Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, July 23-28, 2012, Thessaloniki, Greece
Cambouropoulos E., Tsougras C., Mavromatis P., Pastiadis K. (Editors)

Evaluating Crowdsourcing through Amazon Mechanical Turk as a Technique for Conducting Music Perception Experiments

Jieun Oh[1] & Ge Wang[2]
Center for Computer Research in Music and Acoustics, Department of Music, Stanford University, USA
[1]jieun5@ccrma.stanford.edu, [2]ge@ccrma.stanford.edu

ABSTRACT

Online crowdsourcing marketplaces, such as the Amazon Mechanical Turk, provide an environment for cost-effective crowdsourcing on a massive scale, leveraging human intelligence, expertise, and judgment. While the Mechanical Turk is typically used by businesses to clean data, categorize items, and moderate content, the scientific community, too, has begun experimenting with it to conduct academic research. In this paper, we evaluate crowdsourcing as a technique for conducting music perception experiments by first describing how principles of experimental design can be implemented on the Mechanical Turk. Then, we discuss the pros and cons of online crowdsourcing with respect to subject demography, answer quality, recruitment cost, and ethical concerns. Finally, we address audio-specific factors relevant to researchers in the field of music perception and cognition. The goal of this review is to offer practical guidelines for designing experiments that best leverage the benefits and overcome the challenges of employing crowdsourcing as a research methodology.

I. INTRODUCTION

Web technologies have greatly transformed the ways in which people use web browsers, introducing a new model of computing in the "cloud" through web applications. Among other things, this model provides an online environment for cost-effective crowdsourcing on a massive scale, leveraging human intelligence, expertise, and judgment for various tasks.

There are numerous examples of online crowdsourcing services. Some specialize in translation and exchange of knowledge; others focus on simple tasks, functioning as a micro-task marketplace. Of all such services, Amazon Mechanical Turk[1] (MTurk) is arguably the most well known and the largest in scale. Founded in 2005, Mechanical Turk user base grew to about 400,000 workers ("Turkers") from more than 100 countries by 2009, and there are typically between 50,000 and 100,000 Human Intelligence Tasks (HITs) available for workers to perform at any given time (Ross et al., 2010). MTurk *requesters* create HITs using developer tools provided by Amazon, and pay MTurk *workers* a small amount of money (ranging from $0.01 to a few dollars) for each task completed to the requesters' satisfaction.

While a web-based crowdsourcing is typically used by businesses to clean data, categorize items, and moderate content, the scientific community, too, has begun experimenting with it to conduct academic research. The growth of the MTurk micro-task marketplace has yielded a substantial amount of research that tests the limits of this crowd-powered "artificial artificial-intelligence" system. We utilize these resources, along with a set of recently published studies from other disciplines that employ MTurk as experimental methodology, to evaluate MTurk as a potential platform for conducting music perception studies.

In this paper, we describe how key principles of experimental design can be implemented using MTurk (Section II) and summarize such key concerns as subject demography, answer quality, recruitment cost, and ethical concerns that are important to conducting scientific research on human subjects (Section III). We then take a look at uses of MTurk by the Music Information Retrieval research community, paying attention to audio-specific issues (Section IV). It is the authors' hopes that this review can serve the music perception and cognition community by characterizing the types of research questions for which the crowdsourcing paradigm could open doors to new experimental possibilities.

II. IMPLEMENTING PRINCIPLES OF EXPERIMENTAL DESIGN

Exactly how can an online micro-task marketplace be repurposed into a scientific laboratory? In short, the experimenter, as a Mechanical Turk requester, designs an experiment in the form of a Human Intelligence Task (HIT), and workers who choose to perform this HIT, in effect, serve as subjects who participate in the experiment. Worker responses to a HIT can then be analysed to test some experimental hypotheses.

For instance, a simple HIT designed to conduct a psychophysical experiment on auditory perception could present workers with audio excerpts and have workers perform a certain task in response. Because MTurk allows requesters to host questions on the requesters' own website using the "external question" format, virtually all types of user interactions that can be implemented using web technologies (e.g. HTML5, Ajax, and Javascript) are theoretically feasible as the types of worker responses one can obtain through MTurk. However, for designing less sophisticated tasks (e.g. involving simple audio playback and standard web form responses), it is much easier and more convenient for the experimenter to use the "Requester User Interface" (GUI) provided by Amazon.

