Measuring Crowdsourcing Effort with Error-Time Curves

Justin Cheng
Stanford University
jcccf@cs.stanford.edu

Jaime Teevan
Microsoft Research
teevan@microsoft.com

Michael S. Bernstein
Stanford University
msb@cs.stanford.edu

ABSTRACT

Crowdsourcing systems lack effective measures of the effort required to complete each task. Without knowing how much time workers need to execute a task well, requesters struggle to accurately structure and price their work. Objective measures of effort could better help workers identify tasks that are worth their time. We propose a data-driven effort metric, ETA (error-time area), that can be used to determine a task's fair price. It empirically models the relationship between time and error rate by manipulating the time that workers have to complete a task. ETA reports the area under the error-time curve as a continuous metric of worker effort. The curve's 10th percentile is also interpretable as the minimum time most workers require to complete the task without error, which can be used to price the task. We validate the ETA metric on ten common crowdsourcing tasks, including tagging, transcription, and search, and find that ETA closely tracks how workers would rank these tasks by effort. We also demonstrate how ETA allows requesters to rapidly iterate on task designs and measure whether the changes improve worker efficiency. Our findings can facilitate the process of designing, pricing, and allocating crowdsourcing tasks.

Author Keywords

Crowdsourcing; microtasks; task effort.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous.

INTRODUCTION

Imagine that a requester wants to use Amazon Mechanical Turk to label 10,000 images with a fixed set of tags. How much should workers be paid to label each image? Would labeling an image with twice as many tags result in a task that is twice as much effort? Should the tags be provided in a drop down list or with radio buttons? Answering these questions requires a fine-grained understanding of the amount of effort the task requires. This process today involves trial and error: requesters observe the wait time and quality on test tasks, guess what might have been causing any problems, tweak the task, and repeat. An accurate measure of the effort required to complete a crowdsourced task would enable requesters to compare different approaches to their tasks, iterate toward a better design, and price their tasks objectively. It could also help workers decide whether to accept a task, or even allow systems to offer tasks based on difficulty or time availability.

CHI 2015, April 18–23, 2015, Seoul, Republic of Korea.

ACM 978-1-4503-3145-6/15/04 ...$15.00.
http://dx.doi.org/10.1145/2702123.2702145

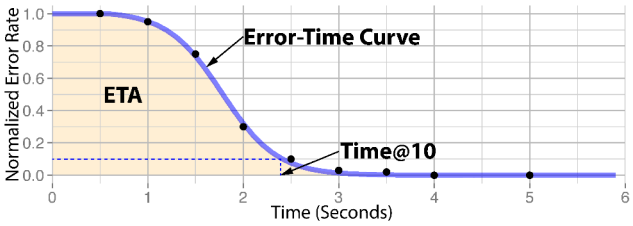

Figure 1. The ETA is defined as the area under the error-time curve.

However, despite its potential value, task effort is challenging to estimate. Workers face cognitive biases in assessing difficulty [21], while requesters cannot easily observe the process and, as experts, categorically underestimate completion times [12]. These limits suggest the need for a behavioral approach to measure effort. One approach might be to let the market identify hard tasks by reacting to the posted price [30]. However, prices cannot easily make fine distinctions in an inelastic market such as Mechanical Turk [14]. Another approach might be to use task duration as a signal of difficulty, but this is unreliable because workers regularly accept multiple tasks simultaneously and interleave work [29]. Measures such as reaction time [32] are not easy to apply to typical crowd tasks: reaction time metrics tend to use simplistic tasks (e.g., shape or color recognition), while others may be too involved for crowd work (e.g., [9]).

In this paper, we propose a data-driven behavioral measure of effort that can be easily and cheaply calculated using the crowd. Our metric, the *error time area (ETA)*, draws on cognitive psychology literature on speed-accuracy tradeoff curves [32], and represents the effort required for a worker to accurately complete a task. To create it, we first recruit workers to complete the task under different time limits. Next, we fit a curve to the collected data relating the error rate and time limit (Figure 1). Last, we compute ETA by taking the area under this error-time curve. Because ETA is calculated using a data-driven approach, task difficulty can be determined with minimal effort and without analytical modeling. Rather than measuring average duration independent of work quality, ETA computes quality as a function of duration and thus can be used to estimate a wage for a task. ETA also allows requesters to compare multiple task designs; for example, we find that tagging an image with an open textbox is less effort than choosing between a fixed list of 16 options, but more effort than choosing between a fixed list of 8 options.

After describing ETA, we explore the metric via four studies:

- *Study 1: ETA vs. other measures of effort*. For ten common microtasking primitives (e.g., multiple choice questions, long-form text entry), we show that the ETA metric represents effort better than existing measures.
- *Study 2: ETA vs. market price*. We then compare ETA as well as other measures to the market prices of these primitives on a crowdsourcing platform.
- *Study 3: Modeling perceptual costs*. By augmenting ETA with measures of perceptual effort, we find we can better model a worker's perceived difficulty of a task.
- *Study 4: Tasks without ground truth*. In order to capture how well people do a task, ETA requires ground truth. We extend the metric to also work for subjective tasks.

We then demonstrate how ETA can be used for rapidly prototyping tasks. ETA makes it possible to characterize tasks in terms of their monetary cost and human effort, and paves the way for better task design, payment, and allocation.

RELATED WORK

Measures of task difficulty or mental workload can be roughly separated into two categories: subjective and objective measures. Subjective measures include multidimensional workload assessment tools such as the NASA Task Load Index (TLX) [10] and time estimates [5]. However, such measures tend to be inaccurate and hard to capture. It is difficult for requesters to accurately estimate task difficulty, as experts categorically underestimate novices' completion times and difficulty [12, 13]. Workers are also inaccurate; subjective metrics collected from workers tend to correlate poorly with each other (e.g., between self-reported effort, self-reported difficulty, and response time [7]), and workers exhibit large variance because they use different ranges of the rating scale [11]. Worker-driven subjective task judgments sometimes appear on web sites such as Turkopticon and mTurk Grind. However, these reviews are limited in number, lag the marketplace by hours, and not available for all tasks. Our own experiments reveal that many subjective metrics, while correlated with effort, are not directly interpretable and cannot differentiate between similar tasks.

