Break It Down: A Comparison of Macro- and Microtasks

Justin Cheng*, **Jaime Teevan**†, **Shamsi T. Iqbal**†, **Michael S. Bernstein***
*Stanford University, †Microsoft Research
{jcccf,msb}@cs.stanford.edu, {teevan,shamsi}@microsoft.com

ABSTRACT
A large, seemingly overwhelming task can sometimes be transformed into a set of smaller, more manageable microtasks that can each be accomplished independently. For example, it may be hard to subjectively rank a large set of photographs, but easy to sort them in spare moments by making many pairwise comparisons. In crowdsourcing systems, microtasking enables unskilled workers with limited commitment to work together to complete tasks they would not be able to do individually. We explore the costs and benefits of decomposing macrotasks into microtasks for three task categories: arithmetic, sorting, and transcription. We find that breaking these tasks into microtasks results in longer overall task completion times, but higher quality outcomes and a better experience that may be more resilient to interruptions. These results suggest that microtasks can help people complete high quality work in interruption-driven environments.

Author Keywords
Crowdsourcing, microtasks, task breakdown, interruptions.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g., HCI): Miscellaneous

INTRODUCTION
Information tasks are often thought of as having a fixed structure. A task may be long and complex, or short and easy to complete, and this is assumed to be a property of the task. However, many large tasks ("macrotasks", e.g., transcribing a speech) can actually be transformed into number of smaller, more easily achievable components ("microtasks", e.g., separately transcribing the speech's individual sentences). There is evidence that information workers already implicitly break larger tasks down: people perceive tasks in segments [13], and mental workload dips at task boundaries, rising during individual subtasks [12]. Additionally, common tasks such as email are accomplished in short bursts of less than five minutes [2]. In fact, microtasking is prevalent in crowdsourcing, where a number of workflows have been developed that decompose large, seemingly complex tasks into microtasks for goals such as as taxonomy creation and copyediting.

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

ACM 978-1-4503-3145-6/15/04 ...$15.00.
http://dx.doi.org/10.1145/2702123.2702146

By explicitly segmenting a macrotask into microtasks, larger, more overwhelming tasks can be done in smaller chunks of time. In crowdsourcing, microtasking has allowed crowd workers greater time flexibility [8]. The rising success of crowd work suggests that traditional information workers may stand to benefit from microwork structure [10], and enable people to complete large tasks in many brief moments when they feel productive but do not have a long, uninterrupted period of time [11].

To explore the trade-offs between microtasking and macrotasking, we conducted an experiment with 110 participants that compared performance on macrotasks with equivalent sets of microtasks. By looking at three simple but common task types (arithmetic, sorting, and transcription), we study how quickly and accurately participants were able to complete the task across macrotasks and microtasks. While breaking a macrotask into microtasks resulted in longer overall task completion times, it also yielded higher quality outcomes and easier work. We also find evidence that microtasks may be more resilient to interruption than macrotasks, suggesting that microtasking could be helpful for information workers, who tend to be interrupted often and have difficulty resuming their tasks afterwards [1]. Overall, our results suggest that microtasks may enable information workers to complete high quality work in short bursts of time that have previously been unproductive.

METHOD
To study the impact of breaking a task down from a macrotask into a set of equivalent microtasks, we manipulated three primary variables: *task type*, *task format*, and the presence of *interruptions*.

Task Type
We studied three types of tasks: receipt arithmetic, line sorting, and audio transcription. We selected these task types because they can be directly decomposed into microtasks, have clear correctness criteria, and are common in information work. Addition is a common activity in personal finance, sorting is implicit in ranking and selection tasks, and transcription is prevalent in crowdsourcing markets.

Arithmetic: Sum the cost of items from a scanned receipt.

Sorting: Sort seven lines of text. Each line contains a list of 10 numbers and should be ordered by the number of odd numbers in each list of numbers.

Transcription: Transcribe approximately 30 seconds of audio from an audiobook of Aesop's fables.

