Do Mechanical Turks Dream of Square Pie Charts?

Robert Kosara Caroline Ziemkiewicz
UNC Charlotte
{rkosara, caziemki}@uncc.edu

ABSTRACT

Online studies are an attractive alternative to the labor-intensive lab study, and promise the possibility of reaching a larger variety and number of people than at a typical university. There are also a number of draw-backs, however, that have made these studies largely impractical so far.

Amazon's Mechanical Turk is a web service that facilitates the assignment of small, web-based tasks to a large pool of anonymous workers. We used it to conduct several perception and cognition studies, one of which was identical to a previous study performed in our lab.

We report on our experiences and present ways to avoid common problems by taking them into account in the study design, and taking advantage of Mechanical Turk's features.

Keywords

Empirical studies, Mechanical Turk.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*Human Factors*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*User-centered Design*

1. INTRODUCTION

Subjecting visualization techniques to empirical studies is the only accepted way of evaluating one's work. Performing such studies is quite tedious and time-consuming, however: the tester has to recruit participants, who are often university students, introduce each participant to the system, and observe for the entire length of all participants' sessions. The choice of participants also favors young people (in their late teens and early to mid-twenties), and is often biased in terms of sex and computer background. The number of participants in such studies is also severely limited by the number of available persons and the amount of time an experimenter can spend on a study. Bias may also result from mixing in participants who work with the experimenter, and such a bias is more likely to have an effect when the number of participants is small.

BELIV '10 Atlanta, GA, USA

In the interest of proper evaluation, large-scale user studies would be needed with hundreds or thousands of participants, from all age groups and backgrounds. Online studies are a way of doing this, but results from these have been mixed [8] and there is general skepticism about their viability and reliability. How do participants get recruited? Who takes the study? Is there a bias from recruiting? How can a study be performed without control over the participants' environment? Etc.

In this paper, we report on experiences gained and data gathered from a number of studies we have conducted using Amazon's Mechanical Turk (MTurk) service [15, 16, 17]. Our results indicate that the data is reliable if the study is designed to give incentives for correctness and speed, and leads to responses that can easily be checked against a ground truth. Perception and cognition studies in particular lend themselves very well to this way of working.

1.1 Lab Studies

Studies in the lab currently are (and will be, for the foreseeable future) the gold standard. They are not flawless, however: a number of problems can easily be identified.

Population Bias. The typical university lab study uses university students as subjects, and often restricts recruiting to one or two disciplines. University students are obviously in a narrow age bracket, have above-average education (by definition), and there is often a gender bias. While the student population in technical fields such as computer science tends to be male-dominated, the humanities (and in particular psychology) often have a majority of female students.

Sample Size. Practical limitations (mostly the time spent with participants during a study) encourage limiting the number of participants, often to less than 20. Each participant is thus required to perform more repetitions of a task, and even small (accidental) population biases can have a large effect on the results.

Workload. The amount of work that is required to conduct a lab study leads to fewer studies being performed. Rather than use formative studies and frequent evaluation as part of a design process, studies are often only used to validate the final design at the end. This produces studies of little value outside the immediate validation itself; formative studies often provide insights that are much more broadly applicable.

Personal Information. Any personal information collected (which at least includes the name on an Informed Consent form) must be kept strictly separate from the measured or recorded data. The personal data is normally of no use to the experimenter, but storing it becomes a liability with potentially severe consequences if it gets lost or compromised.

A large part of these issues can be resolved by using online studies, though these naturally have their own drawbacks.

1.2 Online Studies

Online studies provide fast access to a potentially huge pool of participants, faster turnaround, etc. There are still considerable problems that make them mostly impractical without some kind mediator.

Vote Flooding. A common problem in online polls is that a small number of users will submit a large number of votes to skew the results in a certain direction. In the case of user studies, it is impossible to tell how often a single person has taken part in the study, making statistical analysis problematic and skewing results.

Control. Perception studies usually require a great amount of control over the environment the user is in and at least provide for a period of uninterrupted, focused work. In an online study, the user could be listening to music, cooking dinner, or be otherwise distracted while performing the tasks.

Incentives. Participants in lab studies are usually given some kind of incentive to take part, like a small amount of money or course credit. Lacking an incentive, participants online have little motivation to perform the tasks accurately or quickly, and thus introduce noise. Payment over the Internet is possible in principle, but is too expensive and time-intensive to be a realistic option in practice (unless all participants agree to use a common system).

Recruiting. Finding a good mix of participants is a challenge in any case, and recruiting subjects for an online study can easily lead to bias (e.g., advertising/posting on particular websites). Knowing little about the participants' background means that such a bias might go undetected.