In the sub-sections that follow, we describe how principles of experimental design can be implemented on MTurk. Works by Mason and Suri (2011) and Horton et al. (2010) provide additional information on designing experiments to conduct behavioural research on MTurk.

A. Comparison

An experimental design almost always involves making a comparison between the treatment group(s), affected by an independent variable (at varying degrees), and the control group. Experimenters can design a HIT to make such

[1] https://www.mturk.com

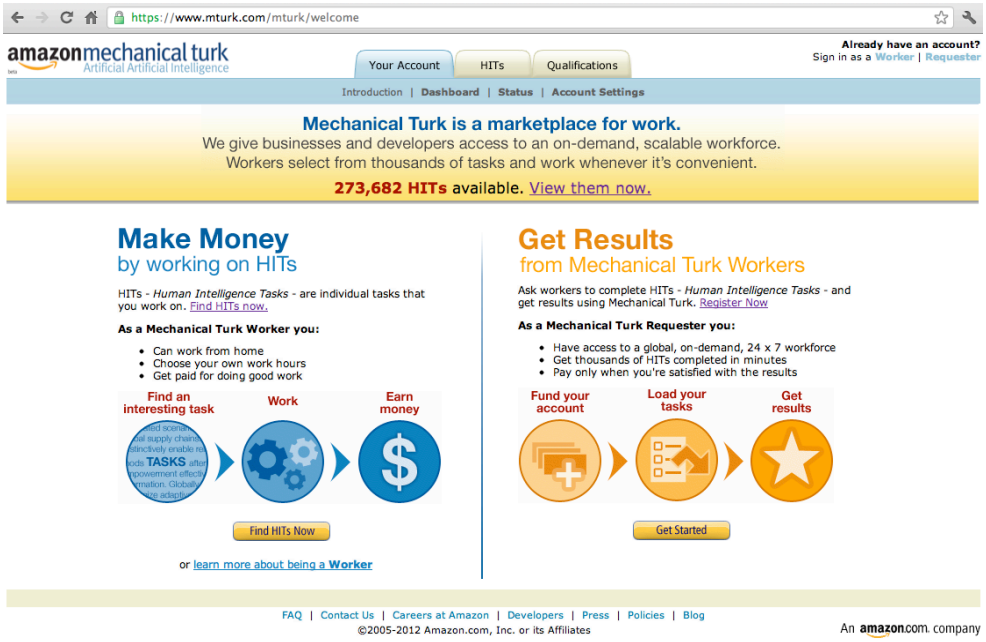

Figure 1. Amazon Mechanical Turk welcome page (accessed April 2012).

comparisons in MTurk. For the purpose of simplifying the illustration, suppose we have a binary comparison: the treatment group (call it *Group T*) and the control group (call it *Group C*). One approach to making comparisons is to present a series of stimuli, some of which are from *Group T* and others from *Group C*, within a single HIT. Consequently, a worker would be subjected to both comparison groups, allowing the researcher to make a within-subject comparison for analysis. Alternatively, *all* stimuli presented in a HIT can be from a single comparison group; this design allows for an across-group comparison.

Unfortunately, MTurk currently does not have robust support for participation assignment to simplify the setup for making across-group comparisons. That is, there is no built-in system of preventing a worker from performing HITs that expose them to stimuli belonging to a different comparison group than the one that the worker is assigned to. A quick work-around may be to temporarily "block" workers who have performed *Group T* HITs from performing *Group C* HITs (and vice-versa), but since blocking in MTurk is intended for spammers and abusive workers, it may bring negative consequences to a worker's reputation and therefore is not recommended. Another option to ensuring that workers perform only one comparison group of HITs is to require workers to go through a "qualification test", which would enable workers to become eligible to perform HITs belonging to just one of the groups. Finally, designing and hosting the HITs on the requester's own website as an "external question", to manually handle the group categorization procedure, is a clean but potentially time consuming approach.

