Shepherding the Crowd Yields Better Work

Steven P. Dow[1], Anand Kulkarni[2], Scott R. Klemmer[3], Björn Hartmann[4]

HCI Institute
School of Computer Science
Carnegie Mellon University
spdow@cs.cmu.edu

Industrial Engineering & Operations Research[2]
Computer Science Division[4]
University of California, Berkeley
anandk@berkeley.edu, bjoern@cs.berkeley.edu

Stanford HCI Group[3]
Computer Science Department
Stanford University
srk@stanford.edu

ABSTRACT

Micro-task platforms provide massively parallel, on-demand labor. However, it can be difficult to reliably achieve high-quality work because online workers may behave irresponsibly, misunderstand the task, or lack necessary skills. This paper investigates whether timely, task-specific feedback helps crowd workers learn, persevere, and produce better results. We investigate this question through *Shepherd*, a feedback system for crowdsourced work. In a between-subjects study with three conditions, crowd workers wrote consumer reviews for six products they own. Participants in the *None* condition received no immediate feedback, consistent with most current crowdsourcing practices. Participants in the *Self*-assessment condition judged their own work. Participants in the *External* assessment condition received expert feedback. Self-assessment alone yielded better overall work than the None condition and helped workers improve over time. External assessment also yielded these benefits. Participants who received external assessment also revised their work more. We conclude by discussing interaction and infrastructure approaches for integrating real-time assessment into online work.

Author Keywords

Crowdsourcing, human computation, feedback system.

General Terms

Experimentation, Design.

ACM Classification Keywords

H5.3. Information interfaces and presentation (e.g., HCI): Group and Organization Interfaces.

INTRODUCTION

Modern crowdsourcing systems enable distributed human problem solving on an unprecedented scale [17,35,36]. On micro-task platforms like Mechanical Turk, people perform short tasks for small amounts of money. This rapid, decontextualized, anonymous work provides few motivational incentives and little time for reflection [16,19,30]. Consequently, workers often expend minimal effort to complete tasks, yielding low-quality results.

CSCW'12, February 11–15, 2012, Seattle, Washington, USA.

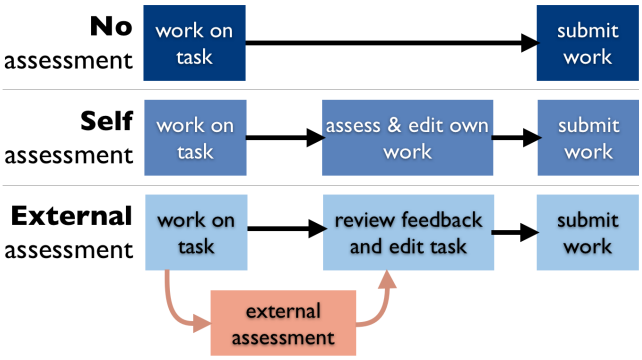

Figure 1: (A) Micro-task workers today receive no feedback. This paper investigates two interventions. (B) Workers reflect and edit their own work and (C) Workers review and edit after reading an external assessment.

How can we increase work quality in micro-task platforms? Increasing payment can improve individual performance, but paying more actually reduces quality after a certain point because it incentivizes speed [12]. Requesters can also increase quality by aggregating answers from multiple workers [2,16,23], filtering spammers by inserting test problems with known solutions [19,37], or using intelligent redundancy to simultaneously filter and aggregate [18]. Moreover, providing context about the rationale for a task can be a powerful motivator [5].

This paper explores the value of providing real-time assessment to help motivate and teach online workers to produce high-quality results. In many micro-task platforms, requesters and workers remain largely anonymous to each other, and little direct interaction occurs between them. We hypothesize that *task-specific feedback will help workers in micro-task markets perform better, learn over time, and persevere longer*. Concrete expert feedback can help people understand and strive toward key metrics in a task domain [28,29,31]. Assessments also provide motivation by making workers cognizant that other people are judging their work [1]. However, external feedback requires the time of someone with sufficient domain knowledge. Alternatively, a well-structured self-assessment may be able to provide guidance without the additional overhead [34]. Self-assessments can help workers see how their work aligns with key performance criteria and may be less discouraging than external assessments [7].

To examine the effects of different kinds of assessment in crowds, we introduce the *Shepherd* system, which provides targeted, real-time assessment as workers complete a series of tasks. In a between-subjects experiment, 105 participants wrote consumer reviews for products they own. This paper compares three scenarios (Figure 1): micro-task work with *No* assessment (the current status quo), work with instant *Self*-assessment, and work with real-time *External* assessment. Participants in both the *External* and *Self*-conditions were encouraged to make edits to their earlier work.

Both external and self-assessment led to significantly more highly-rated work than no feedback at all. There was no overall performance difference between external and self-assessment. External and self-assessment both led participants to produce better work across a series of tasks, indicating a learning effect. Participants in the External condition produced significantly more changes to their work than the other conditions, leading to more overall writing output.

In short, both external and self-assessment helped participants produce better work. External assessments lead workers to submit more work for the same payment, however this incurs the cost of an assessor's time. These results suggest that crowds can be shepherded through the use of concrete rubrics. The following sections elaborate the study's rationale and hypotheses.

Feedback supports community and transfer of expertise

In many communities, senior members help novices learn and stay motivated, often through implicit feedback [21]. Enabling social interaction among peers provides a learning opportunity [1]. (However, biased information can nullify these benefits or have a negative impact [15].) Traditional work environments foster employee development through formal performance reviews and feedback, and informally through apprenticeship, collocated awareness, and observation [21]. Likewise, online communities often provide infrastructure for moderators to review others' content and to encourage the growth of newer members [6,20]. Members choose where to devote resources, and through transparency and reputation systems, the community defines standards and quality control mechanisms [33]. Peer interaction also has motivational benefits [6,14]. LiveOps, a distributed online call center, enabled chat interaction between at-home agents to recreate a "water cooler" setting and to foster cohesion among their workforce [26]. Online freelancing platforms (e.g., oDesk) typically enable requesters and workers to send each other e-mail.

In contrast with traditional labor, peer-production, and freelance systems, micro-task platforms such as Amazon Mechanical Turk typically offer few formal or informal methods for worker-requester communication. In general, these systems provide no way for crowd workers to see what their peers produce, instructions are often the sole means for requesters to communicate to workers during the task, and there is no in-task channel for workers to communicate with requesters. At best, novice workers can only observe expert behavior from a few curated examples provided by the requester. Feedback from external reviewers can complement other quality-improvement efforts such as worker qualifications and clearer instructions. However, real-time communication is difficult to achieve on high-throughput micro-task platforms, where workers perform tasks at a small granularity, on the order of a few seconds to a few minutes.

