

# Confusing the Crowd: Task Instruction Quality on Amazon Mechanical Turk

Meng-Han Wu, Alexander J. Quinn

Purdue University  
West Lafayette, Indiana, USA  
{wu784, aq}@purdue.edu

## Abstract

Task instruction quality is widely presumed to affect outcomes, such as accuracy, throughput, trust, and worker satisfaction. Best practices guides written by experienced requesters share their advice about how to craft task interfaces. However, there is little evidence of how specific task design attributes affect actual outcomes. This paper presents a set of studies that expose the relationship between three sets of measures: (a) workers' perceptions of task quality, (b) adherence to popular best practices, and (c) actual outcomes when tasks are posted (including accuracy, throughput, trust, and worker satisfaction). These were investigated using collected task interfaces, along with a model task that we systematically mutated to test the effects of specific task design guidelines.

## Introduction

Task interfaces are the primary link between requesters and workers on microtask platforms. Without a clear specification of what information is requested, and how to produce it, workers can hardly be expected to produce consistently good quality results. Advice, in the form of "best practices" abounds, but links between the guidelines and actual outcomes have only just begun to be measured (Jain et al. 2017).

Through our other work, we have long experienced the challenges of designing tasks. With early versions of systems, we received subpar results, and sometimes complaints from frustrated workers. In conversations with other requesters and workers, we have heard widespread complaints about poorly designed tasks. However, we are aware of no systematically obtained evidence of this supposed problem.

This paper presents a multidimensional analysis of task instruction quality, including (a) workers' perceptions of task quality, (b) adherence to popular best practices, and (c) actual outcomes when tasks are posted (including accuracy, throughput, trust, and worker satisfaction). The studies were performed with Mechanical Turk, but should apply to any microtask platform supporting freeform task design.

Besides communicating a request, task interfaces also form a semi-formal contract between the worker and requester. On Mechanical Turk, they typically express the criteria by which requesters will decide whether to approve submitted work, or reject it (and deny payment).

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The challenge of writing clear instructions is due, in part, to the diversity of workers' backgrounds, including cultural context, life experience, and educational level (Ipeirotis 2010). Also, unlike in-person workers, web workers normally have no ongoing investment in—or even awareness of—the overarching project the tasks are intended to support. Thus, extra care is necessary to ensure efficiency and clarity. Tasks should not require workers to understand any more than they need to complete the task at hand. This poses an impediment to the attainment of quality results, especially for less experienced users.

We view this as a problem because it impedes users from realizing the potential benefits of delegating data-intensive work to online workers. We hope that with the increasing popularity of crowdsourcing, even novices without expertise in interface design and crowdsourcing can participate in a freely flowing market for labor transfer. Thus, it is important to understand what is bad instruction and how to write instructions that can be efficiently understood by people.

The key contributions of this work are as follows:

1. Evaluations of tasks scraped from Mechanical Turk measured (a) workers' assessment of the task quality, and (b) adherence to established best practices for task design.
2. Actual effects of specific best practices on desired outcomes (e.g., accuracy, worker acceptability, etc.) were measured for a single task by creating systematic mutations and measuring workers perceptions (by inspection) and actual performance (by posting the tasks).
3. Results of the experiments show that (a) adherence to best practices can affect outcome, but that (b) workers are more resilient to flaws in task quality than current popular belief might suggest.

## Background

The impact of interface design (in general) on human performance is well-established (Nielsen 1994; Shneiderman 2010). When building systems that leverage crowd work for new application types, or using new strategies, task design can be a formidable challenge. Despite the importance of task design, research about improving quality of data obtained from crowds has long been dominated by studies of incentives (Ho et al. 2015; Horton and Chilton 2010; Mason and Watts 2010; Rogstadius et al. 2011; Yin, Chen,

and Sun 2013; Shaw, Horton, and Chen 2011, etc.) and optimization algorithms (Ipeirotis, Provost, and Wang 2010; Barowy et al. 2012; Dai, Mausam, and Weld 2010, etc.). The assumption behind these approaches is that workers are lazy, and will submit poor quality work, unless the reward structure incents them to do otherwise. This is at odds with the prevailing belief in usability circles, that design is the responsibility of designers.

## Importance of instructions

Unclear instructions are often cited as a main challenge for crowd workers. Misunderstandings can adversely affect worker's rating by forcing them to abandon a HIT or leave it incomplete (Silberman et al. 2010). A subsequent study by Schulze et al. reported that task clarity impacts workers' acceptance of a task (Schulze et al. 2011). Clarity of instructions has also been found to improve usability for low-income workers in India (Khanna et al. 2010).

Poorly designed instructions affect not only the result quality (Kittur et al. 2013) but the relationship between requesters and workers (McInnis et al. 2016). Long-standing policies of Mechanical Turk give workers little recourse in case of disputes over the correctness of their work. The task instructions are the only explicit expression of the task requirements, and hence serve as an informal contract.