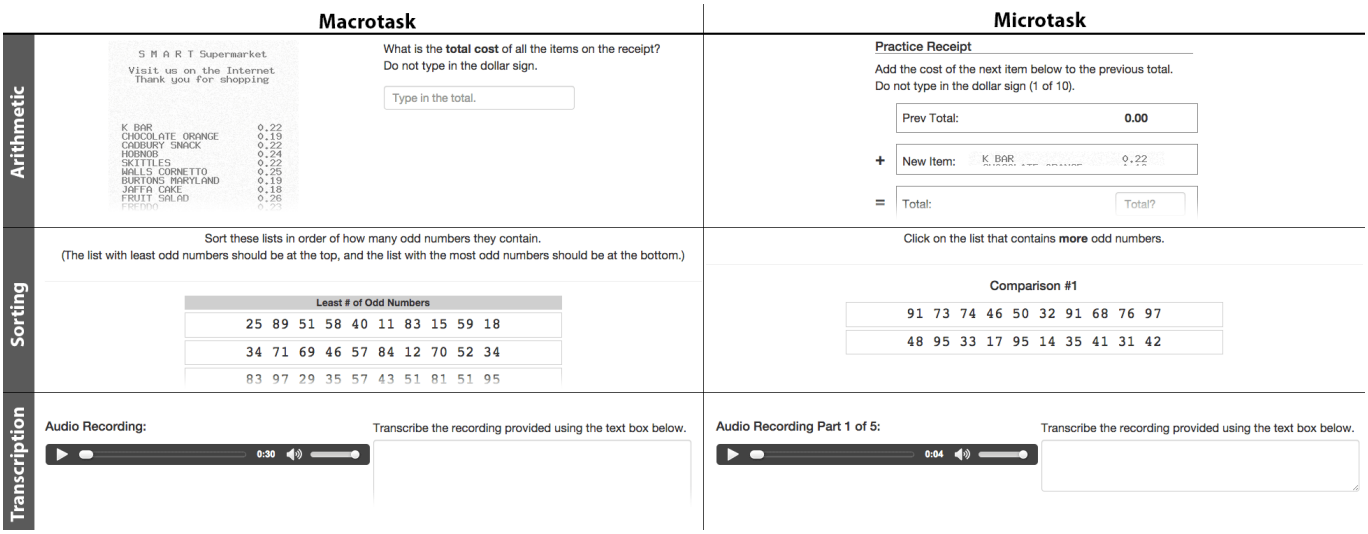

Figure 1. Screenshots of the three task types, showing both the macro- and microtask structure.

Task Format

While the three task types can be performed as complete macrotasks, each has also been explored in the crowdsourcing literature broken down into microtasks (Figure 1).

Arithmetic: In the macro version, participants were shown a receipt listing 10 items and their cost. They were then asked to sum the costs and enter the total. In the micro version, participants were shown a single line at a time and asked to add the cost of the displayed item to a running total.

Sorting: The macrotask showed seven lines of text and asked the participant to sort them by dragging and dropping each line into its new position. The microtask implemented a human-powered quicksort [7]. Participants compared pairs of lines, selecting the line with more odd numbers.

Transcription: The macro audio transcription task involved transcribing a complete 30 second audio clip. The microtask split the clip into five parts using prosodic pauses. Workers were then asked to transcribe each part individually. Similar crowdsourced approaches include Legion:Scribe [5], where different workers capture different parts of a clip in real-time.

Interruptions

In some conditions, we interrupted participants with distractor tasks as they worked. When interruptions occurred, they appeared eight times at random intervals, spaced roughly 10 seconds apart, without regard to where in the underlying task the participant was. The interruption task was modal; it blocked the interrupted task and had to be completed before that task could be resumed.

The interruption task required participants to add a pair of two-digit numbers and select an answer from a set of five possible choices. The choices were designed such that the incorrect options appeared similar to the correct answer. To stop participants from randomly selecting an answer, they were required to select the correct answer in order to continue. A pilot version that required participants to type their answer into a text box revealed similar results. Arithmetic tasks are commonly used in the literature to distract attention when studying interruptions [9, 1].