Self-assessment rubrics capture domain knowledge

Self-assessment activities can help students reflect, learn skills, become more autonomous, and more clearly draw connections between learning goals and evaluation criteria [7,27,31]. Well-designed rubrics can help students align their own work with concrete grading criteria, leading them to better grasp a domain's key principles [7,31]. These benefits diminish when self-assessment rubrics are overly complex or use special domain language difficult for novices to understand [4]. Without clear criteria, learner's performance and their self-assessment often exhibits little correlation, unless the learner believes he/she will be later assessed by an authority [4]. In general, high performers tend to underrate and low-performers tend to overrate – and self-assessment can be used to make learners more aware of gaps between their standards and those of external assessors [25]. Peer assessment can also be valuable, especially when a concrete rubric effectively captures domain knowledge [10,32]. Social forces, such as friendships, can alter peer assessments, but in general, students still learn [24].

Hypotheses: shepherding leads to better work, learning, and perseverance.

This paper hypothesizes that shepherding workers through task-specific rubrics will produce better results, help workers learn, and lead to greater perseverance than current micro-task crowdsourcing practices. Further, we hypothesize that expert rubrics will be more impactful when delivered through an external assessor, rather than assessed by the workers themselves.

> **Hypothesis 1:** *Assessment (both through worker self-assessment and by an external expert) will help crowd workers produce better quality work.*
>
> **Hypothesis 2:** *Assessment will help crowd workers learn and improve their work over time.*
>
> **Hypothesis 3:** *Assessment will motivate crowd workers to produce more output.*
>
> **Hypothesis 4:** *The effect sizes for H1, H2, and H3 will be larger for external assessment than self-assessment.*

To measure work quality, a blind-to-condition expert rated the work results. To examine learning effects, we monitor each worker's progression in terms of expert rating over their set of assignments. To measure perseverance, we look at the percent of modified assignments, the change in number of characters between the original and modified assignments, and the overall amount of text produced.

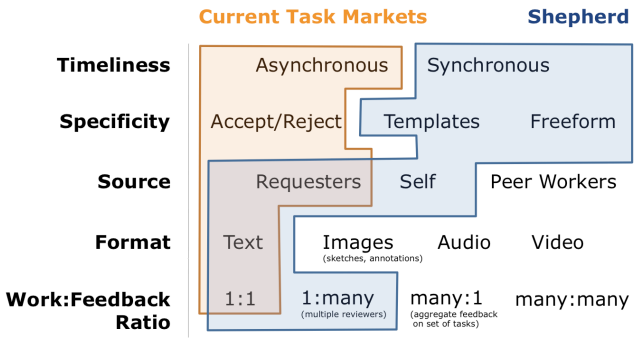

Figure 2: A design space for crowd feedback. Orange: current task markets deliver accept-reject feedback days later. Blue: *Shepherd* explores synchronous feedback with detailed rubrics.

DESIGN SPACE FOR CROWD FEEDBACK

To effectively design feedback mechanisms for micro-task platforms—and to support our goal of improving learning, engagement, and work quality—we generated a possible design space for crowd feedback and analyzed it along several key dimensions (Figure 2).

Timeliness: When should feedback be shown?

In micro-task labor, many workers perform tasks for a short time (seconds to minutes), and then move to other tasks by other employers [12,19]. This implies two timing options: synchronously deliver feedback when workers are still engaged in a set of tasks, or asynchronously deliver feedback after workers have completed the tasks.

Synchronous feedback may have more impact on future task performance since it arrives at a teachable moment, while people are still thinking about the task domain [11]. However, synchronous feedback places a burden on the feedback providers, because they have to review work quickly. Asynchronous feedback gives feedback *providers* more time to review and comment on work. However, *workers* may have forgotten about the task or feel unmotivated to review the feedback or revise their work.

Most current micro-task platforms provide support for asynchronous feedback, often days later (Figure 2, orange). Requesters can provide feedback at payment time, but at that point, workers may care more about getting paid than improving submitted work. More importantly, unless requesters have more jobs available, workers cannot act on requesters' advice. In our *Shepherd* system, we explore the efficacy of synchronous feedback (Figure 2, blue).

Specificity: How detailed should feedback be?

Currently, workers on most micro-task platforms receive a single bit of feedback—acceptance or rejection of their work. (In a few cases, micro-task workers receive feedback on training questions with known answers – but importantly not on actual work [37].) While additional freeform communication is possible, it is rarely used unless workers file complaints. It is feasible a micro-task platform could enable external reviewers to provide detailed and personalized feedback on each micro task. However, this requires external reviewers to spend time authoring open-ended feedback.

Assessment rubrics provide an alternative to freeform feedback. Rubrics codify domain knowledge into pre-authored statements providing an effective and efficient form of feedback [7,31]. However, rubric authors may struggle to completely encapsulate domain knowledge. Moreover, assessors may want to make comments that are not embodied in the rubric. *Shepherd* employs a pre-authored feedback rubric and provides a place for open-ended comments.

Source: Who should provide feedback?

Crowdsourcing requesters post tasks with specific quality objectives in mind; they are a natural choice for assuming the feedback role. However, experts often underestimate the difficulty novices face in solving tasks [13] or use language or concepts that are beyond the grasp of novices [8]. Moreover, as feedback volume increases or feedback becomes more specific, requesters may find it more difficult to complete work assessments in real-time (synchronously).

Alternatively, workers may benefit from assessing their own work [7,10]. By viewing assessment rubrics such as scoring templates, workers become aware of desirable work characteristics and can learn by aligning these characteristics with their own work [31].

As a third approach, workers can be paid to provide feedback to their peers. In theory, peer feedback increases the scalability as more crowd workers can be recruited to handle the volume of feedback needs. Our preliminary trials indicate that workers do perform tasks simultaneously and overlap (Figure 3). Systems like VizWiz demonstrate the feasibility of recruiting workers for tasks in nearly real-time [3]. Such a peer feedback system could have two tiers; more experienced workers could be promoted into the feedback role. This introduces the challenge of identifying and promoting knowledgeable and responsible workers. This paper leaves the possibility of peer feedback for future work. *Shepherd* supports self-assessment and rich synchronous feedback from a singe external reviewer.

