

---

# Exploring Iterative and Parallel Human Computation Processes

**Greg Little**

MIT CSAIL  
32 Vassar St  
Cambridge, MA 02139 USA  
glittle@gmail.com

**Abstract**

Mechanical Turk (MTurk) is an increasingly popular web service for paying people small rewards to do human computation tasks. Current uses of MTurk typically post independent parallel tasks. This research explores an alternative iterative paradigm, in which workers build on each other's work. We run a couple of experiments comparing the efficacy of this paradigm in two different problem domains: image description writing, and brainstorming company names.

**Keywords**

Mechanical Turk, human computation

**ACM Classification Keywords**

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

**General Terms**

Experimentation, Measurement, Performance

**Introduction**

Amazon's Mechanical Turk is a real-time labor market for tasks that pay on the order of cents for completion. Typical uses of MTurk include labeling images and verifying data – tasks which are easy to subdivide into many non-interacting sub-tasks.

---

Copyright is held by the author/owner(s).  
*CHI 2010*, April 10–15, 2010, Atlanta, Georgia, USA.  
ACM 978-1-60558-930-5/10/04.

We are interested in more complicated tasks which *do* have interacting sub-tasks, some of which build on each other.

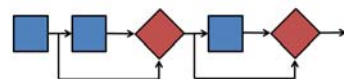
This research explores two such tasks: writing image descriptions, and brainstorming company names. Both examples involve MTurk workers (turkers) generating content. The question we want to answer is: how can we generate the highest quality content?


We compare two models of execution for each task: iterative and parallel. The iterative process has one turker generate content, and then shows this content to another turker so that they can build upon it. The result is then given to a third turker, etc. The parallel process has turkers generate content separately, and may then employ other turkers to select the best content at the end.

### Writing Image Descriptions

This experiment compares the iterative and parallel processes in the context of writing image descriptions. Each process has six turkers write a description for a single image, paying two cents each. Turkers see the instructions shown in Figure 1.

After the first description (□), this description is shown to the next turker to be improved (□). Then five turkers are paid one cent each to vote between the two descriptions (◇). The winning description is shown to the next turker in the following pattern:





- Please describe the image **factually**.
- You may use the provided text as a starting point, or delete it and start over.
- Use no more than 500 characters.

The picture is about a girl holding a yellow guitar. she is pretty and she is wearing a cap.

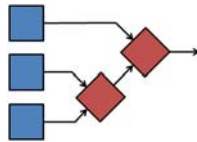
character count: 95/500

**Iterative:** A brown eyed young woman cradles an acoustic guitar in front of her chest. Her head rests against the neck of the guitar. She has long, brown hair and wears a slouchy brown hat, a gray knit long sleeved top, and two silver rings. She is either Caucasian or of Hispanic descent *[sic]*. Her shirt sleeves extend past her wrists. Her hands are holding the neck of the guitar. *(rated 9.1)*

**Parallel:** A young girl poses with her acoustic guitar. She resembles a young Sandra Bullock with dark eyes, full lips and long dark hair. She wears a newsboy-type cap a bit askew on her head, and a long-sleeve gray top; the sleeves are a bit long and extend past her wrists. *(rated 7.6)*

**figure 1.** Turkers are asked to describe an image. Some turkers are shown a previous description to start from. The resulting descriptions from the iterative and parallel processes for this image are shown.

The parallel process is similar, except that all the voting happens after all the descriptions have been written:



To compare the processes, we selected 30 images from [www.publicdomainpictures.net](http://www.publicdomainpictures.net). Images were selected based on having interesting content, i.e., something to describe. We then ran both the iterative and parallel process on each image. For half of the images, we ran the iterative process first, and for the other half, we ran the parallel process first. Turkers were not allowed to participate in both processes for a single image.

In order to compare the results from the two processes, we created a rating task. Turkers were shown an image and a description, and they were asked to rate the quality of the description on a scale of 1 to 10. We obtained 10 ratings for each image description to compute an average rating.

