Workflow Transparency in a Microtask Marketplace

Peter Kinnaird
HCI Institute
Carnegie Mellon University
Pittsburgh, PA 15213
kinnaird@cs.cmu.edu

Laura Dabbish
HCI Institute and Heinz College
Carnegie Mellon University
Pittsburgh, PA 15213
dabbish@cmu.edu

Sara Kiesler
HCI Institute
Carnegie Mellon University
Pittsburgh, PA 15213
kiesler@cs.cmu.edu

ABSTRACT
Interdependent tasks in Mechanical Turk (MTurk) can be managed efficiently with a workflow, a sequence of tasks through which work passes to its completion. We ask if workers should be informed about the workflow, which we call workflow transparency. Transparency could motivate workers or induce social loafing. We describe three experiments to determine the effects of workflow transparency in MTurk. We compared a text description of the workflow, a visualization of the workflow, and the combination of text and visualization with a control condition giving no workflow information. Workflow transparency marginally increased volunteerism on a charity identification task (experiment 1) and significantly increased volunteerism and quality on a business identification task (experiment 2). Results were weaker with a less experienced worker sample (experiment 3). We suggest further research on the design of workflow information to increase workers' motivation.

Categories and Subject Descriptors
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Human Factors

Keywords
Crowdsourcing, workflow, visualization, CSCW, task motivation, productivity, entitativity

1. INTRODUCTION
Employers and researchers are increasingly using microtask platforms such as Amazon's Mechanical Turk (MTurk), Mobile Works, Short Task, or Serv.io to accomplish work that can be separated into portions and distributed to a number of people. Typically, these workers complete brief tasks online such as identifying photos or transcribing text from photos or recordings for piecework pay. This work is a form of crowdsourcing that can sum to significant productive labor.

Fully leveraging these systems remains challenging. Quality can suffer when workers have a fleeting relationship with the employer and low pay [2, 13, 20]. Piecework is associated in some people's minds with sweatshops, undermining motivation [7]. MTurk amplifies these threats to motivation through specific design decisions such as automatic task approval, the absence of a strong reputation management system, and impersonal persistent identifiers. Even simple jobs can employ dozens or hundreds of interchangeable, anonymous workers for a single task, apparently reducing their felt accountability for contributions, task meaningfulness, and perceived instrumentality of effort. Decades of research on groups shows that lack of contribution identifiability and interchangeability significantly reduce trust and motivation and increase free riding [5], social loafing (reductions in individual effort that sometimes occur in group settings) [11], and propensity to withhold effort [4, 12].

Employers can minimize the result of poor worker motivation with carefully designed workflows featuring high task redundancy and internal checks to account for low worker quality. Some schemes address work quality through complex incentive designs including threats of nonpayment for poor or incomplete work (e.g. [17, 19]). MTurk's best practices guide suggests asking multiple workers to complete each HIT.

We wondered if greater transparency might increase worker motivation in MTurk. The majority of work on MTurk requires the coordination of many workers; there are often anywhere from 3-7000 HITs available in a collection of tasks. This situation creates the option of forming the collection of workers into a group. Groups can be more motivating than collections of individuals because group identity enlists loyalty and social feelings among people [16]. We speculated that increasing the transparency of the workflow, that is, the apparent connectivity of tasks and workers, might increase worker motivation. We hoped to increase workers' perceived instrumentality by providing them with a big picture of the work to which they contribute in the crowdsourced setting and information about their role in this work. We accomplished this goal by providing workers with workflow information, that is, information about the division of labor, the overall sequencing of tasks, and where the worker fits in the workflow. Theories of collective effort suggest that this information might enhance worker motivation by increasing their perceived accountability and instrumentality. Further, one could argue that workflow transparency should be provided to workers to accord with the ACM's Code of Ethics sections 1.2 and 1.3.

Nonetheless, the effects of workflow information must be carefully tested because this information could potentially backfire and create harm by revealing to workers that they are only a minute piece of a larger process. Another question is how workflow information should be presented, for example, in a simple text description or through a visual representation that would enable capturing the workflow at a glance.