SHEPHERD: SYSTEM DESIGN

To study the effects of different feedback mechanisms, we created *Shepherd* to visualize work progress and provide real-time work assessments. The *Shepherd* system manages workflows for tasks posted to Amazon Mechanical Turk. Task hosting and data collection occurs on our Web server. We developed *Shepherd* as a web application in PHP with a MySQL database and XMPP for instant notifications.

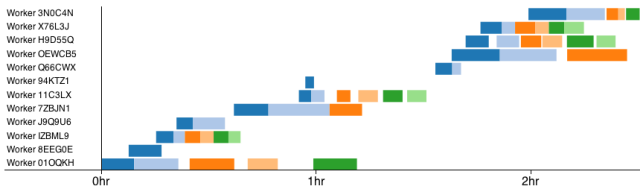

Figure 3: Shepherd's Gantt chart view.

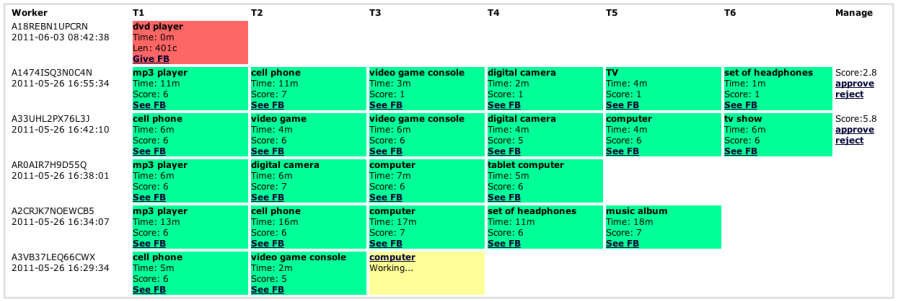

Figure 4: The dashboard interface displays completed tasks (green), in-process tasks (yellow), and tasks that need feedback (red). Requesters are notified via instant messaging.

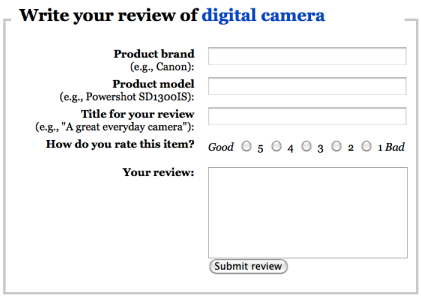

Figure 6: As a study task, participants write reviews for six products they own.

Shepherd's requester interface provides two different real-time views of workers and results. A Gantt chart view shows when workers accept a task, the length of time workers spend on each task, and how many tasks a worker completes within a batch (Figure 3). A dashboard view (Figure 4) displays tasks in each column and workers in each row. Each box represents one worker's single piece of work (e.g., a mobile phone review) and includes status details. The color corresponds to different task states: in progress (yellow), work needs feedback (red), or feedback applied (green). Requesters can monitor incoming work and click on any task to provide feedback using an assessment rubric (Figure 5). The rubric prompts the requester select from a set of feedback statements, to rate the consumer review on a scale from 1 to 9, and to enter freeform comments.

Shepherd supports three feedback modes for the worker: *No* assessment, *Self*-assessment, and *External* assessment. With no assessment, the work environment resembles any other task on micro-task platforms. For self-assessment, a rubric appears after each assignment. For external assessment, an expert assesses one task while that worker completes a subsequent task. The system automatically notifies assessors via instant messages when new product reviews arrive (Figure 7). The assessor follows the instant message link to immediately judge the work. By default, *Shepherd* delivers any new assessments to the worker before he or she begins a subsequent task.

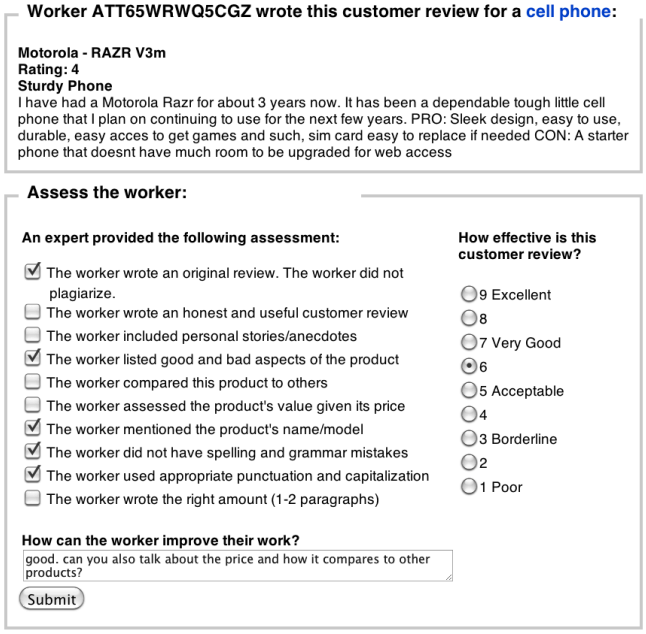

Figure 5: An assessment for the External condition. The Self condition uses the same rubric, but statements are phrased in first person (e.g., "I compared this product to others").

METHOD

A between-subjects study manipulated the writing process for a set of consumer reviews (Figure 6). This task fulfills key criteria: ① The task domain has precedence and relevance to the crowdsourcing community. ② Tasks solutions are open-ended, so expert feedback has the potential to improve results. ③ Performance can be measured.

Study Design

Participants write consumer reviews for six products they own. In the *External a*ssessment condition, an expert reads and judges each consumer review using a grading rubric (Figure 5). The expert performs the assessment while the participant completes the next consumer review. Before the subsequent task, the participant reads the expert assessment and optionally edits his/her consumer review. In the *Self-a*ssessment condition, participants reflect on their own work directly after each consumer review using a grading rubric. The rubric mirrors the external feedback condition's expert grading rubric, but frames each in first person (e.g., "I included personal stories/anecdotes in my review"). Self-assessment participants have the opportunity to edit their reviews. In the *None* condition, participants advance directly from one review to the next; they do not get an opportunity to modify their reviews.

