Economic Utility of Interaction in Crowdsourcing

Michael Toomim
University of Washington, dub Group
http://dub.washington.edu/people/toomim

ABSTRACT
Crowdsourcing, along with much of the Internet, only works when humans find tasks fun, enjoyable, or valuable enough to outweigh the time and effort they require to complete. The more *utility* that humans find in a task and interface, the more "work" they will do. However, we do not yet know how to objectively measure the fun, enjoyment, or value of a user interface applied to a particular task. My research empirically measures the *economic utility*, or "user preference," of a user interface for a task by putting multiple versions of a user interface together with a task on Mechanical Turk and measuring the amount of money required to convince humans to use them.

INTRODUCTION
The ESP game [6] makes image labeling fun. Wikipedia works by making it easy to edit pages and contribute to a body of knowledge. Facebook and Twitter work by making it easy and rewarding to upload your life online for others to see. In general, Crowdsourcing and other human-computer interactions only work when people *like*, *prefer*, and ultimately *choose to use* our interfaces and tasks out of the myriad options available to them.

My research economically measures and quantifies the amount an interface motivates (or demotivates) a user to use it for a task, by putting the interface and task in a crowdsourced labor market and measuring the amount of money you must pay people to use it. Thus, this work relates to crowdsourcing in two ways: (1) it evaluates the most critical quality of a crowdsourcing user interface—its ability to convince a human to use it, and (2) it uses a crowdsourced method to infer this metric.

More specifically, this work operationalizes the *economic* definition of utility, and applies it to Human-Computer Interactions. In Economics, utility is the degree to which a person prefers a particular choice amongst options [7]. We can infer it from user behavior: when a user chooses to use system A instead of B, it is said that *Utility*(A) > *Utility*(B). Utility encompasses all factors of function and usability that affect preference and use, and measures their net impact on user behavior (Figure 1). Economic utility quantifies user preference.

To quantify the utility difference of two interfaces, we vary the wage we offer, and find the amount that compensates for a difference in use between the two interfaces. If a user has no preference between being paid 25¢ for using system A over being paid 50¢ to use system B, then we can describe the difference *Utility*(A)–*Utility*(B) = *Utility*(25¢). That is, system A provides a measure of 25¢ more utility than system B. This is a *money-metric* of utility, a number representing the value of an interface change that can be compared across interfaces and situations, providing a lingua franca for communicating results.

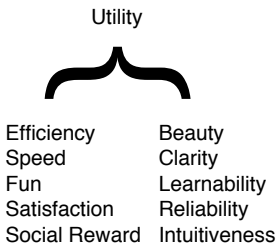

Figure 1. Utility is a summative descriptor of users' decisions of use. Many lower-level factors affect a user's choices of use. Utility aggregates the effects of all lower-level factors into a single number.

CASE STUDIES
Figures 2 and 3 illustrate a study measuring the utility of a standard HCI phenomenon: Fitts' Law [2]. Fitts' Law models the difficulty of target acquisition as a function of target size and distance. To test our methodology, we ran a study online to determine the utility of Fitts' Law—how target size and distance affect user *preference*. We posted the standard Fitts' law task to Mechanical Turk, asking workers to click back and forth between a rectangle that switched sides on the screen. Our experiment manipulated the task's index of difficulty, by changing the size of the rectangle and its distance from the user's cursor (see Figure 2). We expected users to *prefer* easy tasks to difficult tasks, and in fact the data displays this trend (see Figure 3). That is, the degree a Mechanical Turk user prefers a Fitts' law task is inversely proportional to the time it takes to use it. With further experiments, one could learn if this preference result generalizes, and perhaps produce a general law of user preference with respect to task-completion time. Such a law would be useful knowledge for designers. This shows that we can vary an interface and quantify its effect on user preference.

Labor Supply Curves of Interaction Utility
However, if we also vary the price we pay workers on Mechanical Turk, we can view preference economically, and express it in terms of dollars and cents. This is illustrated in Figure 4. In our Fitts' Law study, we varied the amount of money we paid workers along with the index of difficulty. With six prices and three indices of difficulty, we created a total of eighteen experimental conditions. This gives us enough data to induce *labor supply curves* for the interface variations: graphs that show how much work you can expect the average worker on Mechanical Turk to produce with the interface variations, given varying levels of pay.

