A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory

Jon Sprouse
Department of Cognitive Sciences
University of California, Irvine

ABSTRACT

Amazon's Mechanical Turk (AMT) is a web application that provides instant access to thousands of potential participants for survey-based psychology experiments such as acceptability judgment tasks in syntactic theory. However, syntacticians worry that moving formal acceptability judgments out of the experimenter-controlled environment of the laboratory and onto the user-controlled environment of AMT may adversely affect the quality of the data collected. This article reports a quantitative comparison of two identical acceptability judgment experiments each with 176 participants (352 total): one conducted in the laboratory, and one conducted on AMT. Crucial indicators of data quality such as participant rejection rates, statistical power, and the shape of the distributions of judgments are compared between the two samples. The results suggest that aside from slightly higher participant rejection rates, AMT data is almost indistinguishable from laboratory data.

INTRODUCTION

From a purely methodological point of view, syntacticians are interested in identifying the properties of syntactic representations. Over the past 50 years, the dominant method for identifying the properties of syntactic representations has involved comparing two (or more) minimally different representations, and using a behavioral response known as an acceptability judgment to as a proxy for grammatical well-formedness (Chomsky, 1965; Schütze, 1996). Traditionally, these acceptability judgments have been collected using an informal experiment consisting of only a handful of participants (usually the researcher's colleagues), and a handful of experimental items (Marantz, 2005). This informal methodology has worked well because acceptability judgments of linguistic phenomena tend to be strikingly robust, even at very small sample sizes (for a large-scale quantitative evaluation, see Sprouse & Almeida, *submitted*). The success of informal experiments not withstanding, over the past 15 years, a number of syntacticians have argued that formal experimental methods – such as full-scale surveys, large samples, and sophisticated scaling tasks like magnitude estimation – can provide an additional level of detail (usually in the form of statistical models) that can help clarify some theoretical questions in syntactic theory (e.g., Bard, Robertson, & Sorace, 1996; Cowart, 1997; Keller, 2000; Sorace & Keller, 2004; Featherston, 2005a, 2005b; Myers, 2009; Sprouse, 2009; Sprouse & Cunningham 2010; Sprouse, Wagers, & Phillips, *submitted*). Of course, the additional information gained by formal acceptability experiments is offset by the fact that they take considerably more time to deploy than informal acceptability experiments: an informal experiment can be conducted in a matter of minutes, whereas a formal experiments can require several weeks to recruit and run a full sample (e.g., 25-30 participants).

Several free software solutions have been developed to allow acceptability judgments to be collected over the web, and thus reduce some of the collection time, such as *WebExp* (Keller, Gunasekharan, Mayo, & Corley, 2009) and *MiniJudge* (Myers, 2009). Though successful at reducing physical data collection time, these software solutions still require the experimenter to invest time in participant recruitment (and compensation disbursement), which can still take weeks to complete. It has been recently suggested that syntacticians could use the Amazon Mechanical Turk marketplace (henceforth AMT) to completely automate the recruitment of participants, the administration of surveys, and the disbursement of compensation, thus virtually eliminating the time cost of formal experiments (e.g., Gibson & Fedorenko, *in press*). AMT is an online marketplace where companies or individuals (called *requesters*) can post small tasks (called *Human Intelligence Tasks*, or *HITs*) that cannot easily be automated, and therefore require human workers (called *workers*) for completion. These HITs are generally very small in nature (such as identifying the contents of an image), and generally very high in quantity (it is not unusual for requesters to post thousands of tasks in a single batch). Requesters generally pay very little per HIT (e.g., $.02 US), and retain the ability to accept or reject the results of each HIT before Amazon sends payment to the worker. In this way, requesters are able to *crowdsource* (cf. outsource) tasks that would previously have required hours of work by in-house employees at considerably more expensive compensation rates. HITs can be posted using an online interface (www.mturk.com), and results can be downloaded in CSV format. From the point of view of an experimenter, AMT provides instantaneous access to thousands of potential participants, and provides the tools necessary to distribute surveys, collect responses, and disburse payments.

It should be noted that AMT has already proven useful in at least one area of language research, computational linguistics, where it has been used for corpus annotation and evaluation – two tasks that have historically consumed significant time and resources (e.g., see the recent NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk; proceedings available online: http://www.aclweb.org/anthology/W/W10/W10-07.pdf). However, AMT has yet to be widely adopted by syntacticians who run formal acceptability experiments. The primary concern among syntacticians is that moving formal acceptability judgments out of the experimenter-controlled environment of the laboratory and onto the user-controlled environment of AMT may adversely affect the quality of the data collected, and potentially negate the quantitative advantages that motivate formal experiments in the first place. In the laboratory, the experimenter can ensure that all participants are part of the population of interest (such as native speakers of US English); the experimenter can control the environmental distractions; the experimenter can influence the rate of completion ("don't rush"); the experimenter can verify that participants understand the task; and the experimenter can answer any questions that may arise. Before syntacticians can widely adopt AMT, they will need to be reasonably sure that the loss of this control will not affect the quality of the data that is collected. To that end, the goal of this paper is to compare the results of a large-scale laboratory-based experiment (176 participants) with an identical AMT-based experiment (176 participants) along all of the quantitative measures of interest to linguists: time, cost (in money), participant rejection rate, the detection rate of several known effects (both strong and weak) at a range of sample sizes, and the differences in the shape of the distribution of ratings for each condition (peak, dispersion, etc).

EXPERIMENTAL DETAILS

Quantitative validation studies such as this require two large data sets: a reference data set and a target (AMT) data set. Given the relative scarcity of funding in linguistics, it seems unlikely that syntacticians will devote their limited resources to collecting two large data sets simply to validate AMT. However, Sprouse et al. (*submitted*) collected a large data set as part of a theoretically motivated study: 176 participants, 24 different sentence types, 16 different lexicalizations (tokens) of each sentence type, and 4 judgments per sentence type per subject. It is this data set that serves as the reference data for the AMT validation. The details of the experiment are given in the rest of this section.

Method

Participants. 176 (152 Female) self-reported monolingual native speakers of English, all University of California Irvine undergraduates, participated in the laboratory experiment for either course credit or $5. 176 (102 Female) unique AMT workers participated in the AMT experiment for $3.

Materials. There were 24 total sentence types (conditions) tested in this experiment. 16 lexicalizations of each sentence type were created, and distributed among 4 lists using a Latin Square procedure. This meant that each list consisted of 4 tokens per sentence type, for a total of 96 items per list. Two orders for each of the 4 lists were created by pseudorandomizing the items such that related sentence types were never presented successively. This resulted in 8 different surveys.

Procedure. The task for both samples was magnitude estimation of acceptability (Bard et al., 1996; Keller, 2000; Featherston, 2005a; Sprouse & Cunningham, 2010). In a magnitude estimation task, participants are asked to rate experimental items in proportion to a reference item (the *standard*). The standard is pre-assigned a numerical value (the *modulus*). In the example below, the standard has been assigned a modulus of 100. If the participant believes that an experimental item is twice as acceptable as the standard, then she would assign it a value of 200. If the participant believes that an experimental item is half as acceptable as the standard, then she would assign it a value of 50.