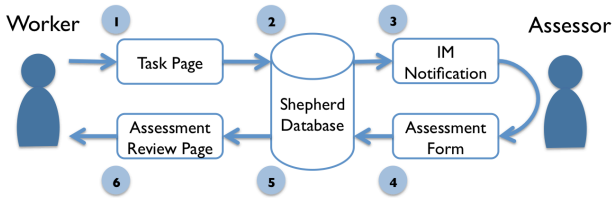

Figure 7: A worker writes a product review (1); when Shepherd receives the completed review (2), the assessor is immediately notified via instant messaging (3). The assessor provides feedback (4), which is again stored in the database (5). The form is shown to the original worker between tasks (6).

Participants

We recruited participants on Amazon Mechanical Turk over seven days in May 2011; 207 workers wrote reviews. Workers were assigned to conditions in round-robin fashion. Their average age was 25; 55% were from the United States, 35% from India, and 10% from other countries.

Procedure

The online experiment includes four main steps. First, participants select six out of twenty possible product categories they wish to review (Appendix A). Second, participants receive instructions on writing effective reviews. Instructions include the basic points of the assessment rubric and examples of high- and low-quality reviews. Third, participants write six reviews – one per chosen category. Fourth, participants fill out a short questionnaire and provide basic demographics. In accordance with current market rates on Mechanical Turk, participants earned $1.50 for their work. Regardless of quality, workers were paid if they completed all six reviews. Moreover, the instructions did not say their payment would be adjusted based on quality.

A member of the research team served as the external assessor in the experiment. To avoid bias, the researcher assessed every review in all three conditions in real-time without knowledge of the assigned condition. The assessor checked for plagiarized consumer reviews through an automated web search on the submitted text. The expert assessment took approximately 60 hours over seven days. Participants who self assessed or got an external assessment could revise their reviews. The assessor only judged the original, unedited version in real-time; the revised versions were judged independently after the experiment.

Dependent Measures

Task Performance

During the experiment, the expert assessor judged consumer reviews as they arrived. Each review received an *expert rating* between 1 (poor) and 9 (excellent). Each review was also scored on a checklist comprising ten features of effective consumer reviews (Figure 5). The number of checked features provides an *expert criteria count*.

After the experiment, all consumer reviews were re-posted to Mechanical Turk for a crowd assessment. Up to five workers judged each review using the same assessment rubric as expert assessors, but without freeform feedback (Figure 5). This yielded two additional performance measures: *crowd rating* and *crowd criteria count*. 615 crowd judges participated and received $0.02 per assessed review. To filter for erratic workers, assessments completed in less than 20 seconds were excluded, resulting in 408 crowd judges and 2229 valid assessments.

Learning

An individual's change in performance over a set of consumer reviews provides an indication of whether participants learned. The amount of learning for a condition was calculated by regressing expert ratings on condition, review order, and worker. Constant scores have a *slope coefficient* of 0; increasing scores have a positive coefficient; decreasing scores have a negative coefficient. This study did not measure long-term learning effects.

Perseverance

Participants could edit their consumer reviews after they self assessed or reviewed external assessment. The amount of revision provides a measure of a participant's willingness to persevere and improve their work. We measured the *ratio of revised reviews* to the total number of reviews. We also examined the length of reviews before and after edits using *overall character length* and *Levenshtein string edit distance* (number of character edits needed to transform the original into the revised review) [22]. As another measure of perseverance, we recorded the *length of time per review*.

To judge the value of revisions, a single independent assessor judged the original and revised versions of reviews, in randomized order and blind to condition. This provides *expert ratings and expert criteria counts for both original and revised reviews*. The expert assessor—a copy editor hired through odesk.com—earned a flat rate of $80 for 380 assessments and used the same rubric form (Figure 5). Likewise, the crowd assessments provide *crowd ratings and crowd criteria counts for both original and revised reviews.*

RESULTS

207 participants wrote product reviews in our experiment: 67 in the None condition, 67 in Self, and 77 in External (Table 1). 71 additional participants signed up but never submitted reviews. We excluded participants from our analysis for three reasons: First, 25 participants plagiarized reviews. The number of plagiarizing participants is not significantly different across conditions (X^2=1.60, df=2, p=0.44). Second, 52 participants were not assessed by the external assessor in time, due to work breaks. Third, 25 participants dropped out early, before experimental manipu-

Condition	None	Self	External	Totals
Participants entered	67	67	73	**207**
Plagiarizers	*-6*	*-8*	*-11*	*-25*
Missed assessments	*-12*	*-27*	*-13*	*-52*
Early dropouts	*0*	*-12*	*-13*	*-25*
Participants analyzed	49	20	36	**105**
Reviews analyzed	240	101	197	**538**

Table 1: Summary statistics

Category: **Game Console**
Title: **Nintendo Wii gets old, fast.**
Rating: **2 / 5**

At first, the Nintendo Wii was great fun, but the appeal quickly faded into boredom. The controls are gimmicky and tiresome. Compared to the other consoles out there, PS3 and XBOX 360, the Nintendo Wii just doesn't have the staying power. I had quite a bit of fun the first month or so of owning the Wii, but lost interest not too long after. I came to realize most of the games were the same concept over and over and over. Very boring. If you're buying this for kids, it may be a good plan, as it is very simple. For adults, I would recommend getting a more complex and versatile system with a better variety of games. I eventually sold my Nintendo Wii with all the games for a measly $75. Don't waste your money like I did.

Figure 8: One participant's consumer review.

lations could affect their work (before two reviews in Self, and three reviews in External). The number of dropouts in Self and External is not significantly different (X^2=1.09, *df*=1, *p*=0.30). The analyzed data set comprises 538 consumer reviews from 105 participants. Figure 8 shows an example review. The average review length was 570 characters (SD=223.1). Most participants were positive towards products, assigning an average rating of 4.29 (SD=0.90) on a scale of 1 to 5.

Both external and self assessment led to better work

Self-assessment participants start assessing themselves after review one; External participants see the first assessment after review two. To gauge whether assessment has an effect, we consider reviews 2-6 in the Self-condition, 3-6 in the External condition, and 1-6 in the None condition.

A single expert judged all reviews on a 9-point scale (1=poor and 9=excellent). Figure 9 (left) shows the distribution of expert ratings. An analysis of variances was performed with condition (External, Self, and None) as a factor, participant (worker ID) as a random effect and expert overall rating as the dependent variable. The External condition (μ=6.01, SD=1.38) and the Self condition (μ=6.35, SD=1.63) outperformed the None condition (μ=5.69, SD=1.19) (F(2,102)=3.02, p<0.05). Pair-wise comparisons with Welch two-sample t-tests indicate a significant difference between *None* and *External* (t(222)=2.48, p<0.05) and between *None* and *Self* (t(110)=3.35, p<0.05). There was no significant difference between *External* and *Self* assessment (t(150)=1.40, p=0.18). Participants produced higher-rated reviews when they self-assessed and when they received external feedback (Figure 10, left).

