Hot Topics Surrounding Acceptability Judgement Tasks

Jana Häussler & Tom Juzek[1]
University of Wuppertal / Nuance Communications[2]
haeussler@uni-wuppertal.de, tom.juzek@googlemail.com

1 Introduction

The aim of this paper is to give an overview of some hotly debated issues in experimental syntax. Section 2 addresses the question whether formal, experimental methods are needed at all for reliable syntactic enquiry. We argue that this is indeed the case, and in Section 3, we discuss selected aspects surrounding the best use of acceptability judgement tasks. These questions concern data reliability and cost efficiency. We do not claim to resolve these issues or to give an exhaustive review; instead we focus on issues we consider most important, in order to stimulate further discussion. Section 3.1 is concerned with two sampling questions: (i) Do participants have to have linguistic experience? (ii) How many participants are needed to get robust results? Arguably, the first question has been answered recently and it is included for completeness. Section 3.2 discusses benefits and potential downsides of running experiments over the internet and using crowdsourcing platforms. Section 3.3 is devoted to the pressing problem of participants not complying with the task in web-based experiments. We review techniques of detecting and preventing such non-cooperative behaviour. Section 3.4 addresses the question of choice of scale. In Section 4, we turn to modelling of experimental findings and grammar architecture. We ask whether grammaticality could be gradient and why it is difficult to experimentally observe fully and unanimously ungrammatical and fully and unanimously grammatical ratings. The paper is concluded by some remarks in Section 5.[3]

2 Why Experimental Work is Needed

Most issues in the present paper address methodological choices concerning acceptability judgement tasks. However, a principal question needs to be addressed first: are experimental methods needed at all? That is, it could be the case that informal, non-experimental methods are sufficiently reliable. This question is at the

[1] Both authors contributed equally to this work.
[2] We thank the two anonymous reviewers for their thoughtful comments.
[3] Smaller sections of this paper partially overlap with Juzek (2016) and Häussler & Juzek (ms). However, in most parts, the present paper departs from these two writings substantially.

heart of a debate on the empirical foundations of syntactic theory, a debate that has increased in intensity after the publication of Schütze's seminal 1996 book and that is fuelled by scepticism about the persistent dominance of informal methods.

Typically, *researcher introspection* is considered as the most common informal method. In researcher introspection, the investigating linguist is his/her own informant. Arguably, though, he/she also takes the views of students, colleagues, and reviewers into account. Researcher introspection is typically contrasted with experimental acceptability judgement tasks, representing the formal methods. There are other experimental methods like eye-tracking studies, fMRI studies, etc., that are equally relevant. However, acceptability judgement tasks are arguably the most common formal method and conceptually the closest to researcher introspection, which is probably why they are used *pars pro toto* for formal methods in this debate.

There are by and large three factions. The first faction consists of those who raise awareness of the issue without necessarily choosing a side (among others, cf. Bard et al. 1996; Schütze 1996; Edelman & Christiansen 2003; den Dikken et al. 2007; Culicover & Jackendoff 2010). The second faction are those who defended the use of researcher introspection. Members of this faction argue that researcher introspection has proven itself as effective and reliable and that there are no reasons to assume that formal methods give better results (Phillips & Lasnik 2003; Bornkessel-Schlesewsky & Schlesewsky 2007; Grewendorf 2007; Phillips 2010; Sprouse & Almeida 2012, 2013; Sprouse et al. 2013). The third faction are those who voice their concerns about the reliability of informal results (e.g. Wasow & Arnold 2005; Featherston 2007; Gibson & Fedorenko 2010, 2013; Gibson et al. 2013).

Sprouse et al. (2013) added to the debate by comparing the results from informal and formal methods in a comprehensive way. One of their motivations was the observation that the phenomena examined in previous papers were handpicked and thus potentially subject to selection bias. To ensure an unbiased selection, Sprouse et al. (2013) created a corpus of all sentences that were discussed in *Linguistic Inquiry* in the years 2001 to 2010. They then randomly sampled 150 marked sentences (marked by a "*", "?", etc.), extracted or created the good counterparts to the marked items, and then tested if informal and formal results matched. In this sense, Sprouse et al. (2013) sampled sentence pairs.

