Not All HITs Are Created Equal: Controlling for Reasoning and Learning Processes in MTurk

Jessica R. Hullman
University of Michigan School of Information
jhullman@umich.edu

ABSTRACT

Challenges of crowdsourcing human-computer interaction (HCI) experiments on Amazon's Mechanical Turk include risks posed by the combination of low monetary rewards and worker anonymity. These include how mirroring task structure across HIT or qualification questions may encourage the learning of shallow heuristics, the difficulty in increasing workers' intrinsic motivation, and how verification questions can interrupt natural reasoning processes leading to a mismatch between experimental and real-world behavior. I discuss how researchers can increase HIT response quality on MTurk despite such challenges by applying findings from experimental psychology related to what increases active processing of information, how to avoid conflicts between opposing reasoning styles, and how to negotiate tradeoffs involved in inducing analytical thought.

MOTIVATION

The development of large-scale platforms for crowdsourcing simple human intelligence tasks such as Amazon's Mechanical Turk (MTurk), CloudCrowd, and Crowdflower has fueled interest in using these platforms for controlled experiments on information interfaces. Crowdsourcing platforms have been used in studies aimed at how individuals process information online at the individual level [10, 14, 17] as well as the community-level [12], and how crowdsourced knowledge compares to that of experts [6, 15, 25].

The asynchronous, distributed and isolated nature of work on a crowdsourcing platform like MTurk differs significantly from the controlled laboratory setting or real world settings like the online social environments being investigated in social computing research. Crowdsourcing studies have led to a list of best practices for leveraging provided features to best insure response quality [10, 15, 25]. Best results occur when MTurk requesters use HITs with bona fide answers, designed such that accurate completion is no more effortful than maliciously providing an invalid answer and including explicitly verifiable questions as part of the task. Similarly, signaling to workers that answers will be scrutinized and detecting suspect answers in multiple ways helps eliminate low-quality responses. Worker ratings and qualification tasks are commonly used to filter users who do not understand task requirements, while learning how to do a HIT is facilitated by using demonstrative examples for each class of HIT.

As an example, consider an MTurk experiment for studying how non-experts' interpretations of information visualizations of economic data compare to expert interpretations. A sample HIT might call for 50 workers to do a sequence of 20 HITs, each with a unique visualization. Interested workers must pass a qualification task where they are first shown an example U.S. map of economic indicators for states along with a paragraph of context. Several questions designed to verify that the worker has closely examined the graph ask for the number of data variables shown and the color in which a particular indicator is represented. Answers to these are provided for the sake of learning. A final target question might require the worker to identify the trend in the visualization, suggesting an appropriate answer achieved using a four-step process of identifying the regions involved, assessing the ranking of the regions according to two different economic indicators, assessing the relationship between the indicators for any region, and then summarizing in several sentences the results of this process. The rest of the qualification task requires the worker to independently complete the same questions for a second visualization. After submitting her responses, the worker is told whether she has passed, and if not how long she must wait before retaking the text. The 20 HITs that become available given a passing qualification mirror the qualification questions, with a unique visualization per HIT.

Despite using best practices, this scenario illustrates several challenges in conducting research on MTurk. It behooves the worker to read the worked examples and carefully respond to questions, yet workers may be motivated by the low monetary rewards to use the qualification to learn the simplest heuristic for generating passing answers on HITs, rather than for gaining real comprehension of task requirements. Similarly, the use of the same style of verification question across similar HITs in a series, combined with the low rewards and motivation, may lead a worker to reuse answers across multiple similar HITs, or the quality of her answers might fall off sharply after the qualification if she suspects that the answers won't be carefully examined. In either case, the researcher must sort

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
Copyright 2011 ACM 978-1-4503-0267-8/11/05....$10.00.

through the results. An alternative possibility is that a worker's response to a verification question primes her subsequent processing of target information in a way that interferes with the desired results. Responses to a linear estimation task, for example, can be unintentionally skewed by information in a verification question [12]. Finally, the results of the study may not accurately reflect the real online visualization interpretations by non-experts as a result of the interruptive aspect of the verification questions to the user's natural interpretative process.

The aforementioned challenges concern the interaction of MTurk features with human reasoning processes. While some recent results [11, 21] suggest that attention and motivation problems on MTurk be no worse than using university subjects and suggest psychological findings on how to guard against satisficing among participants [19], there is room for further exploration of how psychological research could be applied to MTurk HIT design. I apply experimental results from cognitive and educational psychology around how to negotiate task design given the distinctions and tradeoffs between deliberate, analytic reasoning and intuitive, associative reasoning, and between active and passive information processing to designing MTurk experiments involving information interfaces. Specific techniques by which researchers can achieve common goals (e.g. more active processing of HITs, creating enjoyable HITs) are suggested.