For the criteria count, the expert selected up to ten assessment criteria for each review (an excellent review satisfied all criteria; a poor review satisfied none). Figure 9 (right) shows the distribution of criteria counts. A one-way repeated-measures ANOVA was conducted with condition (External, Self, and None) as a factor, participant (worker ID) as a random effect and criteria count as the dependent variable. The ANOVA did not show a significant difference in the criteria count across conditions (F(2,102)=1.78, p=0.17, Figure 10, right).

To examine the crowd assessments, an analysis of variances was performed with condition (External, Self, and None) as a factor, judge (crowd ID) as a random effect, and crowd rating as a dependent variable. Crowd ratings were not significantly different across conditions (F(2,405)=0.93, p=0.40). A second analysis of variances was performed with condition (External, Self, and None) as a factor, judge (crowd ID) as a random effect, and crowd criteria count as a dependent variable. Crowd criteria counts were not different across conditions (F(2,405)=1.43, p=0.24).

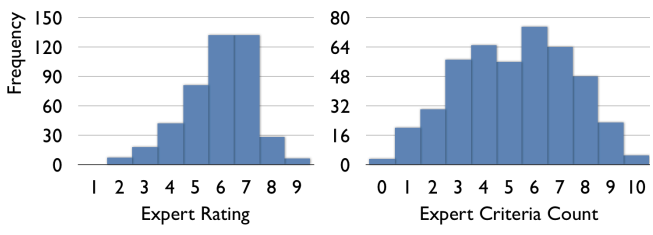

Figure 9: Distributions of expert ratings and expert choice of checkboxes (criteria counts) across all reviews.

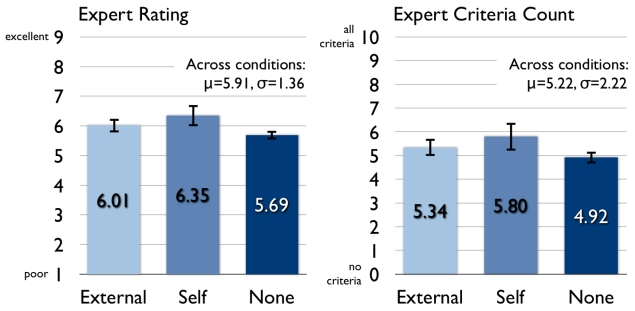

Figure 10: External and Self assessment led to higher overall expert ratings.

In summary, the independent expert assessor rated the reviews in the Self and External conditions higher than the None condition. Self-assessment had the highest scores, but not significantly higher than External assessment. The expert selection of satisfied criteria also favored Self and External assessment, although there were no significant differences across conditions. The crowd ratings and criteria counts did not yield useful data for distinguishing conditions; the discussion section further describes the tradeoffs of using the crowd to generate assessments.

Self assessment helped people learn

An expert rated each review; this analysis considers how those ratings changed over time. A multilevel linear regression was carried out, regressing expert rating on condition (External, Self, and None) and review order (0-5) as fixed effects, and participant (worker ID) as a random effect. We analyze interactions between condition and review order, since we are interested in differential effects of assessment conditions on learning. In the Self-condition, review ratings improved over time (Figure 11). The estimated coefficient with the Markov Chain Monte Carlo method was 0.25 (95% HPD credible interval = [0.09, 0.44]). This effect is significant (p=0.001). Review ratings in the External feedback condition also improved, but the effect was borderline significant (coefficient: 0.10, 95% HPD credible interval = [-0.03, 0.23], p=0.08). For the None condition, reviews did not improve, and order was not significant (coefficient: -0.03, 95% HPD credible interval = [-0.13, 0.06], p=0.38).

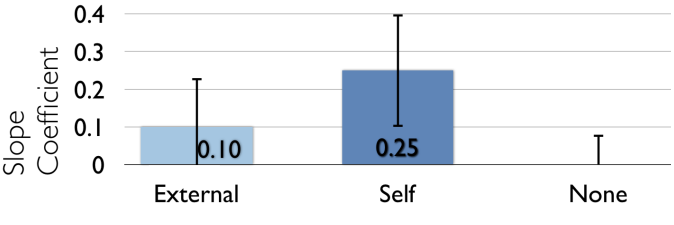

Figure 11: Participants' review ratings in the External and Self condition increased over time; Participants without feedback did not improve. Error bars indicate 95% HPD credible interval.

As a musician, the 3rd generation iPod Touch has helped me immensely in my music studies and in learning to play guitar. The app store is huge, and most of the apps are free or only $1. I was skeptical at first about buying this product because of its size (I like to carry my mp3 in my pocket). This is simply not a problem, even with a rubber case on it. My iPod is a music player, web browser, video player, metronome, synthesizer, video game system, graphing calculator, notebook, and much more. I base the iPod's value on the fact that I saved $100 by getting a graphing calculator app for only $1. I can't believe I kept my iPod Nano for so long when I could've had an iPod Touch. The only problem is the Facebook app, which is often very slow and unresponsive, even with updates. This is no a problem with the iPod, though. I strongly recommend this product for anyone, even my grandma has one!

Figure 12: An External assessment participant responded to feedback by adding text (in red and underlined). The original review received a rating of 5 and a criteria count of 4; it did not fully mention the product name, assess the product's value, and list good and bad aspects of the product. The revised review received a rating of 7 and a criteria count of 7.

In summary, self-assessors improved throughout the series of tasks; external assessment led to only marginal increases. The fact that the expert assessor could see the review order on the Shepherd dashboard presents a potential confound. However, the expert *was* blind to condition, so exposure to ordering cannot explain the observed differential effects.

External assessment encouraged more work revisions

Participants in the External and Self assessment conditions could edit their reviews in the feedback stage. In total, participants edited 184 reviews. 56.5% of External assessment participants changed their review, while only 24.8% of Self-assessment participants changed their review. External assessment led to a significantly larger ratio of revised reviews than Self-assessment (χ^2=47.24, p<0.05). Table 2 summarizes the work effort differences between conditions.