With these curves, we can apply economic analyses to our interface. For instance, we can infer a *money-metric* of utility—quantifying the value or cost of an interface variation in terms of dollars and cents. To do so, we simply fix the number of jobs (the Y axis) to a single number, and measure the distance (in pay, on the X axis) between the two curves for two interface variations. On the other hand, if we fix instead a single value of pay on X axis, we can deduce the amount of work that a change in interface produces.

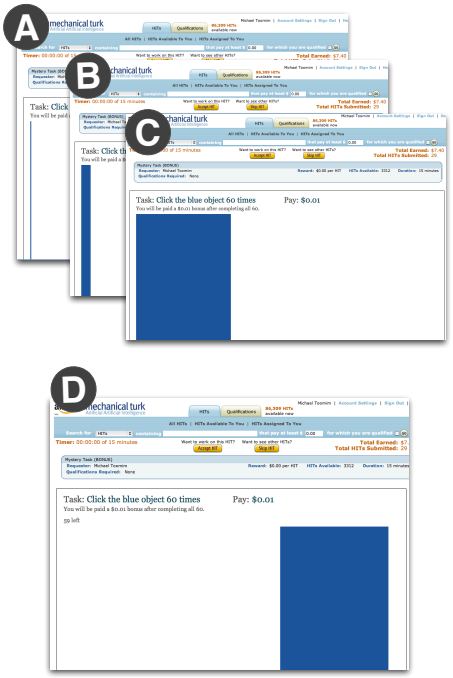

Figure 2. Screenshots of the Fitts' law task. Subjects clicked on a blue rectangle 60 times. We created three variations of bar width and the distance it moved: hard (a), medium (b), and easy (c). Each time they clicked on the bar, it moved to the opposite side of the screen (d).

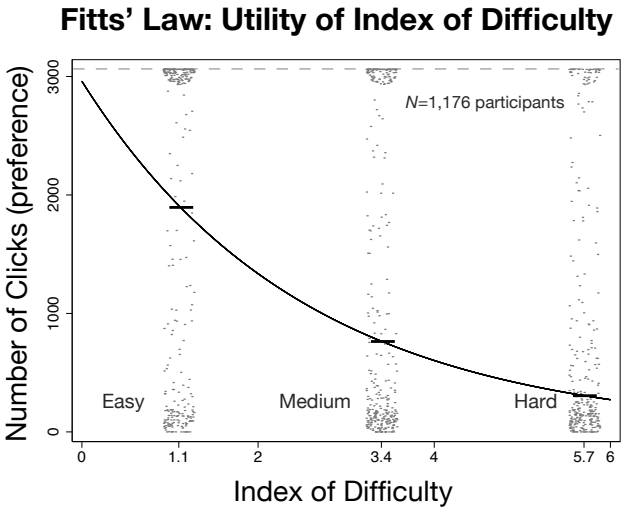

Figure 3. Fitts' law models the *time* required to click a widget of a size and width—our technique can model how much people *prefer to use* a widget. Participants were assigned one of three index of difficulty conditions. Each point is the number of clicks a participant completed before quitting (points jittered to show spread). Participants *preferred* big buttons to small buttons ($p < 0.10$). Participants were allowed a maximum of 3,060 clicks each. The regression line accounts for this maximum using a Tobit analysis.

These calculations are illustrated in the dashed lines of Figure 4. By quantifying the amount an interface is worth in dollars and cents, we can compare results across interfaces, tasks, and situations using the *lingua franca* of utility: money. We can compare the utility of aesthetics in an interface to the utility of efficiency. To summarize, these methods enable us to view human-computer interactions *economically*. We can quantify interface variations in terms of dollars and cents, and empirically model user *preference* and *choice*.