Objective measures of effort include measuring reaction time to a secondary stimulus [32] and dual task performance [33]. Physiological approaches include using EEGs [9], but these approaches are relatively involved. The pricing of tasks on Mechanical Turk can also provide a signal of the effort required for a task; survival analysis empirically models the tradeoff between pricing and rate of completion [8], and measures of how long people work on tasks can characterize a form of interface utility [30]. Still, completion rates are dependent on time of day and the presence of other tasks, and the equilibrium wage in crowdsourcing markets is low and fairly inelastic [14]. Raw task duration is another potential measure, but is a noisy signal as workers tend to switch between tasks or ignore them for periods of time [29]. Accuracy-based measures are subject to ceiling effects, especially on relatively simple tasks typical of crowdsourcing, and are not easily comparable across task types. We employ an objective approach to measuring effort, extending research on accuracy-tradeoff curves [32] to produce a more robust measure that correlates well with several subjective measures.

While our approach is data-driven, measures of effort do not inherently require a task be completed to be calculated. Machine learning can be used to estimate the difficulty of a particular input and the quality of a worker for unobserved tasks (e.g., [6]), but these approaches tend to be fairly complex and require large-scale data collection. Further, they are task-specific (e.g., multiple-choice questions [16]) and not easily comparable across tasks. Our work requires limited data and makes it possible to compare a large variety of tasks.

In addition to measuring the amount of overall effort a task requires, the effort associated with a task can also be broken down into components, for example, by separating its perceptive (e.g., reading a question), cognitive (e.g., formulating the answer), and psychomotor (e.g., writing the answer down) costs [25]. Borrowing from these models of workload, we augment our approach to explore the perceptual load of a task. We do not attempt to exactly differentiate between these component costs, but our initial experiments suggest there is value in accounting for these differences. In human factors research, analytical models for usability evaluation (e.g., ACT-R [3]) have been developed to characterize low-level processes for computer-based tasks, and can be used to predict mental workload. However, these methods require substantial training or metadata annotation to execute, which make them infeasible for many requesters. Rather than model the effort required for a task by analyzing its specific makeup, crowdsourcing allows us to do so by quickly getting workers to perform tasks and observing how they perform.

ETA: A MEASURE OF TASK EFFORT

To quantify the amount of effort a task requires for high-quality results, we measure the impact of time on the number of errors workers make. Taking inspiration from the speed-accuracy tradeoff curves developed in cognitive psychology research [26, 32], we calculate a task's *error-time* tradeoff curve by giving workers varying time constraints to finish the task and measuring the probability they make an error within each time limit. We measure the error-time area under the curve (or, simply, *ETA*) to reduce the curve to a single metric. What follows is a detailed description of this process.

Error-Time Data Collection. Cognitive psychology has demonstrated that the amount of time a person has to do a task impacts their ability to do it well [26, 32]. To calculate how much time is required to perform a crowdsourcing task well, we sample different time limits (e.g., 0.2, 2, or 20 seconds) and ask each worker within-subject to complete the task at each time limit, with the time limits presented in random order. After the time limit elapses, the task is disabled and workers cannot submit their response. By using tasks with known correct answers (gold standard tasks), we can calculate the error rate for a task within each limit. In practice, we recommend at least seven time conditions and 10 workers. For a two cent task, this comes out to under $5. To determine time limits for a task, a task designer could define them evenly spaced around an initial estimate.

The data collected for an example task can be seen in Figure 1, where the x-axis represents the amount of time workers were given, and the y-axis represents the error rate. In this example, workers have a 95% chance of getting the answer incorrect (or incomplete) when given one second to complete the task, but only a 10% chance when given about 2.5 seconds. We later demonstrate how worker consistency can be used to create these curves even without ground-truth data.

Fitting the Error-Time Curve. Next we fit a sigmoid curve to the recorded accuracies. Sigmoids (e.g., logistic curves) are a good model for performance data that undergoes a phase shift. In our data, this shift occurs at the point when the task becomes feasible to complete in the given time limit. An example error-time curve is shown fitted to the data in Figure 1. The range of potential error rates can differ across tasks; for example, a binary choice task has an observed error rate ranging from 0 (always correct) and 0.5 (random). For this reason, we scale the observed error rate to always be between 0 and 1 (and the curve clipped, if necessary, to ensure the predicted error rate is also between 0 and 1).

Calculating the ETA. Prior work exploring the tradeoff between accuracy and time in task performance has primarily focused on characterizing the functions and parameters of these curves (e.g., that they are exponential [32]), or using them to make empirical observations on a small number of very similar tasks (e.g., on different amounts of object rotations [22]). While these plots provide analytical value, they do not facilitate comparison across the wide variety of tasks found on microtask markets. Thus we extend the approach to reduce the error-time curve into a single, cross-task comparable, continuous metric.

With a model fit to the data, we reduce the curve to a single score. Many reductions are possible; we use the area under the curve (AUC). Inspired by the use of ROC AUC in machine learning, the error-time area (ETA) under the curve captures human performance under differing amounts of time. A small ETA suggests the task can be performed correctly almost instantaneously, while a large ETA suggests it is very difficult, although there are some exceptions (e.g., if a task is difficult regardless of the time limit, the ETA will be small as it measures relative, not absolute change in performance). Intuitively, each unit increase in ETA corresponds to a task requiring an additional second to achieve the same work quality. Because error is scaled, the ETA is finite and can be calculated by computing the integral at infinity.

Calculating Work Time and Wage. While ETA is interpretable, other transformations of the curve have more direct interpretations; *Time@10*, or the time it takes to achieve an error rate at the 10th percentile, reflects the amount of time a worker would need to achieve a relatively low error rate. By multiplying this time with a target pay rate (e.g., minimum wage), this metric can be used to price tasks. Such an approach may be preferable to using how long workers spend on the task, since some workers rush through tasks and care little about accuracy, while others spend more time than necessary. Time@10 can be calculated by identifying the point on the x-axis where the error-time curve crosses an error rate of 0.1 (10%) (Figure 1). Alternate thresholds (e.g., Time@5) would represent different tradeoffs between the fraction of workers able to attain the given wage rate and task price.

Code Library. A framework for computing ETA is available open-source at `http://hci.st/eta`. Users of the framework begin by collecting task-relevant data. To do this they include a Javascript library in their existing task HTML file and write a few lines of code specifying the format of the task's input and the time points desired. They then upload the modified task file to their crowdsourcing platform of choice. The library randomizes the order of the time limits, tells the worker how long they have to complete the task, and disables the answer area (but not the submit button) after the time elapses. The task's output can be analyzed using an R code library. It outputs descriptive statistics (e.g., task time), plots the error-time curve, and computes the task's ETA along with other objective and subjective measures.

This code library makes it possible to easily measure the effort a task requires via an initial HIT at the cost of a few dollars. The computed error-time curve can be used in a number of ways. Requestors can iterate on a task's design to reduce its ETA, and multiply Time@10 by an hourly wage to determine the task's payment. The ETA can also be published along with the task to facilitate task selection by workers.