MODES OF REASONING IN DATA INTERPRETATION

Psychologists have long distinguished between two reasoning modes—intuitive, System 1 reasoning and analytical System 2 reasoning [e.g. 8, 13]. A related distinction is often made between active and passive processing. Motivational and engagement factors form a third consideration relevant to the design of MTurk tasks involving information processing.

Reasoning style reliably influences the outcome of a task like a decision. **System 1** reasoning is typically *automatic*, *effortless*, and *intuitive* and *associational* in the sense of being based on prior experience. It is also often associated with more perceptual heuristic processes. Intuitive perceptual heuristics can be an asset, such as human's natural abilities to detect visual patterns in some graphs that may be undetectable via statistics, or they can carry negative effects, as in matching bias leading one to see as relevant information that matches the lexical content in the statement about which one is reasoning, and to neglect the logically relevant information [8]. Conversely, **System 2** processes tend to be *effortful*, *deliberate*, and *analytic*. These involve abstract reasoning and hypothetical thinking, and are thus appropriate for decisions facilitated by mental simulations of future possibilities and judgments not aided by prior knowledge or beliefs. These tasks often involve comparing options, such as rating two web resources according to a list of criteria, yet not all information rich tasks are best solved analytically – some complex choices such as choosing a car or apartment are often best made intuitively, in part to avoid the risks of over-thinking [16]. System 2 processing is constrained by working memory capacity, and correlated with general intelligence, and likely to be activated when people have both the capacity and motivation to engage in effortful processing.

A potential conflict between the two is suggested by research documenting the inhibitory role of System 2 in suppressing default knowledge and belief-based responses [26]. Whether the conflict is resolved to lead to a more accurate judgment depends on task and individual characteristics. As an example of a task that could cue either type, an individual under time constraints might use the number of previous downloads of an album in an online music environment as a signal of its quality, while a less time-pressed user might systematically consult online reviews and ratings of available albums. Recognizing the automaticity of System 1 reasoning even in situations where analytic judgment is more accurate, researchers have investigated the factors that induce a reasoner to overcome the tendency to use System 1 processes. Errors in System 1 are less likely to be corrected when people are under cognitive load or respond quickly, but are more likely to be corrected when people are held accountable for decisions and when the outcome is personally relevant (see [3]). The challenge relevant to researchers using MTurk concerns the difficulty in predicting whether a subject faced with a "reasoning-ambiguous" HIT will successfully reconcile potential conflicts between the two types.

Active and passive processing is a related distinction suggested by cognitive and educational psychologists interested in how to best design learning materials that foster comprehension and engagement. In some cases, requiring more effort of a learner stimulates more active processing of relevant information. Designing a task that induces subjects to explain relevant conceptual relationships, for example, can improve comprehension due to the active construction of knowledge it entails [7].

Engaging a subject tends to increase **intrinsic motivation**, and is thus an intermediate strategy to induce active processing. This can include *aligning info-based cognitive traits with task characteristics* [5], *tailoring or personalization* [18, 26], or increasing *aesthetic appeal*. In MTurk, increasing engagement may help overcome risks imposed by the reward structure.

Tradeoff with Task Difficulty Appraisal

The type of reasoning an individual uses and the relative depth of processing can depend in part on a reasoner's subjective appraisal of how difficult information is to process [3]. Confidence in the accuracy of intuitive judgments appears to depend on the ease with which information is brought to mind [e.g. 9]. If information is processed with difficulty, this cues that intuitive judgments are likely to be inaccurate, activating more elaborate System 2 processing. An example of visual difficulties that improve depth-of-processing can be found in research on hard-to-process (disfluent) fonts that demonstrates how such a font can improve comprehension of target information (see [1]). However, hard-to-process stimuli are often negatively associated, highlighting a tension between a reasoner's response accuracy and subjective enjoyment of a task. We discuss this tension as applies to MTurk below.

IMPLICATIONS FOR MTURK EXPERIMENT DESIGN

Despite features intended to improve requesters' control over response quality, the success of an MTurk experiment depends to an extent on requesters' abilities to attract the appropriate workers, to maintain worker motivation through rewards or other means, and to design HITs that support experimental objectives. There are several ways that design-centered psychology can help address challenges in crowdsourced research, stated below. These are meant to serve as pointers to topics in psychological work that might be explored in future work.

Challenge: Low payments can decrease workers' motivation and likelihood to take work seriously.

