PSIHOLOGIJA, 2010, Vol. 43 (4), 419–428
© 2010 by the Serbian Psychological Association

UDC
DOI:

RESEARCH METHODS AND TECHNIQUES

Using Amazon Mechanical Turk for linguistic research[1]

Tyler Schnoebelen[1], Victor Kuperman[2]
[1] *Stanford University, USA*
[2] *McMaster University, Canada*

Amazon's Mechanical Turk service makes linguistic experimentation quick, easy, and inexpensive. However, researchers have not been certain about its reliability. In a series of experiments, this paper compares data collected via Mechanical Turk to those obtained using more traditional methods One set of experiments measured the predictability of words in sentences using the Cloze sentence completion task (Taylor, 1953). The correlation between traditional and Turk Cloze scores is high (rho=0.823) and both data sets perform similarly against alternative measures of contextual predictability. Five other experiments on semantic relatedness of verbs and phrasal verbs (how much is "lift" part of "lift up") manipulate the presence of the sentence context and the composition of the experimental list. The results indicate that Turk data correlate well between experiments and with data from traditional methods (rho up to 0.9), and they show high inter-rater consistency and agreement. We conclude that Mechanical Turk is a reliable source of data for complex linguistic tasks in heavy use by psycholinguists. The paper provides suggestions for best practices in data collection and scrubbing.

Compared to lab studies conducted with undergraduates, online web experiments can offer broader demographic and educational coverage as well as a cheaper and faster source of data. Yet this practice has not been widely adopted in psycholinguistics. While the advantages of web experiments are widely known (e.g., Birnbaum, 2004; Dandurand, Shultz, & Onishi, 2008; Krantz & Dalal,

Corresponding author: tylers@stanford.edu

[1] This research was supported by the William R. and Sara Hart Kimball Stanford Graduate Fellowship for the first author and the Rubicon grant from the Netherlands Scientific Research to the second author.

2000; McGraw, Tew, & Williams, 2000; Reips, 2002a, 2002b), some practical disadvantages are not discussed: researchers have to code their experiments, host them on a server, and come up with a recruiting strategy (rather than use existing pools of undergraduates). Moreover, Dandurand et al. (2008) argue that web experiments are vulnerable to such problems as a larger number of uncontrolled variables compared to the lab setting; multiple submissions (with participants contributing more than one set of responses to the resulting data pool); and self-selection biases (though this may also affect lab studies that recruit university students). In addition, there is a methodological specter: do web experiments offer the same quality of data as traditional methods? Can we trust the results?

A host of studies comparing online experiments to lab-based studies suggest that online experiments generate similar results for a wide range of phenomena. The classic survey is Krantz & Dalal (2000), which looks at 12 psychological studies comparing Web-based and lab-based results. They find that 10 of these show very strong matches: the two exceptions —Buchanan (1998) and Senior, Phillips, Barnes, & David (1999) —observe weaker relationships between the dependent and independent variables in the web-based data than the lab results. Dandurand et al. (2008) compare online and lab methods for problem-solving and review a number of studies showing that questionnaires, intelligence tests, and tests for biases in syllogistic reasoning have similar results whether conducted in the lab or online. Dandurand et al.'s own study also shows that longer, more demanding tasks can still be done in online environments.

Computer scientists and other researchers have begun tapping in to "crowdsourcing" services where problems are distributed to a broad group. One such crowdsourcing service is Amazon's Mechanical Turk.[2] For questionnaire-based research, Mechanical Turk removes most of the practical disadvantages of web research, such as creating and storing experimental tasks as well as recruiting participants. It's not necessary to download anything to use the service and experiments can be designed using a variety of templates they provide. Amazon hosts these questionnaires and they have a database of thousands of workers who are available to respond to them. Though the focus of this paper is evaluating the service's reliability, we offer an appendix with additional background information and best practices.[3]

Among the most appealing features of Mechanical Turk is, of course, how inexpensive it is for conducting experiments. The rule of thumb for payment is around half-a-penny per question. To give an example, in one of the experiments

[2] While our results are specifically for Amazon's Mechanical Turk service, we believe they will reproduce for other crowdsourcing services—for example, CrowdFlower and clickworker.com. Additional services and resources can be found at http://crowdsortium.org and http://www.crowdsourcing.org.

[3] Both authors have trained others in how to use Mechanical Turk and have been pleased to see how straight-forward our colleagues find it to design experiments and have them running very soon afterwards.

reported below, we asked participants to answer questions about 96 different sentences. We collected responses for 20 people per sentence. This task took under an hour to set-up and was completed by participants within 55 hours. The total cost was $11.60.

Snow, O'Connor, Jurafsky, & Ng (2008) addressed the concern about the comparability of the Mechanical Turk data to laboratory results by report similar findings for five different online tasks. The task most similar to the one we report below was a word similarity task. In this task, Turkers were asked to judge 30 word pairs on a scale of 0-10 for relatedness. These pairs included things like "boy/lad" and "noon/string". Typically, this task gets very high inter-rater agreement. Snow and colleagues compare Turk results to Resnik (1999)'s 0.958 correlation score. With 10 Turkers per word pair, Snow et al. find a correlation of 0.952.

Probably the most difficult of the five tasks that Snow et al. report is affect attribution. Snow and colleagues showed Turkers newspaper headlines and asked them to give a rating from 0-100 for each of six emotions (anger, disgust, fear, joy, sadness, or surprise). The Turkers were also asked to score the headline for overall positive or negative emotion (-100 to 100). They compared the Turkers' results to the "gold standard" of 10 experts. The intra-expert agreement had an average Pearson's correlation of 0.580. The agreement of the Turkers with the expert data was 0.433.[4] That said, pooling the results increased both groups' agreement statistics. Snow et al. found that it took about 4 Turker annotations per example to equal the quality of one expert judgment, though this varied by the emotion being labeled (ranging from 2 to 9).

Snow et al. also had Turkers judge textual entailment tasks to see if, given two sentences, the second can be inferred from the first. Fourth, they also used the TimeBank corpus to perform basic event annotation (was an event strictly before or strictly after another one in a story). Finally, Snow et al. also ran a word sense disambiguation task—did "president" in a context mean a CEO, the chief executive of the United States, or the head of some other country? Across these tasks, they found very strong ($r > 0.9$) correlations between the Amazon Mechanical Turk results and the gold standards provided by the experts offline.