While ETA grows out of a large body of cognitive science literature for measuring effort, it is important to understand how well it works in practice. For this reason, we ran a number of studies to understand when it works, when it does not, and why. Through four studies, we evaluate ETA as a measure of effort. In Studies 1 and 2, we compare ETA to other objective and subjective measures, including market price. In Study 3, we additionally model the perceptual cost of a task to account for instances where ETA does not match workers' perceived difficulty of a task. In Study 4, we extend our approach with a methodology that allows us to measure ETA without ground truth answers. Finally, we demonstrate how ETA can be used to compare effort across different versions of a task.

STUDY 1: ETA VS. OTHER MEASURES OF EFFORT

We begin by comparing ETA and other measures of difficulty (including time and subjective difficulty) across a number of common crowdsourcing tasks. After describing the experimental setup, designed to elicit the necessary data to generate error-time curves and other measures for each task, we show how closely the different measures matched.

Method

Study 1 and all subsequent experiments reported in this paper were conducted using a proprietary microtasking platform that outsources crowd work to workers on the Clickworker microtask market. The platform interface is similar to that of Amazon Mechanical Turk; users upload HTML task files, workers choose from a marketplace listing of tasks, and data is collected in CSV files. We restricted workers to those residing in the United States. Across all studies, 470 unique workers completed over 44 thousand tasks. A followup survey revealed that approximately 66% were female. We replicated Study 1 on Amazon Mechanical Turk and found empirically

Primitive	Example Question
Binary (Binary Choice)	Does this label apply to this image?
Scale (Likert Scale)	How much do you agree with this statement?
Categorize	Which category best applies to this image?
Tag	Label this image with a tag.
Describe	Describe this image.
Math	Add these two numbers together.
Transcribe	Copy this phrase exactly as presented.
Find	Find a spelling error in this sentence.
Fix	Correct the highlighted word in this sentence.
Search	Search for the answer to this question online.

Table 1. Ten primitives common to most crowdsourcing workflows.

similar results, so we only report results using Clickworker in this paper.

Primitive Crowdsourcing Task Types

We began by populating our evaluation tasks with common crowdsourcing task types, or *primitives*, that appear commonly as microtasks or parts of microtasks. To do this, we looked at the types of tasks with the most available HITs on Amazon Mechanical Turk, at reports on common crowdsourcing task types [15], and at crowdsourcing systems described in the literature (e.g., [4]). After several iterations we identified a list of ten primitives that are present in most crowdsourcing workflows (Table 1, Figure 2). For example, the Find-Fix-Verify workflow [4] could be expressed using a combination of the FIND (identify sentences which need shortening), FIX (shortening these sentences), and BINARY primitives (verifying the shortening is an improvement). In many cases, the primitives themselves (or repetitions of the same primitive) make up the entire task, and map directly to common Mechanical Turk tasks (e.g., finding facts such as phone numbers about individuals (SEARCH)).

We instantiated these primitives using a dataset of images of people performing different actions (e.g., waving, cooking) [34] and a corpus of translated Wikipedia articles selected because they tend to contain errors [1].

Experimental Design

We presented workers with a mixed series of tasks from the ten primitives and manipulated two factors: the time limit and the primitive. Each primitive had seven different possible time limits, and one untimed condition. The exact time limits were initialized using how long workers took when not under time pressure. The result was a sampled, not fully-crossed, design. For each worker we randomly selected five primitives for them to perform; for each primitive, three questions of that type were shown with each of the specified time limits. The images or text used in these questions were randomly sampled and shuffled for each worker. To minimize practice effects, workers completed three timed practice questions prior to seeing any of these conditions. The tasks were presented in randomized order, and within each primitive the time conditions were presented in randomized order. Workers were compensated $2.00 and repeat participation was disallowed.

A single task was presented on each page, allowing us to record how long workers took to submit a response. Under

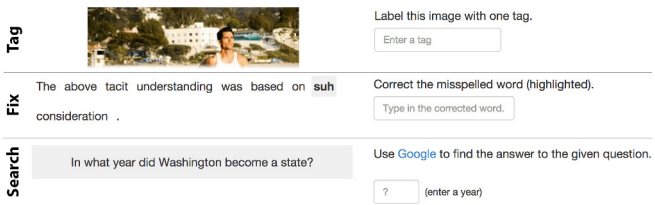

Figure 2. The tag, fix, and search tasks.

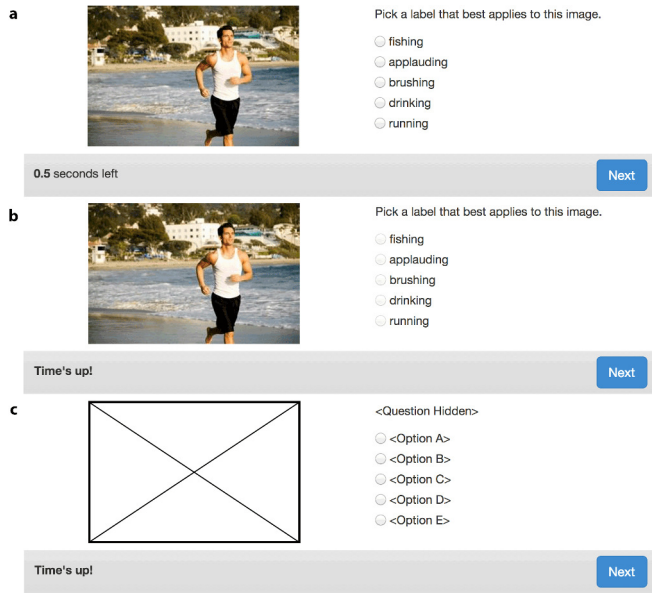

Figure 3. (a) Basic interface showing the categorize task and the countdown timer. (b) To measure effort in Studies 1, 2 and 4, questions were disabled after the timer expires. (c) To measure perceptual costs in Study 3, the question was hidden, but workers could still answer the question.

timed conditions, a timer started as soon as the worker advanced to the next page. Input was disabled as soon as the timer expired, regardless of what the worker was doing (e.g., typing, clicking). An example task is shown in Figure 3.

Measures

The information we logged allowed us to calculate behavioral measures for each primitive:

– *ETA.* The ETA is the area under the error-time curve.

– *Time@10.* We also calculated the time it takes to achieve an error rate at the 10th percentile.

– *Error.* We measured the error rate against ground truth for each primitive. If there were many possible correct responses, we manually judged responses while blind to condition. Automatically computing distance metrics (e.g., edit distance) resulted in empirically similar findings.