Participants who changed their reviews added an average of 119.8 characters (SD=157.8). In the External assessment condition, this was typically in response to expert feedback (Figure 12). An ANOVA was performed with condition (External and Self) as a factor, participant (worker id) as a random effect and review length change as the dependent variable. External assessment (μ=137.1, SD=175.8) led to (weakly significantly) longer revisions than Self-assessment (μ=84.0, SD=104.4) (F(1,81)=3.26, p=0.07).

Levenshtein string edit distance counts the number of characters that change between original and revised reviews. Participants' average string edit distance was 141.42 characters (SD=145.97). An ANOVA was performed with condition (External and Self) as a factor, participant (worker ID) as a random effect and string edit distance as the dependent variable. External assessment (μ=165.4, SD=158.6) led to significantly greater string edit distances than Self-assessment edits (μ=91.8, SD=98.8)(F(1,81)=8.20, p<0.05).

Figure 13: External feedback resulted in longer consumer reviews and more time spent.

Condition	None	Self	External
Num of revised reviews	N/A	60	124
Length change (characters)	N/A	84.0 (104.4)	137.1 (175.8)
Edit distance (operations)	N/A	91.8 (98.8)	165.4 (158.6)
Length before revision	541 (191.8)	577 (210.8)	600 (257.7)
Length after revision	N/A	599 (225.8)	672 (300.0)
Seconds to write original	352 (229.5)	359 (224.6)	422 (236.4)

Table 2: Differences in work effort across conditions: External assessment led to more changes, longer reviews, and more participation time for the same payment. Lengths are reported in characters (with standard deviation in parentheses).

For review length, an analysis of variances was performed with condition (External and Self) as a factor, participant (worker id) as a random effect and overall length as the dependent variable. Condition did not significantly affect the length of the original, non-revised reviews (F(2,102)=1.06, p=0.35). When we also consider revised reviews, the same ANOVA shows a significant difference between conditions (F(2,102)=4.10, p<0.05) (Figure 13). Post-hoc pair-wise Welch two-sample t-tests show that the difference between None and External is significant (t(321)=5.28, p<0.001), while None-Self (t(164)=2.261, p=0.025) and Self-External (t(256)=2.339, p=0.020) differences are weakly significant with respect to the Bonferroni corrected significance level of 0.017.

For the amount of time spent writing the original reviews, an ANOVA was performed with condition (External and Self) as a factor, participant (worker ID) as a random effect and time to write each review as the dependent variable. Condition did not significantly affect the amount of time spent on the original reviews (F(2,102)=1.43, p=0.24).

Revisions resulted in better overall consumer reviews

An expert and the crowd rated the quality of all original and revised reviews. A paired-samples t-test found that expert ratings were significantly higher for revised reviews (μ=5.87, SD=1.36) than for original reviews (μ=5.58, SD=1.34) (t(188)=2.09, p<0.05). A second paired-samples t-test found that expert criteria counts were higher for revised reviews (μ=6.12, SD=1.34) than for original reviews (μ=5.89, SD=1.23), but the difference is only weakly significant (t(188)=1.68, p=0.095) (Figure 14).

To examine the crowd assessments, an ANOVA was performed with version (Original and Revised) as a factor, judge (crowd ID) as a random effect and crowd rating as the dependent variable. Crowd ratings were (weakly signif-

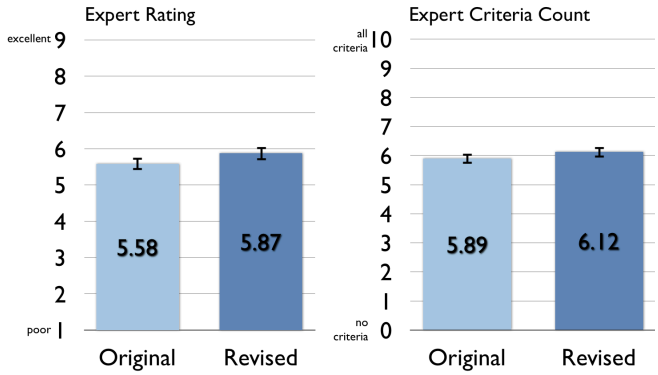

Figure 14: Mean expert ratings and criteria counts for original and revised reviews. Error bars indicate 95% CI.

icantly) higher for revised reviews (μ=6.37; SD=1.76) than for original reviews (μ=6.07; SD=1.90) ($F(1,291)=3.19$, $p=0.08$). Another ANOVA was performed with version (Original and Revised) as a factor, judge (crowd ID) as a random effect and crowd criteria count as the dependent variable. Crowd criteria counts for revised (μ=6.02; SD=2.06) and original reviews (μ=5.74; SD=2.01) were not significantly different ($F(1,291)=0.97$, $p=0.32$). The results suggest that revised consumer reviews receive better expert assessments than the original reviews. Crowd assessments favor revised reviews, but not significantly.

To examine whether revision quality improved differentially across conditions, an analysis of variances was performed with condition (External and Self) and version (Original and Revised) as factors and expert rating as the dependent variable. The ANOVA revealed a main effect for assessment condition ($F(1, 374)=4.55$, $p<0.05$), and a weak main effect for version ($F(1,374)=2.76$, $p=0.099$). There was no interaction effect between condition and version.

DISCUSSION

Both external and self-assessment led to better writing results and helped participants improve over time. These performance advantages were calculated on the original, non-revised versions of work. About 34% of participants in External and Self conditions also revised reviews. Revisions scored better than original reviews. External assessments also led participants to do more work per payment unit than self-assessments alone (although, this did not lead to differentially better quality on the revised versions).

The study confirmed some of our hypotheses about assessment: that feedback leads to better work, helps workers learn over time, and motivates more production. However, our hypothesis that external assessment produces larger effect sizes than self-assessment is not fully supported. While self-assessment did not lead to as many work revisions, it was just as effective as external assessment in increasing work quality in micro-task platforms. From a system evaluation perspective, we provide evidence that *Shepherd* improves crowdsourcing results. This section analyzes the effects of our interventions and discusses implications for future crowdsourcing systems.

Why did external and self-assessment help crowd participants improve their work?

In many contexts, assessment helps people improve performance and learn. However, it was not a foregone conclusion that assessment could improve micro-task work, given that workers perform tasks remotely and anonymously. Workers may not see—or comprehend—expert assessments; or they may not carry out self-assessments carefully. Future work assessments may include verification methods (e.g., [19]) to ensure that workers absorb the feedback.