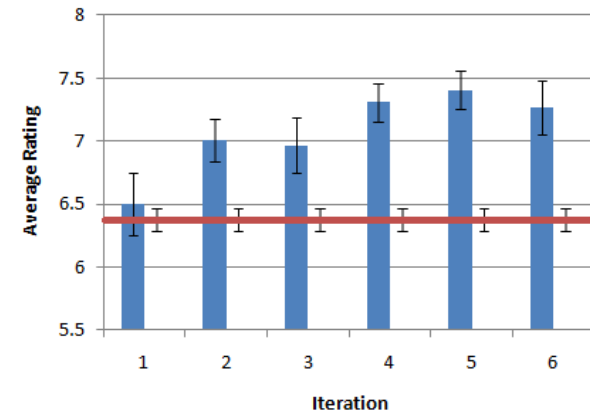
Our hypothesis was that the iterative process would produce better results. We reasoned that turkers would be willing to spend a constant amount of time writing a description, and they could do more with that time if they had a description to start from.

### Results

Figure 1 shows an example image, along with the resulting description from both the iterative and parallel processes. If we average over the final description generated in each process for all 30 images, we get a small but significant difference in favor of iteration (7.7

vs. 7.4, paired t-test  $T_{29} = 2.1$ ,  $p = 0.04$ ). If we average all the descriptions written within each process, the difference is a little bigger (7.1 vs. 6.4, two-sample t-test  $T_{358} = 5.6$ ,  $p < 0.001$ ). This suggests that giving turkers descriptions to start with may help or inspire them to write higher quality descriptions.

Figure 2 shows the average rating of descriptions written in each iteration of the iterative process. The red line shows the average rating of descriptions generated within each parallel process. As expected, the red line is at about the same level as the first iteration of the iterative process, since the first turker in the iterative process is not shown anything to start from. Subsequent iterations appear to grow in quality.



**figure 2.** Average rating given to descriptions written in each of the 6 iterations of the iterative processes. Red line indicates average rating of descriptions from the entire parallel process. Error bars show standard error.

In sum, it appears that iteration has a small positive effect on image description writing, and that the effect increases as workers are shown higher quality descriptions to start with.

**Brainstorming**

This experiment compares the iterative and parallel processes in a different domain—brainstorming company names. Each process has six turkers brainstorm five names each for a single company description. Each turker is offered 2 cents to follow the instructions shown in Figure 3.

We fabricated descriptions for six companies. We then ran both the iterative and parallel process on each company description. As with the previous experiment, we ran the parallel variation first for half of the companies, and the iterative first for the other half. No turkers were allowed to contribute to both the iterative and parallel process of a single company description.

In order to compare the results of these processes, we used the rating technique discussed in the previous experiment to rate each generated company name. Again, we solicited 10 ratings for each company name, and averaged the ratings.

Our hypothesis was that the iterative process would produce higher quality company names, since turkers could see the names suggested by other people, and build on their ideas.

*Results*

Figure 3 shows a fake company description, along with a sorted sample of the names suggested for this company. The best name generated in the parallel

- **Company details:** Our company sells headphones. There are many types and styles of headphones available, useful in different circumstances, and our site helps users assess their needs, and get the pair of headphones that are right for them.
- Please supply 5 new company name ideas for this company.

**Names suggested so far:**

- Easy hearer
- Least noisy hearer
- Hearer of silence
- Silence's hearer
- Sharp hearer

<b>Iterative:</b>	7.3 : 7.1 : 7.1 : 7.1 : 7.0 :	Easy on the Ears Easy Listening Music Explorer Right Choice Headphone Great Sound Headphone ...25 more...
<b>Parallel:</b>	8.3 : 7.4 : 7.0 : 6.8 : 6.4 :	music brain Headphone House Headshop Talkie headphones helper ...25 more...

**figure 3.** Turkers are asked to generate 5 new company names given the company description. Turkers in the iterative condition are shown names suggested so far. The top rated names from both the iterative and parallel processes are shown for this company description.

process is rated 8.3, compared with 7.3 for the iterative process. In fact, the parallel process generated the best rated name in 4 out of the 6 processes.

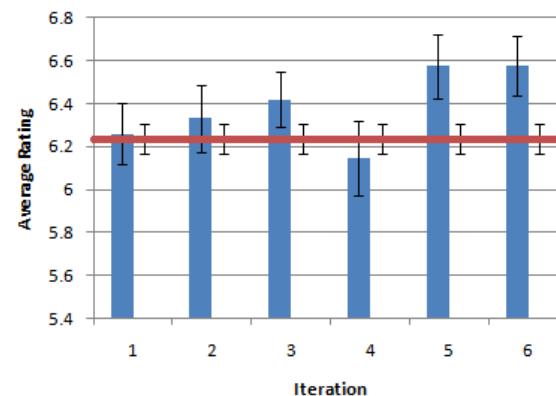
However, if we look at *all* the names generated in each process, we see a small marginally significant difference in favor of the iterative process (6.4 vs. 6.2,

two-sample  $T_{343} = 1.8$ ,  $p = 0.07$ ). This difference is more pronounced in favor of iteration if we only consider names generated in the last iteration of each iterative process (6.7 vs. 6.2, two-sample  $T_{203} = 2.3$ ,  $p = 0.02$ ).