The low monetary rewards and low barriers to participation that make MTurk an attractive platform to researchers and workers, respectively, can nonetheless lead to workers' attempts to game the system through learning the simplest heuristic to generate passing answers [17]. Harnessing intrinsic motivation can lead to work outcomes with levels of quality at least as good as using financial rewards, yet successful examples of volunteer crowd sourcing are difficult to replicate, in part because many arbitrary tasks tend not to be intrinsically enjoyable [17].

Proposed Solution: Increase workers' intrinsic motivation

Psychological literature can make several contributions via generalizable findings around how individual's engagement and motivation for a task can be stimulated. In learning and medical settings, tailoring information resources such as through titles and graphics chosen based on a user's personal features has been shown to increase interest in the information, most likely due to its ability to attract attention [18, 26]. In the aforementioned visualization HIT series, it may be in the interest of the researchers to consider surveying potential workers for preference information, or using available features like geographic location, to dynamically tailor HIT content. Another way in which HIT content might be tailored is through (possibly automated) messaging that provides feedback on already-successfully-completed HITs in order to increase the worker's sense of their work's value.

Engagement (and active processing) could also be increased by designing HITs with that are appropriately challenging given a worker's cognitive characteristics. Need for Cognition is a measurable variable that refers to one's tendency to engage in and enjoy doing difficult cognitive work [5], and can be measured through the NCS scale. In experiments like the visualization example involving sequences of HITs to be completed by single workers, designing HITs that correspond to different levels of Need for Cognition or filtering workers without high levels will likely increase engagement and motivation.

Challenge: Properties of MTurk may encourage shallow understanding based on principle of least effort

Similar to the first challenge, under certain conditions, financial incentives can undermine intrinsic motivation to result in poorer outcomes, lead workers to ignore rational incentives to continue work after accomplishing set targets, or undermine performance such as by focusing workers on only the measured outcomes (see [17]). For example, many workers asked for their own translations of text ignored instructions and cut and pasted machine translations [5].

Proposed Solution: Increase levels of worker understanding and learning to improve HIT responses.

Psychological findings related to depth-of-processing might help explain observations regarding MTurk output. For example, quota systems (e.g. paying a worker per crossword puzzle rather than per word) elicit more work [17], and this may be in part because this pay style increases workers' tendencies to consider the task as a coherent whole and to think more deeply about relationships between components. Findings related to the same theme of cognitive processing quality can guide the design of HITs that work to improve comprehension of relevant information so as to improve responses. In the visualization HIT used as example above, disfluent elements, either perceptually (hard-to-read fonts or more effort-intensive visualization features, such as legends rather than labels [24]), might be incorporated to promote analytical thought when needed.

Similarly, qualification tasks for filtering and training workers might be made more efficient using strategies identified by educational psychologists for how to increase learning by inducing active processing. Opportunities for self-explanation, for example, might be incorporated into qualifications and HITs such as by requiring verification questions that ask the worker to supply their thinking processes for how they arrived at an answer.

Challenge: Analytical / deliberative HIT components override questions designed to solicit beliefs or intuitive responses.

In some cases, a researcher may want to obtain responses to HIT questions that are based on workers' intuitive knowledge and beliefs, such as immediate reactions to a novel social computing application. Through appropriate screening such asking redundant questions split across a sequential survey, it may be possible to contract workers likely to accurately report such judgments. However, verification questions needed to confirm that a worker studies HITs carefully might override her ability to report immediate reactions or intuitive knowledge and beliefs.

Proposed Solution: Clearly distinguish tasks that rely on different types of reasoning in different parts of a task.

By separating as much as possible portions of a HIT that rely on analytic, deliberative thought and those that require intuitive, associational judgments, it is less likely that intuitive responses will be overridden by analytical verification questions. A researcher can offset the risks that workers will not honestly report without verification questions by increasing worker's sense of the value of their work or interest in the content as described above.

Challenge – Risk of disfluency in HIT biasing answers.

As stated above, incorporating hard-to-process perceptual and linguistic elements into a HIT may increase worker's sense that they must think analytically and hence increase answer quality. Yet using disfluent elements poses the risk that the worker's negative association with the difficulty processing the information will bias their responses or decrease their motivation for further work. This is supported, for example, by the finding that people believe that stocks with disfluent names will perform worse than fluently named ones [2]. Because using MTurk hinges on designing HITs in such a way that they will be attractive to workers, this is an important consideration.

Proposed Solution: Design HITs where fluency biases could denigrate results to be balanced by avoiding extremely fluent or disfluent labels or presentation styles.