## Studies of task design

A recent<sup>1</sup> by Gadiraju et al. (2017) investigated task clarity using sample of 7,100 tasks from Mechanical Turk. Workers for rated them for *clarity*, which the authors frame in terms of "goal" (what is needed) and "role" (steps to achieve it). The sampled task instructions were found to be moderately clear. They also tested whether a computational model could predict clarity based on a set of features, including task metadata, type, content, and readability. Using the model to analyze clarity on the site over a 1-year period, they found that no monotonous trend in the overall average task clarity over a year. (In other words, clarity of tasks waxes and wanes over time.) The goals of that work are similar to ours—to better understand the role of task clarity in microtasks. This paper is complementary in that it tests interactions with popular best practices (i.e., by inspection and by a random selection experiment) and with desirable outcomes (i.e., accuracy, worker acceptability, etc.).

Jain et al. recently analyzed the 27 million microtasks from the marketplace. Their examination included a spectrum of facets, including marketplace dynamics, task design, and worker behavior. Their work examined a far greater number of task designs, but using mostly automated methods. For example, they focus on what components in a page—e.g., examples, images, number of inputs—predict agreement (Jain et al. 2017).

Alagarai et al. studied the task design in a psychological way. They modified the *visual saliency* of the target fields and *working memory requirements* of the task to see how

<sup>1</sup>The study by Gadiraju et al. was published in July 2017. We learned of it just prior to the final submission of this paper.

these cognitively inspired features affect workers. The results suggested that the performance of crowd workers can be maximized by using these features for task design (Alagarai Sampath, Rajeshuni, and Indurkha 2014).

Alonso and Baeza-Yates (2011) explored the design and execution of relevance judgments. The *design* here has a broader meaning, which not only refers to the interface design but also includes the task mechanics such as when to post the task and how to filter the workers.

## Complementary approaches

If task design is an acquired skill, then one possible mitigation is to aid novice requesters in producing clearer tasks. Fantastkic was a system developed to test that approach. It uses a wizard interface to elicit task requirements from requesters, and allowed them to preview the task before posting. It led to more consistent results, but was limited to a narrow set of task types (Gutheim and Hartmann 2012).

Instructions can only be effective if they are read. One approach is to use a modified *Instructional Manipulation Check* (IMC) (Oppenheimer, Meyvis, and Davidenko 2009) to check if workers are reading instructions carefully (Kapelner and Chandler 2010; Goodman, Cryder, and Cheema 2013). These are often referred to as "attention checks" in the context of crowdsourcing (Peer, Vosgerau, and Acquisti 2013; Hauser and Schwarz 2016, etc.).

As mentioned above, incentives are often a focus of research on quality in crowdsourcing. Besides incenting good-faith effort, some incentive structures can also increase workers' intrinsic motivation or help to select personality profiles that are well-suited to the task. Studies of worker's intrinsic motivation (Law et al. 2016) found that workers tend to be more productive if tasks are framed as something meaningful. For tasks that benefit from workers with particular motivation profiles, it is possible to design incentives to attract those workers (Hsieh and Kocielnik 2016).

While our study was focused on the task design—the instructions and form fields a worker sees after agreeing to do the task—others have tested the effects of other parameters, such as the number of labels requested per image, number of HITs posted at once, and the reward. Huang et al. generated behavioral models to predict the results of image labeling task from those factors (Huang et al. 2010). They showed that simple models can accurately predict the quality of output per unit task. On the other hand, Grady and Lease tested the effects of a similar set of factors (e.g., query, terminology, pay, bonus) on accuracy, cost, and time spent, for the task of relevance assessment (Grady and Lease 2010). Although their findings were largely inconclusive, they identify obstacles and potential paths for future investigation.

## Evaluation of existing HIT instructions

To assess the status quo of task design on Mechanical Turk, workers were engaged to assess the quality of tasks found on the site, based on screenshots.

## Justification

The goal was to acquire descriptive metrics of existing tasks, and learn how the tasks would perform. Several possible ap-

**1. Is the HIT written with vocabulary that you understand?**  
 Very hard to understand    Somewhat hard to understand    Neutral    Somewhat easy to understand    Very easy to understand    Not applicable    Not sure

**2. Does the HIT clearly specify what the requester wants?**  
*For this question, assume you understand all of the vocabulary.*  
 Very unclear    Somewhat unclear    Neutral    Somewhat clear    Very clear    Not applicable    Not sure

**3. Is the visual design of the HIT layout easy to follow?**  
*For example, can you find where to start, and find the information you need within the HIT? Does the design make you feel comfortable to do the HIT, or is it overwhelming?*  
 Very hard to follow    Somewhat hard to follow    Neutral    Somewhat easy to follow    Very easy to follow    Not applicable    Not sure

**4. How likely would you be to accept this HIT?**  
*Assume the reward amount is reasonable*  
 Very unlikely    Somewhat unlikely    Neutral    Somewhat likely    Very likely    Not applicable    Not sure

**5. How well do you think you would enjoy doing this HIT?**  
 Very much dislike    Somewhat dislike    Neutral    Somewhat enjoy    Very much enjoy    Not applicable    Not sure

**6. How confident are you that if you did the HIT, you could provide correct answers?**  
*In other words, if you did the HIT and did your best, how confident are you that your work would be what the requester wants?*  
 Very unconfident    Somewhat unconfident    Neutral    Somewhat confident    Very confident    Not applicable    Not sure

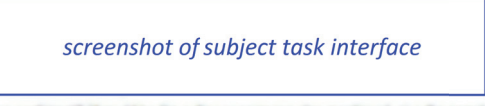
Read the article on the following page and specify the sentiment found for each sentence.  
  