While MTurk cannot solve all these problems, it provides a workable solution to most of them. As we argue below, some of these problems are also non-issues in practice.

1.3 Amazon Mechanical Turk (MTurk)

Amazon's Mechnical Turk [2] service is named after *The Turk*, a 19th-century 'machine' that was supposedly able to play chess, but that was really operated by a person hidden within. Amazon's service was designed to let programs present tasks to humans that a computer cannot – or cannot efficiently – solve. Such tasks include deciding whether an image fits a product description, transcribing podcasts, or finding traffic signs in a series of images taken from a mobile platform. These Human Intelligence Tasks (HITs) are usually easy for a human to solve, and often take only seconds to complete. In addition to these micro-tasks [8], there are also more complex ones that can take longer, like researching company websites or finding a restaurant's opening hours. For both kinds of task, payment is usually small, leading to an average pay of less than $10 per hour.

Mechanical Turk provides an interface for requesters to post HITs to a market where hundreds of thousands of workers (known as *turkers*) can choose to complete them. This enormous pool of anonymous, motivated individuals makes it possible to conduct experiments with a large number of participants in a very short amount of time. While it is theoretically possible to get a large number of participants at once, we have found that during the busier times, we can get about 10-12 participants per hour for user studies.

Some of the key features of MTurk that make its use for studies possible and attractive are the following:

Pay for Performance. Turkers get paid for their work, and they are subject to the usual rules: if their work product is not acceptable, it will be rejected, and they will not get paid for it. Getting many work units rejected limits a worker's access to further HITs, so there is an incentive to not try to game the system too much.

Payment Processing. Payment is handled directly by MTurk, and requires no infrastructure from the experimenter (except for determining the amount). There is a small price to be paid for this (10% of the amount paid, or at least $0.005 per HIT).

Anonymity and Identification. The only information about a worker the experimenter gets is a worker ID. No personally identifiable information of any kind is provided, and MTurk expressly prohibits asking workers for personal information. At the same time, the worker ID serves as an identifier to find out if the same person has taken part in several studies, and it is possible to specify that each person can only perform a HIT once.

Recruiting. The MTurk website provides a constant stream of tasks to workers, and is visited regularly by a large number of people. Further recruiting is therefore not necessary, and may in fact skew the sample much more than the largely random selection of tasks by workers (depending on the time of day, position of the task in the queue, etc.).

Diversity. We know from our studies that turkers cover a much wider age range than college students and have a much higher percentage of women than among computer science students. We also believe that their educational and ethnic backgrounds are broader than that of college students, though we do not have data to support this.

Mechanical Turk is currently only available to requesters in the United States, and the majority of turkers comes from the U.S. However, competing services like CrowdFlower [5] are becoming available that offer their services to requesters from all over the world.

2. RELATED WORK

While user studies are common in the visualization literature [10], online studies are rare. Cawthon and Vande Moere [3] performed a study on the aesthetic properties of different tree visualizations, recruiting mostly readers of Vande Moere's popular website.

Van Ham and Rogowitz [14] built an online study on the Many Eyes collaborative visualization website, where they randomly selected site visitors to take part in the study. Their method is hampered by the small amount of information they were able to collect about their participants (for legal reasons), though, which made it impossible to tell how many different people even took part in the study. They also did not collect any demographic information.

The earliest Mechanical Turk study we are aware of is work by Kittur et al. [8], who asked study participants to read Wikipedia pages to rate them and suggest improvements. Their work is an interesting start, but does not contain any interaction, and also does not present very conclusive evidence about the validity of using this kind of online study. They found a considerable amount of gaming the system, but also had very open tasks that were difficult to validate and they did not attempt to cross-validate between turkers.

More recently, work has been done in judging the quality of drawing-like images [4] and assessing the relevance of the results of information retrieval systems [1]. Studies of turker demographics have been conducted without particular tasks [12]. The most conclusive work so far is the very recent work by Heer and Bostock, who reproduced earlier studies using Mechanical Turk to study its effectiveness for perception studies [7].

One of the earliest (and most unusual) uses of MTurk is Aaron Koblin's *The Sheep Market* [9]. Koblin asked 10,000 turkers to draw "a sheep facing left" for a payment of $0.02. The results are quite fascinating, as are his collection of statistics: the average time per sheep was 105 seconds, which lead to an effective wage of $0.69 per hour.

3. ANATOMY OF AN MTURK STUDY

Mechanical Turk allows the definition of forms that can contain text, images, and embedded websites. There is also a variety of possible responses, like radio buttons or formatted text input. This provides too little control and too little data for the kind of empirical study that is interesting for perception research, however.