(1) An example of magnitude estimation of acceptability

Standard:	Who thinks that my brother was kept tabs on by the FBI?	100
Item:	What did Lisa meet the man that bought?	____

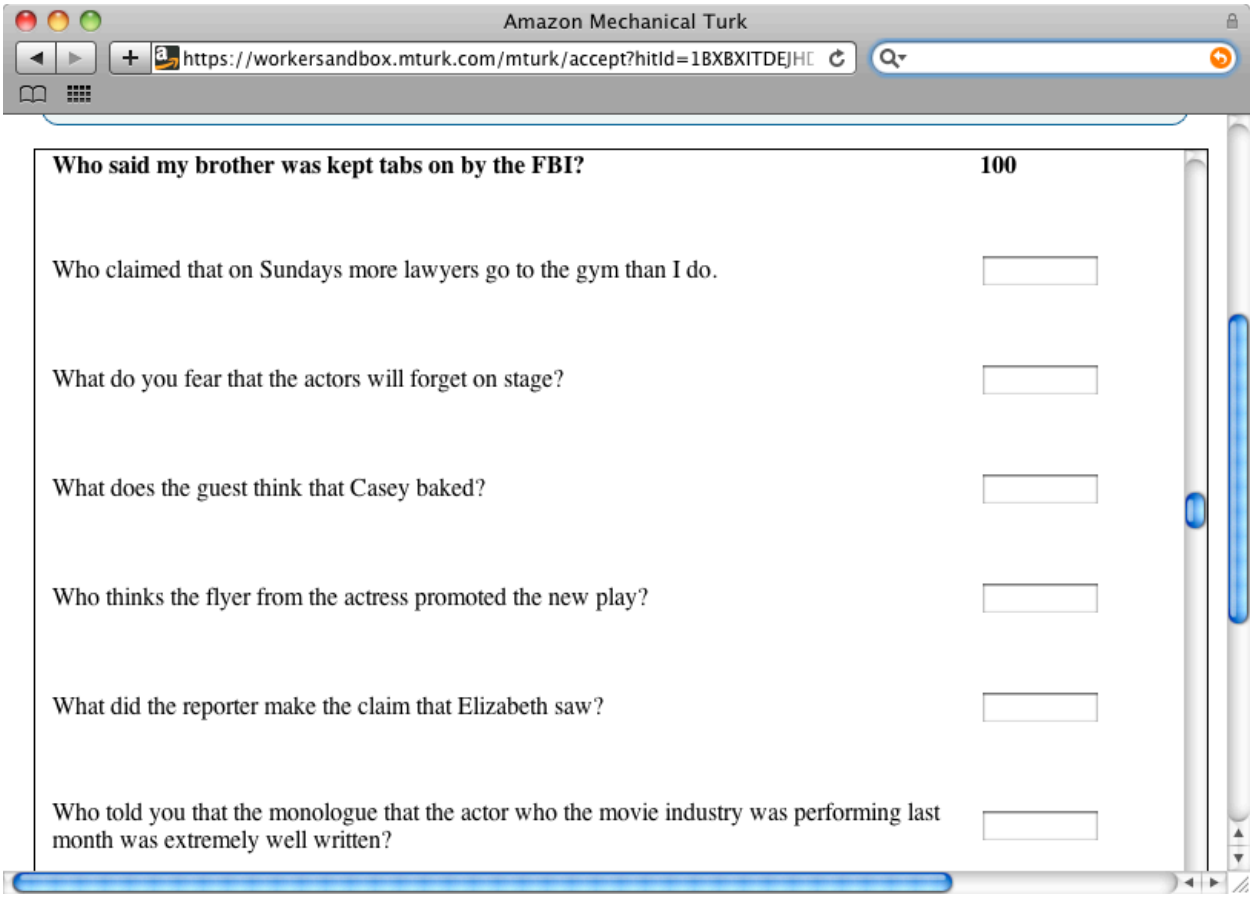
The standard and modulus do not change throughout the experiment. Participants are instructed that they can use any positive number that they feel is appropriate. The standard was identical for all 8 surveys, and was in the middle range of acceptability: *Who said my brother was kept tabs on by the FBI?* The standard was assigned a modulus of 100.

Presentation in the laboratory. The experiment began with a practice phase during which participants estimated the lengths of 7 lines using another line as a standard set to a modulus of 100. This practice phase ensured that participants understood the concept of

magnitude estimation. During the main phase of the experiment, 10 items were presented per page (except for the final page), with the standard appearing at the top of every page inside a textbox with black borders. The first 9 items of the survey were practice items (3 each of low, medium, and high acceptability). These practice items were not marked as such, i.e., the participants did not know they were practice items, and they did not vary between participants in order or lexicalization. Including the practice items, each survey was 105 items long. The task directions are available on the author's website. Participants were under no time constraints during their visit.

Presentation on AMT. The primary difference between the laboratory and AMT presentation was that the AMT survey appeared as a webpage rather than a paper survey. There were no page delineations in the webpage, therefore all of the items appeared as one long page (600px in height) that required that participant to scroll. The standard and modulus were repeated **in bold** every 7 items to ensure that it was always visible on the page. The HTML for the AMT presentation is available on the author's website (www.ling.cogsci.uci.edu/~jsprouse/tools/amt/). All other details were identical.

Figure 1: A screenshot of the magnitude estimation task as it appears on AMT



Preprocessing of responses. The responses to the 9 practice items were removed, and the remaining responses for each participant were z-score transformed prior to analysis. The z-score transformation is a standardization procedure that corrects for some kinds of scale bias between participants by converting a participant's scores into units that convey the number of standard deviations each score is from that participant's mean score.

Case Studies for Analysis

Fourteen of the twenty-four sentence types will be analyzed in this comparison. These 14 sentence types can be paired (one experimental condition and one control) to form 7 theoretically relevant phenomena from the syntactic and sentence-processing literature. The first four phenomena are called *island effects* (Ross, 1967; Huang, 1982; Chomsky, 1986). Island effects are ideal as case studies for AMT as they have many of the properties of other syntactic phenomena: they are discussed in dozens of articles and textbooks, the source of the unacceptability is generally too abstract for naïve participants to identify or correct, and they have been reported to demonstrate a good deal of variability among native speakers (Kuno, 1973; Grimshaw, 1986; Hofmeister and Sag, 2010).

(2) Whether island effect

What do you think that John bought?	(control)
*What do you wonder whether John bought	(violation)

(3) Complex Noun Phrase island effect

What did you claim that John bought?	(control)
*What did you make the claim that John bought?	(violation)

(4) Subject island effect

What do you think interrupted the TV show?	(control)
*What do you think the speech about interrupted the TV show?	(violation)

(5) Adjunct island effect

What do you think that John forgot at the office?	(control)
*What do you worry if John forgets at the office?	(violation)

The next three case studies are contrasts that have historically proven particularly difficult to replicate in acceptability judgment tasks, but are nonetheless detectable with very large sample sizes like those used in this study (Sprouse and Almeida *submitted*). They are the center-embedding illusion (Frazier, 1985; Gibson & Thomas, 1999), the comparative illusion (Phillips, Wagers, & Lau, *in press*) and the agreement attraction illusion (Wagers, Lau, & Phillips, 2009),. The fact that these contrasts are difficult to detect with acceptability judgments is likely because they are not caused by a static property of the syntactic representations, but rather by the way the sentences are processed. Such processing-based effects are generally investigated using measures with high temporal resolution such as reaction times or event-related potentials rather than untimed acceptability judgments; however, these three contrasts have been reported using untimed acceptability judgments, and therefore provide an interesting case study in detection of extremely weak effects using an AMT sample.

(6) Center Embedding Illusion
*The ancient manuscript that the grad student who the new card catalog had confused a great deal was studying in the library was missing a page. (violation)
? The ancient manuscript that the grad student who the new card catalog had confused a great deal was missing a page. (illusion)

(7) Comparative Illusion
*More people have graduated law school than I have. (violation)
?More people have been to Russia than I have. (illusion)

(8) Agreement Attraction Illusion
*The slogan on the poster unsurprisingly were designed to get attention. (violation)
?The slogan on the posters unsurprisingly were designed to get attention. (illusion)

TIME, COST, AND PARTICIPANT REJECTION

There are many aspects of the experimental procedure that could be affected by the change of venue from the laboratory to AMT, such as the time it takes to create and run the experiment, the methods available for ensuring an appropriate sample (e.g., only native speakers of English), and the number of participants that must be removed from the sample prior to analysis. This section provides an in-depth comparison of these pre-analysis aspects of the experimental procedure.

Time

Preparation. Laboratory experiments require the use of experimental software (e.g., *WebExp*, *MiniJudge*) or the creation of paper surveys; AMT experiments require the creation of an HTML survey. It took about 3 hours to explore the AMT documentation (tutorials and discussion threads), and another hour to create the HTML template for the surveys, for a total of 4 hours of initial setup time, which seems comparable to the initial setup of other software options. This is a one-time investment, and the HTML template is reusable, therefore additional experiments will take only a matter of minutes to publish. The HTML template used here can be downloaded for free on the author's website (www.ling.cogsci.uci.edu/~jsprouse/tools/amt/).