B. Randomization

The issue of assigning workers to either the treatment group or the control group (in the context of designing across-group comparisons), naturally takes us to our second principle of experimental design: randomization. In a "true" – as opposed to "quasi" – experiment, subjects are randomly assigned to treatment conditions; true experiments are generally preferred over quasi experiments, as they tend to yield higher internal validity (Sommer, 2006).

Mason & Suri (2011) provides detailed descriptions on how to implement random assignment on MTurk. One technique is to use the provided templates and JavaScript to replace stimuli according to the condition assignment. Another approach is to create external HITs, and log the "HITId", "WorkerId", and "AssignmentId" according to the random assignment handled by the external server; this method also allows the experimenters to check if the worker has already performed a HIT from a different comparison group.

C. Replication

Mechanical Turk facilitates efforts to replicate study results by making HITs reusable. In the Requester User Interface, requesters can copy an existing HIT template of their own to quickly create another similar HIT, or otherwise simply re-use an existing HIT to release to another set of workers.

But probably the more relevant and urgent question for us in practicing the principle of replication is whether results from the MTurk are comparable to the results of previously conducted studies exploring the same research question but run in a traditional controlled laboratory setting. For instance, Heer and Bostock (2010), in a series of MTurk experiments that tested graphical perception, replicated prior laboratory studies on spatial data encodings and luminance contrast. By demonstrating that their results matched previous work and are consistent with theoretical predictions, they were able to infer viability of the crowdsourcing methodology for testing graphical perception. Similarly, Lee (2010) collected human judgments for music similarity evaluation using MTurk and compared the results to that of a more standard methodology, Evalutron6000 (Gruzd et al., 2007), and demonstrated consistency between the two techniques. Given the novelty of MTurk as a platform for conducting scientific experiments,

the authors highly recommend that researchers first check the viability of the crowdsourcing technique on the specific research question by testing the extent to which results of prior laboratory studies are replicable.

D. Blocking (Local Control)

In the statistical theory of design of experiments, blocking is the arranging of experimental units in groups (blocks) that are similar to one another[2], performed in order to reduce known but irrelevant sources of variation between the units and thereby allowing greater precision in the estimation of the source of variation under study.[3] In MTurk, even though requesters are provided with unidentifiable workerIDs, workers can still be grouped according to their geographical locations (inferred from their IP addresses) or by any other demographic or background information they explicitly provide to the requester. Ipeirotis (2010) relies on this technique to conduct a survey comparing the demographic profiles of MTurk workers from India to that of workers from the United States.

III. EXPERIMENTAL FACTORS

Now that we have demonstrated how an online micro-task marketplace can be repurposed into a scientific laboratory, we evaluate the extent to which this new paradigm of experimentation is appropriate for conducting studies in music perception and cognition. Specifically, we take a look at demographics, answer quality, recruitment cost, and ethical concerns, and discuss the pros and cons of employing MTurk with respect to these factors.

A. Demographics

Numerous studies have been conducted to better understand the demographics of Mechanical Turk workers (Ipeirotis, 2010; Ross et al., 2010). Earlier surveys indicated that workers in Mechanical Turk are relatively representative of the population of US Internet users (Ipeirotis, 2010). Then, a new policy announced in May 2007 that allowed for payment in Indian rupees [4] significantly changed the demographic makeup; Ipeirotis found the following through a survey of 1000 workers in February 2010:

> "…approximately 50% of the workers come from the United States and 40% come from India. Country of origin tends to change the motivating reasons for workers to participate in the marketplace. Significantly more workers from India participate on Mechanical Turk because the online marketplace is a primary source of income, while in the US most workers consider Mechanical Turk a secondary source of income. While money is a primary motivating reason for workers to participate in the marketplace, workers also cite a variety of other motivating reasons, including entertainment and education." (Ipeirotis, 2010)