While the results reported in Snow et al. (2008) are encouraging for the use of Amazon's Mechanical Turk for data collection, we believe their investigation needs to be extended in several important respects. Snow et al.'s tasks were relatively simple (e.g., rating the similarity between *boy* and *lad*, as compared to rating the similarity between *give* and *give up*) and hence showed high inter-rater agreement. Furthermore, Snow et al. relied on the gold standard data supplied by a group of experts who showed high internal consistency as well. In this study,

[4] Specifically, Snow et al. calculate the Pearson correlation between each Turker's labels with each of the expert annotators. 0.433 represents the average correlation between each Turker and the group of expert annotators.

we test the reliability of Mechanical Turk as a data collection tool in relatively complex tasks that demonstrate high variability between participants both offline and online. We approach this by replicating online several psycholinguistic experiments conducted in the lab: we make sure to select the tasks that are in heavy use by psycholinguists: the Cloze completion task and the semantic similarity rating task. Correlations between data sets obtained via different methods as well as measures of internal consistency and inter-rater agreement in Mechanical Turk experiments are examined as indicators of Mechanical Turk's reliability If crowdsourced data are reliable, then psycholinguists have a new way of collecting data that is likely to be faster, cheaper, and from a more diverse population than current undergraduate sources.

EXPERIMENT 1: CLOZE SENTENCE COMPLETION TASK

Many psycholinguistic experiments require the reading or auditory comprehension of sentences and larger passages of text. One of the most important norming tasks that accompany these is the Cloze sentence completion task. As psycholinguists build models of sentence processing, they need to understand the effect of the available sentence context on recognition of words in that sentence. One way to gauge the amount of contextual constraint per word was proposed in Taylor (1953): participants are presented with a fragment of a sentence and asked to provide the upcoming word. Researchers use the percentage of responses that match the target word as an index of contextual constraints on the word and as input into their statistical and computational models. By transforming the stimuli of the study in a reading performance into stimuli for a Cloze task, researchers are able to get a sense of the linguistic and non-linguistic biases that are built in to the sentences that they have developed. The results of the Cloze tasks help distinguish words that are highly predictable given their context from those that are not.

Contextual predictability of a word has been repeatedly shown to co-determine behavioral measures of the comprehension effort. For example, Rayner & Well (1996) observed that more predictable words are fixated on for a shorter time and are skipped more often than words with a medium or low level of predictability. While a number of studies ran the Cloze completion tasks on specific words of interest, larger-scale Cloze experiments were conducted to elicit guesses for every word in every stimulus sentence for the Schilling corpus of English (Reichle, Pollatsek, Fisher, & Rayner, 1998; Schilling, Rayner, & Chumbley, 1998) and for the Potsdam Sentence Corpus in German (Kliegl, Grabner, Rolfs, & Engbert, 2004). The robust effects of word predictability led to it being included alongside word length and frequency as a "benchmark" predictor for eye movements in reading. It serves as a parameter in several models of how eye-movement control in reading works (Engbert, Nuthmann, Richter, & Kliegl, 2005; Reichle, Warren, & McConnell, 2009). To date, most

Cloze tasks have to be conducted in the lab. Given the advantages of Amazon's Mechanical Turk outlined in the introduction, it would be good to know whether Cloze tasks performed in Mechanical Turk can achieve similar results. This is the question our first experiment sets out to answer by conducting the Cloze completion task for the Schilling corpus (Reichle et al., 1998).

We also mention here, and investigate below, two other approaches proposed for measuring how predictable words are in their contexts. The goal here is to establish whether these alternatives enter into different relationships with the Cloze judgments obtained in the lab or via Mechanical Turk. One alternative measure of contextual predictability is surprisal, a measure of how unexpected a word is given the syntactic structure of the preceding sentence fragment, measured by its negative log probability (Hale, 2001; Levy, 2008). Boston, Hale, Kliegl, Patil, & Vasishth (2008) reported strong positive effects of surprisal on the eye-movement record in German and English, which would make it a contender for taking the place of the Cloze predictability estimates. However, Boston et al. (2008) observed a lack of correlation between surprisal and Cloze predictability: both were statistically significant and independent predictors of a range of eye-movement measures.

The other approach to measuring how predictable words are in their context is to look at word position in the sentence. This is argued to approximate the contextual constraint on the word in Van Petten & Kutas (1990) and is used as an approximation of contextual predictability, especially in ERP studies. For instance, Dambacher, Kliegl, Hofmann, & Jacobs (2006) found evidence supporting the link between word position in sentence and contextual predictability: the two measures correlated at r=0.41, with words further into the sentence being more predictable. Below we compare both surprisal and word position in the sentence against Cloze predictability norms obtained through Mechanical Turk and in the lab.

Method

In a Cloze completion test as implemented by Reichle et al. (1998) and Kliegl et al. (2004) participants are presented with a fragment of a sentence and are asked to fill in the following word. For example, at the beginning, the subject just sees one word "Margie". They enter a word that they think could come next. Perhaps they fill in "likes" ("Margie likes"). Instead, they see that the next word is actually "moved". Now they need to provide the third word. After they enter a word there, they see the target and are asked for the next word, and so on.

Stimuli 1: *Margie* ___

Stimuli 2: *Margie moved* ___

Stimuli 3: *Margie moved into* ___

In the experiments administered by Kliegl et al. (2004) and Reichle et al. (1998), raters worked through the whole sentence in one sitting. We changed the procedure to accommodate the usual format of Mechanical Turk tasks. Instead of working through each sentence from start to finish, each Turker worked to fill in the blank for a particular word position in a number of different sentences. In the end, both populations evaluated the same words, in the same sentences, with the same amount of "previous context". However, while lab participants were in effect "corrected" as they worked through sentences word-by-word, the Turkers were not.

The first task we launched on Mechanical Turk showed the first word of sentences in the Schilling corpus (Schilling et al., 1998), and asked Turkers to provide the second word. Because we were concerned with overtaxing the Turkers, we only wanted to ask them to fill in the blanks for 12 items at a time.[5]

After this task was completed, we launched a new task showing the first and second word of each sentence and asking for the third word. After that data was collected, we launched the next task, asking for the fourth word, and so on until we had collected words for all sentences in all positions (save the first word). This design made it impossible for someone to fill-in-the-blank for, say, Stimuli 2 if they had already seen it in Stimuli 3. The sentences for each task were randomized separately and Turkers were randomly assigned to them.

There were a total of 50 tasks with 12 stimuli in each task. Each task was completed by 50 Turkers and paid $0.05. Thus, the total cost of obtaining 50 complete protocols for 48 sentences was $130. By comparison, doing the same 48 sentences with 50 students at the usual rate of $8.00 per hour at Stanford University, we would have paid $400 for undergraduate participants. See the Appendix for data scrubbing details and recommendations, but we ended up with approximately 43 judgments per word on average.