Data collection. The primary advantage of AMT is in data collection. The laboratory-based sample took approximately 88 experimenter-hours spread over a 3-month period, whereas AMT returned 170 surveys in 2 hours. That is a rate of 85 participants per hour. Because a few of the participants were excluded during data collection (see *Participant rejection rates*), the total time to collect 176 correctly completed surveys was 4 hours. These rates suggest that a standard sized sample (25-35 participants) could be collected in less than 1 hour using AMT.

Cost

The laboratory-based participants were paid $5 or given course credit for a 30-minute visit to the laboratory. The AMT participants were paid $3 per survey. The $3 compensation rate was chosen based on the other HITs available on AMT: HITs generally pay $.02 per single task, and these surveys required 105 judgments in addition to the reading of detailed instructions. AMT charges a 10% fee in addition to the compensation given to workers, so the total participant compensation cost was $3.30 per participant ($580.80 for 176 participants). The participant compensation cost of AMT is likely to be a concern for linguists without funding.

Whereas laboratory-based experiments can be run at no cost through the use of university subject pools that grant course credit, the AMT system is cash only. At these rates, a standard 30 participant/100 item experiment on AMT will cost approximately $100.

Participant Rejection

Selection. Participant selection criteria will obviously vary from experiment to experiment; however, there are at least two criteria that every experiment will include that can be used as case studies to understand the dynamics of participant selection on AMT:

1. Participants must be native speakers of the language of interest (e.g., US English)
2. Participants must take the experiment only once.

The AMT documentation indicates that requesters can require that workers complete a qualification exam prior to completing HITs. These qualification exams are intended to assess the worker's skill at the particular task. It is theoretically possible to create a qualification exam that will screen out non-native speakers and participants that have already completed a survey. However, workers can re-take qualification exams. This means that a worker who is disqualified for being a non-native speaker can potentially re-take the exam and change her answer. This situation is not ideal, as it potentially encourages misrepresentation. Furthermore, several discussion threads on the AMT forum suggest that qualifications severely decrease participation rates, as many AMT workers routinely ignore HITs that require qualification.

Given the re-take possibility of the qualification exams, it seems that the only option for participant selection is to rely on the truthfulness of the participants and post-collection participant rejection criteria. To that end, the description of the experiment said "You must be a native speaker of US English to participate in this experiment." This description is visible to workers while they are browsing the list of available HITs. Similarly, the first paragraph of the survey instructions explained that this HIT is actually an experiment, and that only native speakers of US English should take it because non-native speakers could contaminate the data. Participants were then told that a native speaker of US English meets the following two criteria, and were asked to choose YES or NO using radio buttons for each criterion:

1. You lived in the United States from birth until age 13.
2. Both of your parents spoke English to you during those years.

Participants were paid $3 regardless of their answers to these criteria. This ensured that there was no incentive to answer untruthfully, and that the responses could be used to reject participants prior to analysis. Only 3 participants answered NO to one or more of the native speaker criteria. These 3 participants were still compensated for their time, so there was a $9.90 loss to self-identified non-native speakers.

To ensure that participants only completed one of the eight surveys that were part of this experiment, a paragraph was placed at the end of the survey (after all of the judgments) that instructed workers not to take any of the seven other HITs available as part of this HIT batch. They were told that they would only be paid for the first survey that they completed, therefore there is no monetary incentive to complete additional HITs in this batch. Because AMT assigns each worker a unique alphanumeric ID number, it is relatively straightforward to search the results for workers that have completed multiple surveys, and reject their later surveys using the

AMT approval/rejection feature. If a worker is rejected through the approval/rejection feature, she is not compensated for that HIT, and that HIT is automatically returned to the list of available HITs to be completed by a different worker. The approval/rejection feature thus ensures that there is no monetary incentive to workers to take more than one survey in a single experiment. One participant submitted three surveys. Only the first was approved; the other two were rejected and returned to the AMT system for completion by other participants.

False submission. Because laboratory experiments are conducted in person, there are generally no false submissions. There can be participants that fail to show for a scheduled appointment, but at many universities there are penalties to dissuade no-shows. On the AMT system, there are no such penalties. Seven participants submitted incomplete surveys. These subjects were rejected using the AMT rejection/approval system, which means that they were not compensated for their time, and their surveys were automatically returned to the AMT system to be taken by other participants. Together with the two repeated surveys mentioned in the previous subsection, this means that 9 out of 176 surveys were rejected using the AMT rejection/approval system and returned to the AMT system (5.1%). Identifying theses 9 surveys took less than 10 minutes of work with no monetary loss.

Rejection. Because acceptability judgments are by definition subjective (there is no external measurement method), there are no universally agreed upon criteria for identifying participants who are not performing the task correctly. One possibility explored by Sprouse & Cunningham (2010) is to plot the mean ratings of each condition in ascending order, and identify a subset of conditions that appear to have a definitive rank order in the sample mean data. The rank order of those items can then be computed for each participant, and compared to the rank order of those conditions in the sample mean data (the "true" ordering) to derive a measure of divergence between each participant's rank order and the sample rank order, such as the tau rank correlation (Kendall, 1938). The tau rank correlation yields a coefficient for each participant between -1 and 1. A perfect match between the two ranks yields a 1, no relation between two ranks yields a 0, and the most dissimilar rank yields a -1.[1] The tau rank correlation coefficients can then be plotted in a histogram to identify any participants whose rank order is qualitatively different from the sample rank order. Crucially, for the purposes of this paper, this procedure does not have to be the best possible outlier identification procedure; it merely has to return results that (i) are logically interpretable, and (ii) allow for a comparison to be made between the two samples.

To derive a baseline rank order for comparison, 8 conditions were chosen that appear to have a reliable set of ordering relations based on the mean ratings of all participants in both samples. In ascending order, these were: (a) Adjunct island violations, (b) Whether island violations, (c) agreement attraction violations, (d) agreement attraction illusions (e) matrix wh-questions with embedded adjunct clauses, (f) long distance wh-questions with embedded that-clauses, (g) matrix wh-questions with embedded complex NPs, and (h) matrix wh-questions with embedded that-clauses.

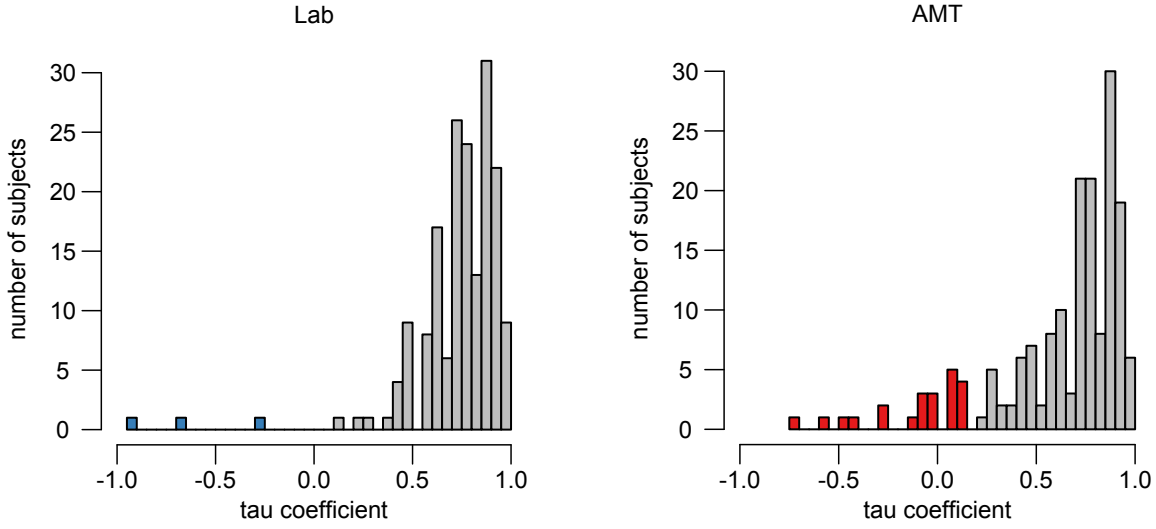
[1] Kendall's tau, which is used in the derivation of the correlation coefficient, is a distance measure between two rank orders based on how many pairwise "flips" of adjacent numbers are necessary to turn one rank order into another.