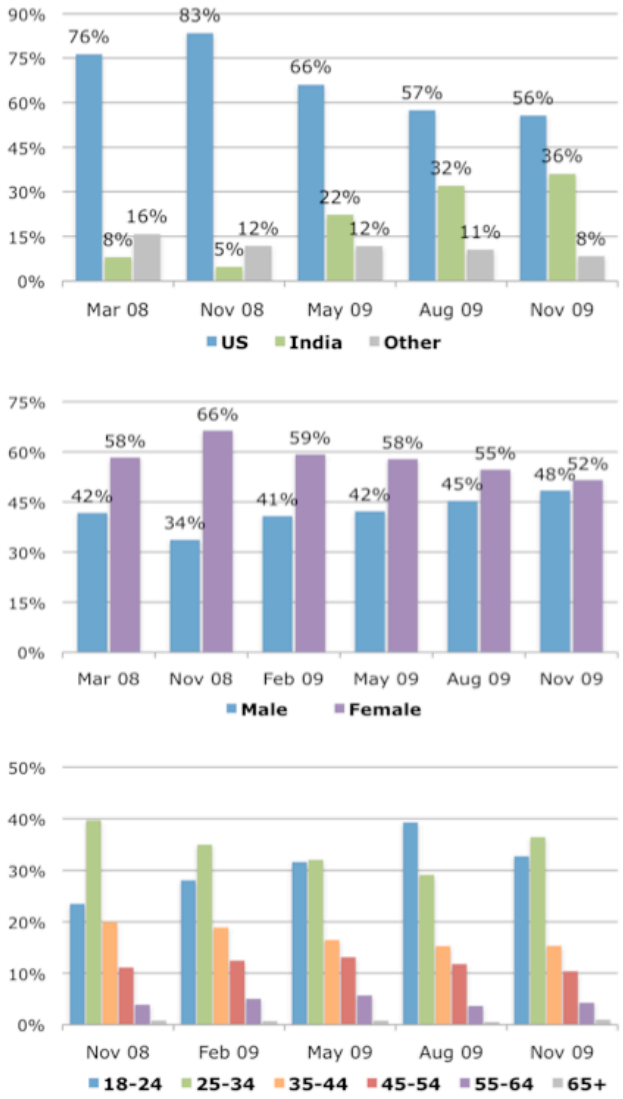

Figure 2. Nationality (top), gender (middle), and age (bottom) of MTurk workers over time. Figure taken from Ross et al., 2010.

Another study by Ross et al. (2010) looked at how the worker population has changed over time from 2008 through 2009, "shifting from a primarily moderate-income, U.S.-based workforce towards an increasingly international group with a significant population of young, well-educated Indian workers".

An advantage of such worker demographics over the traditional methods of subject recruitment is having access to a diverse subject pool. Even though the MTurk worker demographics is still restricted to Internet users who use the MTurk service, MTurk does give researchers quick access to subjects from countries throughout the world, of a wide range of age and income levels (Ipeirotis 2010; Ross et al., 2010), and also quite likely, of diverse musical backgrounds too. Given that HITs can be confined to workers who live in specific countries, researchers can more easily make focused comparisons between subjects from two or more groups (Eriksson & Simpson, 2010). Cross-cultural studies of music as well as studies on how language and music relate can potentially be realized with greater ease on MTurk.

[2] http://en.wikipedia.org/wiki/Blocking_(statistics)
[3] http://en.wikipedia.org/wiki/Design_of_experiments
[4] http://articles.timesofindia.indiatimes.com/2007-05-09/india-business/27871447_1_indian-workers-amazon-web-services-amazon-com

B. Answer Quality

One of the frequent criticisms of using online crowdsourcing services for scientific purposes is the lack of experimental control to ensure answer quality. Yet, according to Paolacci et al. (2010), "there is little evidence to suggest that data collected online is necessarily of poorer quality than data collected from subject pools (Krantz & Dalal, 2000; Gosling et al., 2004)." Because Amazon allows requesters to specify the minimum approval rate a worker must have to be eligible to work on the HIT, workers in general do have the incentives to provide quality work that can be approved.