 (This screenshot is a placeholder for the actual task interface.)

Figure 1: This questionnaire was used to measure workers’ perceptions of task quality for scraped HITs.

proaches were considered.

The metrics of interest (Figure 1) all relate to interactions with the workers (e.g., whether they understand the vocabulary) so any approach would need to involve workers. Thus, a computational analysis (i.e., NLP) would be inappropriate.

Ideally, we could have measured the actual uptake rate (% of workers who accept after previewing), enjoyment, and accuracy. However, many HITs are backed by web applications which could not be easily replicate without access to requesters’ code. Directly engaging with requesters was ruled out, since participation bias would be inevitable.

### Instruction Metrics

The questions used to elicit workers’ assessments of tasks (Figure 1) were based on two types of metrics: descriptive (properties of the task itself) and prospective (workers’ prediction of outcomes related to accuracy, worker acceptability, etc.).

**Descriptive metrics** Although most instructions on Amazon Mechanical Turk are within the scope of procedural instructions, they are different in the way that people are creating crowdsourcing tasks for different purposes. Therefore, common guideline or tips for creating instructions are not applicable to these tasks. We chose three descriptive metrics that are relevant to instructions in general (Wright 1998) and also applicable to crowdsourcing.

- Vocabulary use** in instructions. The difference in domain knowledge influences how people interpret and comply with instructions. Not only technical jargon but uncommon vocabularies can fail to communicate. The vocabulary use in instructions is critical; it is the fundamental element to compose a sentence and the key to helping workers understand the instructions.
- Data specification.** The specification of the desired data is critical for microtask crowdsourcing, because it is widely used for data collection. Workers are at high risk of being rejected if they do not know the criteria of data and understand what they should submit.
- Logical order** of the instruction layout. People form mental model in the process of reading instructions, they aim to find the information that fills the slot in their current mental schema. Any contradiction may lead to confusion and misunderstanding. Also, the order of instructions will be critical especially when people are unfamiliar with the task.

**Prospective metrics** The other three questions were designed to measure workers’ *expected* reactions, if they encountered the HIT. These are based on the outcomes mentioned throughout this paper: accuracy, throughput, trust, and worker satisfaction.

- Confidence** that the worker “could provide correct answers” is a proxy for accuracy. This presumes that confident answers are more likely to be accurate than unconfident ones.
- Enjoyment** is important because we presume that workers who enjoy a task will be more likely to continue it, thus learning to perform the task well and contributing to accuracy and throughput.
- Acceptance** refers to the likelihood that a worker would accept the task after previewing it. This is important for throughput. (In this paper, “acceptance” refers to a single worker, while “uptake” refers to the proportion of all workers who accept a HIT after previewing it. They refer to the same thing.)

The prospective metrics are essentially *expressed preferences*. These were used, despite the inherent limitations, because revealed preferences would be infeasible to observe for any substantial sample of in-the-wild tasks.

### Data collection

The metrics described above were measured using a 5-point Likert scale presented directly above a screenshot of the scraped task design (Figure 1). One task design was presented in each of our HITs. Workers were allowed to rate multiple HITs, but were not required to rate all of them. The evaluation form was available on Mechanical Turk for 3 days. A total of 167 distinct workers rated 135 HITs, with 3 judgments per HIT (405 judgments total). Workers were paid \$0.50 per HIT (average hourly rate: \$10.70 per hour).

**Sampling of scraped HITs** Scraped HITs were sampled to represent the population of *requesters*, rather than the set

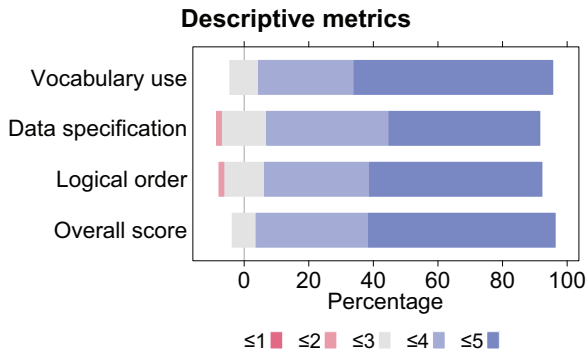


Figure 2: Most HITs were evaluated positively by workers on the descriptive metrics. (Note: Definitions of each factor were given on the previous page. Display order in this figure and Figure 3 are different from the questionnaire in Figure 1.)

Factor	Mean	Std. Dev
Vocabulary use	4.25	0.70
Data specification	3.99	0.78
Logical order	4.07	0.79
Overall score	4.10	0.66

Table 1: Descriptive metrics, mean of 3 ratings  $\times$  135 HITs

of all HITs or HIT groups available at a given time. Top requesters account for more than 30 percent of the overall activity of the market (Ipeirotis 2010). To avoid picking too many HITs from the top requester and neglecting the minority, we categorized the HITs by the requesters name. There were 1300 requesters in total, from which we randomly selected 135 requesters, and one HIT per requester (135 HITs total). Web-scraped HITs that were obviously not answerable, such as totally blank one, consent form and those which hide instructions in preview mode or need to link to an external website for further information, were filtered out. Each screenshot was collected before the task started. While we were collecting screenshots, we also collected requesters’ history of HIT post by using Mechanical Turk-Tracker (Ipeirotis 2010).