We made use of the possibility to embed a Java applet into the HIT page. This applet has to be hosted by the requester, and is downloaded and run when a worker previews or accepts a HIT. There are no limits to what can be done within the applet, but care must be taken to stay within most people's screen dimensions, and not require large amounts of memory or the presence of particular hardware. Limiting the number of people who can perform a HIT slows down the rate at which they are completed and leads to more incomplete submissions with complaints.

The data collected by the applet are not visible to the MTurk system, and do not appear in the dataset provided after a HIT has been completed. Instead, the applet has to implement its own logging facility and communicate its results back to its server (a Java applet can only access the server it was downloaded from).

For each HIT, the applet receives an assignment ID, consisting of a pseudo-random string of letters and numbers. Since there is no discernible order in these, the applet needs an external mechanism for sequencing assignments. Sequencing is crucial for counter-balancing the learning and other effects that would result from using the same selection and sequence of parameters for every study subject. The server that provides the data therefore also needs to record which parameters have been used, and supply the applet with the appropriate information.

After the experiment has been performed, the submitted work units have to be reviewed. MTurk provides two forms of payment: the base payment for the HIT and a bonus. Base payment is often very little, with the bonus accounting for the bulk of a HIT's value. In order to give participants an incentive to answer correctly, we generally make the bonus dependent on the correctness of answers. This needs to be done in a way that is fair to the turkers (i.e., 100% accuracy cannot be expected), but still provides enough motivation to take the task seriously. Depending on the study, getting closer than 10-15% of the correct value is usually a good choice.

HITs can be rejected, which we decided only to do when we did not receive any data or we detected blatant gaming. Workers can add comments when submitting a HIT, and many empty HITs are actually used to report problems with the program.

MTurk's entire requester side is fully automated and accessible through a web services API from most programming languages (and a number of libraries are available to make the task easier). Requesting a user study can easily be done by hand, though we decided to use a program for easier reproducibility of parameters. Reviewing needs to be done automatically for any study of practical size.

4. EXAMPLE STUDIES

We base our findings on experiences on a total of six studies we have conducted using MTurk. This section presents two of the studies in detail, with more information summarizing all the studies in later sections.

4.1 Study I: Metaphors in Visualization

An important step in understanding the use of Mechanical Turk in visualization studies was to test whether the process would produce similar results as recruiting participants in a traditional manner. To answer this question, we extended a study previously performed with 33 students in our lab [15]. The purpose of this study was to examine the effects of compatible visual and verbal metaphors in a user's understanding of two tree visualization methods: a node-link diagram and a treemap [13].

4.1.1 Procedure

The procedure of this extension followed that of the lab study almost exactly, and was driven by the same Java ap-

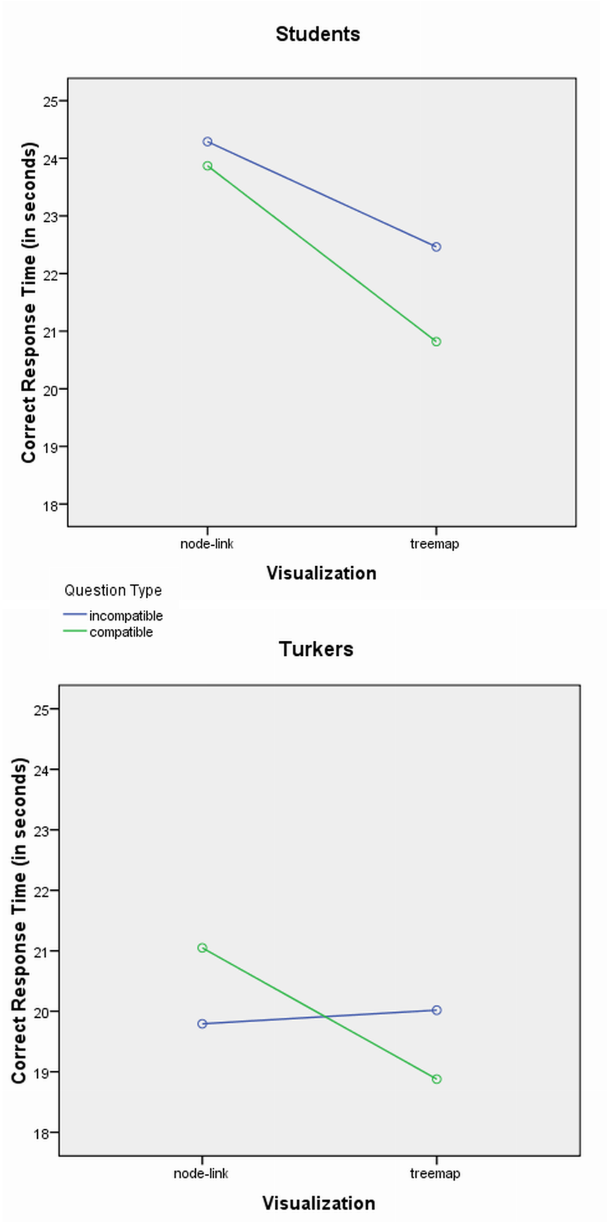

Figure 1: Turkers were overall faster in response time than students in the original metaphors study. They also showed a different pattern in metaphor compatibility effect, and were in fact faster at answering incompatible questions in the node-link condition.