Utility of Aesthetics over Time: Survival Analysis

Our second case study illustrates two new dimensions of interaction utility. First, we show that we can measure the utility of particularly elusive quantities in HCI: *aesthetics* and *feedback*. These quantities are elusive because they do not make an interface slower to use, or otherwise affect the user's actual behavior. They only affect his perception of the interface and his understanding of its internal process. Second, we show how to analyze qualities like these *over time*. To do this, we use *survival analysis* [1], an analytical perspective and toolbox that examines the percentage of workers that continue to use an interface for a task over time, and when they quit. Survival analysis uses a *survival function* $S(t)$ that represents the probability of a user surviving t tasks before quitting.

In our study, we implemented two interface variations for the task of answering CAPTCHAs. One interface had a clear, minimalist design, and the other had gaudy colors, small fonts, and a distracting animated GIF advertisement (Figures 7a & 7b). Both tasks had the same instructions and

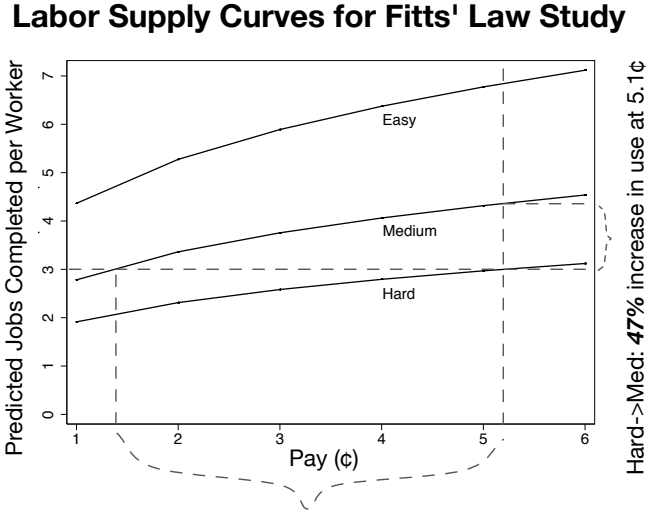

Figure 4. We paid six different prices (1-6¢) for each of the three experimental conditions: a total of eighteen conditions. By regressing on the number of jobs completed in each condition, we estimate these labor supply curves. Holding pay constant, we can quantify the effect of an interface on *use*. Holding the number of jobs constant, we can compute a *money-metric*: the amount of pay required to obtain the same amount of work between two interfaces. For 3 jobs, the utility of Medium over Hard is equivalent to 3.8¢ per 60-click job. (This data is filtered to U.S. workers.)