– *Time.* We measured how long workers took to complete the primitive without any time limit.

After each task block was complete, we additionally asked workers to record several subjective reflections:

– *Estimated time.* We asked workers to report how long they thought they spent on a primitive absent time pressure.

Primitive	Subj.Rank	**ETA** [conf. int.]	**Time@10** [Time@10Cost]	Error	Time			TLX	Cost	
					Time	Est.Time	RSD		Mkt.Price	Est.Cost
Binary	2.50 (1)	1.58 [1.49-1.73] (1)	2.1 [0.7] (1)	0.01 (3)	3.3 (1)	4.1 (1)	0.19 (1)	36.6 (2)	0.33 (1)	4.1 (1)
Scale	3.33 (2)	1.88 [1.73-2.05] (2)	2.7 [0.9] (3)	0.00 (2)	4.2 (3)	4.5 (3)	0.15 (3)	38.1 (3)	0.50 (2)	4.8 (3)
Categorize	3.45 (3)	1.97 [1.87-2.07] (3)	2.4 [0.8] (2)	0.00 (1)	3.9 (2)	4.4 (2)	0.19 (2)	36.2 (1)	1.00 (5)	4.3 (2)
Tag	4.30 (4)	2.91 [2.66-3.46] (4)	4.0 [1.3] (4)	0.10 (6)	5.7 (4)	6.6 (4)	0.14 (4)	44.1 (4)	1.00 (5)	5.8 (4)
Transcribe	6.40 (5)	7.62 [7.14-8.11] (8)	10.0 [3.3] (8)	0.19 (10)	11.6 (8)	10.8 (8)	-0.00 (7)	49.6 (7)	1.00 (5)	10.6 (9)
Find	6.42 (6)	3.88 [3.55-4.34] (5)	5.5 [1.8] (5)	0.13 (7)	10.9 (7)	8.1 (5)	-0.13 (9)	48.3 (5)	1.00 (5)	6.1 (5)
Fix	6.53 (7)	4.31 [3.88-4.66] (6)	6.6 [2.2] (6)	0.17 (8)	10.5 (6)	8.9 (6)	0.00 (6)	49.4 (6)	5.00 (9.5)	7.0 (6)
Add	6.90 (8)	4.92 [4.39-5.38] (7)	7.7 [2.6] (7)	0.09 (5)	9.8 (5)	9.6 (7)	0.03 (5)	51.9 (9)	5.00 (9.5)	8.2 (7)
Describe	7.15 (9)	7.78 [6.96-8.75] (9)	12.2 [4.1] (9)	0.18 (9)	15.4 (9)	11.7 (9)	-0.13 (8)	51.5 (8)	1.00 (5)	8.9 (8)
Search	8.03 (10)	11.7 [10.8-12.5] (10)	16.0 [5.3] (10)	0.04 (4)	18.8 (10)	15.2 (10)	-0.16 (10)	51.9 (10)	3.33 (8)	13.2 (10)
Kendall's Tau Coefficients (*: $p<0.05$, **: $p<0.01$)										
w/Subj.Rank	-	0.87**	0.82*	0.29	0.69*	0.82*	0.69*	0.78*	0.66*	0.78*
w/Mkt.Price	0.66*	0.56	0.51	0.20	0.41	0.51	0.41	0.56	-	0.51

Table 2. Measures of effort for ten crowdsourcing primitives (rank order in each column in brackets). ETA has the highest rank correlation with the subjective rank of a task. All prices are in cents. (Study 1, 2)

Time estimation has previously been used as an implicit signal of task difficulty [5].

– *Relative subjective duration (RSD).* RSD, a measure of how much task time is over- or underestimated [5], is obtained by dividing the difference between estimated and actual time spent by the actual time spent.

– *Task load index (TLX).* The NASA TLX [10] is a validated metric of mental workload commonly used in human factors research to assess task performance. It consists of a survey that sums six subjective dimensions (e.g., mental demand).

A separate experimental design that contained all ten primitives, where each worker completed three untimed practice questions followed by three untimed questions for each primtive (with the primitives presented in random order), was used to obtain the

– *Subjective rank.* Workers considered all of the primitives they completed and ranked them in order of effort required.

As rankings produce sharper distinctions than individual ratings [2], we consider subjective rank to represent our ground truth ranking of the primitives. However, rank would not be a deployable solution for requesters. Ranking means that workers would need to test the new task against at least $log(n)$ of the primitives, incurring a large fixed overhead. Further, ranking is ordinal, and cannot quantify small changes in effort. In contrast, ETA is an absolute ranking, can measure small changes in effort, and only needs to be measured for the target task to compare it with other tasks.

Analysis

60 workers completed Study 1, with 30 performing each primitive. We averaged our dependent measures across all 30 workers, and compared the ranking of primitives induced by each measure to the average subjective ranking (subjective rank was obtained by having 40 other workers rank all ten primitives). We used the Kendall rank correlation coefficient to capture how closely each measure approximated the workers' ranks, with Holm-corrected p-values calculated under the null hypothesis of no association. A rank correlation of 1 indicates perfect correlation; 0 indicates no correlation. Measures that capture the subjective ranking accurately can analyze new tasks types without comparing them against multiple benchmark tasks.

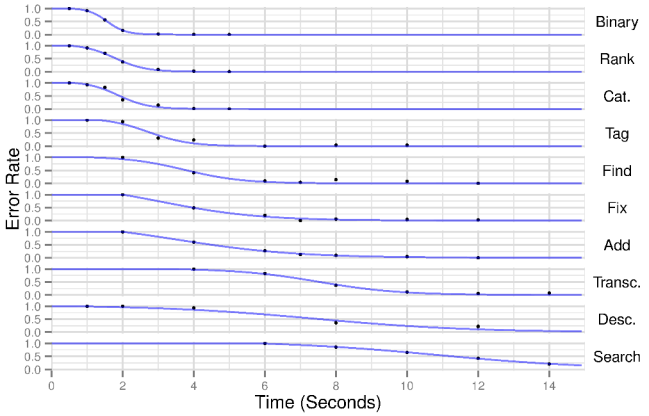

Figure 4. The error-time curves for each primitive show how the errors made decrease as more time is provided to answer a question.

Results

Table 2 reports overall results for each primitive using each measure, as well as the resulting rank ordering. Bootstrap confidence intervals are also reported for ETA; their relatively small range suggests that task ranking remains fairly stable with 30 workers. Raw error-time plots, shown in Figure 4, verify that the data follows a sigmoidal pattern.