## Results

Results were analyzed based on the average of the 3 judgments, with values between 1 and 5. The most positive value (e.g., “very easy to understand”, etc.) was coded as 5.

### Descriptive metrics

Among the three descriptive metrics, “Vocabulary use” had the highest mean score (4.25 out of 5.00). This indicates that the overall vocabulary used in these HITs was easy for workers to understand.

The lowest mean scores were for *Data specification* (3.99) and *Logical order* (4.07), respectively. (See Table 1.)

Most HITs had a mean score between 3.00 and 5.00 for all of the descriptive metrics, with relatively high proportions

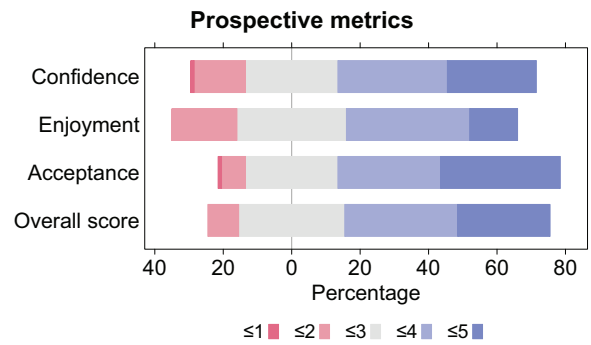


Figure 3: Prospective metrics were widely distributed. The most positive expected outcome was *acceptance* (of a HIT if it were offered). The most negative was enjoyment. This might suggest some degree of willingness to do tasks they do not enjoy.

Expectation	Mean	Std. Dev
Acceptance	3.33	1.07
Enjoyment	3.09	0.92
Confidence	3.60	0.98
Overall score	3.35	0.93

Table 2: Prospective metrics, mean of 3 ratings  $\times$  135 HITs

between 4.00 and 5.00 (Figure 2.) The mean score across all of the descriptive metrics was 4.10. These results suggest that the instructions created by most requesters on Amazon Mechanical Turk may actually be designed adequately, despite the prevalent concerns about poor task design.

### Prospective metrics

The means of each of the three prospective metrics were all over 3.00. The average of the three together was 3.35. This indicates an overall positive assessment with respect to workers’ expected reaction if they encountered these HITs. This is not unexpected. If most workers truly felt that they would not accept most HITs, and would be unconfident of their accuracy and enjoyment when they did accept HITs, the market would not be expected to function productively. However, variation in the response distribution suggests that differences among tasks may affect these metrics. For each prospective measure, significant proportions of the HITs received average scores between 0.00 and 3.00. (Table 2.) For those HITs, workers did not express confidence that they would be accepted, enjoyed, and/or performed correctly.

Using the results in these two parts, we performed a linear regression test to find whether there is a linear relationship between the factors we measured in previous part and workers’ expectation toward instructions. We summarize the results in Table 3.

Except for Logical order v.s. Confidence, the adjusted  $R^2$  of each pair is larger than 0.3, which means, from workers’ view, a portion of the variation in their expected reactions if the HITs were encountered (prospective metrics) is

	Acceptance	Enjoyment	Confidence
Vocabulary use	0.369	0.391	0.339
Data specs	0.429	0.381	0.488
Logical order	0.393	0.427	0.243

Table 3: Adjusted  $R^2$  of each pair from linear regression test,  $p < 0.05$

accounted for by the quality of the instructions (descriptive metrics). If the prospective metrics are predictive of workers' actual reactions, then this would suggest that the quality of instructions do affect the outcomes (accuracy, worker acceptability, etc.).

Instructions are not the only property expected to influence these outcomes (Schulze et al. 2011). Therefore, an adjusted  $R^2$  of 0.3 indicates that, of the variance that can be explained by instructions, these facets of task design quality have an important role.

Causation cannot be inferred from this study, since a random selection design was not feasible for this study with in-the-wild task designs. In the next section, we present a separate, complementary study, which used random selection (of best practices) for a model task design.

### Adherence to best practices

For an objective assessment, all of the scraped HITs were checked for adherence to the following best practices:

1. **BULLETS.** Use bullets for steps, data items, or rules.
2. **MEANING.** Criteria and meaning of inputs should be concretely defined.
3. **FORMAT.** Specify formatting requirements explicitly.
4. **TOOLS.** Specify how the task should be performed, including any tools or methods required.
5. **EXAMPLES.** Use examples to illustrate expectations.
6. **EXPLANATIONS.** Explain criteria for acceptance in detail and clearly state what kind of errors would trigger a rejection of the HIT to avoid unfair rejections.
7. **INDEPENDENCE.** Ensure that every question will be answerable, regardless of the answers to the others.