(2) Examples of the 8 conditions chosen for the rank order analysis
- a. What do you worry if the lawyer forgets at the office?
- b. What does the detective wonder whether Paul took?
- c. The slogan on the poster unsurprisingly were designed to get attention.
- d. The slogan on the posters unsurprisingly were designed to get attention
- e. Who worries if the lawyer forgets his briefcase at the office?
- f. What does the detective think Paul took?
- g. Who made the claim that Amy stole the Pizza?
- h. Who thinks Paul took the necklace?

The R statistical computing environment (R Development Core Team, 2009) was used to compute the order of those 8 conditions for each subject and compare each participant's order to the baseline. The tau correlation coefficients for each sample are presented in figure 2:

Figure 2: Histogram of tau correlation coefficients for each sample

A tau of one indicates perfect agreement between the participant's rank order and the sample rank order; a tau of zero indicates no relationship between the two; and a tau of negative one indicates a perfect reversal of the sample rank order. Participants that were removed from the laboratory sample are colored in blue; participants that were removed from the AMT sample are colored in red.



The tau coefficients for the laboratory sample are much more tightly clustered at the high end of the scale than the AMT sample, which has a much heavier leftward tail. At a practical level, this means that it is much easier to identify outliers in the laboratory sample: the three participants with tau coefficients below 0 are obviously distinct from the primary mass of participants. Furthermore, their negative tau coefficients indicate that their rank order was nearly reverse from the sample rank order. The picture is less clear for the AMT sample. It is still the case that a large majority of the participants have tau coefficients above .5, but there are many more subjects with tau coefficients near or below 0, and there is a less clear separation between the primary mass of participants and the potential outliers. Adopting a cut off criterion that is similar to the laboratory sample (~.15) results in the elimination of 22 participants from the AMT sample and coincides

with a minor mode in the tail of the distribution. However, the fact that this criterion is difficult to establish without a comparison to the laboratory sample raises a potential problem for the use of this method of participant removal with AMT samples; however, for the purposes of this validation study, it provides us with a conservative estimate that is logically comparable to the laboratory sample.

In total, 25 out of 176 participants (14.2%) were excluded from the AMT sample for either self-identifying as non-native (3), or providing results in which the rank order differed significantly from the sample rank order (22). Although the AMT rejection rate appears to compare unfavorably with the 3 rejections for the laboratory sample (1.7%), it should be noted that 14.2% is well within the range of rejection rates for other behavioral methodologies such as self-paced reading and lexical decision, and lower than the rejection rates for electrophysiological methodologies such as EEG and MEG. The minor increase in participant rejections in the AMT sample seems to be more than offset by the 90:1 time advantage. In order to adjust for this slightly higher rejection rate, syntacticians may want to consider adding 15% to the target sample size (e.g., 35 instead of 30). The statistical analyses presented in the following sections were performed on the remaining 173 participants in the laboratory sample, and the remaining 151 participants in the AMT sample.

STATISTICAL POWER

The primary concern of syntacticians is that the noise introduced by the uncontrolled environment of AMT may lead to lower statistical power than traditional laboratory-based experiments. To investigate this concern empirically, resampling simulations were run on each of the phenomena presented in section 2. These resampling simulations were designed to estimate the rate of statistical detectability for each phenomenon for every sample size between 5 and 173 for the laboratory sample, and between 5 and 151 for the AMT sample. In other words, these resampling simulations provide an answer to the questions: How likely am I to detect phenomenon X with a sample size of Y in the laboratory versus AMT?

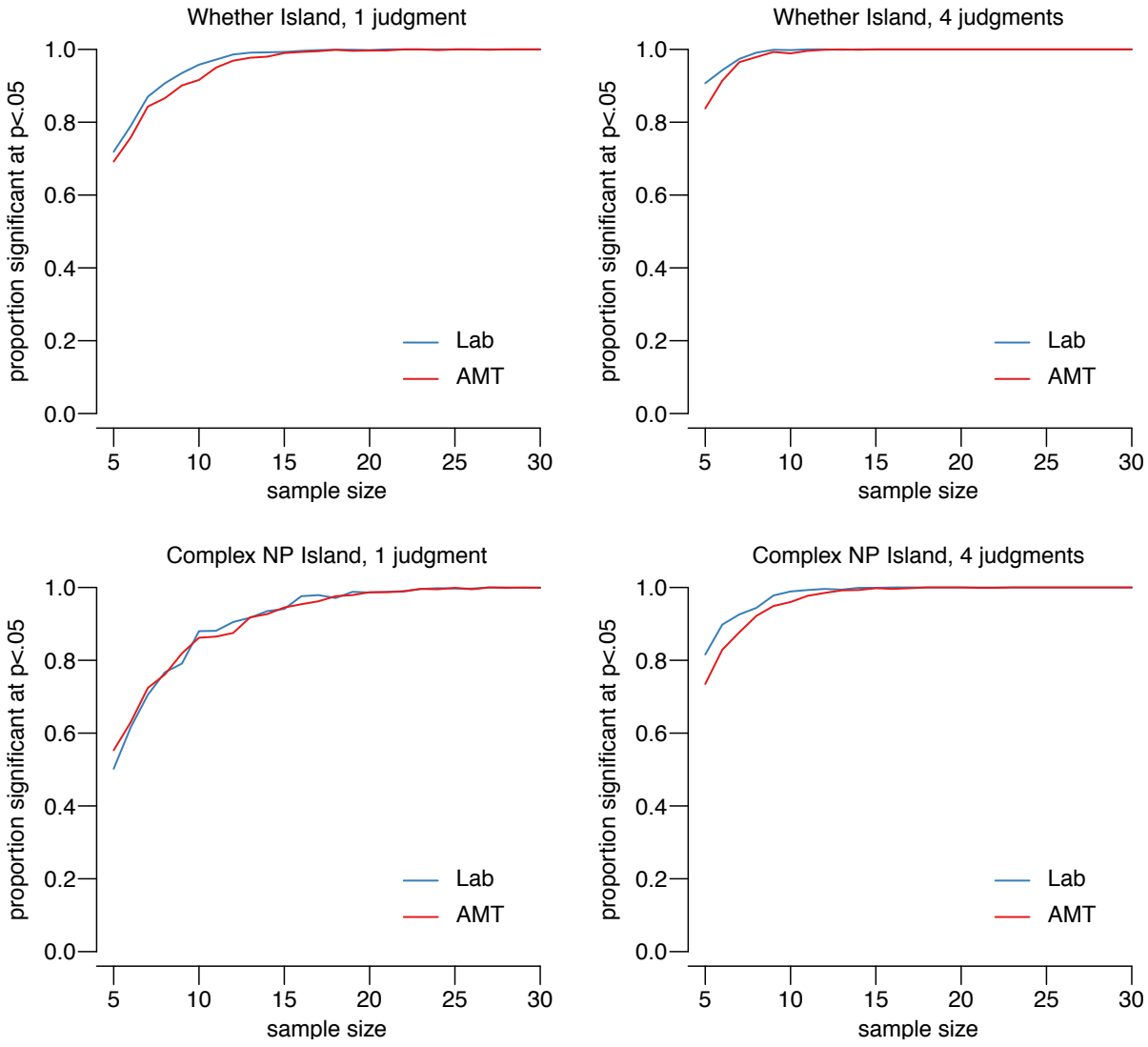
The algorithm for the resampling simulations can be described as follows (see Sprouse & Almeida, *submitted* for more details):