plet. We recruited 86 participants through Mechanical Turk. Participants were shown three hierarchical datasets in one of two tree visualizations: a treemap or a node-link diagram. The visualization type varied between subjects, so that each participant saw only one type of visualization. The three datasets were described to the participants as representing hypothetical file hierarchies.

Participants were initially told that the purpose of the study was to evaluate different types of hierarchy visualization. After an initial training period in which they answered four questions and were given a chance to try again if they answered incorrectly, participants were asked eight questions about each dataset. For each of the eight questions, we prepared two versions: one that reflected a containment metaphor, and one that reflected a levels metaphor. The containment metaphor was considered to be more compatible with the treemap view, and the levels metaphor was considered to be more compatible with the node-link view. We hypothesized that compatible questions would lead to faster response times and greater accuracy.

Verbal metaphor was varied within subjects, in order to study the compatibility effect independently of individual differences in accuracy and response time. During their time with each of the three datasets, a participant saw four questions of the containment type and four questions of the levels type. The set of questions used for each dataset was counterbalanced from subject to subject, and question order was randomized. The result is that each participant, during each session, would answer a series of eight questions that randomly switched between a compatible and an incompatible metaphor relative to the visualization she was using.

For each question, we measured the participant's response time and whether they answered the question correctly. Altogether, participants answered twenty-four task questions. After the three sessions were complete, users were asked to rate the difficulty of the visualization they had used and to describe it. Users were given a base payment of \$0.10, and received a bonus of \$0.05 for each correct response and \$0.02 for each incorrect one.

4.1.2 Results

While the overall response patterns were similar, several differences were apparent between the results of the turker population and the students in the original lab study. There was no apparent difference between the two groups in terms of accuracy; however, a t-test on response time found that the turkers ($M = 20.2s, S.D. = 14.8$) were significantly faster than the students ($M = 23.0s, S.D. = 15.7$), $t(1966) = 3.75, p < 0.001$. This faster response time suggests that turkers may be more motivated to finish the task quickly, perhaps wishing to maximize the amount of money they can make in a particular timespan.

The primary finding of the original study was a correlation between a participant's tendency to answer compatible questions faster and her overall accuracy ($R(33) = .49, p < 0.01$). In the Mechanical Turk extension, this effect was apparent as a trend but did not reach significance ($R(86) = .20, p = 0.07$). Curiously, apart from being generally faster, the turkers showed similar patterns to the students except in

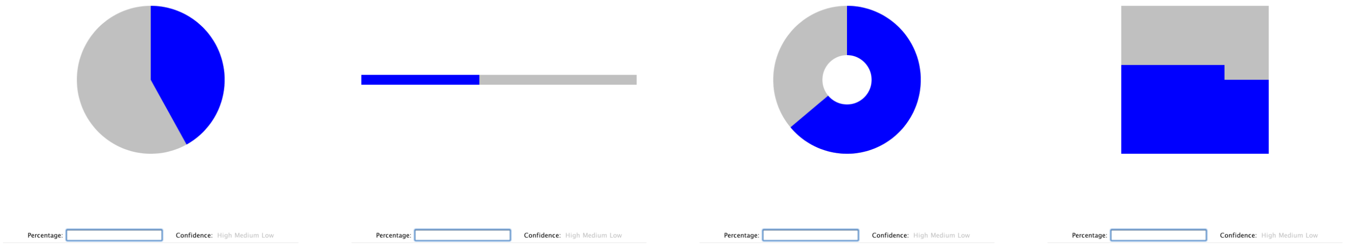
Percentage:
Confidence: High Medium Low
Percentage:
Confidence: High Medium Low
Percentage:
Confidence: High Medium Low
Percentage:
Confidence: High Medium Low

Figure 2: The four visualization types tested in Study II: pie chart, bar chart, donut chart, and square pie chart.

the node-link condition, where they answered incompatible questions faster than compatible ones, suggesting better overall performance for the containers metaphor (Figure 1). While this interaction between participant group, compatibility, and visualization is not significant, it may contribute to the weak replication of the primary finding.