The informal results in Sprouse et al. (2013) are the author judgements extracted from *Linguistic Inquiry*; the formal results were obtained in a crowdsourced acceptability judgement task, using magnitude estimation, a 7-point Likert scale, and a forced-choice scale. Sprouse et al. (2013) used different statistical tests, but their main test was a test of directionality. The test checks whether the experimental rating of any given marked item came out lower than the ratings for its good counterpart. If so, then informal and formal results match, otherwise there is a mismatch. Sprouse et al. (2013) applied this test to all of their marked items and their good counterparts and reported a match rate of informal and formal results of above 95%.

Häussler & Juzek (ms) criticised in particular two aspects of Sprouse et al. (2013). First, their choice to focus on sentence pairs. Second, the fact that Sprouse et al. (2013) created counterparts themselves. Sprouse et al. (2013) did this for cases where the original *Linguistic Inquiry* author did not provide a counterpart; but

they also did this in order to have several lexicalizations per item.[4] With respect to the first point, Häussler & Juzek (ms) argue that while pairwise comparisons are common practise, syntactic argumentation often goes beyond single pairs. Notions like weak and strong islands imply comparison of more than two elements as well as a graded perception of ill-formedness. Whether this gradience reflects the mental grammar is a different issue, which we take up in Section 4. In any case, we need reliable data in the first place. With respect to the second point, Häussler & Juzek (ms) argue that this leaves the door open for implicit bias, i.e. Sprouse et al. (2013) might have – unknowingly – constructed items according to their expectation of the outcome.

Consequently, Häussler & Juzek (ms) revisited the question of data reliability. They, too, created a corpus of *Linguistic Inquiry* items. From their corpus, they randomly sampled 100 items marked with an asterisk and 100 unmarked items; the latter were unrelated to the 100 marked items, i.e. no sentence pairs were sampled. Fig. 1 illustrates the results for the most important subexperiment, i.e. items that (i) come from papers in which authors used more than two judgement categories and (ii) were rated on a 7-point scale in the online experiment. Visual inspection reveals that the author judgements and online ratings do not match well. Instead of observing an approximation of a step function, we see marked and unmarked items "mingle". Some items marked with an asterisk in the corresponding *Linguistic Inquiry* paper receive relatively high ratings in the online experiment, and some unmarked items receive relatively low online ratings. Further, a lot of items receive "in-between" ratings. Häussler & Juzek (ms) used a test that featured as a side note in Sprouse et al. (2013: 234), viz. a threshold test. Häussler & Juzek (ms) report a match rate well below 90%. Though there is no predefined value as to which violation rate is still tolerable, we think that this finding casts serious doubt on the reliability of results obtained through researcher introspection.

[4] For each LI-example selected for their experiments, Sprouse et al. created 8 variants with the same structure but different lexical items (cf. Sprouse et al. 2013: 220).

Fig. 1: From Häussler & Juzek (ms). Items are lined up on the x-axis, ordered by their online ratings (y-axis). Red items were marked by authors in their original Linguistic Inquiry papers, from which items were sampled. Blue items were unmarked

3 Methodological Issues Concerning Acceptability Rating Experiments

3.1 Sampling Participants: Linguistic Expertise and Sample Size

Linguistic expertise. One argument against formal methods is that linguistically naive subjects cannot make the distinctions that a trained syntactician can make. However, critics of this view argue that linguistic expertise leads to potential bias (cf. discussion in Schütze 1996). In fact, several studies found differences between linguists and non-expert participants (e.g. Schütze 1996; Culbertson & Gross 2009; Dabrowska 2010). However, task familiarity seems to be a more important factor than linguistic expertise (Culbertson & Gross 2009).