The literature on fluency effects on judgments offers considerable guidance on stimuli that tend to be considered disfluent. While much work in fonts, for example, has used very hard-to-read fonts (such as **Haettenschweiler**), less difficult but still disfluent variations can help achieve processing benefits without necessarily leading to a strong negative association. Gaining familiarity with some of the more common examples described by psychologists (see [1]) could help researchers using MTurk to avoid designing HITs that might unintentionally bias results.

ACKNOWLEDGMENTS

Thank you to Eytan Adar and Priti Shah.

REFERENCES

1. Alter, A.L. and Oppenheimer, D.M. Uniting the Tribes of Fluency to Form a Metacognitive Nation. *Pers. and Social Psych. Review 13*, 3 (2009), 219-235.
2. Alter, A. L., & Oppenheimer, D. M. Predicting short-term stock fluctuations by using processing fluency. *Proc. of the National Academy of Sciences,* 103, (2006), 9369-9372.
3. Alter, A.L., Oppenheimer, D.M., Epley, N., and Eyre, R.N. Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *J. of Exp. Psych.-General 136*, 4 (2007), 569–576.
4. Bless, H., & Schwarz, N. Sufficient and necessary conditions in dual-process models: The case of mood and information processing. In Chaiken & Trope, *Dual-process theories in social psychology*. New York: Guilford Press (1999). 423-440.
5. Cacioppo, J. T, & Petty, R. E. The need for cognition. *J. of Pers. and Social Psych, 42* (1982). 116-131.
6. Callison-Burch, C. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. *EMNLP '09: 1-1*, (2009), 286–295.
7. Chi, M.T.H., De Leeuw, N., Chiu, M., and Lavancher, C. Eliciting self-explanations improves understanding. *Cognitive Science 18*, (1994), 439-477.
8. Evans, J. In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences* 7, 10 (2003), 454-459.
9. Gill, M.J., Swann, W.B., and Silvera, D.H. On the Genesis of Confidence. *J. of Pers. and Social Psych. 75*, 5 (1998), 1101-1114.
10. Heer, J. and Bostock, M. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. *CHI '10*, (2010), 203–212.
11. Horton, J.J., Rand, D.G., and Zeckhauser, R.J. The Online Laboratory: Conducting Experiments in a Real Labor Market. SSRN eLibrary, (2010).
12. Hullman, J., Adar, E., & Shah, P. The Effect of Social Information on Visual Judgments (forthcoming). *CHI '11*, (2011).
13. James, W. (1950). *The principles of psychology*. New York: Dover. (first published 1890).
14. Kittur, A., Suh, B., and Chi, E.H. Can you ever trust a wiki?: impacting perceived trustworthiness in wikipedia. *CWCW '08* (2008), 477–480.
15. Kittur, A., Chi, E.H., and Suh, B. Crowdsourcing user studies with Mechanical Turk. *CHI '08*, (2008), 453.
16. Klein, G. *Streetlights and Shadows: Searching for the Keys to Adaptive Decision Making*. MIT Press, Cambridge (2009) 76-100.
17. Mason, W. and Watts, D.J. Financial incentives and the performance of crowds. ACM SIGKDD Explorations Newsletter 11, 2 (2010), 100–108.
18. Moreno, R. and Mayer, R.E. Personalized Messages That Promote Science Learning in Virtual Environments. *J. of Educ. Psych.* 96, 1 (2004), 165-173.
19. Novemsky, N., Dhar, R., Schwarz, N., and Simonson, I. Preference Fluency in Choice. *J. of Marketing Research 44*, 3 (2007), 347-356.
20. Oppenheimer, D.M., Meyvis, T., and Davidenko, N. Instructional manipulation checks: Detecting satisficing to increase statistical power. *J. of Exp. Social Psych. 45*, 4 (2009), 867–872.
21. Paolacci, G., Chandler, J., and Ipeirotis, P.G. Running Experiments on Amazon Mechanical Turk. *SSRN eLibrary*, (2010), 411-419.
22. Ross, J., Zaldivar, A., Irani, L., and Tomlinson, B. Who are the turkers? worker demographics in amazon mechanical turk. Technical Report SocialCode-2009-01, University of California, Irvine, 2009.
23. Salganik, M.J. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science* 311, 5762 (2006), 854-856.
24. Shah, P., and Freedman, E. (in press). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science*.
25. Snow, R., O'Connor, B., Jurafsky, D., and Ng, A.Y. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. *Proc. of the Conf. on Empirical Methods in N.L.P.* (2008), 254–263.
26. Skinner, C.S., Strecher, V.J., and Hospers, H. Physicians' recommendations for mammography: do tailored messages make a difference? *Amer. J. of Public Health 84*, 1 (1994).
27. Stanovich, K.E. (1999) Who is Rational? Studies of Individual Differences in Reasoning. Lawrence Erlbaum Associates.