A Kendall's tau test reveals that ETA is significantly more correlated with subjective rank (τ=0.87) than other measures such as Time or TLX. It correctly lists Binary as the easiest task (ETA=1.58) followed by Scale (1.88) and Categorize (1.97), with Search being the most difficult (11.7), suggesting that the ordering of task difficulty under ETA matches workers perceptions of relative task difficulty relatively well. In fact, we find that ETA only misranks TRANSCRIBE — it was perceived as easier than other tasks that took more time.

In an upcoming section, we demonstrate how the high perception cost in this task may account for why it seemed like less effort. To understand the robustness of ETA with respect to the number of workers requested, we conducted a bootstrap analysis by subsampling from our worker pool, varying n between 1 and 30, then computing the 10th and 90th percentile of the estimated ETA for each n. We find that ETA can be reasonably estimated with a small number of workers On average, ten workers were required to obtain an ETA where 90% of the values are within 20% of the true value; with 20 workers, 95% of values are within 10% of the true value. Assuming each task costs 2 cents, it costs about $5 to generate the ETA for a task using ten workers.

The next best metrics after ETA were Time@10 (a transformation of the same error-time curve used for ETA), and estimated time. But while the relative rankings provided by Time@10 and estimated time are the same, Time@10 additionally encodes work quality. Corroborating prior research, our results suggest that time (and estimated time) can be misleading. For example, FIND takes 10.9 seconds with no time limit (and workers estimate it takes 8 seconds), but the task can be completed well in as little as 5.5 seconds (Time@10).

Previously-validated approaches were less highly correlated with subjective rank. Relative subjective duration struggled because the relative error increases as tasks take less time, making it difficult to use as an estimate across tasks of different time classes. NASA-TLX could not discriminate easy tasks well and did not order these correctly.

STUDY 2: ETA VS. MARKET PRICE

Study 1 shows that ETA is highly correlated with workers' subjective ranking of task difficulty. Measuring task difficulty is also useful because it can be used to price tasks. Thus, Study 2 attempts to estimate the market price of each task, which we then compare against our measures from Study 1.

Method

To extract the market price for our 10 primitives, we launched a set of 10 pricing tasks, one for each primitive. For each pricing task, we held the task's price constant at 10 cents, but manipulated the number of primitive tasks required to complete the pricing task. This was conducted as a between-subjects study, with each worker assigned in a round-robin fashion to a primitive and fixed number of primitive tasks to complete. This number varied between 1 and 40, with six levels used for each primitive. The exact range and values for each level depended on the primitive's difficulty.

We measured each pricing task's acceptance rate as the ratio of task accepts to task previews. When workers previewed the task, they were shown a sample primitive task and told the number of such tasks they would have to complete to get paid. Market behavior should result in more workers accepting the task if they only have to complete one repetition than if they have to complete 40. All pricing tasks remained on the market for the same 18-hour period. A mean of 20.0 (σ=3.82) people viewed each condition, with 17.3 (σ=4.38) completing a task after viewing it.

– *Market price.* We estimated the market price of each primitive by identifying the maximum number of primitive tasks where the pricing task's acceptance rate was above 90%. The mean unit cost of that primitive was then calculated by dividing 10 cents by that number.

Additionally, we consider two other measures of cost:

– *Time@10 cost.* Using Time@10, we estimated how much requesters should pay to complete each primitive, assuming continuous work and an hourly wage of $12.

– *Estimated cost.* As a rough, subjective estimate of a primitive's reservation wage, we also asked workers how much they would want to be paid for performing it once.

Results

Table 2 reports all three measures of cost, with Time@10 cost correlating best with subjective rank. While we find that market price provides some signal of effort, it is ultimately noisy and expensive to obtain. For one, the estimated market price for many primitives ended up being the same, indicating market inelasticity, making it hard to accurately rank the tasks. Further, the estimate was noisy — obtaining greater precision may require significantly more workers and conditions. With six conditions and 20 workers per condition over 18 hours, pricing data is already relatively expensive ($12 per task) and time-consuming to gather. In a paper measuring economic utility, over 20,000 jobs costing almost $1000 were run to compute the market price of three tasks (where price was varied between 1 and 6 cents) [30]. Also, market prices fluctuate; tasks that are cheap to perform at one moment may wind up being expensive later.

A Kendall's tau test reveals that only subjective rank, our previous ground truth, is significantly correlated with market price (0.66), supporting its use in Study 1. While the lack of significance with other metrics is unsurprising given the coarse-grained nature of ranking by market price, we nonetheless find that ETA and TLX are the next most highly correlated measures (0.56). These results also add an ethical dimension to market pricing as a dependent variable, as it allows requesters to estimate the wage they are paying for quality work.

STUDY 3: MODELING PERCEPTUAL COSTS WITH ETA

Study 1 suggests that ETA can accurately capture task effort, and Study 2 suggests that market price, while noisy, is correlated with subjective rank and ETA. In Study 3 we develop a model that also measures the perceptual cost of a task, to help account for the one ranking error that ETA made relative to workers' rankings. In this model, we note that effort has two components. The first, perceptual and cognitive effort, is the effort workers need to process the provided input (e.g., an image or text). The second, motor effort, is the effort spent translating their decisions into physical action (e.g., typing or clicking). Informally, workers may value their perceptual and cognitive time more than they value their motor time. By isolating the perceptual component of the task relative to the whole and estimating the time workers need to perceive the task (as opposed to the time required to complete the task), we can further improve our estimates of effort.

ETA predicted that TRANSCRIBE required more effort than workers felt it did. One important reason for the difference may lie in the relative balance of perceptual and cognitive effort. In other words, TRANSCRIBE relies heavily on low-level perception, whereas other primitives such as ADD rely on more cognition (or System 1 vs. System 2 thinking [19]). For instance, we may expect to spend a larger proportion of time reading the text provided for TRANSCRIBE as compared to SEARCH, where more time is instead spent searching for the answer than reading the provided question.

Method

For Study 3 we followed a similar design to Study 1. However, instead of randomly choosing a time limit for the entire task, we randomly chose a limit for the amount of time workers could *see* the stimulus. When the time limit expired the stimulus was hidden but workers still had to complete the task (Figure 3c). Workers were only limited in the time they could spend perceiving the task. For example, with BINARY, a worker might be shown the source image for 0.5 seconds and asked to answer whether the tag "bird" applied to the image. Workers had unlimited time to submit their response. Using the results with 60 workers, we again generated a error-time curve by fitting an exponential curve to the data, computing three metrics based on it.