These guidelines are recommended by several guides and blogs written by experienced requesters. Our primary source was the Amazon Requester Best Practices Guide<sup>2</sup>, since its authors presumably have a particularly strong interest in helping requesters produce task instructions that will help the market function productively. We also surveyed several others guides, including ProPublica Guide to Mechanical Turk<sup>3</sup>, Mturk blog<sup>4</sup>, Crowdsourcing.com Do's and Don'ts<sup>5</sup>,

<sup>2</sup>[https://mturkpublic.s3.amazonaws.com/docs/MTURK\\_BP.pdf](https://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf)

<sup>3</sup><https://www.propublica.org/article/propublicas-guide-to-mechanical-turk>

<sup>4</sup><https://blog.mturk.com/>

<sup>5</sup><https://www.crowdsourcing.com/blog/2012/07/hit-layouts-dos-and-donts-part-1>

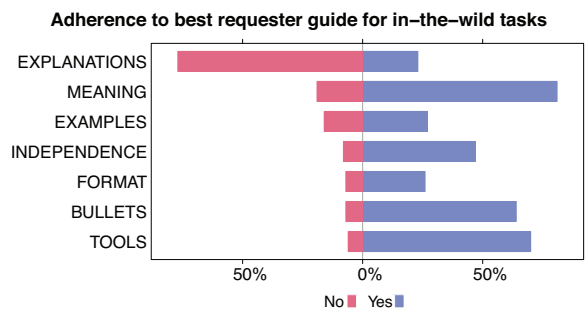


Figure 4: The 135 task instructions were categorized by an independent rater to determine which of the guidelines were followed. Tasks for which a particular guideline would be irrelevant or could not be evaluated were marked as N/A, and are excluded from the given column. These guidelines are explained in the text.

a guide by Edwin Chen<sup>6</sup>, detailed forum posts found at Quora.com<sup>7</sup> and TurkerNation<sup>8</sup>.

The seven guidelines below were widely repeated across these sources. Note that we excluded guides that focus on compensation or ethical considerations (e.g., WeAreDynamo guidelines, etc.).

These guidelines do not serve as golden rules for creating instructions. Some of the guidelines could not be applied to certain kinds of tasks. For example, answering a survey does not need examples for inputs or specify tools or methods. Instructions that can be clearly explained in one sentence do not need to use bullets for writing. Therefore, before the rating, we set up rules to separate these cases out and do the measures for those instructions which apply. An expert rated the 135 instructions in 5 hours, and the result is in Figure 4. From the results, except "Explain accept/reject criteria and approval process," in other criteria, the proportion of requesters who used it was more than the percentage of the requester who did not. Up to 81% of the requesters clearly explained the criteria and meaning of input, while 77% of requesters did not explain the accept/reject criteria. These results show that the sampling instructions follow most of the sorted guidelines, which adds the validity for "negative" results in previous sections.

### Discussion

We set out to document the problems with task design that have been widely presumed to be prevalent. We evaluated the instructions in three different ways: (a) descriptive metrics rated by workers, (b) prospective metrics rated by workers, and (c) adherence to best practices.

**Descriptive metrics** were more positive than expected, across all three. For 92.6% (125/135) of the HITs, scores were between 3.00 and 5.00, indicating that from the workers' perspective, most instructions were properly designed.

<sup>6</sup><http://archive.is/tBmxM#selection-21.0-32.0>

<sup>7</sup><http://qr.ae/Tbc2qa>

<sup>8</sup><http://turkernation.com/showthread.php?21300-How-to-improve-the-poor-task-instruction>

One possibility is that our metrics did not include some other metrics that bother workers. It is possible that additional factors might emerge only in the actual performance of the tasks, a situation which we were unable to measure. Another possibility is that confusing tasks cause memorable annoyance to workers that is disproportionate with the actual problem. Given the productive functioning of the market, we find this explanation plausible.

**Prospective metrics** revealed much greater variation, compared with the descriptive metrics. In this case, the possibility of missed metrics would not be a factor, since the prospective metrics are related to desirable outcomes (accuracy, worker acceptability, etc.). One possible explanation for the difference in variation between prospective and descriptive metrics is that workers' response to the HITs might be influenced by exogenous factors (e.g., interest in the topic) that are unrelated to the quality of the task design.

**Adherence to best practices** was relatively high. The most violated guideline was *Explanations*; many requesters do not explicitly explain their criteria for acceptance of results.

The positive assessment of the scraped task instructions contradicts our observations and experiences, and those of many other crowdsourcing researchers with whom we have discussed this. One possible explanation is that as researchers, the interfaces we need to create are inherently more demanding, and thus our experiences are not representative.

Another possible factor is that in our evaluation, workers only inspected static screenshots of task interfaces, and did not perform them under organic conditions. Thus, they may not have encountered the problems that would cause frustration due to poor instruction quality. In other words, our evaluation measured expressed preferences rather than revealed preferences. Measuring workers' actual behavior while doing these tasks would normally be impossible.

Since usability is well-known to impact performance and satisfaction, and learning to design interfaces takes some combination of experience and possibly aptitude, then those requesters who are unable to achieve good results in their initial attempts may be expected to leave, and try other avenues. The type of task may also affect this, since some types of tasks are much easier to design for than others, in our experience. When requesters with such tasks receive poor results, they may also leave. What would *remain* would be only the requesters with the resources to create good quality tasks, and with tasks that are amenable. Therefore, the assessment of the scraped task instructions is biased by requesters who "survived" and are still active on the platform.

This theory can be framed in terms of *survivorship bias* (Brown et al. 1992), a type of selection bias in experiments. The stereotypical example of survivorship bias is a pharmaceutical trial in which participants who do not benefit from a treatment are allowed to exit the study, leading to a final evaluation based on participants who disproportionately benefited. In our case, we are supposing that requesters who did not benefit from the help of crowd workers—perhaps because the requesters' task designs were inadequate and led to poor data—chose to stop using the platform, leaving only

proficient task designers.