Experimental Design

The study followed a 3 (task type) $\times$ 2 (task format) $\times$ 2 (interruptions) design. The experiment was between-subjects with respect to task type and within-subjects with respect to task format and the presence of interruptions. Each participant was assigned a task type and completed four tasks of that type, performing the macro and micro versions, with and without interruptions. The order of these tasks was randomized. To familiarize participants with their assigned task type and the interruption task, participants initially completed a practice task with interruptions. After completing all trials, participants were asked if they preferred the macro- or microtask structure, and completed a NASA Task Load Index (TLX) assessment [3] to measure subjective mental workload. The entire process took under 30 minutes.

Measures

For each condition, we measured the total time it took to complete the task excluding the time spent on interruptions, and the average amount of time it took to complete each interruption task. To further control for external interruptions, we also tracked any loss of browser window focus during each task and excluded participants who left a task for more than ten seconds. We also measured the quality of work performed. For the receipt arithmetic task we measured the probability that the final answer was incorrect. For the sorting task, we measured the Kendall's tau distance of the submitted ranking from the correct one, and for audio transcription we measured the word error rate.

To calculate significance we used a linear mixed-effects model, with participant as a random effect. p-values were calculated using an F-test with a Kenward-Roger correction. A mixed-design ANOVA resulted in empirically similar results. Error bars in figures depict standard errors.

	Coef.	SE	df(num/den)	F	p-value
Task Time ($R^2 = 0.77$)					
(Arithmetic Macrotask)	55.2	11.8			0.00**
Task Type			2/181	59.8	0.00**
Sorting	29.7	17.6			0.00**
Transcription	134	16.2			0.00***
Microtask	18.9	9.30	1/377	13.3	0.00***
Interrupted	12.8	9.30	1/377	6.00	0.01*
Type × Micro			2/377	4.12	0.02*
Type × Interrupted			2/377	0.42	0.61
Microtask × Interrupted	-3.30	13.2	1/377	2.64	0.10^
Type × Micro × Interrupted			2/377	0.46	0.63
Task Error Rate ($R^2 = 0.37$)					
(Arithmetic Macrotask)	0.10	0.04			0.00**
Task Type			2/143	9.38	0.00**
Sorting	0.23	0.06			0.00**
Transcription	0.04	0.05			0.00***
Microtask	-0.05	0.05	1/377	15.3	0.00***
Interrupted	0.03	0.05	1/377	0.07	0.78
Type × Micro			2/377	0.79	0.46
Type × Interrupted			2/377	0.29	0.75
Micro × Interrupted	-0.02	0.06	1/377	0.01	0.93
Type × Micro × Interrupted			2/377	0.13	0.88

Table 1. A linear mixed-effects model for both task time and task error rate (^: $p<0.1$, *: $p<0.05$, **: $p<0.01$, ***: $p<0.001$). Some non-significant effects elided for space.

Participants

A total of 110 people participated in the experiment, with 42 in the arithmetic condition, 28 in the sorting condition, and 40 in the transcription condition. The experiment was conducted using an in-house microtasking platform that outsources crowd work to vendors, similar to CrowdFlower. We used the Clickworker vendor, with all participants from the United States. Empirically similar observations were obtained when testing each condition on Mechanical Turk. Participants were compensated $3.00 and repeat participation was disallowed.

RESULTS

Completing a task via microtasks took longer overall than completing it as a macrotask, but the microtask format led to fewer mistakes, an easier experience, and greater stability in the face of interruption. Table 1 shows the strength of the impact of task type, task format, and the presence of interruptions on task time (in seconds) and task error rate. The intercept corresponds to the arithmetic task in a macrotask format without interruptions.