We take that the only crucial difference in procedures between the undergraduate lab participants and the Turkers is that the former saw full sentences in one sitting, whereas the Turkers gave us point-wise estimates. In the next section we discuss the demographic differences.

RESULTS

The Turk results came from 488 Americans, ranging from age 16-80 (mean: 34.49): see Appendix for data scrubbing procedures. The Turkers were from all over the US (about 25% each from the East and Midwest, about 31% from the South, the rest from the West and Alaska). They represented a range of education levels, though the majority had been to college: about 33.8% had bachelor's degrees, another 28.1% had some college but without a degree. By contrast, the lab data was gathered from 20 participants, all of whom were undergraduates at the University of Massachusetts at Amherst in the mid-1990's (Reichle et al., 1998). Further demographic information is not available. Both populations provided judgments on the same 488 words in the same 48 sentences.

Predictability was calculated by looking at the percentage of participants who filled in the blank with the target word that was actually used in the Schilling corpus sentences. For example, the sixth word of Sentence 1 is "apartment". 15 of the 20 lab participants filled in "apartment", so the predictability of this word in this position is 0.75 for the lab participants. Meanwhile, 17 of 34 Turkers put "apartment", so the Mechanical Turk predictability is 0.50.

[5] After they completed a batch, Turkers could go back and do more. At the maximum, Turkers would see all 48 sentences—but no repeats. Very few Turkers did this, but their experience is closer to the Student participants who were asked about all 48 sentences.

Sentence 1 (target): *Margie moved into her new apartment at the end of the summer.*

Sentence 1, word position 6 as seen by participants: *Margie moved into her new* _____

Recall that the ultimate purpose of predictability measures is for use in explaining sentence processing. Predictability—along with word frequency, length, and other factors—serves as a key input for psycholinguistic models of reading time and eye fixation experiments (Kliegl et al., 2004; Kliegl, Nuthmann, & Engbert, 2006; Rayner & Well, 1996).

We set out to see how close the Cloze predictability estimates were for our two data sets across all words in all sentences. Looking at the basic statistics in Table 1, the Turk population did not seem to match the Schilling target words as often as the student participants. However, there is a fairly strong correlation between the two datasets. Since our data is not normally distributed, we use Spearman's rank correlation. It shows that the correlation for the lab and Turk data is 0.823 ($p<0.0001$).

Table 1. Descriptive statistics for lab-based Reichle et al. (1998) vs. Mechanical Turk raters in the Cloze sentence completion task for the Schilling corpus.

	Lab	Mechanical Turk
Mean predictability	0.332	0.187
S.D.	0.365	0.246
Min predictability	0.000	0.000
Median	0.15	0.075
Max predictability	1.000	1.000
Number of "full misses" (not a single match between response word and Schilling target word)	163 / 488	153 / 488

Note that 124 out of the 488 words had predictability scores of 0 across both groups (meaning no one matched them even once). This implies that most words with a zero predictability score overlapped across the two data sets. Only about 18% of zero responses for the Turk data set (29 out of 153) were unique to the Turk data and 23% of zero responses for the Schilling corpus (39 out of 163) were unique to it. We take this as a sign that the lab participants and Turkers are consistent in which words they failed to predict. We also take it to be more legitimate to drop out those words that have zero predictability in both data sets (see Table 2). Since these zero ratings have perfect correlation, it is more conservative to throw them out and focus on the data where there is variance between the two data sets. We therefore ran our correlation for a subset of the data that excluded items that had zero predictability in both data sets. In that case, rho=0.759 (p<0.0001), see Figure 1. The lab participants still have higher overall predictability scores, but the correlations are still strong.

Table 2. Descriptive statistics for lab-based Reichle et al. (1998) vs. Mechanical Turk raters in the Cloze sentence completion task for the Schilling corpus: shared zero ratings removed.

	Lab without shared zeros	Turk without shared zeros
Mean predictability	0.446	0.250
S.D.	0.358	0.256
Min predictability	0.000	0.000
Median	0.400	0.156
Max predictability	1.000	1.000
Number of "full misses" (not a single match between response word and Schilling target word)	39 / 364	29 / 364

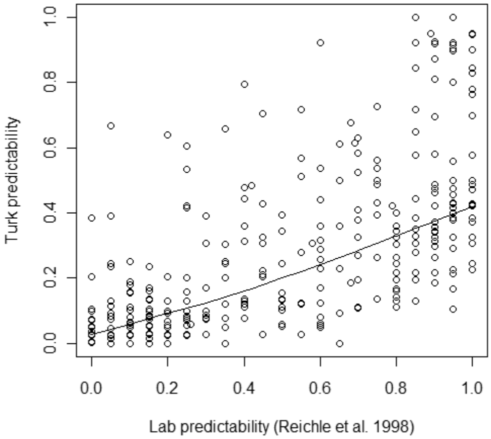


Figure 1. Cloze predictability estimates by Turkers vs. Cloze predictability estimates by lab participants (shared zeros removed). The solid trend line is obtained using the locally weighted scatterplot smoothing lowess function.

Is there evidence beyond correlations that the data from the Mechanical Turk and lab participants will perform similarly? Though it is beyond the scope of this article, we welcome psycholinguists to use our data in their model of choice and see whether the best fit to behavioral data is afforded by Cloze predictability ratings observed in the lab or by using those obtained using the Amazon Mechanical Turk interface.

To summarize, the results of the Cloze sentence completion task provided by Turkers are an excellent approximation of those obtained in the lab setting by Reichle et al. (1998)—in the most conservative measurement, the correlation is rho=0.759 (p<0.0001). The two populations overlap in which of the Schilling target words they consistently fail to match. In addition, the percentage of their correct guesses, defined as the Cloze predictability, appears to converge for the same set of sentences.

Comparing Cloze Predictability To Structural Surprisal

In this section, we compare predictability to "surprisal" values, as estimated by a syntactic parsers. Predictability taps into the range of human linguistic expectations conditioned on a particular context. Surprisal is a measure of how unexpected a word is given the syntactic structure of what has come before. We would anticipate these to be highly correlated because syntactic expectations constitute part of what people are using in filling in the blanks when they take part in a Cloze task.