These assumptions are supported by the historical data from MTurk-Tracker (Difallah et al. 2015), a site, which monitors activity on Mechanical Turk. Of the 135 requesters covered by our sample, 128 had posted HITs in the past. If task design is an acquired skill, then inexperienced requesters would be expected to produce less clear instructions than experts. More evidence is needed to test this theory.

## Systematically mutated task instructions

To see how instructions influence workers, we systematically mutated instructions based on the same best practices used in the previous study. The goal was to understand how these influence actual outcomes (e.g., accuracy, worker acceptability, etc.). We also measured the mutated tasks using the questionnaire from the previous study, though this part was inconclusive.

## Methods

A single model task was used (Figure 5). It was chosen because it met a few necessary criteria: (a) Ground truth is known. (b) An accurate response depends on careful reading of the instructions. (c) All of the best practices are directly applicable, and can be applied independent of one another.

The task asked workers extract information from a picture and provide the right information based on the given constraints. Note that the purpose of this experiment is to study how the design principles affect workers' actual performance. Since wide variation in the task would make this study infeasible, we kept the task contents static for all workers (except for the variations in adherence to best practices). Although the results would be more generalizable if more task types were tested, we believe this study design still advances understanding of how task design guideline affect outcomes.

To effect a random selection experiment, we selectively followed one of the design principles at a time, while deliberately violating the rest. For example, we can choose whether to use bullet when writing instructions.

To validate that the mutations matched our study design, an expert rater (graduate student from outside our lab) reviewed all of the variations and confirmed that they meet or do not meet the best practices, as we intended. (In other words, the code for mutating the tasks works correctly.)

We measured workers' performance in the following ways: workers' uptake (i.e., meaning the percentage of workers who accept versus those who viewed); workers' enjoyment of doing the task; time spent on the task; and the accuracy of the results.

Each version of the task adhered to one of the principles while deliberately violating the others. For the control, we used a version that deliberately violated all of the best practices. Thus, eight versions of the instructions (1 control + 7 experimental)—all requesting the same work—were used. To measure workers' enjoyment, we added a question at the bottom of the instructions and asked them to report their feeling. Each worker was only allowed to do the task once.

### Deliberate violation of all guidelines

(bad)



Find the person in this image who meet the following criteria: 1. Wearing jewelry 2. Blonde hair 3. Successful actor/actress. Put the name of the person who matches the criteria in input box A.

Search on internet. Put the URL of the wiki page of the person who meets these criteria in input box B. All URLs have to be shortened.

A. Put the name here

B. Put the shortened URL of the wiki page of the person who meets these criteria here.

BULLETS

MEANING

FORMAT

TOOLS

EXAMPLES

EXPLANATIONS

INDEPENDENCE

### Compliance with all guidelines

(good)



1. Find the person in this image who meet the following criteria

- Wearing jewelry
- Blonde hair
- Academy award (Oscar) winner or nominee in any year

2. Search on google. Put the shortened URL of the wiki page of the person who meets these criteria in input box A.

3. All URLs have to be shortened by using Google URL Shortener  
[www.exampleurl.com](http://www.exampleurl.com) → <http://goo.gl/...>

4. Approval for this HIT will be completed with 24 hours, and will be based on the following criteria:

- The person chosen must actually meet the specified criteria, according to information in the photo and/or on their Wikipedia page

Submissions that do not meet these criteria may be rejected.

A. Put the shortened URL of the wiki page of the person who meets these criteria here.

Figure 5: Task used in the last part. Instructions in the left is the control group. We use the guidelines to modify the task and create 7 different instructions. Each version of instructions only applied with one design principle.

### Results for systematically mutated instructions

A total of 120 unique workers performed the task. Each of the 8 versions of the task (7 best practices + 1 control) was performed by 15 workers, who were paid \$0.75 each (average hourly rate: \$9.59 per hour).

We performed a one-way between-subjects ANOVA to test whether there exist significant differences between each group. For time spent ( $p = 0.457$ ) and satisfaction ( $p = 0.865$ ), there are no significant differences between each group ( $p > 0.05$ ). This was not unexpected since there are many other factors involved. Time spent may be influenced by a worker's environment or overall capability to perform the task. Satisfaction may be influenced by a worker's interest in the task content, or the adequacy and importance of the reward (Kaufmann, Schulze, and Veit 2011; Rogstadius et al. 2011). We believe this experiment may not have been powerful enough to determine if there was an effect. As for uptake rate and accuracy, we found significant differences among the groups ( $p < 0.05$ ).

We use Dunnett's test as post-hoc test to find which practice affects uptake rate. The results are reported in Table 4. From the table, only the absolute difference of the instructions with explained accept/reject criteria is larger than Critical Difference, which means applying this practice significantly affects workers' uptake rate. However, it also means that applying this practice lowers workers' willingness to accept the task, because the control group has the highest

Bullets	Meaning	Format	Tools	Examples	Exp	Ind
0.195	0.262	0.209	0.182	0.108	0.454	0.199

Table 4: Uptake absolute differences between control and experiment groups. Critical Difference = 0.283. "Exp" = "Explanation" (of acceptance criteria); Ind = "Independence" (of questions in the HIT).  $p < 0.05$ .

uptake rate among all groups. To understand the issue, we count the words in each group and find that the instructions with accept/reject criteria have 199 words, which is the largest word count among the group and is much larger than the average word count of the remaining groups (90.7). Based on this observation, the significance might be caused by the length of the instructions. Workers are likely to skip the content-heavy task and thus lower the uptake rate. On the other hand, the control group only has 89 words, which is nearly close to the average word count (90.7). The differences might be too subtle to affect a typical worker's impression of the task or explain the results, which echoes prior findings that workers prefer short and straightforward task (Schulze et al. 2011).