Sample size. Another question surrounding the choice of participants is how many participants one should have in an experiment. While a sample size of 30 to 40 participants is common in psychology, linguistics, and related fields, there is research asking whether there is a minimum sample size for *robust* results and a maximum sample size for *meaningful* results. Since sample size, power, significance level and effect size are interrelated, we can estimate the required sample size for a given power level and effect size (for guidelines see Cohen 1988).[5] Thus, if the researcher knows roughly which effect size is to be expected, e.g. from related research, it could be an option to calculate the approximate sample size required for the experiment in advance.

[5] A common significance level in linguistics and other behavioural studies is $\alpha = 0.05$. A reasonable power level is $1-\beta = 0.8$.

Mahowald et al. (2016) showed that as few as 7 participants can be sufficient for robust results for an acceptability judgement task. However, we need to be sure that these 7 participants provide reliable data. Small samples are particularly prone to outliers. A single participant not complying with the task but giving random or dishonest responses can spoil the data. This issue is particularly pressing when collecting data over the internet, though a small-scale experiment will typically be run in the lab. For strategies to ensure data quality see Section 3.3.

At the same time, large sample sizes can make standard statistical tests, such as t-tests, come out positive even when only small differences are present. Such differences are still significant in a technical sense, but the question is whether or not they are *meaningful* (e.g. Runkel 2012; see Lin et al. 2013, for related research in the area of information systems). While there is no research into the question of how many participants are too many in a syntactic acceptability judgement task, we suggest testing 6-8 participants per condition as a rule of thumb.

3.2 Offline vs Online and Conventional vs Crowdsourced

Once one has decided to obtain acceptability judgements in a formal experiment, methodological decisions have to be made. These include the question whether the experiment should be run in the lab, classroom, or over the web. For the latter, several free software solutions are available. Some of them are specifically designed to run (psycho-)linguistic experiments: WebExp (http://groups.inf.ed.ac.uk/webexp/, Keller et al. 1998; Keller et al. 2009), MiniJudge (http://www.ccunix.ccu.edu.tw/~lngproc/Mini_Judge.htm, Myers 2009) and Ibexfarm (http://spellout.net/ibexfarm/).

Running an experiment over the web is convenient for researchers and participants alike, since it sets them free from time and locational constraints. Participants can complete an online questionnaire anytime and anywhere they like. Moreover, several participants can do this simultaneously. Handing out a questionnaire to students in a classroom also allows for fast data collection, but requires additional expenditure of time for transcribing the data. Collecting the ratings on a computer saves time, but many linguists will not have access to a lab with several computers for running more than one participant at a time. And even the few who have this possibility still need to invest time for running the lab, be present during the experiment, and arrange for participants to come to the lab.

Participant recruitment is another time consuming part of data collection. Outsourcing this task to the web using crowdsourcing platforms speeds up the whole process enormously (Sprouse 2011; Mason & Suri 2012). Amazon's Mechanical Turk (MTurk) is one of the biggest players among the crowdsourcing platforms.[6] MTurk is an online job market – requesters post jobs and workers choose jobs and get payed via Amazon. Most of these HITs (*human intelligence tasks*) are micro jobs requiring only a few minutes to complete. Payment is typically also at the micro level, which might invite gaming the system (see Section 3.3).

[6] By Amazon's own account the participant pool comprises 500,000 workers from more than 190 countries. Yet, the pool of actually active workers might be considerably smaller (Stewart et al. 2015) and dominated by US residents (Berinsky et al. 2012).

MTurk can also be used for surveys and other scientific experiments (for an overview see Mason & Suri 2012; for guidelines how to gather acceptability ratings via MTurk see Gibson et al. 2011; Sprouse 2011; Erlewine & Kotek 2016). However, MTurk has two main restrictions: the majority of workers are US residents, and researchers willing to use MTurk need an US credit or debit card and a valid US social security number. Alternatives are Prolific Academic (www.prolific.ac), which is specifically designed for research, and Clickworker (www.clickworker.com), which is particularly interesting for researchers working on German because Germans make up a quarter of Clickworker's crowd.[7]

The combination of web-based surveys and crowdsourcing makes data collection comparatively cheap and strikingly fast. It enables data collection from hundreds of participants within a few hours. In addition, crowdsourcing studies reach a broader population than most lab studies. This is particularly attractive for linguists working on languages not readily available within their local community.