To create a measure of structural surprisal, we fed all of the stimuli sentences through a state-of-the-art English part-of-speech (POS) tagger developed by Tsuruoka & Tsujii (2005) that is reported to have 94% accuracy.[6] The resulting POS tags were fed into the dependency parser DepParse 2.5 (Boston, 2010) and surprisal values were computed by the Stacks algorithm for each word in each sentence. (For an overview of surprisal, see Hale (2001) and Levy (2008), for detailed explanation of how it was calculated here see Boston et al., (2008); Boston, J. Hale, R. Kliegl, & Vasishth (2008); Boston, J. Hale, Vasishth, & R. Kliegl (in press).

Next, we compared surprisal values from the parser with predictability estimates for the lab participants and the Turkers. Using Spearman's rank correlation, the correlation between the lab participants and the parser surprisal is rho=0.256 (p<0.0001); the correlation between the Turkers and the parser surprisal is rho=0.220 (p<0.0001).

These correlations drop off even more if zeros are removed (as described above—words in sentences that were not matched by participants in either study). When zeros are removed, the correlation between the lab participants and the parser's surprisal estimate is only rho=0.0966 (p=0.0656); the correlation between the Turkers and surprisal is rho=0.0747 (p=0.155). The lack of correlation between predictability and surprisal supports the finding of Boston et al. (2008), who looked at the German Potsdam Sentence Corpus and found that structural surprisal computed over sequences of POS tags is <u>not</u> a good approximation of human predictability judgments.

If predictability is not surprisal, what is it? We are pursuing the implications elsewhere since they are beyond the scope of the current paper, but briefly, the

[6] The tags for commas were removed, as were apostrophe-s since neither dataset has judgments for them. We also removed the period after "Mr." since it was misunderstood as marking the end of a sentence.

results suggest one of three interpretations: (i) human judgments reflected in Cloze rates are not syntactically informed, (ii) syntactic predictability is largely overridden by non-syntactic dimensions of predictability, or (iii) the Cloze method is too crude to capture syntactic expectations and is primarily driven by non-syntactical information (lexical, pragmatic, word knowledge, etc.). We take (i) and (ii) to be extremely unlikely in light of the extensive literature on syntactic biases and people's sensitivity to it, and given the independent predictive roles of surprisal and Cloze predictability in Boston et al. (2008).

Comparing Cloze predictability to word position

To understand the difference between the offline and online populations of participants, we also examined the correlations between the datasets in terms of the predictability of particular word positions (Figure 2). As the participants have more context—that is, when they are getting a word at the end of a sentence—predictability goes up. The correlation is stronger (and the trend line is higher in Figure 2 below) for the lab participants, whose task actually builds sentences word by word, as in Stimuli 1-3 above.

Turkers were only asked to evaluate sentence fragments of the same length. A typical Turker would opened the survey on Monday and saw six questions. Each had one word followed immediately with a blank (for example, Stimuli 1). It is possible that subjects from Monday could come back on Tuesday and see the next word in the sentence (Stimuli 2), but the data logs from Mechanical Turk showed that this happened very rarely.

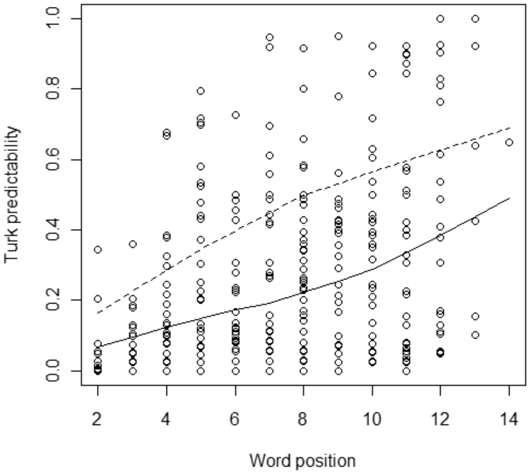


Figure 2. Mean Cloze predictability ratings per word as a function of the word's position in the sentence. The dashed (lab participants) and solid (Turkers) trend lines are obtained using the locally weighted scatterplot smoothing lowess functions. Zeros have been removed.

Table 3 reveals that the relations of both the Turk and the lab Cloze scores to alternative measures of contextual predictability are qualitatively similar. Both types of Cloze scores show no correlation with the surprisal measures computed over POS-tags of the sentence fragment preceding the critical word. Also, both versions show correlations of a similar magnitude with the word position in sentence, supporting the claim of Van Petten & Kutas (1990) that contextual constraints may be assessed by the latter measure.

Table 3. Numbers above the diagonal show correlation coefficents' Spearman rho; numbers below the diagonal report respective p-values. Shared zeros have been removed.

	Lab predictability	Turk predictability	Surprisal	Word position
Lab predictability		0.759	0.087	0.346
Turk predictability	<0.001		0.076	0.388
Surprisal	0.096	0.150		0.281
Word position	<0.001	<0.001	<0.001	

To sum up, the lab and the Turk versions of the Cloze sentence completion task show convergent results in at least three respects: (i) the words that both populations of participants failed to match are a largely overlapping set; (ii) for the rest of the words, the ratio of correct matches to total trials (Cloze predictability ratings) is highly correlated (rho = 0.759), and (iii) both sets of results enter into qualitatively similar relationships with alternative measures of contextual predictability: weak non-significant correlations with surprisal and reliable positive correlations with word position in sentence. Thus, for the researcher who typically uses lab participants, Mechanical Turk offers more data in a faster time period and for less money. These advantages are gained without losing performance.

Experiments 2-6: Semantic similarity judgments

Phrasal verbs are expressions in which a verb and a particle are used together to express a verbal meaning, for example *give in* and *pull up*. Some of these expressions are **transparently** made up of the two parts: when you *lift up a piece of paper* you are both lifting the paper and it physically goes up. But when you *bring up children* you are not really using the central meanings of *bring* or *up*—the relationship between the meaning of the parts and the meaning of the whole is rather **opaque**.[7] Semantic relatedness of the verb and/or the particle

[7] It turns out that transparency has consequences for how the phrasal verbs are used—as Bannard (2002) showed, opaque (or "idiomatic") phrasal verbs are less likely to allow

to the meaning of the phrasal verb as a whole has been demonstrated to affect both the choices of the particle placement before or after the object (cf. *lift up the paper* vs. *lift the paper up*) and the production and comprehension latencies in producing phrasal verbs (Bannard, 2002; Bolinger, 1971; Fraser, 1974; Gries, 2003; Kennedy, 1920; Lohse, Hawkins, & Wasow, 2004; McCarthy, Keller, & Carroll, 2003).