We conducted three experiments to examine the influence of workflow information on worker motivation. The results suggest that providing workflow information can cause greater volunteerism and higher quality work but that positive effects may be limited to particular combinations of tasks and worker populations.

GROUP'12, October 27–31, 2012, Sanibel Island, Florida, USA.

2. HYPOTHESES

Almost all crowdsourced tasks present the worker with information only about their own small task even when it is embedded within a larger workflow and may be interdependent with others' work. For example, workers may be asked to transcribe part of a conversation that will be combined with other transcriptions, proof read, then corrected. Workers typically do not know how many other workers have already contributed to the task, how many will come after them, or even that they are working with others towards a common goal. In MTurk, however, workers are often implicitly working with others through one coordination scheme or another since HITs are commonly replicated or strung together in workflows [1, 14, 15]. In addition to social and institutional isolation, tasks are typically not challenging. Many crowdsourced tasks are repetitive or boring, forcing employers to motivate the work entirely from pay [19].

Numerous studies have shown that suggesting to workers that they are redundant will reduce their effort and motivation [4, 6, 8, 10, 11, 18], the de facto context in MTurk. Research on job design suggests that providing employees with information about the importance of their work to the organization can change the amount of effort they expend on their tasks. In a test of these ideas, Hackman and Oldham [6] found that task identity and task significance enriched work, and influenced employees' internal motivations to perform their jobs effectively. Given the implicit use of workflows in MTurk a priori, it is unclear what effect showing workers their role in a workflow might have on their motivation. We formulate two hypotheses that the motivational effects of seeing work as a contribution to a larger goal will outweigh the inducement of social loafing behaviors.

H1*: Workers with workflow information will complete more work more accurately than those without workflow information.*

Workflows can be explained in text or they can be visualized as flow charts, trees, or as directed acyclic graphs. Gestalt theory suggests a number of perceptual features of visualizations which may improve entitativity resulting in higher motivation. For example, objects that are close together are perceptually grouped together. Likewise, similar visual elements become perceived groups. Visual elements that are connected or continuous tend to be grouped together, and symmetrically arranged objects are more likely to be perceived as a whole [3, 21]. Therefore, we anticipate that the effects of workflow transparency will be greater for visualization conditions than for the text condition.

H2: *Workers with graphical visualizations of workflow information will complete more work more accurately than those provided with text information about the workflow.*

3. METHOD

To test our hypotheses about the impact of a transparent workflow, we conducted two between-subjects experiments on Mechanical Turk and then replicated these studies with a third experiment. We limited participation to the United States to reduce the impact of language understanding on the results. Workers could participate in just one of the experiments. Participants were paid 10 cents for completing a HIT (worker task). All experiments employed the same procedure (see below). After participants accepted the HIT and completed the minimum required work they were able to advance to a brief survey.

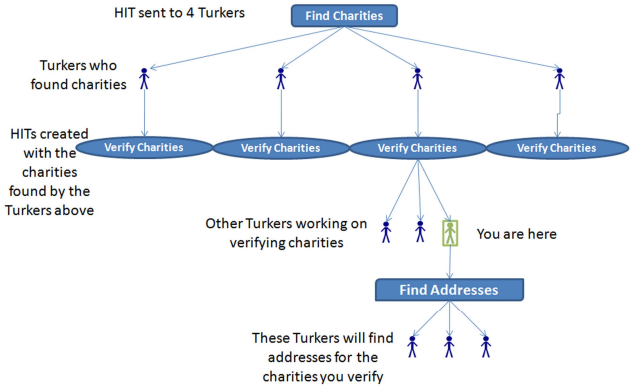

Figure 1. The workflow visualization as shown to Turkers.

3.1 Tasks and Independent Variables

In experiment 1, participants were presented with a list of 20 charities of which half were fictional (e.g. "Citizens for Animal Projection"). The worker's task was to check the correctness of the charities' names. In the second experiment, we replicated the task but in this case participants were given 20 businesses, half of which were fictional (e.g. "Duff Beer"). In both experiments, workers had to check on the validity of at least three of the listed organizations. The third experiment replicated the first two with random assignment to business or charity task.