Occasionally, however, there are "malicious" users, who provide nonsense answers to questions to make more money and spend less time (Kittur et al., 2008). But more frequently, there are "lazy" users – not necessarily with bad intentions – who put in minimal effort needed to perform the instructed tasks (Bernstein et al., 2010). And it is also possible that "good" users get distracted from their external environment, which is beyond the experimenter's control in unsupervised, online settings like the MTurk (Paolacci et al., 2010; Oppenheimer et al., 2009). Finally, there may be inherent quality issues with the worker's hardware (such as graphic cards, speakers, or slow network connection) that may adversely affect the experiment experience; Section IV Part C discusses how researchers may test for such issues when running music perception experiments.

Even though MTurk requesters can choose to not reward workers who provide low quality answers, it still takes time and resources finding, removing, and rejecting unusable responses. Nonetheless, it is possible to predict answer qualities and detect malicious or lazy workers. The simplest of these techniques is to look at the duration of time from the point at which a worker accepted a HIT to the point at which a worker submitted a HIT. (Amazon provides this information to the requesters, along with other meta-data including the worker's approval rate.) Work time that is too short may suggest that the worker rushed through the questions. Submissions with work duration that is shorter than the amount of time it should have taken the worker to listen through all of the audio stimuli presented in the HIT should probably be disregarded. On the other hand, work time that is too long may suggest that the worker became distracted in the middle and came back to complete the HIT; experimenters should also be mindful of noting such instances.

Another technique for detecting poor answer quality is through catch trials (with verifiable questions), consistency checks, and other heuristics. Catch trials are questions with obvious answers to which any worker should be able to answer correctly. Speck et al. (2011), in conducting a musical mood annotation study, occasionally inserted identical audio excerpts to serve as "verification clips". Similarly, Heer and Bostock (2010) designed a qualification test with answer choices that are grossly wrong except for the one correct choice. Lee (2010), in collecting music similarity judgments, performed consistency checks to predict work quality: "the same candidate was included twice in a single HIT, once towards the beginning, and again towards the end of the list of candidates. The expectation here was that the Turker should provide the same response for both instances since they are the same candidate". Finally, Mandel et al. (2010), in collecting audio tags, came up with a set of heuristics to detect spammers: HITs were automatically rejected if, for instance, (1) they had fewer than 5 tags, (2) a tag had more than 25 characters, or (3) less than half of the tags were found in a dictionary of Last.fm[5] tags.

Even though some of the problems in answer quality are inherent to the unsupervised online nature of MTurk, experimenters should not remain passive about this issue because a significant portion of problems that arise with answer quality can actually be mitigated by careful design of the HITs. A work by Kittur et al. (2008) demonstrates this by redesigning an MTurk experiment that led to significant improvements on answer quality. They concluded, "it is advantageous to design the task such that completing it accurately and in good faith requires as much or less effort than non-obvious random or malicious completion". We also recommend that researchers explicitly articulate their expectations on answer qualities on HIT instructions, as prior studies have shown that simply clarifying the expectations can increase data quality (Oppenheimer et al., 2009).

Finally, amidst concerns of answer quality, there are various ways in which the MTurk setting can actually yield results of improved quality over the traditional laboratory methods. Because MTurk workers can complete experiments without interacting with experimenters, possibly without even knowing that they are in an experiment, MTurk "avoids concerns of experimenter bias (Orne, 1962), subject crosstalk (Edlund et al., 2009) and reactance" (Paolacci et al., 2010). Horton et al. (2010) provides a further discussion on the validity of experiments.

C. Recruitment Cost

Crowdsourcing experiments can reduce recruitment cost, in terms of both monetary costs and time requirements.

1) Money. Setting aside for now the ethical concerns of cheap labor (which are discussed later in "Ethical Concerns"), MTurk can be an attractive option for researchers that significantly lowers the subject payment cost. In MTurk, most workers accept compensation of less than $2 per hour (Chilton et al., 2009), with the average HIT paying about $5 per hour (Ipeirotis, 2010). Paolacci et al. (2010) were able to replicate classic studies in judgment and decision-making at approximately $1.71/hour per participant and obtain results comparable to the same studies conducted with undergraduates in a laboratory setting (Mason & Suri, 2011). Heer and Bostock (2010) claimed that their crowdsourced studies on graphical perception realized a cost savings of factor of 6, compared to paying a typical compensation using the same number of subjects as laboratory studies. And beyond lower cost, the built-in payment mechanism and administration in the MTurk alleviate the logistical hassle of compensating subjects for experiment participation.