We collected judgments on the connection between particles and verbs for 96 phrasal verbs. Our main methodological question was whether these judgments would be the same regardless of who was judging them. While individual differences were expected, if we indeed measured the transparency of these phrasal verbs, groups as a whole should have had roughly similar patterns of responses, whether they were Stanford students or people who found and completed the task online via Amazon's Mechanical Turk service. We also tested how the presence/absence of the sentence context affected similarity ratings for both the lab participants and Turkers. For Turkers, we additionally tested the role of the experiment duration (the number of stimuli in the experimental task) in the similarity judgments

Method

Two offline experiments (2 and 3) were performed using Stanford University undergraduates. Experiment 2 involved a questionnaire asking participants to rate the semantic transparency of 96 phrasal verbs. Experiment 3 consisted of a paper questionnaire with the phrasal verbs in context. That is, the first group of Experiment 2 ("StudentLong") participants rated the similarity of "cool" to "cool down" on a scale 1-7:

cool cool down _______________

The "StudentContext" participants in Experiment 3 performed the same basic task but saw each verb/phrasal verb pair with an example of the phrasal verb in context:[8]

The fan will cool down the engine when it starts to heat up.

cool cool down _______________

With Mechanical Turk, we had three experiments. Experiment 4 ("TurkLong") was a replication of the first context-less questionnaire and its 96 questions. In Experiment 5 ("TurkShort") the 96-questions of the context-less questionnaire were randomized into batches of 6. Thus, some participants ended up giving responses to all phrasal verbs, while others only gave 6, 12, 18, etc responses. Experiment 6 ("TurkContext") was a variation of

the noun and the particle to alternate positions. Knowing nothing else, we would expect *lift a piece of paper up* to occur more frequently than *bring children up*. This has consequences for syntactic theory as well as in practical applications—phrasal verbs are notoriously difficult for parsers to detect in natural language processing tasks (Baldwin & Villavicencio, 2002; Roland, Dick, & Elman, 2007).

[8] StudentContext participants used a computer to fill in an Excel spreadsheet with the values—some people filled the questionnaire out in the lab, while others did it at home and e-mailed their judgments later.

the "StudentContext" task—participants were given examples of the phrasal verbs, though as with "TurkShort", they were only asked to rate 6 phrasal verbs at a time.

StudentLong and TurkLong were nearly identical in their methods. However, the two populations have different expectations about experiments and we wanted to understand whether this had an effect. Turkers tend to expect shorter projects, while students generally comply with even long tasks. Having students only see six questions was prohibitively expensive (in cost and time), but it was possible to see whether questionnaire length had an effect by comparing TurkLong and TurkShort.

We are also interested in what happens when phrasal verbs are shown as part of sentences—StudentContext and TurkContext test the effects of context on the two populations. Finally, the variation in experiments can demonstrate whether there is consistency between three different sets of Turkers giving responses to three different methods.

RESULTS AND DISCUSSION

All student participants were native English speakers who were recruited from the joint linguistics-psychology experimental participant pool of students at Stanford University. For the Mechanical Turk experiments, only data from monolingual native English speakers was used (see Appendix for data scrubbing procedures as well as best practices). The results reported here represent judgments by 215 people.

Table 4. Descriptive statistics of five experiments on similary ratings in phrasal verbs.

	Number of participants under analysis	Number of total ratings	Mode of verbs rated	Median number of verbs rated	Average number of verbs rated
TurkLong	29	2,783	96	96	96
TurkShort	66	2,372	6	18	35.94
TurkContext	81	1,834	6	6	22.64
StudentContext	27	2,592	96	96	96
StudentLong	18	1,728	96	96	96

To begin with, let's look at the correlation between ratings that were given (Figure 3).

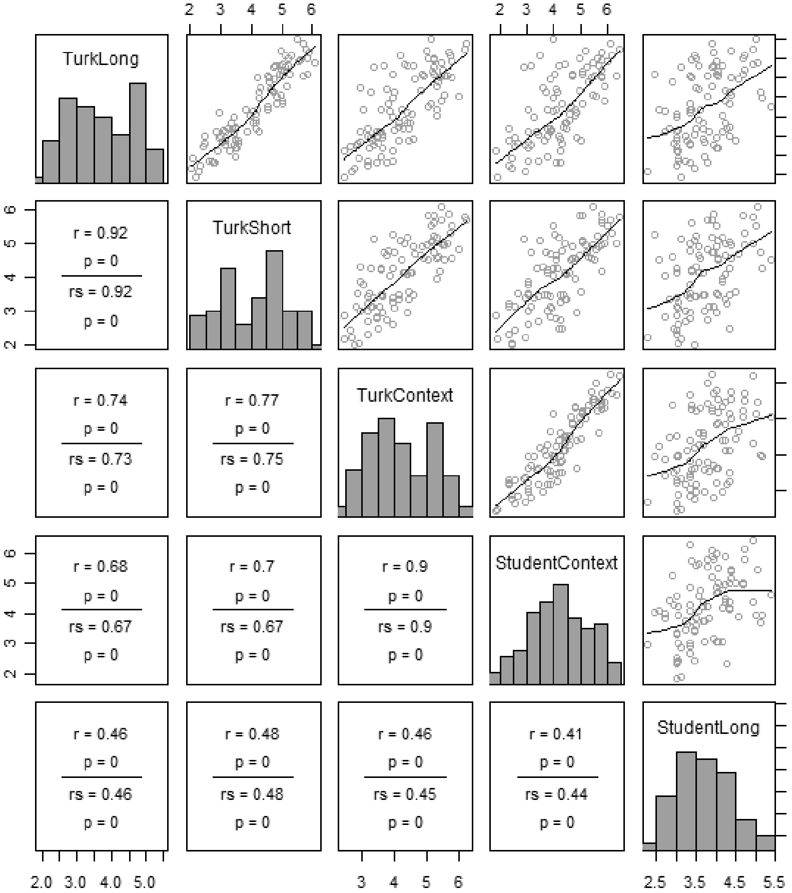


Figure 3. Panels at the diagonal report histograms of distributions of ratings across populations of participants in respective experiments; panels above the diagonal plot the locally weighted scatterplot smoothing lowess functions for a pair of correlated variables; panels below the diagonal report correlation coefficients (the “r” value is Pearson’s r, the “rs” value is Spearman’s rho) and respective p-values. In general, the correlations are quite high, with the exception of the StudentLong participants (who do not even correlate particularly well with the other participants in StudentContext).

We find a split into relatively high and low correlations. All Mechanical Turk tests correlate very well with one another (all rhos > 0.7), although the tasks and raters are different. The correlation between the student participants who were given sentence contexts and the Turkers who saw context is especially high (0.9).