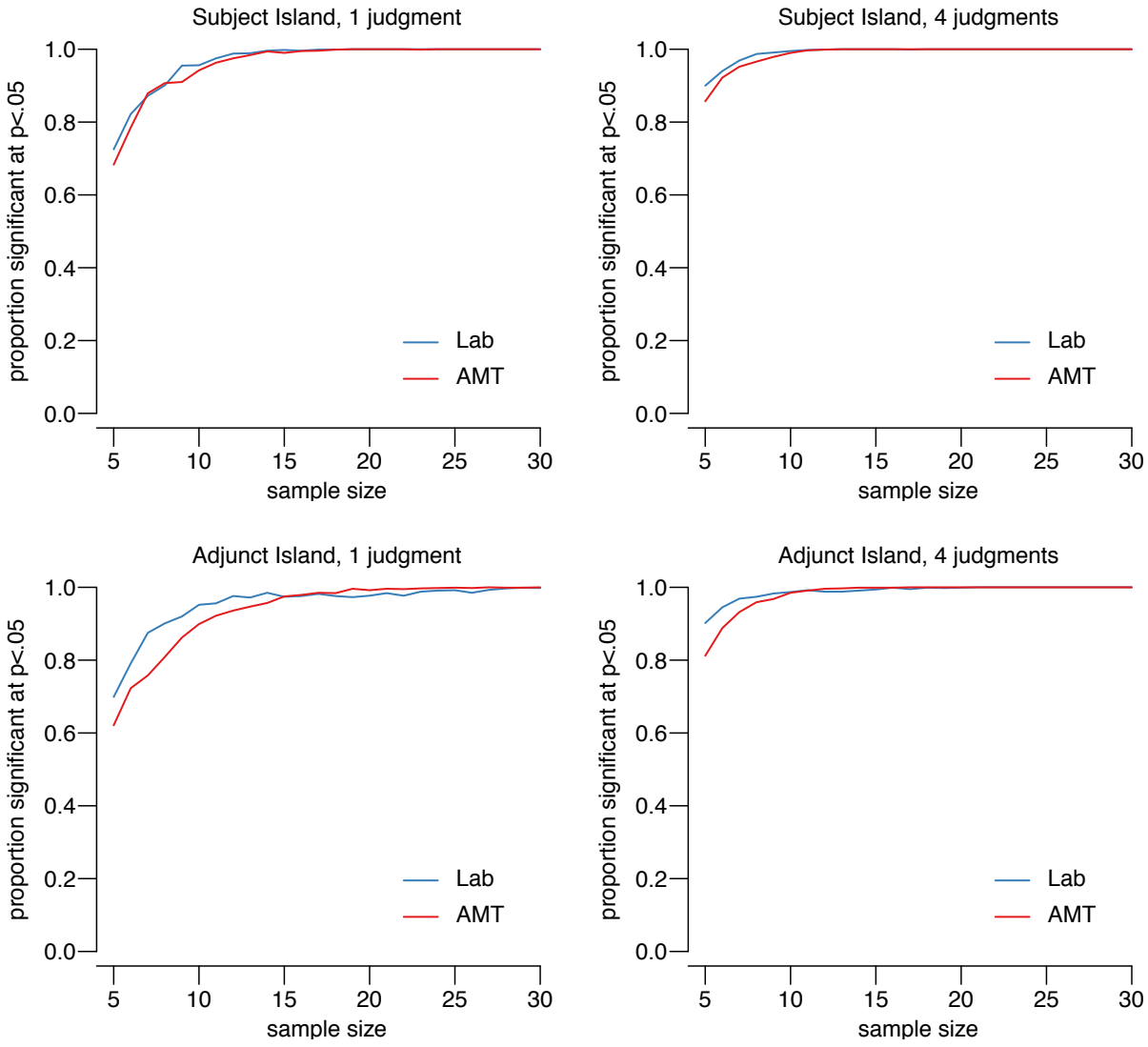
1. Choose one of the two samples (laboratory or AMT)
2. Choose a sample size (e.g., 5)
3. Randomly sample (with replacement) a number of participants equal to that size (e.g., 5 subjects) from the full data set
4. Randomly choose 1 judgment for each condition from each of the participants in the sample
5. Run a paired *t*-test on the sample
6. Repeat steps 3-5 1000 times
7. Calculate the proportion of significant results ($p<.05$) out of those 1000 samples; this is an estimate of the detection rate at that sample size.
8. Repeat steps 2-7 for all of the other possible sample sizes (5-173 for the laboratory sample, 5-151 for the AMT sample)
9. Repeat steps 2-8 for every possible number of judgments per participant per condition (in this case, 1-4)
10. Repeat steps 2-9 for the other sample (laboratory or AMT)

It should be noted that sample sizes below 5 were not tested because paired *t*-tests are not necessarily computable for sample sizes smaller than 5. Only graphs for 1 judgment per participant per condition and 4 judgments per participant per condition are presented here, as these were the upper and lower bounds made possible by the design of the experiment. Because all the island effects tested asymptote at 100% detectability at relatively small samples, figure 3 only presents the detectability estimates for sample sizes up to 30.

Figure 3: A comparison of the estimated detectability rate of island effects

The x-axis represents every possible sample size for the laboratory (5-173) and AMT (5-151) samples. The y-axis represents the proportion of random samples at that size that returned a significant t*-test result (*p*<.05). The blue line represents the detectability rate for the laboratory sample. The red line represents the detectability rate for the AMT sample.*