4.2 Study II: Visualizing Percentages

A number of widely used chart types are often criticized for being difficult to read accurately when showing percentages. This includes the ever-present pie chart, as well as its variation, the donut chart (which is a pie chart with a circular hole in the center). An alternative is the square pie, or waffle chart, which is a square divided into $10 \cdot 10 = 100$ fields, which makes it quite easy to read a value with an accuracy of a single percent. A stacked bar chart was also included.

4.2.1 Procedure

The study used a within-subject design, with 20 questions for each visualization type, using numbers from 1 to 99 that followed a Gaussian distribution ($\mu = 50, \sigma = 25$). Values outside that range were clipped, and consecutive values had to differ by at least three percent points. Subjects ($n = 48$) entered estimated percentage shown in the chart and indicated their confidence as *low*, *medium*, or *high* (Figure 2). All input was collected through the keyboard, with participants typing in the number and one letter (l, m, or h) for the confidence, and hitting the return key to advance to the next question.

Participants were paid a base rate of \$0.20, with a bonus of \$0.02 for answers that were within two percent points of the actual value, and \$0.01 for each answer outside that range (i.e., a total of \$1.00 to \$1.20).

4.2.2 Results

A Pearson's Chi-Square test of independence finds a significant interaction between chart type and confidence, $\chi^2(6, N = 3840) = 293.4, p < 0.001$. Examining the crosstabulation shows that confidence is more likely to be high when using a square pie and more likely to be low or medium when using a bar chart. Confidence in the donut condition is the least likely to be rated as low.

An ANOVA on the difference between a participant's estimate and the actual value found a significant main effect of chart type, $F(3, 3836) = 20.24, p < 0.001$. Follow-up tests using a Tukey HSD found that the square pies lead to significantly lower deviations from the actual value ($M = 1.52, S.D. = 0.13$) than all other chart types. Donuts ($M = 2.17, S.D. = 0.13$) and pie charts ($M = 2.23, S.D. = 0.13$) did not have significantly different means from one another, while bar charts ($M = 2.93, S.D. = 0.13$) lead to significantly higher deviations from the true value than all other chart types.

A participant's confidence is also a highly effective predictor of the deviation from the true value. An ANOVA on the difference between a participant's estimated value and the actual value found a significant main effect of confidence, $F(2, 3837) = 775.88, p < 0.001$. The mean differences for low confidence ($M = 3.96, S.D. = 0.36$), medium confidence ($M = 2.69, S.D. = 0.09$), and high confidence ($M = 1.56, S.D. = 0.09$) suggest a good match between a participant's self assessment and her actual performance.

5. POPULATION DIFFERENCES

Turkers differ considerably from the students who are often the subject for lab studies. We collected different information from our six completed MTurk studies to compare to our own lab study and data from the literature. These studies include the two previously described as well as a study on individual differences in visual metaphor use [17] and a study on the effect of design on semantic responses to data [16]. Data on age and gender was not collected from the Visualizing Percentages study. While our demographic information is largely self-reported and unverifiable, participants had no motivation to deceive and our results were consistent across the different studies.

5.1 Age

The Mechanical Turk population is clearly much older and represents a much wider range of ages than studies with university students usually do. Our lab study ($n = 33$) had a median age of $M = 23.4$, with a standard deviation $\sigma = 5.11$ (min = 18, max = 40). The MTurk participants' self-reported ages over the three studies in which we collected this information ($n = 366$) has quite a different distribution, with $M = 32.0$, $\sigma = 9.6$, min = 18, max = 64 (Figure 3). Our turkers are older and more evenly distributed than the ones reported elsewhere [12], which had 40% of their samples in the 18–24 range.

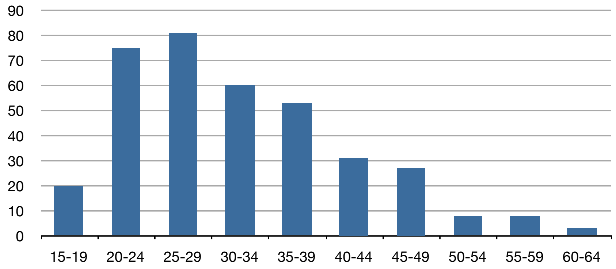

Figure 3: Age histogram for participants in three MTurk studies.

The maximum age of 40 in the lab study is an outlier, while the MTurk population had 19 (5%) participants who were 50 or older.

5.2 Gender

There are vast gender gaps between the participant populations in the different studies. Our own lab study had 73% males and 27% females; as a comparison, the validation study for the International Personality Item Pool (IPIP) [6] personality test which we used in one of our studies reports almost the exact opposite (73.3% female, 26.7% male). Both are taken from student populations, but our own study was mostly computer science students, whereas the other study seems to have included mostly the humanities.