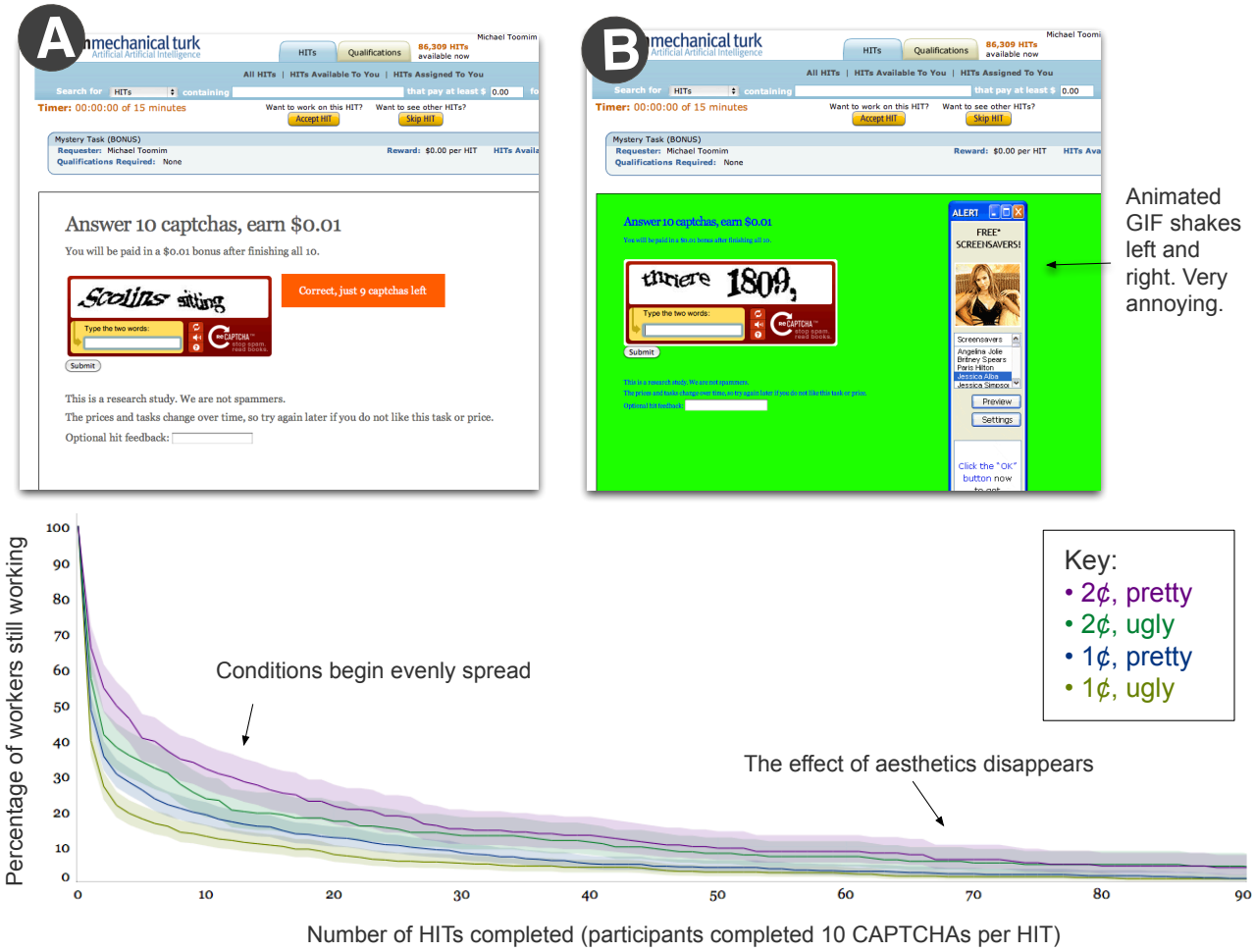

Figure 7. *Survival graph for the Aesthetics & Feedback study.* We made two interfaces for answering CAPTCHAs: one "pretty" (a), one "ugly" (b), but identical in behavior. The *survival graph* shows how many workers made it through how many tasks, for each of our four experimental conditions. The shaded regions are 95% confidence intervals. At the far left, 100% of these workers looked at the task, but only 10% to 40% completed 10 tasks (100 CAPTCHAs). Note that the *pretty* and *ugly* lines are separated at the left, but converge toward the right. This suggests either that the utility effect of aesthetics fades over time, or that the types of users who complete many CAPTCHAs are more concerned with pay than aesthetics.

wording, required 10 CAPTCHAs to be completed per job, and took the same amount of time to complete. The *pretty* condition implemented an elegant animated countdown reminding the user how many CAPTCHAs they had left, and the *ugly* condition only told them when they had completed all 10.

The survival graph for the CAPTCHA experiment is shown in Figure 7. The confidence intervals for each line are shaded. The survival analysis shows how use changes over time. We can see that all four conditions are spaced apart roughly equivalent for the first 20 tasks, but for work done at 80 tasks, the top two lines (2¢) and bottom two lines (1¢) converge. This means that price dominates the utility for workers who acquire more experience with the task, and aesthetics is primarily important for those who are inexperienced. Or, those who stick with the task are more resilient to aesthetic quality. We cannot distinguish between these two competing hypotheses, however their difference is intangible from our perspective, since they predict the same result in *use*.

