right through until the fifth task position ($p = 0.030$). Remarkably, even after four intervening tasks, the effect of having performed different initial tasks shifts worker's responses to the fifth test task by 12.4% total variational distance.

Intertask effects and concept activation.

To gain further insight into the nature of intertask effects, we investigated the vocabulary that workers used to label test tasks. One might expect that, within a given experiment, those workers exposed to food (whether through framing or initial tasks) would label test tasks using food-related words more often. To test this we developed an automated approach to labeling food-words based on the WordNet knowledge base [13].

WordNet provides hyponym and hypernym relations between words. A hypernym is a generalization (for example, "bread" is a hypernym of "pumpernickel"), while a hyponym is a specialization. We took the set of food-words to be all those

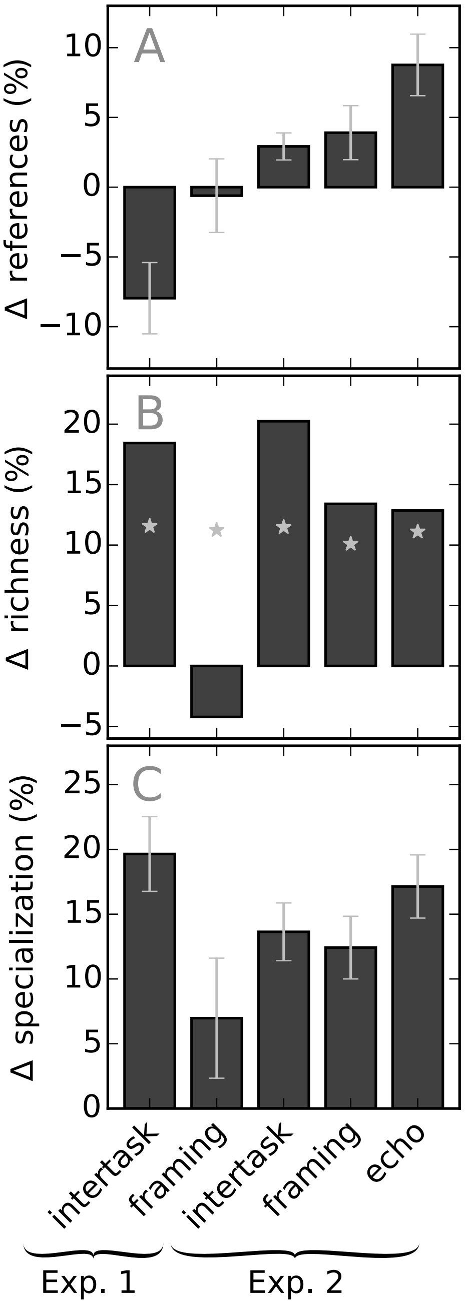

Figure 5. Workers exposed to food (either in framing or initial tasks) showed a different tendency to focus on food when labeling test tasks, and their vocabulary in reference to food was affected. In all three plots, positive values indicate a larger quantity for workers from the food treatment. (A) Exposing workers to food significantly changed the fraction of food references they provide, but did not necessarily *increase* it. (B) The number of unique food-related words (richness) was greater for food-exposed workers, except in the case of the framing sub-experiment of Experiment 1 (stars indicate the threshold for a significant deviation, $\alpha = 0.05$). (C) Food-exposed workers used more specialized words to refer to food. Standard error bars are shown in (A) and (C).

Exp. 1 intertask		Exp. 1 framing		Exp. 2 intertask		Exp. 2 framing		Exp. 2 echo	
spicy	26	indian	11	coffee	38	bread	18	apple	24
sauce	17	banquet	8	meal	34	wine	18	cheese	23
indian	15	spicy	7	cheese	34	cheese	16	wine	15
buffet	14	asian	6	apple	32	apple	14	coffee	14
exotic	12	variety	6	dessert	21	oil	12	oil	7
festival	-11	delicious	-6	cup	-30	table	-9	knife	-24
offering	-12	meat	-7	glass	-45	meal	-10	dinner	-26
statue	-15	festival	-7	table	-70	candle	-12	fork	-27
india	-20	spice	-7	candle	-74	dinner	-13	candle	-35
food	-56	food	-9	food	-80	food	-32	food	-55

Table 2. The five words whose frequencies as labels for the first test task increased (or decreased) the most between treatments of the given sub-experiments. Values indicate the absolute change in number of occurrences of the word, and positive values indicate that the food-exposed treatment used the word with higher frequency. Note that the word "food" is always the most suppressed among food-exposed workers.

reachable through a chain of hyponym relations from the word-senses `food.n.01` and `food.n.02`, in other words, all words denoting a specific kind of food (a total of 3590 words). To improve the coverage of words for less common foods, we augmented this set with all words discovered while crawling the World Food section of `allrecipes.com` which were not already in WordNet.