All correlations with StudentLong are relatively low, but this is actually true for StudentLong vs. StudentContext, too (rho = 0.44), even though both groups are Stanford undergraduates and hence supposedly form a more homogeneous group than the Turkers . The correlations are low even if we restrict the Mechanical Turk groups to the same age and education level as the Student participants.

Moreover, the results suggest that results of StudentLong are the least internally consistent as well as the least externally consistent. A possible explanation for the pattern of results is that the StudentLong participants assigned similarity ratings within a much smaller mid-range of the 1-7 scale than participants in other experiments who used the entire scale. This is perhaps best seen in the per-experiment values of the coefficient of variation for the similarity scores defined as the ratio of the standard deviation and the mean. Column 4 in Table 5 demonstrated that the judgments for the StudentLong participants are the least spread out. This suggests that – if anything – that it is Stanford students that diverge in their performance from the consistent patterns in other participant populations.

Table 5. Distributional characteristics for similarity judgments across Experiments 2-6.

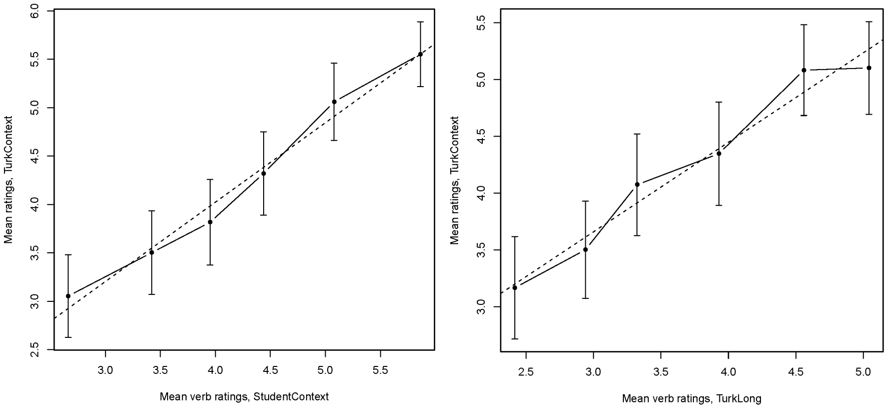
	Mean	Median	Coefficient of variation
TurkLong	3.702	3.649	0.252
TurkShort	4.0921	4.173	0.254
TurkContext	4.224	4.111	0.233
StudentContext	4.230	4.204	0.260
StudentLong	3.740	3.694	0.179

As a next step, we tested how similar the distributions of similarity ratings were across tasks. Because the data are not normally distributed, a non-parametric measure is called for—the two-sample Wilcoxon test. If the Wilcoxon tests show a significant difference in distributions, it indicates that participants regard the tasks differently and it is necessary to rethink whether student-based experiments and Mechanical Turk experiments are equivalent. However, the results actually confirm our hypothesis that we have a similar probability of getting rating *x* using either Mechanical Turk or our offline experiments. Table 6 demonstrates that StudentLong and TurkLong have similar distributional characteristics, as do StudentContext and TurkContext: the Wilcoxon tests showed the p-value above the 0.05 threshold. These results also suggest that providing context really does shift the distribution compared to just asking for judgments on context-less phrasal verbs—the ratings are higher and the correlation between Turkers and students is much stronger when context is present.

Table 6. P-values of the two-sample Wilcox tests show that there is no significant difference between StudentLong and TurkLong, nor between StudentContext and TurkContext.

	TurkLong	TurkShort	TurkContext	StudentContext	StudentLong
TurkLong					
TurkShort	p=2.797e-12				
TurkContext	p=2.606e-09	p=0.0443			
StudentContext	p=5.025e-08	p=0.101	p=0.869		
StudentLong	p=0.572	p=0.001	p=6.389e-06	p=4.063e-05	

The main question has been whether the two populations consistently describe opaque phrasal verbs as opaque and transparent phrasal verbs as transparent. Another way to look at this is through comparing mean similarity ratings across experiments. For example, in the first graph in Figure 4, phrasal verbs were binned into six groups of 16 each based on the increasing order of their mean StudentContext ratings. The mean rating for each bin in StudentContext was then compared to the mean TurkContext rating for the verbs in that bin. If transparency and opacity were understood the same way, then mean ratings for the bins should be very similar. Figure 4 shows four of the ten comparisons—all of the comparisons involving StudentContext and the Turk experiments demonstrate consistent ratings (see upper panels of Figure 4). The three Turk experiments also match very closely. Only pairs that involve StudentLong (bottom panels of Figure 4) deviate—including StudentLong vs. StudentContext and StudentLong vs. TurkLong. Again, we observe that the Turker population of data shows more consistency than the student-based population.



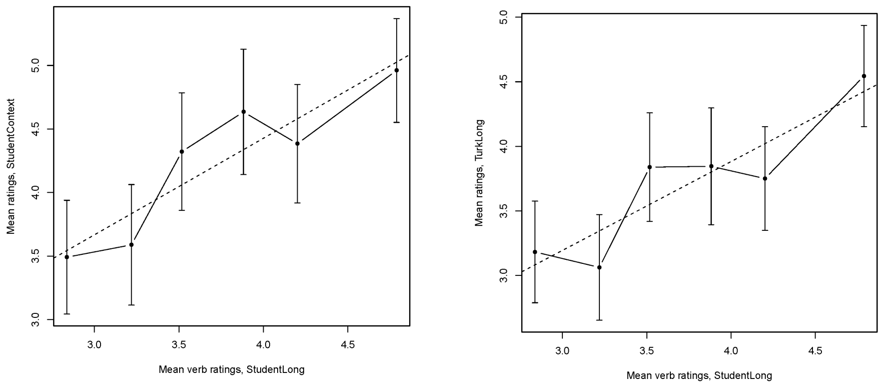


Figure 4. (top left) Mean ratings in TurkContext experiment per bins of StudentContext; (top right) Mean ratings in TurkContext experiment per bins of TurkLong; (bottom left) Mean ratings in StudentContext experiment per bins of StudentLong; and (top right) Mean ratings in TurkLong experiment per bins of StudentLong. Plots in bottom panels deviate from the straight line.

Finally, we examined the intra-class correlation coefficient (ICC), which tells us about the amount of agreement among participants. We find quite high agreement between the StudentContext and TurkContext raters, but we again find that the StudentLong raters are fairly different than all other groups. In fact, there is an exceedingly low amount of agreement *within* the StudentLong results. The ICC consistency among just the StudentLong participants is 0.0934, and the ICC agreement is just 0.0854.