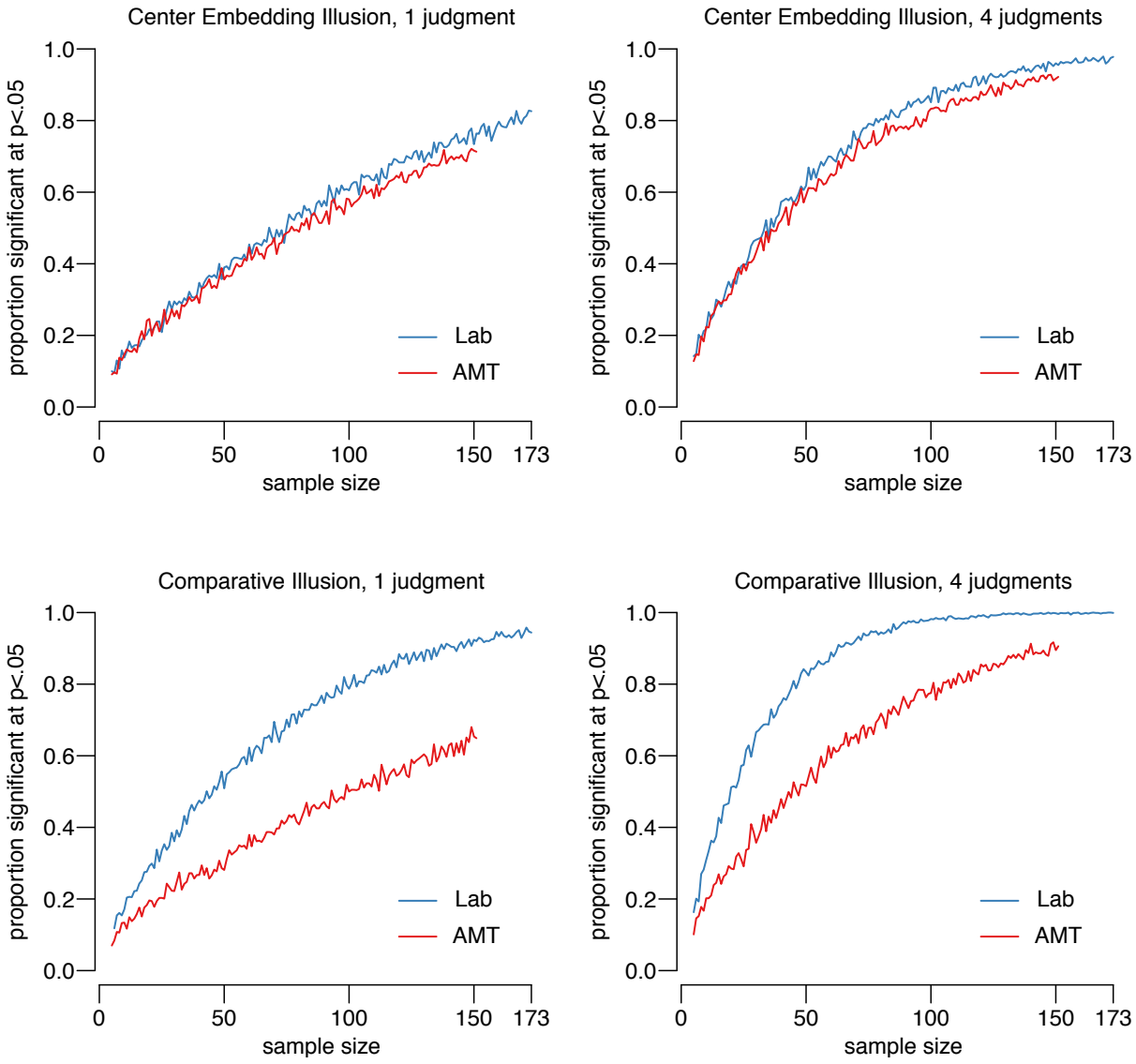


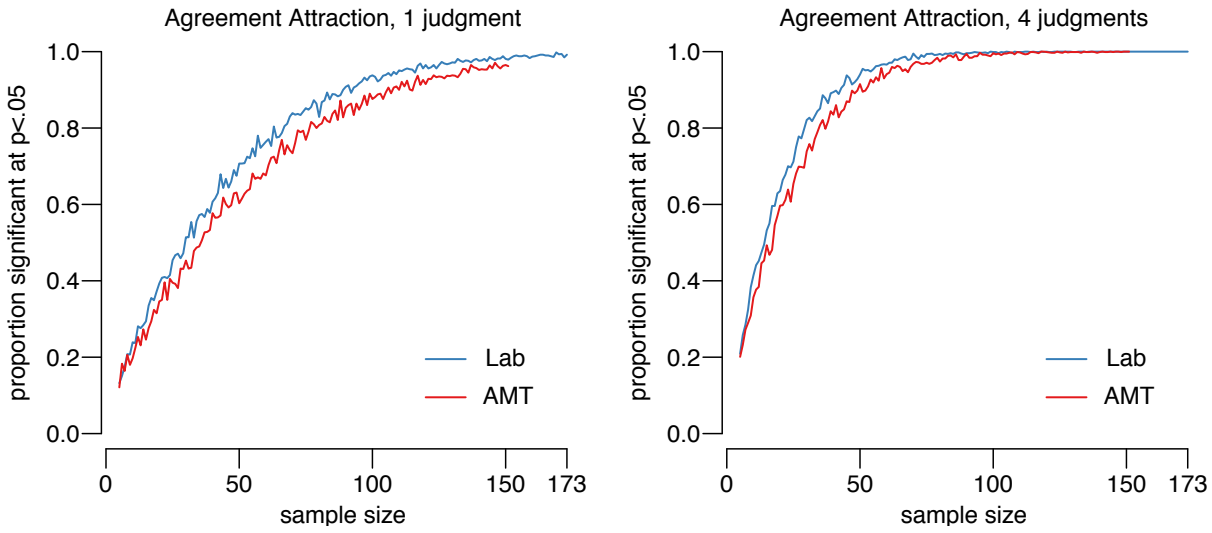


There does appear to be a slight loss of statistical power in the AMT sample: by and large, the AMT sample requires 3 or 4 more participants than the laboratory sample to reach 100% detectability. This suggests that any concern that syntacticians may have about AMT can be alleviated by slightly increasing the sample size. It should also be noted that both the laboratory sample and the AMT sample reached 100% detectability with fewer than 20 participants in the relatively underpowered 1 judgment analysis. Given that the standard sample size in formal acceptability judgments is 25-30, and that it is standard to give each participant more than one judgment per condition, syntacticians may not even notice the slight power loss. In short, these results suggest that AMT is well suited to detect standard syntactic phenomena without any noticeable loss in statistical power.

Figure 4: A comparison of the estimated detectability rate of extremely weak effects

The x-axis represents every possible sample size for the laboratory (5-173) and AMT (5-151) samples. The y-axis represents the proportion of random samples at that size that returned a significant t*-test result (*p<*.05). The blue line represents the detectability rate for the laboratory sample. The red line represents the detectability rate for the AMT sample.*





The three weak phenomena presented in figure 4 have historically been difficult to detect with standard acceptability judgment experiments, likely because they are not caused by static properties of the final syntactic representation, but rather dynamic properties of the way these sentences are processed. Nonetheless, these effects are detectable with extremely large samples, as demonstrated in figure 4. This makes them an ideal test case for the ability to detect extremely weak effects using AMT.

For the Center Embedding and Agreement Attraction effects, the AMT sample once again appears to yield slightly lower detectability rates than the laboratory sample: the AMT sample requires 10 additional participants to reach detectability rates that are comparable to the laboratory sample. This does not appear to pose a significant problem for the use AMT given the ease with which an additional 10 participants can be recruited. However, the Comparative Illusion detection rate in the AMT sample is potential cause for concern: the AMT sample appears to require 50 additional participants to reach detectability rates that are comparable to the laboratory sample. Given that two of the three extremely weak effects were detected within the AMT sample at rates comparable to the laboratory sample, it seems likely that the lower detection rate for comparative illusions says more about comparative illusions than it does about the use of AMT. In fact, as we shall see in the next section, the distributions of the comparative illusion data suggest that the fewer AMT participants were fooled by the illusion, which suggests that the lower detectability of the effect in the AMT sample is actually indicative of more accurate judging by the AMT participants. Taken together with the fact that none of these effects are well suited to investigation using (non-speeded) acceptability judgments in the first place, these results strongly suggest that syntacticians need not worry about the statistical power of AMT samples for syntactic phenomena.