On the other hand, web-based studies, and in particular crowdsourcing, have raised a number of objections and worries. In Section 3.3, we address a major thread for data quality, namely unreliable participants. Further concerns regarding reliability are related to technical problems like multiple submissions from the same participant, timing accuracy (for WebExp see Keller et al. 2009), compatibility issues, and so on. We will not discuss them here.

A yet different scepticism concerns the composition of the crowd and potential sampling biases. This issue is widely discussed in the social sciences, political sciences, and related disciplines, but hardly acknowledged within the linguistic community. In sociolinguistics and psycholinguistics, on the other hand, convenience sampling is less accepted (for obvious reasons).

The typical sampling population for offline experiments are students. Crowdsourcing studies reach a broader population – in terms of age, education, profession, location etc. Large samples obtained by crowdsourcing will thus be more representative and allow for identifying groups of speakers. However, the typical linguistic experiment involves too few participants to distinguish random variation between speakers and systematic variation between groups of speakers. In this situation, the diversity turns into a disadvantage since it increases the noise in the data.

In addition, crowdsourcing studies have a population bias too. Older people and in particular people who do not use the internet are underrepresented. These people do not participate in our offline experiments very often, but they can be reached with offline questionnaires. Another problem is self-selection. We consider self-selection bias a minor problem for linguistic studies given that lab studies typically apply convenience sampling, i.e. they draw from that part of the population that is close to hand, viz. students, often from their own department. By and large, participants in crowdsourcing studies are comparable to participants in linguistic lab studies. They are only slightly older and not all of them have an academic degree (cf. Ipeirotis 2010a; Berinsky et al. 2012; for linguistic studies see Schnoebelen & Kuperman 2010; Gibson et al. 2011).

[7] For further platforms and a comparison see Vakharia & Lease (2015).

3.3 Participant Reliability and Non-Cooperative Behaviour

Though several crowdsourcing studies replicated results from lab experiments (for linguistics see Munro et al. 2010; Schnoebelen & Kuperman 2010; Sprouse 2011; for other domains see Mason & Suri 2012; Paolacci et al. 2010; Krantz & Dalal 2000; Dandurand et al. 2008), there are still caveats. A common caveat on web-based studies concerns participants' motivation and reliability. Can we rely on their cooperative behaviour? Do they comply with the task? In principle, non-cooperative behaviour can occur in a lab study as well. But the observer effect (also known as Hawthorne effect) in face-to-face situations makes non-cooperative behaviour less likely to occur. Furthermore, a participant voluntarily taking the effort of coming to the lab is most likely intrinsically motivated and therefore less prone to non-cooperative behaviour.

Non-cooperative behaviour might be particularly virulent in crowdsourcing studies. Several studies provide evidence that the worries are justified (e.g. Downs et al. 2010; Zhu & Carterette 2010; Kazai et al. 2011). Kazai et al. (2011), for instance estimate that up to 57% of their participants did not comply with the task. The number of non-cooperative participants depends on how "non-cooperative" is defined exactly and on properties of the study itself, including payment (see Sorokin & Forsyth 2008; Kazai 2011)[8] as well as the type of task and task design (Eickhoff & de Vries 2013). Repetitive click jobs are particularly prone to cheating since they are easy to complete without actually engaging in the task. Participants can minimise their effort by simply "clicking through". For linguistic rating tasks, there are only few data available. Schnoebelen & Kuperman (2010) report an overall rejection rate of about 25% (for several tasks and including rejection based on language background). Sprouse (2011) reports a rate of 11–16% non-cooperative participants. The exact number depends on whether we count non-natives and participants submitting incomplete surveys as non-cooperative. In our own work, we had up to 14.1% non-cooperative participants (Häussler & Juzek 2016).[9]

3.3.1 Detecting Non-Cooperative Behaviour

Non-cooperative behaviour jeopardises data quality beyond creating noise. Data from non-cooperative participants do not only increase the variance but also affect mean ratings (Häussler & Juzek 2016). Therefore, it is important to identify non-cooperative behaviour and exclude the data. Common detection strategies make use of gold standard tasks, attention checks, inter-worker agreement, and time spent on a task.