– *PerceptionETA.* Similar to ETA, this measures the area under the error-time curve, but where time now corresponds to how long the stimulus was shown.

– *Perception@10.* We estimated how long the stimulus needed to be shown in order to achieve an error rate at the 10th percentile.

– *PerceptionRatio.* This measure computes the ratio of perception time, Perception@10, to total required time, Time@10. As workers have unlimited time to answer questions, Perception@10 can exceed Time@10. Nonetheless, this rarely occurs in practice.

To analyze the data, we compared the three measures of perception with the workers' subjective ranking from Study 1. We also hypothesized that ETA and measures of perception could be combined to produce an improved ranking model, and use ETA, PerceptionETA and PerceptionRatio as features in an ordered logistic regression (OLR) model.

Results

Our results in Table 3 show that ADD requires significantly less perception time (0.95) than TRANSCRIBE (5.36). This suggests that workers spend a larger proportion of time processing the perceived information for ADD. In other words, the cognitive load for ADD may be much higher than for TRANSCRIBE, albeit for a shorter amount of time. This increase may explain why ADD was perceived as more difficult than TRANSCRIBE, even though it took significantly less time to complete.

However, perception metrics alone are a poor predictor of subjective rank or market price (ranging from 0.10 to 0.47, n.s.). In other words, proportion of time spent perceiving may only have a secondary effect on the perceived difficulty of a task. But by using a combined ETA-Perception OLR model, we obtain a rank correlation of 0.90 ($p<0.01$), higher than any individual metric. Models that used combinations of subjective or time-based measures were not as performant. TRANSCRIBE now appears in the right rank order, with the only confusion being that now FIND is estimated as easier than TAG and TRANSCRIBE. This same model also produces a rank correlation of 0.67 with a primitive's market price, higher than any other individual metric.

Primitive	Perception...			**OLR**
	...ETA	...@10	...Ratio	
Binary	0.37 (4)	0.63 (4)	0.30 (4)	1 (1)
Scale	0.40 (5)	0.65 (5)	0.24 (5)	2 (2)
Categorize	0.86 (8)	2.28 (9)	0.95 (1)	3 (3)
Tag	0.09 (1)	0.14 (1)	0.04 (8)	4 (5)
Transcribe	3.87 (10)	5.36 (10)	0.54 (2)	6 (6)
Find	1.44 (9)	2.21 (8)	0.40 (3)	3 (3)
Fix	0.42 (6)	0.67 (6)	0.10 (7)	7 (7)
Add	0.50 (7)	0.95 (7)	0.12 (6)	8 (8)
Describe	0.16 (2)	0.31 (2)	0.03 (10)	9 (9)
Search	0.31 (3)	0.49 (3)	0.03 (9)	10 (10)
Kendall's Tau Coefficients (*: $p<0.05$, **: $p<0.01$)				
w/Subj.Rank	0.07	0.11	0.47	0.90**
w/Mkt.Price	0.10	0.10	0.20	0.67*

Table 3. Measures of perceptive effort for the same primitives, as well the rankings produced by an OLR model that takes into account both ETA and these measures. (Study 4)

In summary, while ETA alone is a reasonable predictor of overall task effort, it can be augmented with measures of perceptual cost to attain even better predictions, at the cost of performing an additional experiment. Perceptual cost can also indicate along what dimension a task is easy or hard.

STUDY 4: MEASURING ETA WITHOUT GROUND TRUTH

To evaluate worker performance and calculate ETA we have thus far assumed that ground truth labels for a task exist. These labels are used to establish the y-axis of the error-time curve — the probability of an error when given the time limit on the x-axis. Ground truth labels may also be difficult to generate for tasks with subjective answers.

In Study 4 we create error time curves for tasks without using a ground-truth measure of quality. We explore two alternative measures of performance for subjective tasks:

– *Internal consistency.* If a worker submits the same response under a time limit as they give with no time limit, then the task can be deemed correct.

– *Between-subject variation.* If different workers submit the same responses with a given time limit, then the task can be deemed correct.

Measures of internal consistency and between-subject variation make different assumptions about a task: consistency assumes that workers will reproduce similar answers for a particular question, while variation assumes that some answers to a question are more likely than others. Both approaches can be applied to objective tasks, but it is cheaper

Primitive	Subj.Rank	Consistency		Variation		Time		TLX	Est.Cost	Perception ETA	OLR
		ETA	**Time@10**	**ETA**	**Time@10**	Time	RSD				
Obj. Scale	2.63 (1)	1.57 (1)	2.10 (1)	1.75 (1)	2.30 (1)	3.51 (2)	-0.13 (2)	29.2 (1)	10.65 (2)	0.41 (2)	1 (1)
Obj. Categorize	2.74 (2)	1.80 (2)	2.40 (2)	2.07 (2)	3.10 (2)	3.51 (1)	0.14 (1)	30.1 (2)	10.72 (3)	1.13 (4)	2 (2)
Subj. Scale	3.22 (3)	1.96 (3)	3.00 (3)	2.35 (3)	4.00 (4)	4.11 (3)	-0.25 (4)	30.8 (4)	10.40 (1)	1.25 (5)	3 (3)
Subj. Categorize	3.89 (4)	2.77 (6)	4.20 (6)	2.71 (4)	4.20 (5)	4.74 (4)	-0.31 (6)	30.4 (3)	10.79 (4)	1.55 (6)	4 (4)
Obj. Tag	3.96 (5)	2.61 (5)	4.10 (5)	3.34 (6)	4.40 (6)	5.27 (5)	-0.21 (3)	32.6 (5)	12.76 (6)	0.77 (3)	5 (5)
Subj. Tag	4.56 (6)	2.50 (4)	3.50 (4)	2.94 (5)	3.50 (3)	6.14 (6)	-0.28 (5)	35.9 (6)	11.72 (5)	0.36 (1)	6 (6)

Table 4. Task difficulty can also be measured in the absence of ground truth data. While ETA alone accurately captures the relative ranking of objective tasks, perception measures can help explain differences for subjective tasks. (Study 4)

to use ground truth when available. We use them to see if we can replicate replicate the results from Study 1 with the same primitives, and to quantify the effort required for new, subjective versions of the primitives.

Method

We focused on three primitives in Study 4: CATEGORIZE, SCALE, and TAG. For each primitive we created alternatives that were objective (i.e., identical to those in Study 1) or subjective. For example, objective CATEGORIZE asked workers which of five action labels applies to an image, while subjective CATEGORIZE asked workers which of five emotions best represented their impression of an image. Objective SCALE asked how applicable an action was, while subjective SCALE asked how applicable an emotion was to an image. Objective TAG asked workers to tag the image with an action being performed, while subjective TAG instead asked the workers for a word describing how they felt about the image.