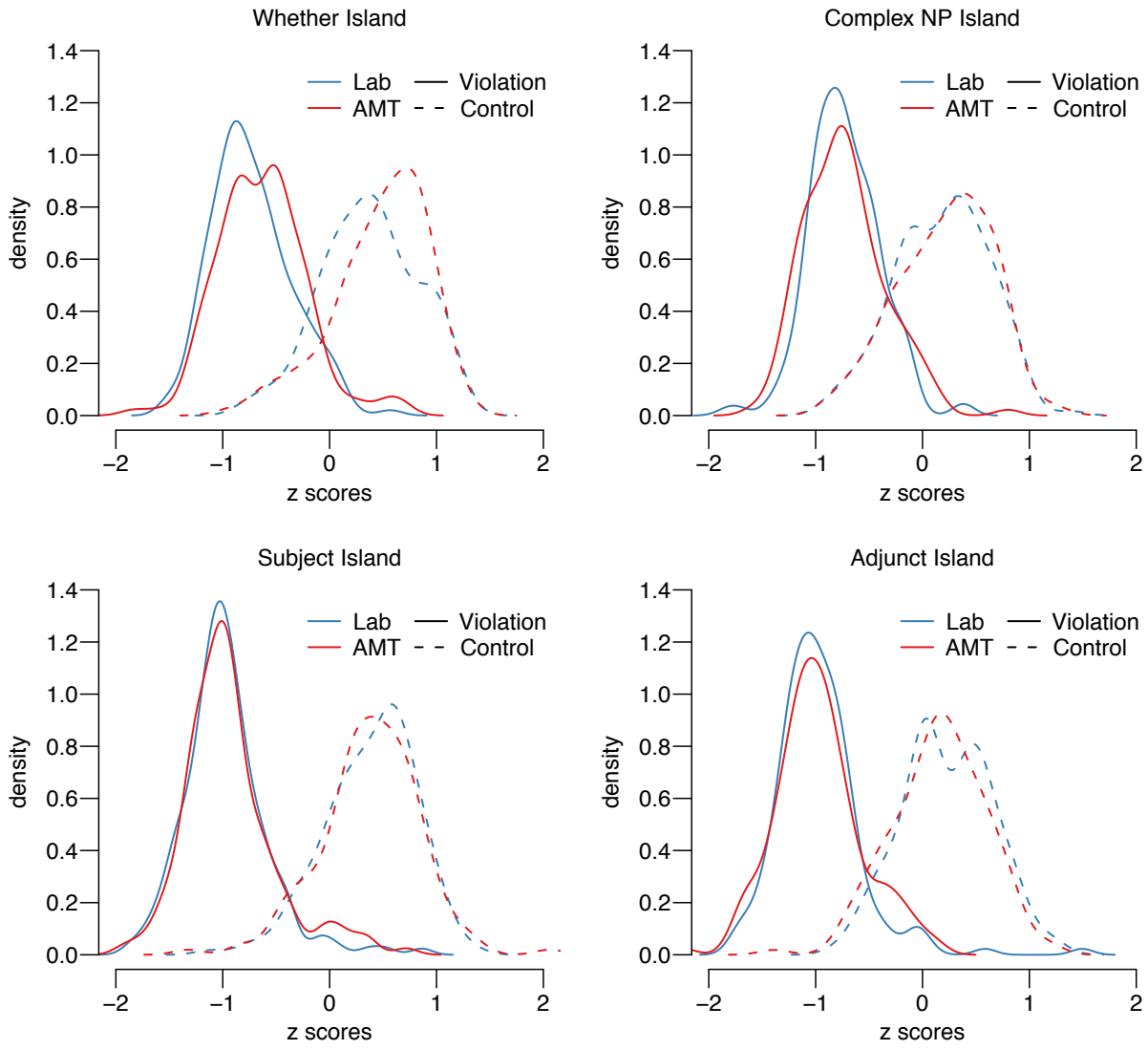
THE SHAPE OF THE DISTRIBUTIONS

One final analysis that may be of interest to syntacticians considering the use of AMT is a direct comparison of the shape of the distributions of each condition in the laboratory and AMT samples. Whereas the resampling simulations in section 4 confirmed that differences between condition means arise at approximately the same rate in each sample, the direct comparison of the distributions can confirm that the source of the difference between condition means are

identical for each sample (i.e., the location of the peak (mode) versus the heaviness of the tail). To aid in the visualization of the distributions, density curves for each condition were calculated using the function density in the base statistics package {stats} in R. These density curves are plotted in figure 5.

Figure 5: The distribution of judgments for the island effect conditions

Density curves for each condition of the island effects. The x-axis represents the judgments after a z-score transformation. The y-axis is density. The grammatical control conditions are plotted in dashed lines. The island violation conditions are plotted with solid lines. The laboratory sample is in blue. The AMT sample is in red.

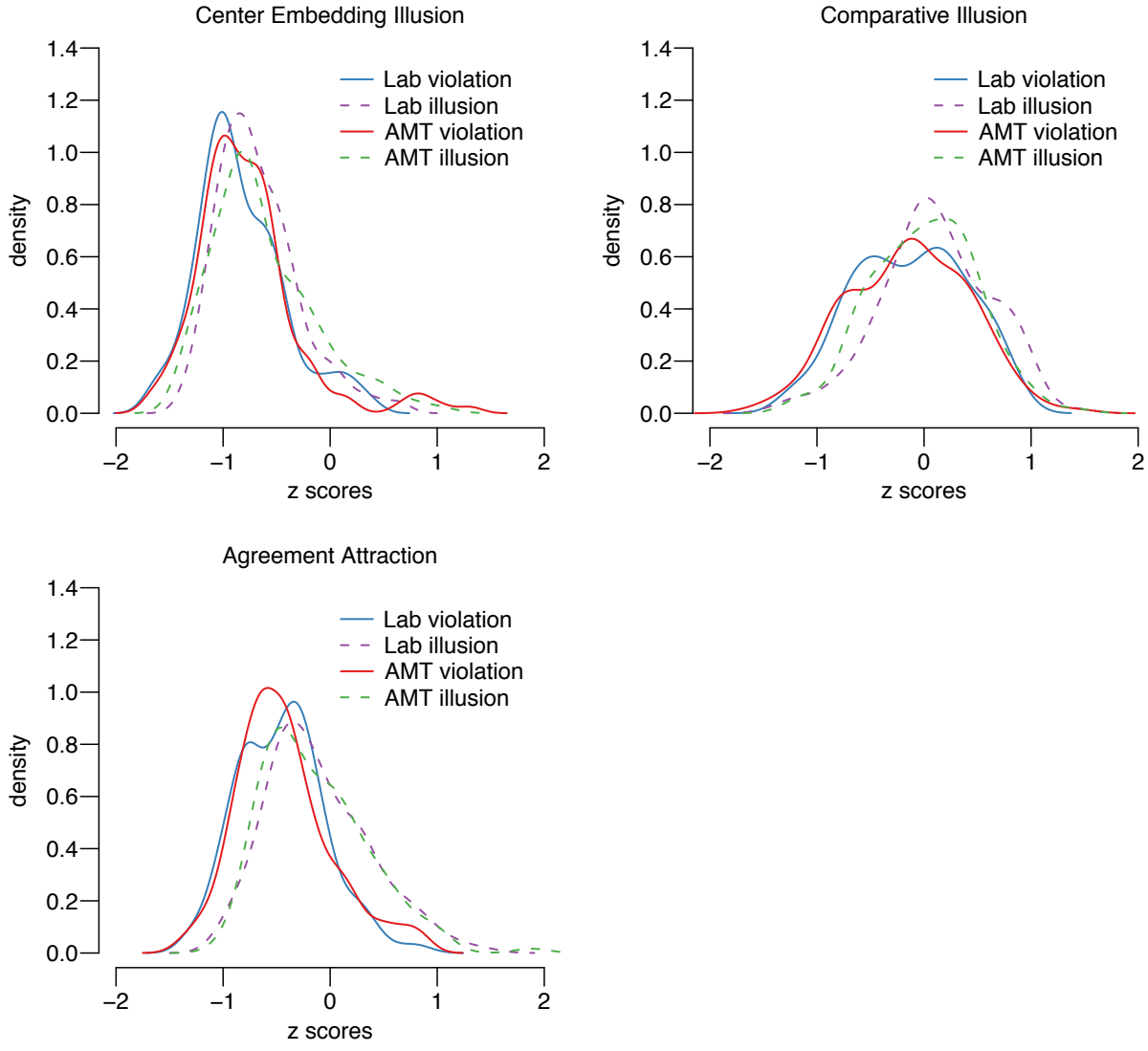


The distributions of the two samples are very similar for each of the conditions relevant to the island effects: the peaks (modes) are approximately equal in location and frequency, and the overall shape and width of the distributions are approximately equal. It does appear that the

rightward tail of the AMT distributions is slightly heavier than the rightward tail of the laboratory distributions, which may account for the marginal power difference between the two samples. But overall, the variation between the distributions appears to be well within the bounds of normal variation between samples.

Figure 6: The distribution of judgments for the extremely weak effects

Density curves for each condition of the extremely weak effects. The x-axis represents the judgments after a z-score transformation. The y-axis is density. The control violations are plotted in solid lines. The illusion conditions are plotted with dashed lines. The laboratory sample is in blue and purple respectively. The AMT sample is in red and green respectively.



The first point to note about the illusions in figure 6 is that the mean differences are not driven by as clear of a peak (mode) separation as the island effects; instead, the differences between the control violations (solid lines) and the illusions (dashed lines) appear to be driven by both a small shift in the location of the distribution along the x-axis, and small changes in the shape of the

distribution. Nonetheless the shape of the laboratory and AMT distributions for each condition again appear to be relatively similar. It should be noted that the reason for the discrepancy between the two samples with respect to the detectability of the comparative illusion may be visible in the density curves in figure 6: although the peaks of the illusion conditions appear to be equal in the two samples, the laboratory illusion condition appears to have a slightly heavier right side than the AMT illusion condition. This suggests that fewer AMT participants were fooled by the illusion, which would result in the lower detectability rates of the comparative illusion in the previous section. This raises the interesting possibility that the AMT sample included more accurate participants than the laboratory sample, at least for the comparative illusion. Of course, additional research on the comparative illusion itself is necessary to better understand the differences between the two samples.