Participants in all experiments were randomly placed into one of four conditions: no workflow information (control) condition, text-only description of the workflow, tree visualization of the workflow (see Figure 1), or both the text and the tree. The third experiment used a 2 x 4 design incorporating random task assignment in addition to random workflow condition assignment. In the text workflow condition, the text said:

You are working with a group of other Turkers to help identify real charities [businesses]. Your job is to determine whether the charities [businesses] listed below are real or not. Other Turkers have built up this list, but we need you to verify that they are real. Once you've verified the charities [businesses], other Turkers will be asked to find their addresses and phone numbers.

3.2 Dependent Measures

Participants were instructed that were only required to work on three charities [businesses] to receive pay. Thus any work beyond three items represents volunteer work. We used volunteerism

Table 1. Results of experiment 1: The effect of workflow information worker motivation in a charity validation task.

Dependent Measures	Control Mean (SE)	Text Mean (SE)	Vis Mean (SE)	Text & Vis Mean (SE)	Planned Contrast (Workflow vs. None)
Number of items completed	9.33 (0.88)	11.93 (0.88)	11.60 (0.91)	10.41 (0.93)	F [1, 284]=3.7, p=.05
Accuracy	0.81 (0.02)	0.80 (0.02)	0.80 (0.02)	0.80 (0.02)	n.s.
Net contribution	5.31 (0.61)	6.47 (0.61)	6.43 (0.64)	5.56 (0.64)	n.s.
Entitativity	3.07 (0.13)	3.26 (0.14)	3.29 (0.14)	3.27 (0.14)	n.s.

Table 2. Results of experiment 2: The effect of workflow information worker motivation in a business validation task.

Dependent Measures	Control Mean (SE)	Text Mean (SE)	Vis Mean (SE)	Text & Vis Mean (SE)	Planned Contrast (Workflow vs. None)
Number of items completed	6.31 (2.07)	11.38 (1.87)	11.24 (1.81)	12.92 (2.07)	$F [1,55] = 5.5, p < .02$
Accuracy	0.56 (0.06)	0.67 (0.06)	0.71 (0.06)	0.77 (0.06)	$F [1,55] = 4.1, p < .04$
Net contribution	1.38 (1.17)	4.50 (1.05)	4.06 (1.03)	5.85 (1.17)	$F [1,55] = 6.6, p < .01$
Entitativity	2.84 (0.28)	3.42 (0.24)	3.82 (0.24)	3.67 (0.25)	$F [1,50] = 6.6, p < .01$

(total number of items completed) and quality of work (accuracy rate) as behavioral measures of worker motivation. We also created a measure of net contribution which is calculated as the sum of correct responses minus the sum of incorrect responses. Large positive values for net contribution indicate a strong worker contribution to the overall subtask of which they are a part. Negative values mean that they have harmed the overall process more than they have helped it. This measure is, arguably, the most important for employers primarily concerned with productivity.

We adapted an entitativity questionnaire from Postmes et al. [9] to measure how much workers felt part of a larger group. This perception has been associated with lower social loafing [9, 11]. Items were presented on 5-point Likert scales. Two additional questions measured pertained to Turker tenure (number of prior HITs the workers had completed duration of activity on MTurk). In experiment 3 we introduced a manipulation check to ensure that workers were aware that they were only required to complete 3 items. 74% of workers passed the check.

4. RESULTS

4.1 Experiment 1: Charity Identification

The charity task had a 22% dropout rate, that is, workers signed up but did no work, leaving 288 participants. These workers were evenly spread across conditions (p = 0.31). We observed generally high output and quality, suggesting that the task was intrinsically interesting or easy. Despite our requiring participants to complete only three items to receive pay, more of them completed all 20 items (n = 100) than completed the minimum required (n = 75). The mean net contribution was 5.92 items and overall accuracy mean was 80.26% correct items.