**Lesson:** Compactness has the greatest impact on uptake, among all of these guidelines. To reduce job

Bullets	Meaning	Format	Tools	Examples	Exp	Ind
0.033	0.100	0.100	0.300	0.200	0.033	0.200

Table 5: Accuracy absolute difference between the control group and experiment groups. Critical Difference = 0.320. “Exp” = “Explanation” (of acceptance criteria); Ind = “Independence” (of questions in the HIT).  $p < 0.05$ .

turnaround time, keep tasks short.

For accuracy, we adopt the same method to compare experiment groups with the control group. The results show that there is no significant difference between the control group and experiment groups because all the absolute differences are less than Critical Difference, see Table 5. We then use Tukey’s method to make pairwise comparisons for all possible pairs. The critical difference for this method is 0.372. Results show that there are significant differences between Tools and Explanations (difference = 0.4 > 0.372), Tools and Independent (difference = 0.5 > 0.372), Examples and Independent (difference = 0.4 > 0.372). In our task, we ask workers to provide shortened URLs. However, many workers did not shorten the URL as requested. From the results above, we find that two guidelines—Tools and Examples—are related to the shortened URLs. These principles may help workers enter the correct results.

**Lesson:** For accurate results, explicitly state the tools needed for the task, and give concrete examples.

In this experiment, we systematically modified the instructions based on the seven design principles. Eight different sets of instructions with the same task type were generated. From the results, we find that the design principles do not affect worker’s time spent on a task and enjoyment. However, they directly or indirectly affect worker’s uptake and accuracy. For example, applying these design principles changes the number of words in the instructions and affect workers’ uptake. Workers are likely to skip the content-heavy instructions even when the task type is the same.

On the other hand, applying these principles directly affect the accuracy of the results. For example, the common mistakes made by workers were that they did not shorten the URL. In that case, we find a significant accuracy improvement between the instructions applied with design principles that are related to how to shorten URL and those which did not apply. Based on the result above, it implies that the design of instructions does have an influence on worker’s behavior. However, it does not mean that using certain principles guarantees the quality of the result. Requesters should consider the requirements of the task and appropriately apply the principles in instructions design based on different situations.

**Ratings.** As before, each HIT was evaluated by 3 workers. We compared HITs on the average of 3 ratings. Workers were paid \$0.75 (average hourly rate: \$4.90 per hour).

To test for an effect between the ratings and actual performance, we used a two-way ANOVA on the enjoyment data, which is the common data we collected in the first study and the mutated instructions, to see if there exists a difference between self-report workers and workers who did the task. We found no statistically significant effect ( $p=0.0556$ ).

## Conclusion

We examined the relationship between task design best practices, worker perceptions of quality, and actual outcomes.

1. Collected HITs were assessed by workers, and checked by an expert for adherence to best practices.
2. A model HIT was systematically mutated to learn how best practices affect outcomes.

The first study found that, contrary to our expectations, tasks created by requesters currently using the market were positively viewed by most workers. We also found widespread adherence to most of the best practices, except that many tasks fail to explain criteria for acceptance. We suspect survivorship bias may be a factors. Requesters who get poor results from workers may either improve their task designs or stop using the platform. This would suggest that these results are higher than what would be expected if all requesters were novices.

The second study showed substantial variation in the effects of the best practice guidelines on outcomes. One insight was that worker uptake—the proportion of workers who accept the task, relative to those who viewed it—is higher for short tasks. This illustrates a tension: Although more details can increase trust (e.g., clarity criteria) and accuracy (e.g., clear steps, format specifications), it may also makes tasks less appealing to workers, leading to longer delays to receive final results.

## Acknowledgement

We gratefully acknowledge Gaoping Huang and Hsuan Hsieh for thoughtful feedback on the experiment design, Yu-Hsuan Lai for evaluating workers’ instructions, and the 597 workers who contributed throughout all phases of this work.

## References

- Alagarai Sampath, H.; Rajeshuni, R.; and Indurkha, B. 2014. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’14, 3665–3674. New York, NY, USA: ACM.
- Alonso, O., and Baeza-Yates, R. 2011. Design and implementation of relevance assessments using crowdsourcing. In *Advances in information retrieval*. Springer. 153–164.
- Barowy, D. W.; Curtsinger, C.; Berger, E. D.; and McGregor, A. 2012. AutoMan: A Platform for Integrating Human-based and Digital Computation. In *Proceedings of the ACM International Conference on Object Oriented Programming Systems Languages and Applications*, OOPSLA ’12, 639–654. New York, NY, USA: ACM.