CONCLUSION

Data Quality

The quantitative comparison of these two large-scale samples suggests that Amazon Mechanical Turk is a viable alternative to laboratory-based acceptability judgment experiments. AMT provides impressive time-savings (the collection rate is about 85 participants per hour) without any meaningful disadvantage on the measures of concern to syntacticians:

- The participant rejection rate is less than 15%, which is well within the normal bounds for behavioral experiments.
- There is no evidence of a meaningful power loss for syntactic phenomena, and only a slight power loss for extremely weak (processing-based) effects.
- There is no evidence of meaningful differences in the shape or location of the judgment distributions.

Limitations

The most obvious limitation of AMT is the cost: AMT is a payment only marketplace, and therefore requires research funding (e.g., $3.30 per participant for a 105 item survey). Although these sums are relatively small, they do lead to a significant increase over the (free) university participant pools that syntacticians are accustomed to. In addition to cost, there are also other less obvious limitations imposed by the AMT environment that syntacticians should keep in mind as they switch from laboratory based experiments to online AMT experiments:

- The online only interface means that there is no way to ensure that the participants understand the task. This may contribute to the increased participant rejection rate over laboratory-based experiments.
- There is similarly no way to debrief subjects after the experiment to identify potential problems with the design, instructions, responses, etc. The only option is to include debriefing questions as part of the survey itself, which limits the ability to follow-up based on the participant's responses.

- The increased participant rejection rate suggests a need for standard participant rejection criteria. Unfortunately, at present there are no standard participant rejection methods in the acceptability judgment literature.
- The HTML foundation of AMT means that audio and visual stimuli may be used instead of text (as long as web browsers support the multimedia file type). However, Amazon provides no mechanism for uploading multimedia files. Instead, the researcher must store the multimedia files on their own web server, and link to the files in the HIT itself. An example template for audio files (an auditory acceptability judgment task) is included on the author's website.
- The AMT system provides no mechanism for the collection of reaction times. The only time recorded by the AMT system is HIT completion time (the time from acceptance of the HIT to submission of the HIT), which can be used for participant rejection. If reaction times are crucial to the acceptability judgment experiment, one could use an independent experimental platform (such as *WebExp*) and use AMT to recruit participants and direct them to the independent experimental platform.
- The AMT system does not include functions to aid in experimental design (as is common in dedicated experimental platforms). For example, AMT cannot automatically randomize the order of presentation in a survey. Instead, the experimenter must create randomized versions of the surveys by hand. If the experimenter does not create a novel randomization for each participant, then several participants will see the same randomization (as in this experiment). This adds some time to the construction phase of the experiment.
- At present, the AMT worker pool is primarily comprised of residents of the US (46.8%) and residents of India (34%) (Ipeirotis, 2010). The composition of the worker pool is a direct reflection of Amazon's payment system, which is currently set-up to pay in US dollars and Indian rupees. The composition may change in the future as Amazon's payment system expands; however, at present the lack of geographic diversity will likely affect the collection rates for languages other than English and Hindi, potentially limiting the benefit of AMT for cross-linguistic studies.

Recommendations

In addition to being aware of the limitations discussed above, I would also strongly recommend the following practices to help control the unique properties of the AMT environment:

- Any questions about native speaker ability should be informational only, and crucially not lead to non-payment. This discourages misrepresentations, so that the answers can be used during data analysis.
- Researchers should run some sort of participant rejection or outlier removal process prior to analysis, as the AMT outlier rate is higher than the laboratory rate (14.2% vs 1.7%).
- Target sample sizes should be increased by 15% to accommodate the higher participant rejection rate.

- If extremely weak effects are being investigated (i.e. effects that require sample sizes of 100 or more), 10 additional participants should be added to accommodate the slightly lower statistical power of the AMT sample.

Supplemental Materials

HTML templates for five different acceptability judgment tasks (magnitude estimation, 7-point scale, yes-no, forced-choice, and auditory) can be found on the author's website (currently: www.ling.cogsci.uci.edu/~jsprouse/tools/amt/). This webpage also includes links to R-scripts that may aid in the analysis of data collected using AMT, and an online tutorial offered by Amazon about using the AMT website.

Author Note

This research was supported in part by National Science Foundation grant BCS-0843896. I'd like to thank Diogo Almeida for helpful comments, and Jessamy Norton-Ford for assistance in the early stages of this project, and two anonymous reviewers for their thoughtful comments. Correspondence concerning this article should be sent to Jon Sprouse, 3151 Social Science Plaza A, University of California, Irvine, CA 92697-5100 (email: jsprouse@uci.edu).

References

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, **72**, 32-68.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

Chomsky, N. (1986). *Barriers*. Cambridge: MIT Press.

Cowart, W. (1997). *Experimental Syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage.

Featherston, S. (2005a) Magnitude estimation and what it can do for your syntax: some wh-constraints in German. *Lingua*, **115**, 1525-1550.

Featherston, S. (2005b). Universals and grammaticality: wh-constraints in German and English. *Linguistics*, **43**, 667-711.

Frazier, L. (1985). Syntactic complexity. In D. Dowty, L. Karttunen and A. Zwicky (Eds.), *Natural language processing: psychological, computational and theoretical perspectives* (pp. 129-189). Cambridge: Cambridge University Press.

Gibson, E. & Fedorenko, E. (in press). The need for quantitative methods in syntax. *Language and Cognitive Processes.*

Gibson, E., & Thomas, J. (1999). Memory limitations and structural forgetting: the perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, **14**, 225–248.

Grimshaw, J. (1986). Subjacency and the S/S' Parameter. *Linguistic Inquiry*, **17**, 364-369.

Hofmeister, P. & Sag, I. (2010). Cognitive constraints and island effects. *Language*, **86**, 366-415.

Huang, C.-T. (1982). Move wh in a language without wh-movement. *The Linguistic Review*, **1**, 369-416.

Ipeirotis, Panos. (2010). Demographics of Mechanical Turk. *Center for Digital Economy Research working papers*, **10**. http://hdl.handle.net/2451/29585

Keller, F., Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing Accuracy of Web Experiments: A Case Study Using the WebExp Software Package. *Behavior Research Methods*, **41**, 1-12.

Keller, F. (2000). *Gradience in grammar: experimental and computational aspects of degrees of grammaticality* (Doctoral dissertation). University of Edinburgh.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, **30**, 81-89.

Kuno, S. (1973). Constraints on internal clauses and sentential subjects. *Linguistic Inquiry*, **4**, 363-85.

Marantz, A. (2005). Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review*, **22**, 429–445.

Myers, J. (2009). The design and analysis of small-scale syntactic judgment experiments. *Lingua*, **119**, 425-444.

Phillips, C., Wagers, M., & Lau, E. (in press). Grammatical illusions and selective fallibility in real-time language comprehension. *Language and Linguistics Compass*.

R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.R-project.org

Ross, J. (1967). *Constraints on variables in syntax* (Doctoral dissertation). MIT, Cambridge.

Schütze, C. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. The University of Chicago Press.

Sorace, A., & Keller, F. (2004). Gradience in Linguistic Data, *Lingua*, **115**, 1497-1524.

Sprouse, J. (2009). Revisiting Satiation: Evidence for an Equalization Response Strategy. *Linguistic Inquiry*, **40**, 329-341.

Sprouse, J., Fukuda, S., Ono, H. & Kluender, R. (in press). Grammatical operations, parsing processes, and the nature of wh-dependencies in English and Japanese. *Syntax*.

Sprouse, J. & Almeida, D. (*submitted*). A quantitative defense of linguistic methodology.

Sprouse, J. & Cunningham, H. (in press). Evaluating the assumptions of magnitude estimation of linguistic acceptability. *Language*.

Sprouse, J., Wagers, M., & Phillips, C. (submitted). A test of the relation between working memory capacity and island effects.

Wagers, M., Lau, E., & Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *Journal of Memory and Language*, **61**, 206-237.