To validate the resulting set of food words, three independent annotators labeled 500 words selected from workers' responses as either food or non-food, and these were compared to the labels derived using the automated approach. Among those words, 26% were deemed to be food by the majority of annotators. Taking the majority label among annotators to be the correct one, the automatically-identified food-words had 88% correspondence to the human annotators. Treating the automated identification of food words as another annotator, there was an 82.4% agreement between annotators.

Using the automated labeling of food words, we found that, contrary to expectations, workers exposed to food (via prior tasks or framing) did not necessarily use more food-words when labeling images in the test tasks. In the intertask treatment of Experiment 1, workers from the food treatment actually used significantly[10] *fewer* food-related words during test tasks (Figure 5A). This finding rules out a seemingly-simple idea that workers emphasize content seen in earlier tasks. Seeing given content significantly influences workers' propensity to refer to it in subsequent tasks, but it does not necessarily *increase* it.

To deepen our understanding, we investigated workers' lexical richness in reference to food, that is, the number of *unique* food-related words used. Even if workers provide an abundance of food-related words, there can be less diversity, if, for example, workers repeat generic references to food. The *intertask* sub-experiments from both Experiments 1 and 2 showed that workers from the food treatments had greater lexical richness, in reference to food, than their counterparts (as much as 20% more) (Figure 5B). This is particularly noteworthy for the intertask sub-experiment in Experiment 1, because there, workers from the food treatment made fewer total references to food. Thus, although prior task exposure does not necessarily increase the propensity to identify content in later tasks, it does appear to activate vocabulary pertaining to content in the prior tasks.

To test the significance of the difference in sizes of food lexicons, we created a null model for each experiment, by assuming that responses for both treatments were in fact drawn from the same distribution. This was accomplished by pooling responses for both treatments of a given experiment, then drawing two bootstrap samples of 119 responses and calculating the difference in the size of their food lexicons. We repeated this 1000 times and took the 2.5th and 97.5th percentiles as the critical values for the rejection of the null hypothesis (the latter shown by stars in Figure 5B).

The above observations regarding lexical richness suggest that initial tasks might influence workers to use more *refined* or *specialized* words, when referring to aspects of content that had been present in the initial tasks. To test this directly, we used the hypernym and hyponym relations in WordNet to operationalize the notion of word specialization. Within each sub-experiment, we determined the relative specificity of the food-words, between the treatments using the following equation:

$$S(P,Q) = \frac{\sum_{w\in P}\sum_{v\in Q}\left(\mathbf{1}_{[w>v]} - \mathbf{1}_{[v>w]}\right)}{\sum_{w\in P}\sum_{v\in Q}\left(\mathbf{1}_{[w>v]} + \mathbf{1}_{[v>w]}\right)}, \qquad (8)$$

where P and Q are sets of words associated to different experimental treatments, and $\mathbf{1}_{[w>v]}$ evaluates to 1 if word w is more specific than (i.e. is a hyponym of) word v. The relative specificity lies within $[-1, 1]$; we report it as a signed percentage. In computing this quantity between two treatments, we first computed the relative specificity for the treatments separately for each test task, and averaged the results obtained across the five test tasks. To establish significance, we again used a null model, based on bootstrapping.

In all sub-experiments, workers from the food treatment used more specialized words, in reference to food (about 15% more; $p < 0.05$) (Figure 5C). Except in the case of the framing treatment of Experiment 1, these between-treatment differences in word specialization were significant ($p < 0.05$).

[10] Frequencies of food references within a treatment were calculated as an average across workers, and modeled as normally distributed.

It is interesting that such substantial increases in both the lexical richness and specialization of food-related words held for the intertask sub-experiment of Experiment 1, where, as mentioned, we observed that food-exposed workers made *fewer* references to food overall. These observations point to countervailing factors: one factor tending to activate the more specialized and less common food-related words (yielding greater lexical richness and specialization), and the other tending to suppress certain, presumably more common and generic words (yielding fewer food-related words in total).

This hypothesis is corroborated when we look at those words whose frequencies changed the most from one treatment to another (Table 2). The word "food", which is the most generic possible food-related word, was always *suppressed* among food-exposed workers, along with other very generic food references being suppressed too, such as "meal" and "dinner". In fact, for all experiments, "food" was the *most suppressed* word. Meanwhile, the most activated food-related words, among food-exposed workers, tended to refer to specific foods, like "apple", "cheese", and "bread".

DISCUSSION

Our results show that intertask effects exist, are very strong, and are on par with the effect of an explicit statement of a task's purpose reinforced by requiring the worker to repeat that purpose. This very surprising result immediately raises follow-on questions for future work: what are the psychological mechanisms at play? How are other kinds of tasks (i.e. other than image labeling) impacted by intertask effects? How can intertask effects best be used to produce optimal task design? We expect these questions will be fruitfully explored in future work.