Cohen's kappa measurement is another way to measure how well raters agree, weeding out chance agreements. In general, anything about a score of 0.2 has "fair agreement", while those above 0.4 have "moderate agreement". Those with kappa statistics above 0.6 have "substantial agreement" and those above 0.8 have "almost perfect agreement" (Landis & Koch, 1977, p. 165). We also find that the various Mechanical Turk experiments have very high agreement. But the highest agreement of all is between StudentContext and TurkContext (κ = 0.823). The inter-class correlation coefficient and Cohen's kappa are summarized in Table 7.

Table 7. The StudentLong population continues to be an outlier (although we still find “fair” agreement). Most of the other pairs of populations have substantial to almost perfect agreement.

	ICC consistency	ICC agreement	Weighted kappa
StudentLong/TurkLong	0.432	0.434	0.384
StudentLong/StudentContext	0.364	0.320	0.396
StudentLong/TurkShort	0.439	0.408	0.436
StudentLong/TurkContext	0.423	0.365	0.419
StudentContext/TurkContext	0.899	0.900	0.823
StudentContext/TurkShort	0.696	0.692	0.641
StudentContext/TurkLong	0.673	0.595	0.596
TurkContext/TurkLong	0.735	0.642	0.628
TurkContext/TurkShort	0.768	0.763	0.746
TurkShort/TurkLong	0.913	0.848	0.822

SUMMARY

In two complex and very common types of psycholinguistic tasks (the Cloze sentence completion task and semantic similarity judgments), we found that Amazon’s Mechanical Turk service provides data that are comparable in a number of parameters to the data obtained in the lab with the conventional pools of students. Importantly, Turkers demonstrated even better agreement between participants than student data, which is remarkable given a broader demographic, socio-economic and educational coverage offered by the crowdsourcing services like Mechanical Turk. We conclude that the use of Amazon’s Mechanical Turk allows us to move beyond testing only the language capabilities of university undergraduates, without compromising the quality of measurements and with a set of advantages that include high internal consistency, low price, and speed.

REFERENCES

Baldwin, T., & Villavicencio, A. (2002). Extracting the unextractable: a case study on verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning* (Vol. 20, pp. 1-7). Association for Computational Linguistics.

Bannard, C. (2002). Statistical techniques for automatically inferring the semantics of verb-particle constructions. LinGO Working Paper No. 2002-06.

Birnbaum, M. (2004). Human research and data collection via the Internet. *Annual Review of Psychology*, *55*(1), 803-832. doi:10.1146/annurev.psych.55.090902.141601

Bolinger, D. (1971). *The phrasal verb in English*. Cambridge, MA: Harvard University Press.

Boston, M. (n.d.). *DepParse 2.5*. Retrieved February 21, 2010, from http://conf.ling.cornell.edu/Marisa/papers.html

Boston, M., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1), 1-12.

Boston, M., Hale, J., Kliegl, R., & Vasishth, S. (2008). Surprising parser actions and reading difficulty. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers* (pp. 5–8).

Boston, M., Hale, J., Vasishth, S., & Kliegl, R. (in press). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*.

Buchanan, T. (1998). Internet research: Self-monitoring and judgements of attractiveness. Presented at the The 28th Annual Convention of the Society for Computers in Psychology, Dallas, TX.

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, *1084*(1), 89-103. doi:10.1016/j.brainres.2006.02.010

Dandurand, F., Shultz, T., & Onishi, K. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, *40*(2), 428-434. doi:10.3758/BRM.40.2.428

Engbert, R., Nuthmann, A., Richter, E., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777-813. doi:10.1037/0033-295X.112.4.777

Fraser, B. (1974). The phrasal verb in English by Dwight Bolinger. *Language*, *50*(3), 568. doi:10.2307/412224

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, *68*(1), 1-76. doi:10.1016/S0010-0277(98)00034-1

Gries, S. (2003). *Multifactorial analysis in corpus linguistics: A study of particle placement*. Continuum Intl Pub Group.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL*, *2*, 159–166.

Ipeirotis, P. (2008, March 19). Mechanical Turk: The demographics. *A Computer Scientist in a Business School*. Retrieved February 21, 2010, from http://behind-the-enemy-lines.blogspot.com/2008/03/mechanical-turk-demographics.html

Kennedy, A. (1920). The Modern English verb-adverb combination. *Language and Literature*, *1*(1), 1-51.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1-2), 262-284. doi:10.1080/09541440340000213

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, *135*(1), 12-35. doi:10.1037/0096-3445.135.1.12

Krantz, J., & Dalal, R. (2000). Validity of Web-based psychological research. *Psychological experiments on the Internet*, 35–60.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159. doi:10.2307/2529310

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126-1177. doi:10.1016/j.cognition.2007.05.006

Lohse, B., Hawkins, J., & Wasow, T. (2004). Domain minimization in English verb-particle constructions. *Language*, *80*(2), 238-261. doi:10.1353/lan.2004.0089

McCarthy, D., Keller, B., & Carroll, J. (2003). Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18* (p. 80).

McGraw, K., Tew, M., & Williams, J. (2000). The integrity of Web-delivered experiments: Can you trust the data? *Psychological Science*, *11*(6), 502-506.

Rayner, K., & Well, A. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, *3*(4), 504–509.

Reichle, E., Pollatsek, A., Fisher, D., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*, 125–157.

Reichle, E., Warren, T., & McConnell, K. (2009). Using E-Z Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*(1), 1-21. doi:10.3758/PBR.16.1.1

Reips, U. (2002a). Theory and techniques of conducting Web experiments. In B. Batinic, U. Reips, & M. Bosnjak (Eds.), *Online Social Sciences* (pp. 229-250). Seattle: Hogrefe & Huber.

Reips, U. (2002b). Standards for Internet-based experimenting. *Experimental Psychology (formerly "Zeitschrift für Experimentelle Psychologie")*, *49*(4), 243-256. doi:10.1026//1618-3169.49.4.243

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, *11*, 95-130.

Roland, D., Dick, F., & Elman, J. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, *57*(3), 348-379. doi:10.1016/j.jml.2007.03.002

Schilling, H., Rayner, K., & Chumbley, I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory and Cognition*, *26*, 1270-1281.

Senior, C., Phillips, M., Barnes, J., & David, A. (1999). An investigation in the perception of dominance from schematic faces: A study using the World-Wide Web. *Behavior Research Methods, Instruments, & Computers*, (31), 341-346.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. (2008). Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 254-263). Honolulu, Hawaii: Association for Computational Linguistics.

Taylor, W. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, (30), 415-433.

Tsuruoka, Y., & Tsujii, J. (2005). Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 467-474). Vancouver, British Columbia, Canada: Association for Computational Linguistics.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. Memory & Cognition, 18(4), 380–393.