Gold standard tasks are tasks with a known correct answer. Accuracy on these items can be used to measure participants' accountability based on their performance. For acceptability ratings, designing a gold standard task faces at least two challenges: (i) Defining/measuring accuracy and (ii) constructing appropriate

[8] Fair payment is not only a matter of data quality but also an ethical issue. Note that Prolific, unlike MTurk and many other crowdsourcing platforms, defined a minimum wage ("ethical reward") of £5 per hour.

[9] Munro et al. (2010) had rejection criteria but do not report the number of participants that were rejected due to non-cooperative behaviour. Other papers do not mention non-cooperative participants at all.

items. As to the first challenge, acceptability ratings are by definition subjective, thus there is no right or wrong, even for supposedly clear-cut cases like the examples below.

(1) a. *There is music in the air.*
b. **There music in the air is.*

There is no correct value on a Likert scale to choose. On a scale ranging from "1" (worst) to "7" (best), ratings as 1 and 2 are equally appropriate for a sentence like (1b). Depending on the other items in the materials, even a 3 might be OK. The problem gets even worse when we turn to magnitude estimation or thermometer judgements. One can circumvent the problem by defining an expected range of correct ratings rather than an exact value. This, however, decreases the chance of detecting non-cooperative behaviour since it increases the odds for a judgement to be in that range. The criterion for deeming a participant non-cooperative can be based on the number of violations (ratings outside the expected range) or by comparing the mean ratings for supposedly grammatical and ungrammatical control items. The mean rating for ungrammatical items should not be higher than the mean rating for grammatical items. This latter criterion can also be used for z-scores in magnitude estimation and thermometer judgement studies.

For acceptability ratings, constructing appropriate items is not a trivial task. There is no established set of such "booby trap" items (aka control items or simply filler items though filler items serve other purposes as well, mainly distracting from the experimental manipulation). In principle it is easy to come up with perfectly natural sentences like *Peter, Paul and Mary are sitting in the kitchen* and clearly unacceptable sentences like *There music in the air is*. Probably, many linguists working experimentally have a set of such items that they regularly use as control items and for calibration. But these are not designed to trap non-cooperative participants. In most cases, they will deviate from the experimental items in one way or another. But booby trap items are powerful guards against non-cooperative participants only when they cannot be differentiated from regular items. This is because experienced cheaters watch out for such "got-you"-questions.

Items like *Repeat your previous judgement* or *Click on the upper left corner* are effective against bots, but easily identifiable for humans and especially experienced workers. Still, they can serve as a simple test for attention. Comprehension questions are another means for checking attention. They are commonly used in psycholinguistic reading studies but rarely so in acceptability judgement studies (unless used to guarantee judgements under some interpretation). An exception is Gibson et al. (2011), who strongly recommend the inclusion of comprehension questions in studies using MTurk. Gibson et al. (2011) do not explicitly mention non-cooperative participants, but it is likely that non-cooperative and inattentive participants motivated this additional task.

Another means to check for non-cooperative participants is inter-worker agreement using kappa statistics or majority votes. We consider this questionable, because it implicitly assumes that non-cooperative participants are outliers whereas in fact non-cooperative participants may make up a substantial portion of the sample. Inter-worker agreement will fail to identify non-cooperative participants when there are many of them. Moreover, inter-worker agreement is problematic for

subjective judgements, including acceptability judgements, because in that case, the judgements of both cooperative and non-cooperative participants will vary.