2) Time. Quick access to a large subject pool is arguably one of the biggest strengths of Amazon's Mechanical Turk services. Consequently, subject recruitment is generally very fast. Heer and Bostock (2010) describe how "in just a few days we were able to run studies that normally would have taken two weeks due to recruiting and scheduling." Paolacci,

[5] http://www.last.fm/

Chandler, and Ipeirotis (2010) describe that it took just three weeks to collect 1000 subjects. A recent study by Bernstein et al. (2011) devises the technique of "retainer model" to achieve an even faster recruitment for near real-time participation from the subjects. Such a novel technique has the potential to enable experiments involving multiple subjects working together simultaneously to explore social aspects of music performance and listening.

With regards to monetary costs, time requirements, and answer quality, raising the reward for each HIT increases the quantity of individual responses but not the quality of the work performed (Mason & Watts, 2009). The implication is that experiment results can be obtained faster by increasing the payment (Heer & Bostock, 2010).

D. Ethical Concerns

As with any studies involving human subjects, researchers using MTurk should make sure that their subjects are treated ethically. Even though IRB is more likely to treat studies in MTurk as exempt from reviews (Paolacci et al., 2010), researchers must submit a protocol to conduct any "research" involving "human subjects", per their precise definitions – as defined by United States Department of Health and Human Services (DHHS)[6], for instance.

Because MTurk and other crowdsourcing services are relatively new paradigms in scientific research, the exact recommendations given out by Institutional Review Boards may vary. See Felstiner (2010) for detailed information on ethical issues related to crowdsourcing, and Barchard & Williams (2008) for issues that apply to online experimentation. Mason & Suri (2011) provide helpful suggestions on how informed consent and debriefing may be incorporated into the design of a HIT.

As for ethical concerns with regards to unfair compensation of workers, it is recommended that the researchers pay subjects at a rate comparable to what is paid in traditional laboratory settings; unfortunately, it is often difficult to estimate the work time of a HIT, and the variation of work time across workers is quite large. Legally, however, the workers on MTurk are considered "independent contractors", and therefore fall outside the minimum wage laws (Mason & Suri, 2011). Ipeirotis's finding (2010) that many workers actually do *not* consider MTurk as their primary source of income and, in fact, the motivation for participating in MTurk is often that it "is a fruitful way to spend free time and get some cash (e.g., instead of watching TV)", could further justify lower payment rates.

IV. AUDIO-SPECIFIC FACTORS

A. MTurk Usage in MIR Research

Although using web-based crowdsourcing services has not been widely explored by researchers in music perception, the field of Music Information Retrieval (MIR), in recent years, have begun experimenting with this technique. We briefly summarize such studies by Lee (2010), Mandel et al. (2010), and Speck et al. (2011).

Lee (2010) employed the crowdsourcing methodology using the Mechanical Turk to collect music similarity judgments. Lee used the Yahoo! Media player in the MTurk version of the interface to playback pairs of audio samples, and workers were instructed to rate them as "not similar", "somewhat similar", or "very similar". Workers were paid $0.20 for a HIT, comprising of 15 pairs of stimuli for comparison. Lee demonstrated that results were comparable to similarity judgments collected from Evalutron6000 (E6K) and concluded that MTurk may be "a useful method for collecting subjective ground truth data."

Mandel, Eck, and Bengio (2010) collected song descriptor tags using MTurk. They asked workers to listen to a song excerpt (of 10-second duration) and to describe its characteristics using 5 to 15 words, and paid workers between $0.03 and $0.05 per clip, on which workers spent about a minute. The songs were mp3 files having at least 128 kbps under 10 megabytes in size. Mandel et al. concluded that MTurk "was a viable means of collecting ground truth tag data from humans."