APPENDIX: BEST PRACTICES

Taking advantage of Mechanical Turk

When Amazon's Mechanical Turk service was opened to the public in 2005, it was to help solve problems that computers *ought* to be able to solve but cannot—tagging photographs with keywords, for example, or recognizing duplicate web pages. Since computer programs cannot solve these programs (efficiently), they are crowd-sourced out to legions of people who are willing to quickly and cheaply perform the tasks in exchange for money. Our use of Mechanical Turk for linguistic tasks is different than the site's original purpose ("artificial artificial intelligence"), but suits the workers just fine—if your task looks interesting or lucrative, people will sign up to do it, whether you are asking them to tag a photo, write a review, tell you where you lost your keys, or give you a linguistic judgment.

Amazon does not publish demographic information, so the most thorough demographics have been collected by Ipeirotis (2008). We have not found any effects for demographics in our experiments, but the fact that such data can be collected fairly easily allows for experiments that do require broad demographic coverage.

Indeed, getting a population that is more diverse than those enrolled in lower division psychology/linguistics courses is one of the main advantages to web-based research and Mechanical Turk does offer a broader range of subjects than are available on most campuses (Birnbaum, 2004, p. 820). With the large numbers and low cost, it is also possible to see different slices of the population—when the numbers are large enough, you can compare groups to see if the general trend holds for each subgroup. These techniques should give psycholinguists more generalizability.

Pilots also give you the opportunity to see if there are any self-selection biases, although there are reports that these are usually unimportant in cognition (Reips, 2002a, p. 247). Still, as we ask more socially-oriented linguistic questions, self-selection biases is well-worth keeping in mind. Who is it that participates within Mechanical Turk and who among the Turkers sees your experiment and decides to participate?

Another typical concern with web-based experiments is multiple submissions. Reips (2002: 250) reports that this typically is not a problem, but the infrastructure for Mechanical Turk makes it even less likely. Each Turker is given a "worker ID"—but most do not seem to be aware of this. What that means is that you can filter out multiple submissions by the same person. The only thing you'll miss is if the Turker has registered two different accounts and signs out and back in to take your task while it is still posted. For a variety of reasons including how payment works, this seems unlikely. Nevertheless, there are further controls possible by collecting IP address information in the background.

Setting up the task

It isn't necessary to download anything to use Mechanical—you just need to register for an account at http://mturk.amazon.com. Once you sign in as a "requester", you should be able to get started right away. The Mechanical Turk service supports a number of different formats—you can use drop-downs, check boxes, text boxes and the like. It is also quite possible to use Turk to direct people to a web experiment that you host on your own servers. That is one way to give you even greater control, though for a great deal of tasks, the templates that Amazon provides should be adequate.

Rather than launch any experiment immediately, the best advice is to run a small pilot. Consider creating a very short version of the experiment and adding questions to the pilot about whether the participants understand it or not. One of the main advantages to web-based research is that it can take out (a) experimenter bias during interactions with the participants, and (b) it can bring the experiment to the subject rather than putting the subject in an unfamiliar environment. The flip side of this, however, is that no one is around to answer questions.

We also recommend that you keep tasks relatively short. This will not increase your cost and it will increase the chances of people paying full attention and completing the task. Drop-out rates are always a concern, given that they obscure true results. Birnbaum (2004, p. 816) gives an example of assessing an SAT prep class—imagine the class actually decreases performance. If people who do poorly in the "final practice test" are told not to take the SATs at all, then the results may suggest that the class is beneficial when it's not. It's a little harder to imagine how this is related to linguistic tasks, but imagine a study on complicated syntax.

The reporter who the senator who John met attacked disliked the editor (Gibson, 1998)

If participants are given sentences like (1), you could well imagine that some will just drop-out because it's too difficult. That may well change the distribution of judgments.

Scrubbing the data

Scrubbing the data is a crucial part of the pre-analysis work. It goes without saying that one should pre-define what will exclude a data point/participant before running the test so that one does not contaminate the results with cherry-picking. In this section, we discuss best practices for cleaning up Mechanical Turk data. These are the practices we followed with our own data, which took under half an hour to implement for.

- We find that it is useful to pay just about anyone, whether you're going to use their data or not. You do not want to create incentives to lie. If we had only been interested in men, we could have stated that, but since the experiment is paid and anonymous, such a request may incentivize people to lie. Instead, if you're collecting information on a subset population, give everyone the chance to participate and get paid and then narrow down to the subgroup of interest.
- Our definition of "English speaker" was quite strict—we asked for languages that they grew up speaking, and if they included any language other than English, we excluded their data. We also restricted ourselves to people who grew up in the US.
- We judged that participants who did not fill in all of the demographic information to be not taking the experiment seriously and their data was excluded.
- Sometimes it's quite clear what an "acceptable" answer is. In our semantic transparency test, we removed ratings that were outside of the 1-7 scale we asked for. Not only that, but we removed all of that Turker's other "within scale" responses, too.
- Recall that the Cloze task is to fill-in-the-blank with one word—responses with more than one word were thrown out.

- Using the "Worker ID" tag, we could keep track of Turkers who saw the same phrasal verb in different experiments (e.g., both TurkShort and TurkContext). We removed data from all but the first experiment they participated in.
- For most of the experiments reported here, Turkers only saw about 6-8 items at a time. That meant they could keep taking the test until they had seen all items (only once). Because of this structure, we collected redundant demographic information from these participants. Whenever the demographic information was contradictory, we eliminated all of the Turker's responses from consideration.
- Mechanical Turk offers the amount of time that participants spend on the task. For each Mechanical Turk experiment, we removed results that were more than two standard deviations from the mean of the experiment.[9]
- New scripts that have become available since our experiments make it possible to detect information "behind the scenes", such as location, operating system, browser language and the like. These can be used to restrict to people of interest—and potentially to test whether participants are answering questions honestly or not (that is, you can ask them something you have independent knowledge of).

Across all our experiments, we end up scrubbing away about 25% of the individual Turkers (about 37% of the data points). If that seems high recall that we paid $11.60 to get valid responses from 20 different people per sentence for 96 sentences—the cost of the people we had to scrub away was about $4.00.

[9] Additionally, we estimated that each question required at least 2 seconds to complete without context and 3 seconds to complete with context (the demographics section was estimated to take at least 5 seconds), so responses for the TurkLong test that took less than 2*96+5=197 seconds were also excluded as not having really performed the task (i.e., they were entering nonce values to complete the experiment and get paid).