H1 predicted that the presence of workflow information would improve the worker's motivation, operationalized as total items completed, item accuracy, and net contribution. To test H1, we ran an ANOVA across all conditions and then ran a planned contrast comparing the control condition to the other conditions. Participants completed marginally significantly more work in the workflow conditions as compared to the control condition (see Table 1). There were neither differences in accuracy, entitativity, or perceived task importance between the workflow conditions and control condition, nor between the text information and the visualization conditions. H2, therefore, was not supported

4.2 Experiment 2: Business Identification

In experiment 2, the method was the same but participants were asked to identify whether business listed were real or fake business. We again observed a 22% dropout rate with zero items completed with no significant differences across conditions, leaving 59 participants. However, once workers began, the task appears to have been less attractive or possibly more difficult than the charity validation task used in experiment 1. Overall net contributions mean was just 3.98 (SD = 4.4) versus 5.92 in experiment 1, and mean accuracy mean was just 67% versus more than 80% in experiment 1.

In H1 we predicted that the presence of workflow information would improve workers' motivation. In this study, participants completed significantly more work, more accurate work, higher net contribution, and perceived significantly greater entitativity in the presence of workflow information. These results also are shown in Table 2.

4.3 Experiment 3: Putting It All Together

In experiment 3, participants (N = 267) were randomly assigned to workflow information condition and to either the charity or business task. The dropout rate for experiment 3 was not significantly different from the first 2 experiments. However, Turkers in experiment 3 were significantly less experienced in MTurk as compared with Turkers in experiments 1 and 2, as measured by HITs completed (mean = 308 [SD = 373] vs. mean = 452, [SD = 428], $p < .004$) and days using MTurk (mean = 225 days [SD = 435] vs. mean = 290 days [SD = 442], $p < .0001$). We think this phenomenon occurred because we did not allow the more frequent users of MTurk, who would have participated in the first experiments, participate in experiment 3. As a result, a larger proportion of Turkers in experiment 3 were relative newcomers. In general, the results were in the same direction as experiments 1 and 2 but they were not statistically significant overall. However, we were able to demonstrate with our random assignment of charity vs. business tasks that the charities task was indeed more motivating (or less difficult) than the business task. Turkers in the charities task were more accurate than those in the businesses condition by about 13% ($p < .0003$) and their net contribution was about 58% higher ($p < .0003$).

The comparatively low experience level of this sample prompted us to examine statistical interactions between condition, task type, and experience. In doing so, we could determine whether workflow transparency and task assignment varied with worker experience. We found that more experienced workers perceived greater entitativity on the charities task but less entitativity on the businesses task relative to less experienced workers ($p < .03$). More experienced workers also completed significantly more items ($p < .02$) and made a higher net contribution ($p < .01$) to the charities task ($p < .02$) than the business task.

The findings above suggest that the effects found in experiments 1 and 2 may hold only for more experienced workers. Perhaps workflow information simply confuses newcomers. The three-way interaction ($p < .05$) of workflow information, task, and worker experience allows us to examine this possibility. We found a significant interaction for overall accuracy ($p < .007$) and net contribution ($p < .05$). Workflow information improved accuracy ($p < .01$) and contributions (but not significantly) for experienced workers given the business task, replicating experiment 2. However, workflow information only helped inexperienced workers on the charities task ($p < .01$).

5. DISCUSSION

In all 3 experiments, workers who completed more than 3 items were volunteering for additional work. The tasks elicited a surprising amount of voluntary work regardless of condition.

The first experiment we conducted provided just marginal support for our hypothesis that workflow information will increase motivation (H1) and no support for our hypothesis that this information would have a stronger influence on motivation when presented visually (H2). A limitation of this experiment is that we might have experienced an important ceiling effect. With 31% of participants performing at 100% accuracy, the likelihood that a manipulation would significantly increase accuracy was low.

The second experiment used the same setup as the first but showed much stronger results. The move from the charity to business identification task apparently reduced workers' intrinsic interest in the task, or perhaps checking the existences of businesses was harder than checking charities. We found that workflow information increased the number of items completed, accuracy, the entitativity (group feeling), and net contribution (productivity) improved by 248% (mean of three visualization conditions against the control).

Finally, experiment 3 highlighted the importance of Turkers' relative experience. Workflow visualizations lead to significantly greater work in a number of scenarios, but, in one scenario, actually reduced the number of items completed by inexperienced workers in the business task. Perhaps inexperienced workers were confused by the task or visualization, while more experienced workers felt like part of a team or group and thus motivated to contribute to a larger goal.