Practicality of Utility Measurement

Our experimental method is mostly automated, with a software framework that automatically posts tasks to Mechanical Turk at different prices and interactive conditions, logs user interactions, and analyzes and visualizes the results. A major advantage of this approach to evaluating human-computer interactions is that studies are dramatically easier to run than with traditional methods, and results can be reproduced at the click of a button, using the same labor market, and often many of the same workers. This engenders a scientific process. Subsequent researchers can alter and re-run an experiment, by altering and re-running source code. Traditional studies in HCI are rarely reproduced. The studies in this paper deployed 15,000–22,000 jobs on Mechanical Turk, recruiting 1,100–1,200 workers, took 5-10 hours to complete, and cost $300–$1,000. With methodological improvements we believe we can achieve similar results for $50–$200. We believe this makes it practical to study this important topic—the factors that convince a user to *choose to use* an interface for a task—with a

quantitative, empirical measure, and build the study into a science.

Our current methods also have a broad set of limitations, particularly due to evaluating use in an artificial labor market instead of a real system. We discuss these limitations in detail in our paper at this conference [4]. In that paper, we also discuss how these limitations might be overcome by extending our methods to real systems.

THE FUTURE OF CROWDSOURCING

The first crowdsourced systems produced exciting results, but often raised questions of how crowdsourcing will scale. A game with a purpose may help us label images, but what happens if the gaming populace becomes bored over time and no longer derives utility from the game? Do games with a purpose scale over time?

Or from another perspective, how do crowdsourcing systems scale over space—to larger datasets and larger markets of humans? Is Mechanical Turk cheap because it is tapping into a narrow subsegment of the population who is willing to do work for cheap? As crowdsourcing becomes more mainstream, and more of our societal institutions (*e.g.* news, media, government) are opened up to crowdsourcing, will this small market of cheap Internet labor be quickly tapped out?

These unknowns illustrate the larger general questions of the capabilities and limits of crowdsourcing in general. What is possible to build? How do we evaluate a system's ability to scale over time or space? From the perspective of Computer Science (and, in turn, Human Computation), we might try to characterize the amount of data we can process in terms of resources of time and space. We might try to characterize the "computability" or "complexity" of crowdsourcing problems analogously to how we do in Computer Science.

However, in crowdsourcing, these questions involve markets, incentives, values, and utility—they are economic in nature. Thus, in order for us to analyze the capabilities and limits of a crowdsourced system, we must learn to merge our understanding of computational complexity with Economics. My research investigates this intersection, relating the two approaches at the place where humans meet computers—the user interface. By developing a quantitative, empirical, objective measurement for the utility of human-computer interactions, and by learning to think of interactions economically, I believe the field of HCI will found a scientific inquiry into the capabilities and limits of crowdsourcing, the performance characteristics of different approaches, and transform the design of crowdsourced systems and games from a black art into an empirical engineering practice.

BIO

I am a PhD student in Computer Science and Engineering at the University of Washington. I primarily research Human-Computer Interaction. I have also used crowdsourcing in my previous published research: (1) to evaluate a security system by paying Mechanical Turk workers to try to break it [5], and (2) as a "neutral" third party to select an unbiased sample of webpages to evaluate coverage of a machine learning tool on [3]. I also have unpublished work using crowdsourcing to generate media, such as chain stories and a television series.

REFERENCES

1. Klein, J.P. *Survival Analysis*. Springer, 2003.
2. MacKenzie, S. Fitts' Law as a Research and Design Tool. *in Human-Computer Interaction (HCI)*, 1992 7(1), 91–139.
3. Toomim, M., Drucker, S. M., Dontcheva, M., Rahimi, A., Thomson, B., and Landay, J. A. Re-forming the Internet with its End-Users. *in Proc. CHI 2009.* ACM Press (2009).
4. Toomim, M., Kriplean, T., Pörtner, C. and Landay, J.A. Utility of Human-Computer Interactions: Toward a Science of Preference Measurement. *in Proc. CHI 2011.* ACM Press (2011).
5. Toomim, M., Zhang, X., Fogarty, J. and Landay, J.A. Access Control by Testing for Shared Knowledge. *in Proc. CHI 2008.* ACM Press (2008).
6. Van Ahn, L. and Dabbish, L. Labeling Images with a Computer Game. *in Proc. CHI 2004.* ACM Press (2004)
7. Varian, H.R. *Microeconomic Analysis*. Norton, 1984.