While our setup does not allow us to tell what psychological mechanism(s) are responsible for intertask effects, we suspect that priming is involved. As mentioned, we believe microtasks create conditions conducive to priming, and this is what first lead us to suspect intertask effects might exist. Priming occurs when a prior stimulus causes a person to respond to a subsequent task with increased speed or accuracy, or with the ability to recognize briefer or noisier stimuli [36, 21, 18]. In the case of image labeling, pre-activation of vocabulary relating to prior tasks, through a priming mechanism, could lead to more rapid retrieval of those words in subsequent tasks. A worker may be more likely to enter those words which enter her mind first, so priming could lead to certain words being preferred. As the worker completes many tasks with a common theme, the worker's vocabulary relating to that theme will be increasingly activated, enabling more specialized words to be retrieved quickly, and therefore to be submitted as responses.

However, as we alluded to above, there appears to be countervailing factors: while richness and specialization of food-related words always increased (or were not significantly changed) among workers exposed to food in initial tasks, the total number of references to food significantly *decreased* in one case. Here we suspect *negative* priming to be at play. Negative priming occurs when, after exposure to a stimulus considered to be non-salient, subsequent recognition of the stimulus is inhibited [29]. In the case of our image labeling tasks, repeated exposure to images depicting food might cause the worker to stop regarding the basic presence of food as salient, and to direct her attention instead to the features that are specific to the food in a given image. This would explain the suppression of very generic references to food observed in Table 2. The balance between the inhibition of generic food-related words, and the activation of specialized ones, could lead to more or less food references overall, so this dual priming mechanism is consistent with our observations. Direct experimentation is needed to determine whether priming mechanisms, positive or negative, can explain intertask effects.

Prior to this study, it was not realized that earlier microtasks could have such strong effects on later ones. But our findings show this is an important design factor that should be considered by anyone using microtask platforms. The commonly employed practice of randomizing task ordering probably introduces a significant amount of noise due to intertask effects: even chains of two or three similar tasks, which will not be reliably eliminated in random permutations, could lead to the levels of influence we observed in our experiments. Preferably, a deeper understanding of intertask effects might allow them to be properly controlled.

But perhaps more importantly, our findings show that intertask effects might be leveraged obtain greater quality and reproducibility in crowdsourcing. A consistent goal in human computation is the elicitation of expert-level judgments from non-expert workers [24]. This has been achieved in some applications [32, 30, 35]. The distinction between experts and novices is partly attributable to specialized knowledge and heuristics. But experts also simplify tasks by more efficiently directing their focus toward salient features [20]. Using strategic task exposure during training, it might be possible to guide workers' focus and salience attribution, enabling expert-level judgment in a wider variety of crowdsourced applications.

Intertask effects are a new basic discovery in the quest to design reliable and reproducible tasks for human computation. We anticipate future work will yield techniques to control intertask effects, to reduce unwanted bias, and to tune the focus, diversity, and specificity of worker responses.

REFERENCES

1. P.D. Adamczyk and B.P. Bailey. 2004. If not now, when?: The effects of interruption at different moments within task execution. In *Conference on Human Factors in Computing Systems - Proceedings*. 271–278.

2. T. Batu, L. Fortnow, R. Rubinfeld, W.D. Smith, and P. White. 2013. Testing closeness of discrete distributions. *Journal of the ACM (JACM)* 60, 1 (2013), 4.

3. Henry K Beller. 1971. Priming: Effects of advance information on matching. *Journal of Experimental Psychology* 87, 2 (1971), 176.

4. A.J.a Berinsky, G.A.b Huber, and G.S.c Lenz. 2012. Evaluating online labor markets for experimental

research: Amazon.com's mechanical turk. *Political Analysis* 20, 3 (2012), 351–368.