Many workers dropped out before completing all reviews (and getting paid). We observed higher attrition rates in the Self (78%) and External (61%) than in the None condition (47%). Poor performing workers may have dropped out upon realizing they would be assessed and would need to put in extra effort. The fact that Self and External assessment led to higher quality work may be due to a combination of driving poor performers to drop out and actually helping the remaining workers learn the task. Future work should disentangle these two effects experimentally.

For workers who attempted to write consumer reviews in good faith, the assessment rubric served as an expert checklist [9]. Notably, there were no performance differences between workers who self-assessed and those who got external feedback. The concrete rubric enabled participants to understand key success criteria without external involvement. Including self-assessment in micro-task workflows provides a practical and scalable way to improve results.

Why did external assessment encourage more work?

Participants in the External assessment condition revised more of their consumer reviews, resulting in more text written and more time on task. There are several possible reasons for this increase in production over self-assessment. The fact that someone else is paying close attention may motivate workers to take more responsibility for their work. The power status of an external "expert" assessor may have led workers to make obligatory changes to their work out of fear of payment forfeiture. Finally, workers received open-ended comments from external experts, but self-assessors did not. Comments included suggestions for improvement that were more concrete than the general grading criteria.

Can we trust self-assessed ratings?

Self-assessing participants produced better work than those who receive no assessment. Can self-assessment scores inform filtering algorithms or guide the promotion of workers into leadership roles? In our study, self-assessors rated themselves 1.8 points higher (μ=7.9; SD=1.01) than experts' ratings of the same work (μ=6.1; SD=1.49). Criteria counts were also much higher (by 3.2 points) for self-assessors. The significance of the respective paired t-tests ($t(86)=9.96$, $p<0.05$ and $t(86)=13.02$, $p<0.05$) suggests that self-assessors systematically over-rate the quality of their own contributions. There is a moderate positive correlation between self-ratings and expert ratings (Spearman's rho=0.20, $p=0.057$). While self-ratings do not serve as an accurate measure of performance, concrete assessment cri-

teria help people reflect and make appropriate adjustments to current and subsequent work. Future work should investigate whether the difference in self-assessed and expert-assessed scores—the amount of score inflation—can help identify undesirable workers.

Can peer workers be effective shepherds?

If crowd workers can effectively assess each other, they could realize the benefits of external assessment at a much larger scale, without requester intervention. There are two key challenges for real-time peer assessment: scheduling and variance. Timely feedback is important for engaging workers before they complete a set of assignments. Participants spent an average of 6 minutes and 8 seconds per product review (SD=232.9) (Figure 15). Crowd workers hired after the experiment took an average of 58 seconds (SD=55.3) to assess a review (Figure 16). Therefore, real-time feedback by crowd workers is plausible if assessors can be recruited within this five-minute difference.

Bigham *et al.*'s success with real-time crowds [3] suggests that short recruitment times are achievable. In addition, we already observed simultaneous work occurring during our experiment: Figure 15 shows a Gantt chart of a subset of workers reviewing products. A fraction of these simultaneous workers could perform assessor roles. With deeper analysis of the arrival time data, we hope to develop algorithms for recruiting assessors within a desired time frame.

The high variance of peer assessment is also problematic. Our expert and the oDesk editor had moderate agreement on their ratings (Cohen's Kappa=0.41 using squared weights, Spearman's rho=0.39). The expert and crowd workers had lower agreement (treating the crowd as a single aggregate rater for comparison: Cohen's Kappa=0.20 using squared weights; Spearman's rho=0.28). Even when collected redundantly and averaged, crowd ratings can vary widely from an expert's opinion. To improve consistency among the crowd assessors, our filtering heuristic eliminates workers who completed the assessments too quickly, although more sophisticated techniques to infer worker quality exist [16,18]. Such techniques often rely on comparing answers to ground-truth values, but such "gold standards" may not exist for inherently creative tasks. An alternative approach would be to recruit only *experienced* workers who have been previously rated favorably by others. Future work should explore approaches for identifying expert workers suitable for the assessment role.

What tradeoffs do crowdsourcing platforms present for controlled experiments?

This study recruited participants through Amazon Mechanical Turk and paid $1.50 per task for an average of 27 minutes and 46 seconds of participation. This yields an effective hourly rate of $3.26; which is consistent with current crowd market rates [17] and much less than paying a freelance writer (currently $10-$15 per hour on oDesk.) From a crowdsourcing business perspective, a corpus of 881 consumer reviews (total non-plagiarized data) cost about $220 to generate. Despite the low costs, using crowd

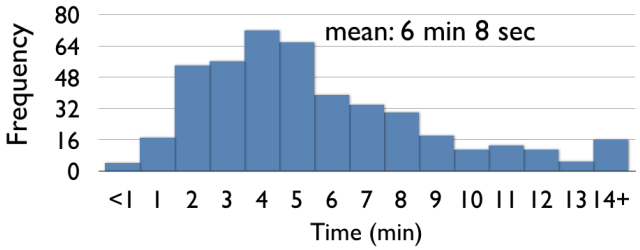

Figure 15: Distribution of time to complete reviews.

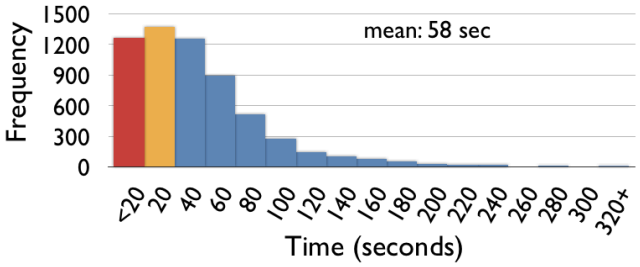

Figure 16: Distribution of time to assess reviews. Workers in red were excluded from our analysis; workers in orange are suspect as well.

participants created two main challenges: many workers dropped out early and many submitted plagiarized work.

CONCLUSIONS AND FUTURE WORK

This paper found that online workers produce better results when they self-assessed or received external feedback. External assessment led to more work output than self-assessment, but not necessarily better overall work. The results suggest a practical implication for requesters on crowd platforms: enumerate concrete criteria for the work output and then ask workers to self reflect on their prior work along those criteria. For task domains with more complex success criteria, a real-time expert assessor may also prove effective. Alternatively, the two approaches can be combined. Prompting for frequent self-assessment and providing intermittent expert feedback may deliver the benefits of both approaches at acceptable extra cost.