Taken together, our three MTurk studies show a slightly more balanced image, with 57.4% women and 42.6% men out of 366 total participants. Ross et al. [12] found a similar distribution, with 55% women and 45% men.

5.3 Personality Traits

In our individual differences study, we asked participants to fill out a reduced International Personality Item Pool ("Mini-IPIP") survey [6] that assesses five personality traits: agreeableness, conscientiousness, extraversion, neuroticism, and openness. We compared the 92 surveys we received to a study performed with 329 students at an American college [6]. That sample is comparable to students in our lab experiments, with 68.8% first- or second-year students, though its gender skew is the opposite of our usual population. The MTurk participants who filled out these surveys were 60.2% female.

The results, summarized in Figure 4, show a significantly lower agreeableness (defined as cooperativeness and compassion), slightly lower conscientiousness (self-discipline and aim for achievement), significantly lower extraversion (seeking the company of others), significantly higher neuroticism (emotional instability), and significantly higher openness (imagination, curiosity). All significant differences are at $p < 0.005$.

While these differences may not have broad implications for visualization research specifically, they should be kept in mind when comparing results between MTurk and lab studies. In particular, our research found an effect of openness on the ability to switch between visual metaphors [17], so the higher scores of the turkers on this measure may be especially noteworthy for visualization researchers.

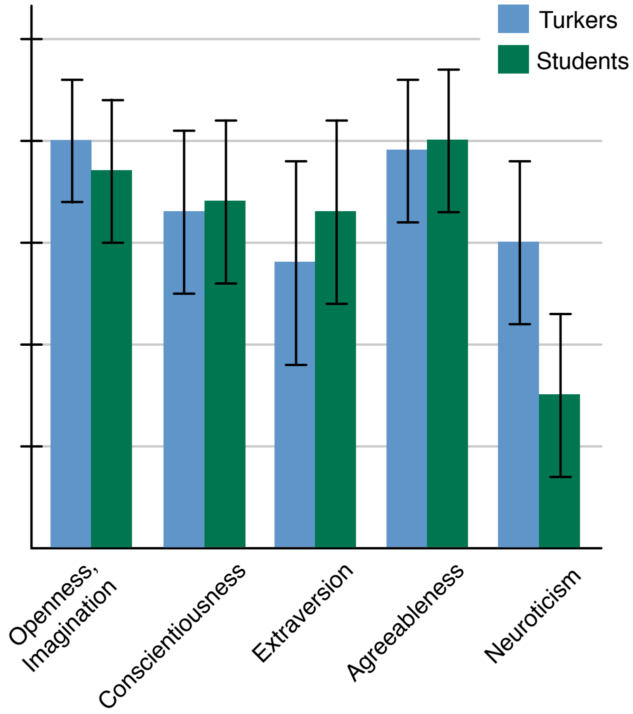

Figure 4: Comparison of personality traits between MTurk participants and students. The differences in agreeableness, extraversion, neuroticism, and openness are significant ($p < .005$).

5.4 Geography and Time

While the vast majority of MTurk participants (73%) came from the U.S., there were participants from every continent (except Antarctica). Our analysis shows participants from 30 other countries, including Canada, Australia, UK, Russia, Spain, India, Israel, Chile, Korea, and Tunisia.

Almost 10% of IP addresses could not be mapped to countries, and there is some disagreement between different location services for particular IP addresses. Within the US, the uncertainty is even higher, with 45% of IP addresses in unknown locations. Among those identified, the majority was found in California, with New York, Texas, and Virginia close behind. Turkers were located in 36 states.

Looking at the distribution of HITs over the day reveals a similar picture (Figure 5). Most of the work is being done during the day in U.S. Eastern Time (UTC-5), with the night hours from 2am to 8am being much less active than the rest of the day. There is a slight increase in the early afternoon and from about 6pm to midnight. These results are based on our first four studies running over 16 days total and 260 submitted HITs.

6. DISCUSSION

Our experiments with MTurk were quite successful, but we still learned a number of interesting and surprising lessons.

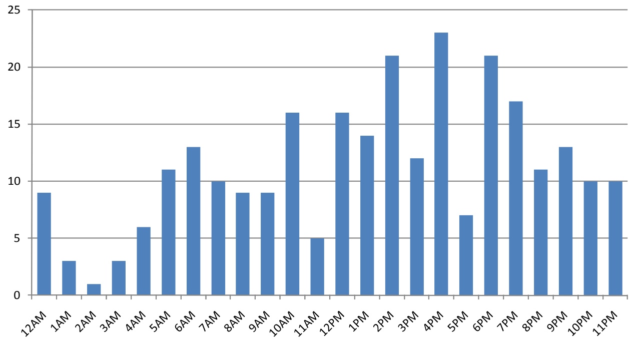

Figure 5: Time histogram for HITs completed by study participants in four MTurk studies; times are in U.S. Eastern Time (EST).