Time

For the three task types that we studied, overall, completing the task via microtasks took longer than doing it as a macrotask ($p<.001$). For instance, the uninterrupted arithmetic task took approximately 42 seconds longer when broken down than whole (Figure 2). This was expected, as each task type requires some additional work when broken into microtasks

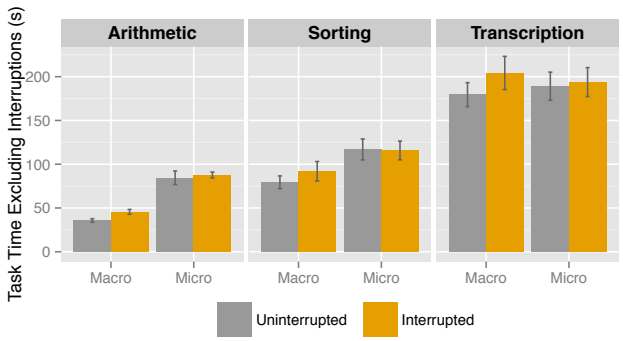

Figure 2. Microtasks incur a significant additional fixed cost. Interruptions cause an increase in task time for macrotasks but not microtasks.

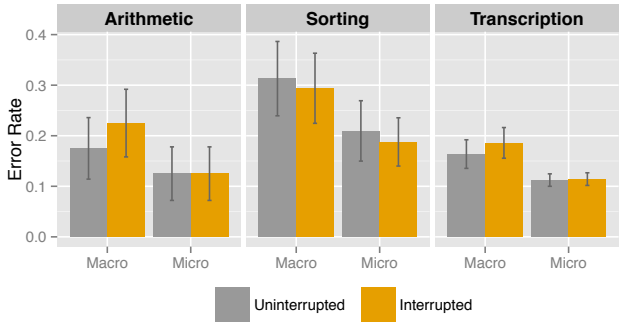

Figure 3. Microtasks significantly reduce task error rates for a task. The presence of interruptions does not impact task quality.

to support the added structure (e.g., in the arithmetic task, participants enter many additional numbers to keep a running total in the microtask condition, but only a single number in the macrotask condition), and corroborates prior work showing that introducing breaks between microtasks increases overall completion time far beyond a continuous microtask workflow [6].

However, while arithmetic and sorting took significantly longer as microtasks, this was not the case for transcription. We hypothesize that users may have been implicitly breaking the transcription macrotask similarly to how the microtasks were structured (i.e. listening in parts): participants started, paused or stopped the audio a median of 5.3 times in the macrotask, but only 10.2 times in the microtask condition, where the minimum number of clicks is ten (to start and stop each of five segments).

Quality

The microtask structure produced higher quality work for all three task types, with participants making fewer errors when doing a task as a series of microtasks than as a macrotask ($p<.001$). The gray bars in Figure 3 represent the number of errors made for each task type in the uninterrupted condition. One explanation for this is that the additional work required to support the microtasks helped externalize some of the mental processing necessary to complete the tasks, which improved performance. In the case of transcription, splitting audio at shorter segments of about six seconds at prosodic pauses may have been be a good balance between keeping subtasks relatively short, while still maintaining adequate context.

In summary, with macrotasks, people can use any strategy they want, but these strategies vary in effectiveness, and tend to be faster but less accurate. With microtasks, the strategy is set and enforced by the system. If this strategy is similar to how a person would break the task down (e.g., transcription), task time is likely to be similar.

Preference

In addition to producing higher quality results, the microtask format also made each task type easier to complete. When asked to rank the task structure in terms of difficulty, an average of 77% of participants preferred the microtask format to the macrotask format ($p<.01$). Likewise, on the NASA TLX scale, the microtask format was consistently rated as easier than the macrotask format, with the microtask score being 6.2% lower for the arithmetic tasks, 9.2% lower for the sorting task, and 5.7% lower for the transcription task ($p<.01$).