One direction for future work is to study whether crowd workers can effectively assess other workers. It may be prudent to recruit promising assessors based on prior task performance and domain knowledge by interspersing short test questions. In a related question, how does the quality of feedback affect work production and results? Follow-up studies may explicitly manipulate how good and bad feedback propagates within a peer assessment system. Further studies are needed to understand whether feedback provides advantages in different task scenarios and how feedback compares to other incentives for improving performance, such as adjusting worker pay based on quality measures.

ACKNOWLEDGEMENTS

We thank Truc Nguyen and Brie Bunge for early prototypes and CMU's Social Computing Lab for feedback. Funding support provided by the Hasso Plattner Design Thinking Research Program, Intel, and Google.

REFERENCES

1. Annett, J. Feedback and human behaviour: the effects of knowledge of results, incentives, and reinforcement on learning and performance. Penguin Books, 1969.
2. Bernstein, M.S., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D.R., Crowell, D., and Panovich, K. Soylent: a word processor with a crowd inside. *Proc of ACM Symposium on User Interface Software and Technology* (2010), 313–322.
3. Bigham, J.P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C., Miller, R., Tatarowicz, A., White, B., White, S., and Yeh, T. VizWiz: nearly real-time answers to visual questions. *Proc of ACM Symp. on User Interface Software and Technology* (2010), 333–342.
4. Boud, D. Sustainable Assessment: Rethinking assessment for the learning society. *Studies in Continuing Education* 22, 2 (2000), 151.
5. Chandler, D. and Kapelner, A. Breaking monotony with meaning: Motivation in crowdsourcing markets. *University of Chicago mimeo,* (2010).
6. Cheshire, C. and Antin, J. The Social Psychological Effects of Feedback on the Production of Internet Information Pools. *Journal of Computer-Mediated Communication* 13, 3 (2008), 705-727.
7. David, B. Enhancing Learning Through Self-Assessment. Routledge, 1995.
8. Ericsson, K.A., Charness, N., Feltovich, P.J., and Hoffman, R.R. The Cambridge Handbook of Expertise and Expert Performance. Cambr. Univer. Press, 2006.
9. Gawande, A. The Checklist Manifesto: How to Get Things Right. Metropolitan Books, 2009.
10. Hanrahan, S.J. and Isaacs, G. Assessing Self- and Peer-assessment: The students' views. *Higher Education Research & Development* 20, 1 (2001), 53.
11. Havighurst, R.J. Human Development and Education. Longmans, Green and Co, 1955.
12. Heer, J. and Bostock, M. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. *Proc of ACM conf on Human factors in computing systems* (2010), 203-212.
13. Hinds, P. The Curse of Expertise: The Effects of Expertise and Debiasing Methods on Predictions of Novice Performance. *Journal of Experimental Applied Psychology* 5, (1999), 205-221.
14. Horton, J.J. Employer Expectations, Peer Effects and Productivity: Evidence from a Series of Field Experiments. *SSRN eLibrary*, (2010).
15. Hullman, J., Adar, E., and Shah, P. The impact of social information on visual judgments. *Proc of conf on Human factors in computing systems,* ACM (2011), 1461–1470.
16. Ipeirotis, P.G., Provost, F., and Wang, J. Quality management on Amazon Mechanical Turk. *Proc of ACM SIGKDD Workshop on Human Computation*, (2010), 64–67.
17. Ipeirotis, P.G. Analyzing the Amazon Mechanical Turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students* 17, 2010, 16–21.
18. Karger, D., Oh, S., and Shah, D. Budget-optimal Crowdsourcing using Low-rank Matrix Approximations. *Proc. of the Allerton Conf. on Communication, Control, and Computing*, (2011).
19. Kittur, A., Chi, E.H., and Suh, B. Crowdsourcing user studies with Mechanical Turk. *Proc of SIGCHI conference on Human factors in computing systems, ACM* (2008), 453-456.
20. Lampe, C. and Resnick, P. Slash(dot) and burn: distributed moderation in a large online conversation space. *Proc. of the SIGCHI conference on Human factors in computing systems*, ACM (2004), 543–550.
21. Lave, J. and Wenger, E. Situated Learning: Legitimate Peripheral Participation. Camb. University Press, 1991.
22. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10, 8, 707-710.
23. Little, G., Chilton, L.B., Goldman, M., and Miller, R.C. Exploring iterative and parallel human computation processes. *Proc. of the ACM SIGKDD Workshop on Human Computation, ACM* (2010), 68–76.
24. Masten, A.S., Morison, P., and Pellegrini, D.S. A revised class play method of peer assessment. *Developmental Psychology* 21, 3 (1985), 523-533.
25. Mattheos, N., Nattestad, A., Falk-Nilsson, E., and Attstrom, R. The interactive examination: assessing students' self-assessment ability. *Medical Education* 38, 4 (2004), 378-389.
26. Musico, C. There's No Place Like Home. *destinationCRM.com*, 2008.
27. Orsmond, P., Merry, S., and Reiling, K. A Study in Self-assessment: tutor and students' perceptions of performance criteria. *Assessment & Evaluation in Higher Education* 22, 4 (1997), 357.
28. Sadler, D.R. Formative assessment and the design of instructional systems. *Instructional Science* 18, 2 (1989), 119-144.
29. Shute, V.J. Focus on Formative Feedback. *Review of Educational Research* 78, 1 (2008), 153 -189.
30. Silberman, M.S., Ross, J., Irani, L., and Tomlinson, B. Sellers' problems in human computation markets. *Proc ACM SIGKDD Workshop on Human Computation,* (2010), 18–21.
31. Taras, M. Using Assessment for Learning and Learning from Assessment. *Assessment & Evaluation in Higher Education* 27, 6 (2002), 501.
32. Taras, M. To Feedback or Not to Feedback in Student Self-assessment. *Assessment & Evaluation in Higher Education* 28, 5 (2003), 549.
33. Viégas, F., Wattenberg, M., and Mckeon, M. The Hidden Order of Wikipedia. *In Online Communities and Social Computing.* 2007, 445-454.
34. Zimmerman, B.J. Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary Educational Psychology* 11, 4 (1986), 307-313.
35. http://en.wikipedia.org/.
36. http://www.thejohnnycashproject.com/.
37. Crowdflower. Crowdflower.com/.

APPENDIX A:

Twenty product categories for writing reviews

MP3 player	Refrigerator	Digital camera	TV
Mobile phone	Microwave	Computer	Headphones
DVD player	Vacuum	Printer	TV show
Video game	GPS device	Tablet	Music album
Game console	Camcorder	Washer/dryer	Set of speakers