We did not find a difference between a text description of the workflow and the graphical visualization. Thus we cannot say whether a different graphical design would be more effective or whether visualizations are not necessary to impart workflow information.

6. IMPLICATIONS FOR DESIGN

These results demonstrate the effectiveness of including workflow information when designing the presentation of some crowdsourced tasks. We also contribute evidence that MTurk workers with varying levels of experience respond differently to experimental manipulations. Future work will examine alternative positioning of the worker in the workflow, interactive and real-time visualizations which might depict information about other workers' contributions, and alternative visualization paradigms.

7. ACKNOWLEDGMENTS

This work was supported by a National Science Foundation grant, CNS-1040801 and the Center for the Future of Work at Carnegie Mellon University's Heinz College

8. REFERENCES

[1] Bernstein, M.S. et al. 2010. Soylent: a word processor with a crowd inside. *UIST* (2010), 313–322.

[2] Downs, J.S. et al. 2010. Are Your Participants Gaming the System? Screening Mechanical Turk Workers. *Science*. (2010), 2399-2402.

[3] Fekete, J.-D. et al. 2008. Information Visualization. A. Kerren et al., eds. *Springer-Verlag*. 1-18.

[4] George, J.M. 1992. Extrinsic and Intrinsic Origins of Perceived Social Loafing in Organizations. *The Academy of Management Journal*. 35, 1 (Mar. 1992), 191-202.

[5] Granovetter, M.S. 2005. The Impact of Social Structure on Economic Outcomes. *J.Econ.Persp.*19,1(2005),33-50.

[6] Hackman, J.R. and Oldham, G.R. 1976. Motivation through the design of work: test of a theory. *Org. Behav. and Hum. Perf.* 16, 2 (1976), 250-279.

[7] Harris, M. 2008. Email from America. *Sunday Times*.

[8] Huberman, B.A. et al. 2009. Crowdsourcing, attention and productivity . *J.of Info.Sci*. 35,6(Dec. 2009),758-765.

[9] Jans, L. et al. 2011. The induction of shared identity: The positive role of individual distinctiveness for groups. *Pers. soc. psych. bulletin*. 37, 8 (2011), 1130-1141.

[10] Jones, G.R. 1984. Task Visibility, Free Riding, and Shirking: Explaining the Effect of Structure and Technology on Employee Behavior. *Academy of Management Review*. 9, 4 (Oct. 1984), 684-695.

[11] Karau, S.J. and Williams, K.D. 1993. Social Loafing: A Meta-Analytic Review and Theoretical Integration. *J. of Pers. and Soc.Psych.*. 65, 4 (1993), 681-706.

[12] Kidwell, R.E. and Bennett, N. 1993. Employee propensity to withhold effort: A conceptual model to intersect three avenues of research. *Acad. of Mgmt* 1993.

[13] Kittur, A. et al. 2008. Crowdsourcing user studies with Mechanical Turk. *CHI 08*. 08, April 5-10 (2008), 453.

[14] Kittur, A. et al. 2011. CrowdForge: Crowdsourcing Complex Work. *HCI* (2011), 1801-1806.

[15] Kulkarni, A. et al. 2012. Collaboratively crowdsourcing workflows with turkomatic. *CSCW* 2012.

[16] Lakens, D. 2010. Movement synchrony and perceived entitativity. *J. of Exp.Soc.Psych.* 46, 5 (2010), 701-708.

[17] Mason, W. and Watts, D.J. 2010. Financial incentives and the "performance of crowds." *ACM SIGKDD Explorations Newsletter*. 11, 2 (2010), 100.

[18] Roland E. Kidwell, J. and Bennett, N. 1993. Employee Propensity to Withhold Effort: A Conceptual Model to Intersect Three Avenues of Research. *The Academy of Management Review*. 18, 3 (Jul. 1993), 429-456.

[19] Shaw, A.D. et al. 2011. Designing Incentives for Inexpert Human Raters. *CSCW*. 2011.

[20] Snow, R. et al. 2008. Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. *Emp.Meth.in Nat Lang.Proc.* 2008.

[21] Ware, C. 2000. *Information visualization: perception for design*. Morgan Kaufmann Publishers Inc.