6.1 Reliability of Data

Kittur et al. [8] found a considerable amount of people trying to game the system by providing nonsensical or otherwise low-quality reviews of pages. The difference between their tasks and the typical perception of cognition study task is that the latter usually have a ground truth or correct answer that can be used to measure correctness. MTurk's bonus system can be used to reward turkers that take the tasks seriously. By providing a low base amount and either only pay for correct responses, or at least considerably higher pay for correctness, the study participants have an incentive to honestly try to do their best.

Of course, few responses will be 100% correct, so a fair reward system needs to be realistic. In the percentage estimation study, we considered responses within 10% of the correct value to be correct. Most responses were correct by that metric, so the average payout was very close to the maximum possible.

In our studies, we have only found one instance of obvious gaming, where a worker apparently got bored halfway through the study and simply started checking the same answers for all remaining tasks. Having HITs returned is quite common, and likely the result of the workers losing interest or the task taking too long for the pay it promises. Returning a HIT does not carry any penalty for the worker, whereas getting one's work rejected can bar him or her from subsequent tasks that require a minimum level of accepted work (typically between 75% and 90%).

Given the incentives for trying harder, the potential for checking correctness, and the disincentives against getting one's work rejected, we believe that our results reflect honest attempts at answering correctly.

All the demographic data we present in this paper is self-reported by turkers, with no way for us to verify it. While it is certainly possible that some of it is inaccurate, there is no obvious reason for lying. And while it would be possible to verify age and sex in a lab study, we did not verify either in our initial metaphors lab study.

MTurk's user agreement and our ethics guidelines require that workers and study participants are at least 18 years old. One participant in our studies reported his age as 16, which we take as validation that there is no incentive for lying (the worker had no reason to believe that we would report him to MTurk). We did not consider that participant's data in our analysis, however.

While learning effects can be controlled easily within each subject, and the HITs can be designed to only allow each user to complete one, there might still be a learning effect from a user completing several studies. We have found about a dozen users who have taken more than one of our studies, but did not consider that fact in our data analysis for each. If this might be a problem (in the case of highly related studies, for example), MTurk offers the possibility of using qualifications to filter eligible turkers. These could be used to filter out workers that have taken part in earlier studies.

6.2 Technical Issues

Given the potentially different screen and window sizes, it is important to make sure that the entire applet is seen by the user. In one study, we included a simple calibration step at the beginning that required the use to click Xs that popped up in the corners of the applet. If the user could not see the Xs, she was not able to click them, and could not continue the study. Similar safeguards might be used to check for a minimum contrast setting (e.g., by displaying text in gray on a background that differs only little), color blindness, etc.

In one of the studies, the participant's response consisted of a single key press, but without providing an input field. We had not seen any problems with that in our lab study, but this turned out to cause some confusion – especially in conjunction with input focus issues reported by a number of participants. The lesson from this is to stick to known user interface paradigms as much as possible for the study input, and to provide the user with some feedback about his or her response. Improved instructions might also have helped, as would a help function that paused the study and showed the instructions again.

An interesting problem that was due to a bug on Amazon's part was that after running a reviewer program repeatedly, we found that several of the workers had received bonuses several times. We tried to track down the bug in our program, and eventually discovered that there was an issue in the reported numbers from Amazon: all payments were correct, they were just reported wrong on the results page.

What we learned from this was that we needed complete and gapless audit trails for everything the reviewer program does, including why it does it. Amazon provides a detailed statement that shows the base and bonus payments for each worker, but the actions of the reviewer program that makes the payments are not covered by this. Bugs in most academic prototypes are not very critical, but in a program that makes payments, such problems are potentially much more severe.

6.3 Other Lessons Learned

New HITs are picked up very quickly, and slow down the older they get. We assume that the workers tend to look at the top of their list and pick from the HITs presented

Study	Participants	Days to Completion
Metaphors in Visualization	86	5
Individual Differences	84	6
Visualizing Percentages	48	3
Design Elements	42	2
Unpublished 1	47	<1 (1 hour)
Unpublished 2	65	<1 (17 hours)

Table 1: Duration of six recent MTurk studies.

there. Once a task slides off the first page, there is a considerable slow-down in the number of results reported back. Our metaphors pilot study with 10 participants took only about 30 minutes, while the remaining 100 we added later took almost a week (Table 1). Anecdotal evidence suggests that Monday morning or lunch time (EST) makes it more likely for studies to be completed quickly. One recent study with 47 participants was completed within an hour.