Interruptions

Finally, we see a trend that completing a task via microtasks may be more robust to interruption than completing it as a macrotask. The amount of time it took to complete a task when it was broken down into microtasks was very similar regardless of whether the participant experienced interruptions or not (Figure 2). In contrast, in many cases it took longer to complete a task in the macrotask format when the task was interrupted than it was not. The weak significance of the "Format $\times$ Interruptions" interaction effect (Table 1) suggests that overall, interruptions may slow down task completion when a task is performed as a macrotask, but not as a series of microtasks ($p<.10$). Still, a separate analysis found that interruptions significantly slowed down the arithmetic macrotask ($t(41) = 3.67$, $p<.001$).

One possible explanation is that because task boundaries are clearly delineated for microtasks and the individual tasks themselves are short, participants regularly found themselves at a task boundary when interrupted using the microtask structure. While interruption costs can also be reduced by scheduling interruptions at subtask breakpoints [4] or using goal reminders [1], the inherent structure of microtasks results in more explicit breakpoints, making them more interruption-friendly than larger tasks. Across all conditions, the interruptions themselves took similar amounts of time to complete (μ=7.4s), and do not appear to impact the number of errors made. In both cases, we observe no significant differences.

CONCLUSION

We explored the impact of breaking macrotasks down into microtasks. For three common task types, we find it tended to take longer to perform a task using microtasks than macrotasks. However, the structured microtasks enabled participants to produce significantly higher quality work than they did using macrotasks, and participants found the microtasks easier to complete. Additionally, microtasks may be more robust to interruption, as there was some evidence that interruptions impacted the time it took participants to complete the macrotasks but not the microtasks. If differences exist even at this task size boundary, benefits may increase as the macrotask increases in size.

This paper presents a first step at studying the effect of task decomposition. Future work includes better understanding when this is beneficial. For example, while sorting seven lines of text is easier when done as a series of microtasks, this may not be the case for sorting only three lines where the content of each line can be held in short term memory. Additionally, many complex tasks do not have obvious structure that can be used to support microtasking. Cognitive modeling techniques (e.g., GOMS) may provide finer insight into how a particular task decomposition affects performance. Our findings suggests a future where microtasks could enable people to complete tasks in interruption-driven environments in a structured way that requires less cognitive effort, and in spare moments of time.

REFERENCES

1. Cutrell, E., Czerwinski, M., and Horvitz, E. Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. In *INTERACT* (2001).
2. González, V. M., and Mark, G. Constant, constant, multi-tasking craziness: managing multiple working spheres. In *CHI* (2004).
3. Hart, S. G., and Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Adv. Psychol.* (1988).
4. Iqbal, S. T., and Bailey, B. P. Effects of intelligent notification management on users and their tasks. In *CHI* (2008).
5. Lasecki, W., Miller, C., Sadilek, A., Abumoussa, A., Borrello, D., Kushalnagar, R., and Bigham, J. Real-time captioning by groups of non-experts. In *UIST* (2012).
6. Lasecki, W. S., Marcus, A., Rzeszotarski, J. M., and Bigham, J. P. Using microtask continuity to improve crowdsourcing. Tech. rep., 2014.
7. Little, G., Chilton, L. B., Goldman, M., and Miller, R. C. Turkit: human computation algorithms on mechanical turk. In *UIST* (2010).
8. Martin, D., Hanrahan, B. V., O'Neill, J., and Gupta, N. Being a turker. In *CSCW* (2014).
9. Sakai, K., Rowe, J. B., and Passingham, R. E. Parahippocampal reactivation signal at retrieval after interruption of rehearsal. *J. Neurosci.* (2002).
10. Teevan, J., Liebling, D., and Lasecki, W. Selfsourcing personal tasks. In *CHI* (2014).
11. Vaish, R., Wyngarden, K., Chen, J., Cheung, B., and Bernstein, M. Twitch crowdsourcing: Crowd contributions in short bursts of time. In *CHI* (2014).
12. Wickens, C. D. Multiple resources and performance prediction. *Theor. Issues. Ergon.* (2002).
13. Zacks, J. M., Tversky, B., and Iyer, G. Perceiving, remembering, and communicating structure in events. *J. Exp. Psychol. Gen.* (2001).