- Brown, S. J.; Goetzmann, W.; Ibbotson, R. G.; and Ross, S. A. 1992. Survivorship bias in performance studies. *Review of Financial Studies* 5(4):553–580.
- Dai, P.; Mausam; and Weld, D. S. 2010. Decision-theoretic Control of Crowd-sourced Workflows. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, 1168–1174. Atlanta, Georgia: AAAI Press.
- Difallah, D. E.; Catasta, M.; Demartini, G.; Ipeirotis, P. G.; and Cudré-Mauroux, P. 2015. The dynamics of micro-task crowdsourcing: The case of amazon mturk. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, 238–247. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.
- Gadiraju, U.; Yang, J.; and Bozzon, A. 2017. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, 5–14. New York, NY, USA: ACM.
- Goodman, J. K.; Cryder, C. E.; and Cheema, A. 2013. Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making* 26(3):213–224.
- Grady, C., and Lease, M. 2010. Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, 172–179. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Gutheim, P., and Hartmann, B. 2012. Fantasktic: Improving quality of results for novice crowdsourcing users. *Masters thesis, EECS Department, University of California, Berkeley*.
- Hauser, D. J., and Schwarz, N. 2016. Attentive Turk-ers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods* 48(1):400–407.
- Ho, C.-J.; Slivkins, A.; Suri, S.; and Vaughan, J. W. 2015. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, 419–429. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.
- Horton, J. J., and Chilton, L. B. 2010. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM Conference on Electronic Commerce*, EC '10, 209–218. New York, NY, USA: ACM.
- Hsieh, G., and Kocielnik, R. 2016. You get who you pay for: The impact of incentives on participation bias. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW '16, 823–835. New York, NY, USA: ACM.
- Huang, E.; Zhang, H.; Parkes, D. C.; Gajos, K. Z.; and Chen, Y. 2010. Toward automatic task design: A progress report. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, 77–85. New York, NY, USA: ACM.
- Ipeirotis, P. G.; Provost, F.; and Wang, J. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, 64–67. New York, NY, USA: ACM.
- Ipeirotis, P. G. 2010. Analyzing the amazon mechanical turk marketplace. *XRDS* 17(2):16–21.
- Jain, A.; Sarma, A. D.; Parameswaran, A.; and Widom, J. 2017. Understanding workers, developing effective tasks, and enhancing marketplace dynamics: A study of a large crowdsourcing marketplace. *Proc. VLDB Endow.* 10(7):829–840.
- Kapelner, A., and Chandler, D. 2010. Preventing satisficing in online surveys. In *Proceedings of 2010 CrowdConf*.
- Kaufmann, N.; Schulze, T.; and Veit, D. 2011. More than fun and money. worker motivation in crowdsourcing—a study on mechanical turk. In *AMCIS*, volume 11, 1–11.
- Khanna, S.; Ratan, A.; Davis, J.; and Thies, W. 2010. Evaluating and improving the usability of mechanical turk for low-income workers in india. In *Proceedings of the First ACM Symposium on Computing for Development*, ACM DEV '10, 12:1–12:10. New York, NY, USA: ACM.
- Kittur, A.; Nickerson, J. V.; Bernstein, M.; Gerber, E.; Shaw, A.; Zimmerman, J.; Lease, M.; and Horton, J. 2013. The future of crowd work. In *Proceedings of the 16th Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '13, 1301–1318. New York, NY, USA: ACM.
- Law, E.; Yin, M.; Goh, J.; Chen, K.; Terry, M. A.; and Gajos, K. Z. 2016. Curiosity killed the cat, but makes crowdwork better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 4098–4110. New York, NY, USA: ACM.
- Mason, W., and Watts, D. J. 2010. Financial incentives and the "Performance of Crowds". *SIGKDD Explor. Newsl.* 11(2):100–108.
- McInnis, B.; Cosley, D.; Nam, C.; and Leshed, G. 2016. Taking a hit: Designing around rejection, mistrust, risk, and workers' experiences in amazon mechanical turk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 2271–2282. New York, NY, USA: ACM.
- Nielsen, J. 1994. *Usability engineering*. Elsevier.
- Oppenheimer, D. M.; Meyvis, T.; and Davidenko, N. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45(4):867–872.
- Peer, E.; Vosgerau, J.; and Acquisti, A. 2013. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods* 46(4):1023–1031.
- Rogstadius, J.; Kostakos, V.; Kittur, A.; Smus, B.; Laredo, J.; and Vukovic, M. 2011. 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, ICWSM '11*, 321–328.

Schulze, T.; Seedorf, S.; Geiger, D.; Kaufmann, N.; and Schader, M. 2011. Exploring task properties in crowdsourcing—an empirical study on mechanical turk. In *19th European Conference on Information Systems, ECIS '11*. Atlanta, GA: AISeL.

Shaw, A. D.; Horton, J. J.; and Chen, D. L. 2011. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, CSCW '11*, 275–284. New York, NY, USA: ACM.

Shneiderman, B. 2010. *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education India.

Silberman, M. S.; Ross, J.; Irani, L.; and Tomlinson, B. 2010. Sellers' problems in human computation markets. In *Proceedings of the ACM SIGKDD Workshop on Human Computation, HCOMP '10*, 18–21. New York, NY, USA: ACM.

Wright, P. 1998. Printed instructions: Can research make a difference. *Visual information for everyday use: Design and research perspectives* 45–66.

Yin, M.; Chen, Y.; and Sun, Y.-A. 2013. The effects of performance-contingent financial incentives in online labor markets. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, AAAI '13*, 1191–1197. AAAI Press.