Speck, Schmidt, Morton, and Kim (2011) used MTurk to collect musical mood annotation in order to compare results against MoodSwings[7], an online collaborative game that collects dynamic mood ratings within the two-dimensional arousal-valence (A-V) representation of emotion. The dataset consisted of 240 15-second clips, which were extended to 30 seconds. Speck et al. utilized their own website and server to host the HIT because the task required a special graphical user interface for workers to make the annotations. On the MTurk website, workers had to enter a 6-digit verification code, which was revealed to them on the external website upon finishing the task, as proof of completion. Workers were paid $0.25 per HIT. They found that MoodSwings and MTurk data[8] produced statistically consistent results.

B. Handling Audio Playback, Recording, and Accessing Other Hardware

Audio playback is an essential component in music perception and cognition experiments. If HITs are designed and hosted in external websites, it is possible to play high-quality audio in the format of Wave or Ogg, as supported by modern web browsers[9]. Essentially, all user interactions that are possible through standard web technology can theoretically be implemented as a HIT, and the user experience of such web-based experiments should become more consistent as browsers adhere to the HTML5 standard.

Audio (and video) recording may also occasionally be necessary in conducting experiments in music perception and cognition. Even though there is currently no web-native support for accessing the computer's hardware (such as the microphone or video camera), this support is currently being standardized by the World Wide Web Consortium (W3C)[10]. Also, by utilizing Flash[11], Java Applets[12], or other browser

[6] http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html

[7] http://moodswings.ece.drexel.edu/

[8] http://music.ece.drexel.edu/research/emotion/moodswingsturk

[9] http://www.w3schools.com/html5/html5_audio.asp

[10] http://www.w3.org/2011/07/DeviceAPICharter

[11] http://www.adobe.com/software/flash/about/

[12] http://java.sun.com/applets/

plugins, it is possible to access the user's microphone or a video camera from the browser. Note that this requires an additional step from users, to install or otherwise approve such programs to run in the browser. When recording and storing users' audio and video feed, researchers should be careful about privacy and data confidentiality.

In addition, researchers may also desire to have subjects input from an external hardware, such as a MIDI keyboard. This is not impossible, but faces similar challenges as implementing the recording functionality as the browser needs to access the hardware connected to user's computer. A custom browser plugin would be necessary to, for instance, send MIDI messages to Javascript.

With the shift in paradigm from computing using personal computers to cloud computing via web applications, web technologies are developing rapidly. The authors anticipate that in the coming years there will be a better browser support for accessing various input and output devices.

C. Testing Equipment Quality and Noise Level

In order to conduct music perception experiments using the subjects' own computers and speakers, it is crucial to check the equipment to ensure that they are adequate for the types of perception tasks that must be performed. Researchers can utilize Mechanical Turk's "qualification test" to gauge the quality of the equipment. For instance, to check the functioning of a stereo playback, a simple word can be spoken on the left channel, followed by a different word on the right channel, and the subject can be instructed to type the words they heard in each channel. Or, to determine equalization of the speakers, pairs of tones (at varying frequency ranges) can be played at different levels, and the subject would need to decide the extent to which one tone was louder than the other.

Using a similar approach, researchers may determine whether the worker's work environment is sufficiently quiet by asking the workers to make a recording of the room atmosphere. However, this requires extra effort from users that may discourage participation.

V. CONCLUSION

Web-based crowdsourcing services, such as MTurk, seem to be an under-explored platform for conducting music perception studies. Among other things, MTurk offers researchers a more diverse subject pool (in age, geography, language, and culture), and subjects can participate in the experiments at their convenience.

Various limitations may render web-based crowdsourcing services inappropriate in many situations. However, ongoing improvements in web technology, new techniques to ensure quality of crowdsourced responses, and better understanding of proper experimental design are heightening the potential of MTurk to serve the scientific communities at large.

REFERENCES

Barchard, K. A., & Williams, J. (2008). Practical advice for conducting ethical online experiments and questionnaires for United States psychologists. *Behavior Research Methods*, 40(4), 1111– 1128.