The potential significance of iteration becomes more clear in Figure 4, where we show the average rating of names generated in each iteration of the iterative process. The red line indicates the average rating of names in the parallel process—the iterative process is close to this line in the first iteration, where turkers are not shown any names.

The average rating seems to steadily increase as turkers are shown more and more examples. The notable exception to this is iteration 4. This appears to be a coincidence—3 of the contributions in this iteration were considerably below average. Two of these contributions were made by the same turker (for different companies). A number of their suggestions may have been marked down for being grammatically awkward: “How to Work Computer”, and “Shop Headphone”. The other turker suggested names that could be considered offensive: “the galloping coed” and “stick a fork in me”.

Overall, iteration appears to have a positive effect on the average quality of brainstorming ideas, but it is unclear whether the net effect is positive, since the best names were usually generated in the parallel process. It is possible that iteration increases the average, but reduces the variance. This is a question to answer in future work.



**figure 4:** Average rating given to names generated in each of the six iterations of the iterative brainstorming processes. Red line indicates average rating of names generated in the parallel brainstorming processes. (See the text for a discussion of iteration 4, which appears below the red line.)

### Related Work

One challenge in writing human computation algorithms is motivating humans to do work. One approach is Games With a Purpose [1] [2] [3], where humans perform useful computation as a byproduct of playing computer games, or reCAPTCHA [4], where humans perform computation to prove they are human. User-generated content websites such as Wikipedia use human computation to generate content, and this content along with social factors seem to motivate future contributions [7]. Bryant [5] makes observations about how people begin contributing to Wikipedia, and what tools expert contributors use to manage and coordinate their work. MTurk provides a platform for performing Human Intelligence Tasks (HITs) where humans are motivated by money. This platform has

been adopted for a variety of uses, both in industry and academia. Kittur [6] discusses how to run user studies on MTurk, while Sorokin [8] uses MTurk to label images. Thus far, the typical usage pattern for MTurk involves generating all the HITs that need to be completed, posting them to MTurk, and later downloading all the results. Several websites focus on managing HITs that fit this template (e.g. HIT-builder). It is currently rare, however, to automatically generate new HITs based on the results of previous HITs.

### Conclusion

This research breaks down a couple of tasks into automated human computation processes that can be executed on Mechanical Turk. We compare the effectiveness of two models of computation on a couple different problem domains. Future work will focus on refining these models, and testing them in more problem domains. The ultimate goal is to build a foundation of models and techniques that can be used to construct more elaborate and effective human computation processes.

### Acknowledgments

I would like to thank everyone who contributed suggestions and ideas to this work, including Lydia B. Chilton, Max Goldman, Robert C. Miller, Thomas W. Malone, Robert Laubacher, and members of the UID group. This work was supported in part by the National Science Foundation under award number IIS-0447800, by Quanta Computer as part of the TParty project, and by the MIT Center for Collective Intelligence. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the sponsors.

### REFERENCES

- [1] Luis von Ahn. Games With A Purpose. IEEE Computer Magazine, June 2006. Pages 96-98.
- [2] Luis von Ahn and Laura Dabbish. Labeling Images with a Computer Game. ACM Conference on Human Factors in Computing Systems, CHI 2004. Pages 319-326.
- [3] Luis von Ahn, Shiry Ginosar, Mihir Kedia and Manuel Blum. Improving Accessibility of the Web with a Computer Game. ACM Conference on Human Factors in Computing Systems, CHI Notes 2006. pp 79-82.
- [4] Luis von Ahn, Ben Maurer, Colin McMillen, David Abraham and Manuel Blum. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science, September 12, 2008. pp 1465-1468.
- [5] Susan L. Bryant, et al. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. GROUP 2005.
- [6] Kittur, A., Chi, E. H., and Suh, B. 2008. Crowdsourcing user studies with MTurk. CHI 2008.
- [7] Kittur, A. and Kraut, R. E. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. CSCW '08. ACM, New York, NY, 37-46
- [8] Sorokin, A. and D. Forsyth, "Utility data annotation with Amazon MTurk," Computer Vision and Pattern Recognition Workshops, Jan 2008.