5. P.J. Bickel and E. Levina. 2004. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* 10, 6 (12 2004), 989–1010.
6. S Chan, I. Diakonikolas, P. Valiant, and G. Valiant. 2014. Optimal Algorithms for Testing Closeness of Discrete Distributions.. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*. 1193–1203.
7. Dana Chandler and Adam Kapelner. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization* 90 (2013), 123–133.
8. J. Cheng, J. Teevan, and M. S. Bernstein. 2015. Measuring Crowdsourcing Effort with Error-Time Curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1365–1374.
9. L.B. Chilton, R.C. Miller, J.J. Horton, and S. Azenkot. 2010. Task search in a human computation market. In *Workshop Proceedings - Human Computation Workshop 2010, HCOMP2010*. 1–9.
10. Peng Dai, Jeffrey M Rzeszotarski, Praveen Paritosh, and Ed H Chi. 2015. And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 628–638.
11. J. Davis, J. Arderiu, H. Lin, Z. Nevins, S. Schuon, O. Gallo, and M.-H. Yang. 2010. The HPU. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*. 9–16.
12. S. Dow, A. Kulkarni, S. Klemmer, and B. Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. 1013–1022.
13. C Felbaum. 1998. *Wordnet, an Electronic Lexical Database*. Cambridge: MIT Press.
14. A. Finnerty, P. Kucherbaev, S. Tranquillini, and G. Convertino. 2013. Keep it simple: Reward and task design in crowdsourcing. In *ACM International Conference Proceeding Series*.
15. Avniel S. Ghuman, Moshe Bar, Ian G. Dobbins, and David M. Schnyer. 2008. The effects of priming on frontal-temporal communication. *Proceedings of the National Academy of Sciences* 105, 24 (2008), 8405–8409.
16. T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. 2009. *The elements of statistical learning*. Vol. 2. Springer.
17. E. Huang, H. Zhang, D.C. Parkes, K.Z. Gajos, and Y. Chen. 2010. Toward automatic task design: A progress report. In *Workshop Proceedings - Human Computation Workshop 2010, HCOMP2010*. 77–85.
18. D.E. Huber. 2008. Immediate Priming and Cognitive Aftereffects. *Journal of Experimental Psychology: General* 137, 2 (2008), 324–347.
19. G. Kazai, J. Kamps, and N. Milic-Frayling. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval* 16, 2 (2013), 138–178.
20. P.J. Kellman and P. Garrigan. 2009. Perceptual learning and human expertise. *Physics of life reviews* 6, 2 (2009), 53–84.
21. S. T. Kempley and John Morton. 1982. The effects of priming with regularly and irregularly related words in auditory word recognition. *British Journal of Psychology* 73, 4 (1982), 441–454.
22. A. Kilgarriff. 1996. Comparing word frequencies across corpora: Why chi-square doesn't work, and an improved LOB-Brown comparison. In *ALLC-ACH Conference*.
23. P. Kinnaird, L. Dabbish, and S. Kiesler. 2012. Workflow transparency in a microtask marketplace. In *Proceedings of the ACM 2012 International Conference on Support Group Work*. 281–284.
24. A. Kittur, B. Smus, S. Khamkar, and R.E. Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 43–52.
25. W.S. Lasecki, J.M. Rzeszotarski, A. Marcus, and J.P. Bigham. 2015. The Effects of Sequence and Delay on Crowd Work. *CHI – Human Factors in Computing Systems* (2015).
26. John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. 2010. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*. 21–26.
27. Sheena Lewis, Mira Dontcheva, and Elizabeth Gerber. 2011. Affective computational priming and creativity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 735–744.
28. W. Mason and D.J. Watts. 2009. Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation, HCOMP '09*. 77–85.
29. Susanne Mayr and Axel Buchner. 2007. Negative priming as a memory phenomenon: A review of 20 years of negative priming research. *Zeitschrift für Psychologie/Journal of Psychology* 215, 1 (2007), 35.
30. J.M. Mortensen, M.A. Musen, and N.F. Noy. 2013. Crowdsourcing the verification of relationships in biomedical ontologies. *AMIA Annual Symposium proceedings* (2013), 1020–1029.

31. D. Powers. 1998. Applications and explanations of Zipf's law. In *Proceedings of the joint conferences on new methods in language processing and computational natural language learning*. Association for Computational Linguistics, 151–160.

32. Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 254–263.

33. P.H. Thibodeau and L. Boroditsky. 2013. Natural language metaphors covertly influence reasoning. *PloS one* 8, 1 (2013), e52961.

34. G. Valiant. 2012. *Algorithmic Approaches to Statisitical Questions*. PhD Thesis. University of California at Berkely.

35. S.C. Warby, S.L. Wendt, P. Welinder, E.G.S. Munk, O. Carrillo, H.B.D. Sorensen, P. Jennum, P.E. Peppard, P. Perona, and E. Mignot. 2014. Sleep-spindle detection: Crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nature Methods* 11, 4 (2014), 385–392.

36. Clive Warren and John Morton. 1982. The effects of priming on picture recognition. *British Journal of Psychology* 73, 1 (1982), 117–129.

37. S.M. Wolfson and M. Lease. 2011. Look before you leap: legal pitfalls of crowdsourcing. *Proceedings of the American Society for Information Science and Technology* 48, 1 (2011), 1–10.

38. Ming Yin, Yiling Chen, and Yu-An Sun. 2014. Monetary interventions in crowdsourcing task switching. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

39. H. Zhang. 2004. The optimality of Naive Bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*, Vol. 2. 562–567.

40. G. K. Zipf. 1949. Human behavior and the principle of least effort. (1949).