Workers in Study 4 were shown tasks for one primitive, and first completed three practice tasks to minimize practice effects. They were then exposed to seven timed conditions, with three tasks per timed condition, sampled so that all primitives were shown the same number of times in each condition across all workers. After workers performed the tasks in the timed conditions, the exact same set of tasks was shuffled and presented under an untimed condition.

We calculated internal consistency by comparing the worker's answer on the timed categorization, scale, or tag task to the untimed answer afterwards. We defined consistency as the probability of a worker's answers being identical in both cases. Defining a distance metric for each primitive (e.g., edit distance for the tag task) results in empirically similar findings. While we find that these measures were sufficient in practice, other robust methods of measuring worker error also exist (e.g., [18]). We calculated the between-subjects variation for each task by counting the number of unique answers in the timed conditions across all workers. For both metrics, we can again generate error-time curves — in this case consistency-time curves. They are, as before, sigmoidal. We use these curves to calculate ETA and Time@10.

Similar to Study 1, we recorded task completion time, subjective cost estimates for the task, and NASA-TLX scores. Also replicating Study 1, to understand how workers subjectively ranked the primitives, we ran a separate task containing all six primitives, showing only the untimed condition. We then sorted the six tasks by workers' summative rank and compared it to the other measurements using Kendall's tau. As in Study 3, we also measured the perceptual cost of these tasks, slightly modifying the timed condition to hide the stimulus rather than disable input. 40 workers completed Study 4.

Results

As expected, the more time a worker was given to complete a task, the more likely their initial answer would be consistent with the second, untimed one ($p<.001$). In the case of objective CATEGORIZE, the error rate drops from a mean of 0.82 when workers were given half a second, to 0.03 when given five seconds. With between-subjects variation, the number of unique responses also decreases as more time is given ($p<.001$). For objective CATEGORIZE, the mean number of unique answers for an image decreased from 3.5 with half a second, to 1.2 with five seconds.

We find that the relative ordering of ETA and Time@10 for the three objective primitives is the same as when we use ground truth data in Study 1, and the absolute values of these metrics are also similar (Table 4). This suggests that both internal consistency and between-subjects variation are reasonable substitutes for ground truth data for objective tasks. However, between-subjects variation has a few advantages. It only requires each question to be asked once, and thus is significantly cheaper to run. Additionally, the rank correlation of ETA with subjective rank was better using between-subjects variation (0.87, tying Duration and TLX) than internal consistency (0.6). While the rank ordering of the objective tasks using other metrics such as RSD are also similar to what we observed in Study 1, their absolute values are substantially different — though they were correlated with effort overall, they appear difficult to interpret individually.

The perceptual cost of a task also seems to play a significant role in the case of the subjective primitives we studied. For instance, subjective TAG had the lowest perceptual cost, suggesting that while its ETA is lower than that of objective TAG, workers may perceive it as more difficult. In fact, an OLR model that combines ETA and PerceptionETA results in a completely rank-correlated prediction.

EXAMPLE APPLICATION: TASK EVALUATION

Using four studies, we showed that ETA is an accurate data-driven measure of task effort, and that it enables robust comparison across similar tasks. An accurate measure of the effort required to complete a task can enable requesters to compare different approaches to their tasks, iterate toward a bet-

		Subj.Rank	ETA	Time@10 [Time@10 Cost]	Time
Categorize	2	1.75 (1)	1.63 (1)	2.3 [0.8] (1)	3.74 (1)
	4	2.60 (2)	1.84 (2)	2.5 [0.8] (2)	4.11 (2)
	8	3.50 (4)	2.53 (3)	3.6 [1.2] (3)	5.71 (3)
	16	4.15 (5)	3.20 (5)	5.0 [1.7] (5)	6.42 (4)
Tag		3.00 (3)	3.08 (4)	4.2 [1.4] (4)	6.86 (5)

Table 5. As the number of alternatives increases, categorization tasks increase in difficulty. With 16 options, instead providing a direct tag is easier, even though it requires more time in an uncontrolled setting.

ter design, and price their task objectively. We have discussed how the measure might be transformed into price using a simple transform. We now provide an example of how it can also be used to prototype and evaluate task designs, using a similar approach to previous work that compared user interfaces by offering them for different prices on Mechanical Turk [30]. Using this same approach, task designers can iterate and make data-driven decisions about their own tasks.

In our example, we consider several alternatives to an image tagging task. There are many ways such a task could be designed. For example, workers could be asked to choose from among a number of existing tags or to enter the text of a new tag. We use ETA to decide when it is better to ask them to select tags versus enter text by comparing the TAG primitive ("Generate a label that applies to this image,") with the the CATEGORIZE primitive ("Pick the label that applies to this image,") where workers were shown 2, 4, 8, or 16 labels.

For each potential task design we collected the data necessary to calculate ETA from 20 workers, paying $2. For the purpose of this example, we also collected a subjective ranking from each worker. Relative performance information, however, is not necessary in practice to iterate on the design.

As the number of options increases, the time taken to generate a correct answer also increases (Table 5). ETA seems to increase sigmoidally, suggesting that answer strategies may change as the list of options gets longer. We also find an interesting discontinuity: asking workers to generate a tag from scratch requires more effort than asking workers to select from a list of 8 labels, but less effort than selecting from 16 options. Although workers took longer in the uncontrolled setting to generate a tag (6.9s vs 6.4s), they were able to provide the right tag within less time (4.2s vs 5.0s). In addition to the higher perceptual costs of processing more options, the paradox of choice may also partially explain the increased difficulty of selecting from more options [17].

Using other measures resulted in slight ranking differences: workers ranked tagging as easier than selecting from 8 options, but wished to get paid most for tagging. As noted previously, using subjective ranking makes comparison with other tasks difficult, and time or cost estimates tend to be noisy.

DISCUSSION

ETA represents a first step towards effectively modeling task effort. Future directions involve extending the measure and exploring alternatives, understanding the impact of monetary incentives, and exploring additional uses.

While we believe that ETA is broadly applicable, the metric has limitations. For one, it requires gold-standard responses or limited response variability. For tasks where this is not possible (e.g., written editorials) an additional crowdsourcing step could manually evaluate each response. Next, ETA measures the effort required for a task in the average case. However, perceived effort differs between workers depending on their intrinsic ability (e.g., math aptitude) [27]. Measuring ETA with respect to particular groups or individuals could account for differing expertise and help workers better select ability-appropriate tasks, as would alternative models that account for multiple dimensions of expertise (e.g., [31]). Additionally, task effort is moderated by the content of a task, not just its format. While in our experiments we focused primarily on image-based tasks, a reading comprehension multiple-choice question intuitively seems harder than choosing the right color for an image from a list for instance. Thus, in addition to differentiating primitives in general, ETA could be extended to understand different users and types of content.