Bernstein, M. S., Brandt, J. R., Miller, R. C. & Karger, D. R. (2011). Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, UIST '11* (pp.33-42), Santa Barbara. http://doi.acm.org/10.1145/1866029.1866080

Bernstein, M., Miller, R.C., Little, G., Ackerman, M., Hartmann, B., Karger, D.R., & Panovich, K. (2010). Soylent: a word processor with a crowd inside. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, UIST '10* (pp.313-322), New York.

Chandler, J., Mueller, P., & Paolacci, G. (working paper). Methodological concerns and advanced uses of crowdsourcing in psychological research. http://venus.unive.it/paolacci/chandler%20mueller%20paolacci.pdf

Chilton, L. B., Horton, J. J., Miller, R. C., & Azenkot, S. (2009). Task search in a human computation market. *HCOMP '09: Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp.77–85), New York.

Edlund, J. E., Sagarin, B. J., Skowronski, J. J., Johnson, S. J., Kutter, J. (2009). Whatever happens in the laboratory stays in the laboratory: The prevalence and prevention of participant crosstalk. *Personality and Social Psychology Bulletin*, 35, 635–642.

Eriksson, K., & Simpson, B. (2010). Emotional reactions to losing explain gender differences in entering a risky lottery. *Judgment and Decision Making*, 5, 159–163.

Felstiner, A. L. (2010). Working the crowd: Employment and labor law in the crowdsourcing industry. http://works.bepress .com/alek felstiner/1/

Gosling, S., Vazire, S., Srivastava, S., & John, O. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, 59, 93–104.

Gruzd, A. A., Downie, S. M., Jones, M. C., & Lee, J.H. (2007). Evalutron 6000: Collecting music relevance Judgments. *JCDL 2007*, Vancouver.

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *Proceedings of ACM Conference on Human Factors in Computing Systems*, CHI 2010, Atlanta.

Horton, J., Rand, D., & Zeckhauser, R. (2010). The online laboratory: Conducting experiments in a real labor market. NBER Working Paper w15691.

Ipeirotis, P. (2010). Demographics of Mechanical Turk. Working Paper CeDER-10-01, New York University, Stern School of Business. http://hdl.handle.net/2451/29585

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of ACM Conference on Human Factors in Computing Systems*, CHI 2008, Florence.

Krantz, J. H., & Dalal, R. (2000). Validity of web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 35–60). New York: Academic Press.

Lee, J. (2000). Crowdsourcing music similarity judgments using mechanical Turk. In *Proceedings of 11th International Society for Music Information Retrieval Conference, ISMIR 2010,* Utrecht.

Levy, M. (2011). Improving perceptual tempo estimation with crowd-sourced annotations. In *Proceedings of 12th International Society for Music Information Retrieval Conference, ISMIR 2011*, Miami.

Mason, W. & Suri, S. (2011). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1), 1-23. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1691163

Mason. W. A., & Watts, D. J. (2009). Financial incentives and the "performance of crowds". In *KDD-HCOMP*, Paris.

Mandel, M. I. , Eck, D. , & Bengio, Y. (2010). Learning tags that vary within a song. . In *Proceedings of 11th International Society for Music Information Retrieval Conference, ISMIR 2010* (pp. 399–404), Utrecht.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867–872.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17, 776–783.

Paolacci G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk." In *Judgment and Decision Making*. 5(5), 411-419.

Ross, J. Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B (2010). Who are the crowdworkers? Demographics in Mechanical Turk. In *Proceedings of Proceedings of ACM Conference on Human Factors in Computing Systems*, CHI *2010* (pp. 2863 – 2872). Georgia.

Rzeszotarski, J. M., & Kittur, A. (2011). Instrumenting the crowd: Using implicit behavioral measures to predict task performance. In Proceedings of *the 24th Annual ACM Symposium on User Interface Software and Technology, UIST '11* (pp.13-22), Santa Barbara.

Sommer, B. (2006). *Types of Experiments*. http://psychology.ucdavis.edu/SommerB/sommerdemo/experiment/types.htm

Speck, J. A., Schmidt, E. M., Morton, B. G., & Kim, Y. E. (2011). A comparative study of collaborative vs. traditional musical mood annotation. In *Proceedings of 12th International Society for Music Information Retrieval Conference, ISMIR 2011*, Miami.