The lesson learned from this is that if a pilot seems to work, it is important to add additional assignments to the HIT as quickly as possible. Adding assignments does not push the HIT to the top of the list again, so the later this is done, the fewer people see it on the first page of available HITs. Adding to the same HIT ensures that every worker can only participate in the study once, which is more difficult to achieve otherwise. Also, to really make use of the power of the crowd, providing a large number of assignments right away will lead to more work being done in parallel.

While turkers are less extroverted than students according to the Mini-IPIP, there was a surprising number of comments. Most of these were positive (if we disregard reports of errors that were mostly due to keyboard focus issues in the applet), with many of them expressing curiosity about how well they had done. One participant even included her email address and asked to be sent information once the study was complete.

One problem we found was that a number of workers hit the submit button on the HIT even though they had not completed the study. Some of them reported problems, but some did not. We had not anticipated this, and had to come up with a consistent policy. We decided to still pay them as if all their responses had been wrong, but given the number of such responses we ended up with, we should have decided to not pay them, and communicate that from the very beginning.

7. CONCLUSIONS AND FUTURE WORK

Online studies provide the means of reaching a large number of study participants in a short amount of time. While there are some additional technical requirements, the time savings from not having to supervise the study in a lab and the parallel nature of the work are significant. The population that can be reached with these studies is also much more diverse than the student population usually used in studies.

While it may appear easier to game the system and collect bad data, we believe that we can eliminate this problem by narrowly focusing the tasks and providing a clear incentive structure for good work. Blatant gaming can also typically detected very easily in the analysis.

We continue to run studies using MTurk and are also exploring other platforms. In addition to refining our own software (which we intend to publish soon), we are also looking into other toolkits like TurKit [11].

In addition to our findings, there are many more questions to be explored to optimize studies. One would be to find the best day of the week and time of day to launch studies, so they are picked up and completed in the least amount of time. Another is to experiment with the number of questions and study duration per worker to minimize the number of returned HITs (which decrease parallelism).

8. REFERENCES

[1] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[2] Amazon mechanical turk. http://mturk.com/.

[3] N. Cawthon and A. V. Moere. The effect of aesthetic on the usability of data visualization. In *Proceedings of the 11th International Conference on Information Visualisation (IV)*, pages 637–645. IEEE CS Press, 2007.

[4] F. Cole, K. Sanik, D. DeCarlo, A. Finkelstein, T. Funkhouser, S. Rusinkiewicz, and M. Singh. How well do line drawings depict shape? In *Proceedings SIGGRAPH*, pages 1–9. ACM, 2009.

[5] CrowdFlower. http://crowdflower.com/.

[6] M. B. Donnellan, F. L. Oswald, B. M. Baird, and R. E. Lucas. The mini-ipip scales: Tiny-yet-effective measures of the big five factors of personality. *Psychological Assessment*, 18(2):192–203, 2006.

[7] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings CHI*, 2010.

[8] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings CHI*, volume Data Collection, pages 453–456. ACM Press, 2008.

[9] A. Koblin. The sheep market. http://www.thesheepmarket.com/.

[10] R. Kosara, C. G. Healey, V. Interrante, D. H. Laidlaw, and C. Ware. Thoughts on user studies: Why, how, and when. *Computer Graphics and Applications (CG&A), Visualization Viewpoints*, 23(4):20–25, July/August 2003.

[11] G. Little, L. B. Chilton, R. C. Miller, and M. Goldman. Turkit. http://groups.csail.mit.edu/uid/turkit/.

[12] J. Ross, A. Zaldivar, L. Irani, and B. Tomlinson. Who are the turkers? worker demographics in amazon mechanical turk. http://www.ics.uci.edu/˜jwross/pubs/SocialCode-2009-01.pdf.

[13] B. Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, 11:92–99, 1992.

[14] F. van Ham and B. E. Rogowitz. Perceptual organization in user-generated graph layouts. *Transactions on Visualization and Computer Graphics*, 14(6):1333–1339, 2008.

[15] C. Ziemkiewicz and R. Kosara. The shaping of information by visual metaphors. *Transactions on Visualization and Computer Graphics*, 14(6):1269–1276, 2008.

[16] C. Ziemkiewicz and R. Kosara. Design elements and the perception of information structure. In *InfoVis Posters/VisWeek DVD*, 2009.

[17] C. Ziemkiewicz and R. Kosara. Preconceptions and individual differences in understanding visual metaphors. *Computer Graphics Forum (Proceedings EuroVis)*, 28(3):911–918, 2009.