Other uses of the error-time curve may reveal new dimensions of task effort. Characterizing the curve's shape could reveal subtler differences in task effort. Beyond measuring the perceptual effort of a task, ETA could also be used to capture motor costs. This would allow us to understand how effort is split between perceptive, cognitive, and motor function.

We could also study whether the ETA components associated with each primitive were additive. If the total effort required for a task is simply the sum of its constitutent primitives' ETAs, it would be easy to use the approach to iterate on complex tasks. However, tasks may have interaction effects with each other. For example, repeating a task may lead to learning effects or complacency. Interleaving different tasks may increase task time because of switching costs but potentially reduce boredom [28], and it would be valuable to extend ETA to capture these nuances.

Finally, our analysis focused on instances where workers are paid for their work. However, monetary incentives can influence task completion; paying more motivates workers to complete more tasks [24], while not paying at all tends to lead to fast but low-quality work [23]. Providing variable pay based on speed of completion could induce different points on the error-time curve, despite the fact that we saw that crowd markets were generally price-inelastic. Different approaches may be required to capture task effort using volunteers.

CONCLUSION

Crowdsourcing envisions an effective marketplace that connects interested requesters to on-demand human intelligence. However, that marketplace today is plagued with problems of quality and pricing. The Mechanical Turk market for lemons [14] may be the result of both a large number of poorly designed tasks (and hence poor quality crowd work [20]), as well as requesters' systematic underestimation of the effort required to complete tasks [12].

The error-time area (ETA) metric, which measures the effort required for crowdsourced tasks, can help improve the status quo. It accurately recovers workers' subjective rankings of

task effort, and can be cheaply computed with simple instruments. The metric can be used in pricing tasks, and enables rapid task prototyping and evaluation. Given a crowdsourced library of tasks and their associated measures, requesters can benchmark their tasks against known templates. This may make it possible to identify, using ETA, task transformations that use the primitives that require the least effort. With a platform that automatically measures ETA for tasks, workers can also use these scores to figure out whether a task is worth their time. By providing requesters and workers with improved signals of effort, tools such as ETA can pave the way towards a better crowdsourcing experience.

ACKNOWLEDGMENTS

This work was supported in part by NSF award IIS-1351131.

REFERENCES

1. Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles. `http://goo.gl/4SfnUi`.
2. Alwin, D. F., and Krosnick, J. A. The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly* (1985).
3. Anderson, J. R., Matessa, M., and Lebiere, C. ACT-R: A theory of higher level cognition and its relation to visual attention. *HCI* (1997).
4. Bernstein, M. S., et al. Soylent: a word processor with a crowd inside. In *UIST* (2010).
5. Czerwinski, M., Horvitz, E., and Cutrell, E. Subjective duration assessment: An implicit probe for software usability. In *IHM-HCI* (2001).
6. Dai, P., et al. Decision-theoretic control of crowd-sourced workflows. In *AAAI* (2010).
7. DeLeeuw, K. E., and Mayer, R. E. A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *J. Educ. Psychol.* (2008).
8. Faradani, S., Hartmann, B., and Ipeirotis, P. G. What's the right price? pricing tasks for finishing on time. *Human Computation* (2011).
9. Gevins, A., and Smith, M. E. Neurophysiological measures of cognitive workload during human-computer interaction. *Theor. Issues. Ergon.* (2003).
10. Hart, S. G., and Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Adv. Psychol.* (1988).
11. Herlocker, J., Konstan, J. A., and Riedl, J. An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. *Inform. Retrieval* (2002).
12. Hinds, P. J. The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *J. Exp. Psychol.-Appl.* (1999).
13. Impara, J. C., and Plake, B. S. Teachers' ability to estimate item difficulty: A test of the assumptions in the angoff standard setting method. *J. Educ. Meas.* (1998).
14. Ipeirotis, P. G. Mechanical Turk, low wages, and the market for lemons. `http://goo.gl/uoXA6Y`.
15. Ipeirotis, P. G. Analyzing the Amazon Mechanical Turk marketplace. *XRDS* (2010).
16. Ipeirotis, P. G., Provost, F., and Wang, J. Quality management on Amazon Mechanical Turk. In *KDD* (2010).
17. Iyengar, S. S., and Lepper, M. R. When choice is demotivating: Can one desire too much of a good thing? *J. Pers. Soc. Psychol.* (2000).
18. Joglekar, M., Garcia-Molina, H., and Parameswaran, A. Evaluating the crowd with confidence. In *KDD* (2013).
19. Kahneman, D. *Thinking, fast and slow*. 2011.
20. Kittur, A., et al. The future of crowd work. In *CSCW* (2013).
21. Kruger, J., and Dunning, D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Pers. Soc. Psychol.* (1999).
22. Lohman, D. F. The effect of speed-accuracy tradeoff on sex differences in mental rotation. *Percept. Psychophys.* (1986).
23. Mao, A., et al. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *HCOMP* (2013).
24. Mason, W., and Watts, D. J. Financial incentives and the performance of crowds. *SigKDD Explorations* (2010).
25. McCracken, J., and Aldrich, T. Analyses of selected LHX mission functions: Implications for operator workload and system automation goals. Tech. rep., 1984.
26. Pew, R. W. The speed-accuracy operating characteristic. *Acta Psychol.* (1969).
27. Robinson, P. Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Appl. Linguist.* (2001).
28. Rzeszotarski, J. M., et al. Inserting micro-breaks into crowdsourcing workflows. In *HCOMP* (2013).
29. Rzeszotarski, J. M., and Kittur, A. Instrumenting the crowd: using implicit behavioral measures to predict task performance. In *UIST* (2011).
30. Toomim, M., Kriplean, T., Pörtner, C., and Landay, J. Utility of human-computer interactions: Toward a science of preference measurement. In *CHI* (2011).
31. Welinder, P., et al. The multidimensional wisdom of crowds. In *NIPS* (2010).
32. Wickelgren, W. A. Speed-accuracy tradeoff and information processing dynamics. *Acta Psychol.* (1977).
33. Wickens, C. D. *Processing resources in attention, dual task performance, and workload assessment*. 1981.
34. Yao, B., et al. Human action recognition by learning bases of action attributes and parts. In *ICCV* (2011).