A final group of detection strategies is based on time. At least a subgroup of non-cooperative participants, so-called spammers, aims for efficiency by fast and mindless clicking through the survey. As a result, they are identifiable by fast response times. A simple implementation of this idea would be to check for completion time of the whole task. In Häussler & Juzek (2016), we suggest a more refined approach checking response times for each individual judgement. Spammers have unrealistically fast response times. But why shouldn't that be visible in overall completion time as well? Clever cheaters do several jobs in parallel switching between them back and forth (Buchholz & Latorre 2011). This behaviour results in pauses and gives the impression of normal or even long overall completion times. This is also the reason why we recommend to check median response times rather than mean response times.

Technically, pauses are outliers increasing the mean while leaving the median unaffected. Median response times from non-cooperative participants are extremely fast too. As a result, they decrease the mean of median response times and increase the standard deviation. Therefore, the rejection criterion should be rather strict when based on these measures. Standard-deviation based approaches to outlier detection commonly set a threshold 2 or 3 standard deviations below the mean (e.g. Baayen & Milin 2010). Rejecting participants with a median response time 2 standard deviations below the mean of median response times misses many non-cooperative participants as demonstrated in Häussler & Juzek (2016). The median absolute deviation from the median (MAD) suggests itself as an alternative because it is more robust to outliers. Its breakdown point is at 50%, i.e. it can handle highly contaminated data (Huber 1981: 108). Yet, there is no established threshold criterion for outlier rejection based on the MAD (but see Leys et al. 2013), not to speak of a criterion that works for non-cooperative participants.

In summary, it is sensible to check performance on control items and median response times to filter out non-cooperative participants. Yet, excluding non-cooperative participants post experiment is a sub-optimal strategy. It is not cost-efficient to first collect these data, only to dump them afterwards. Ideally, we would prevent non-cooperative behaviour in the first place. In the next two sections we describe filtering strategies to fend off non-cooperative participants, as well as methods to discourage non-cooperative behaviour when it occurs nevertheless.

3.3.2 Fending off Non-Cooperative Participants

Rather than rejecting non-cooperative participants after data collection, we want to bar them from participation at the outset. As mentioned above particular types of tasks attract particular worker personalities (Eickhoff & de Vries 2013). Non-repetitive and challenging tasks are less attractive for participants who try to minimise their effort while maximising their financial gains.[10] However, there is little we can do about the task when we aim to collect acceptability judgements.

[10] Note that we are not judging here. For people who try to make their living based on HITs, this is a reasonable strategy given the low wages for most HITs if adjusted to pay-per-hour. One might want to argue that such "full-time workers" should be excluded through a pre-

Participant selection can also be used to fend off non-cooperative participants. Apart from criteria regarding language background, participant selection criteria are hardly used in linguistic studies: none of the crowdsourcing-based studies we cite in this paper applied other selection criteria than language background. Filtering can be based on previous performance and/or on demographic properties, e.g. age, gender, education. Previous studies on non-cooperative behaviour have shown that the typical non-cooperative participant is a young male in his twenties (Downs et al. 2010). Hence, one could ban young men from online surveys. Yet, this strategy has a major drawback. The resulting sample will not be representative (though this might be a minor problem for most acceptability rating studies). It is also an ethical issue as it is prejudging prospective participants. Using selection criteria based on prior performance does a bit more justice to the prospective participants but still results in a biased sample, e.g. due to banning participants new to the platform. Approval rates are a widely used measure for filtering participants, though they are no robust predictor of participant accountability and can be manipulated (Ipeirotis 2010b; Eickhoff & de Vries 2013).

A more robust strategy is the use of a qualification task (e.g. Soleymani & Larson 2010).[11] Candidates have to successfully complete a task, before being eligible to participate in the actual study. A gold standard task would be a suitable task for this purpose. Though being effective, such a two-step design has its downside. The additional task increases the costs both in terms of time and money and decreases the number of potential participants as many of them will avoid the effort and look for other jobs without this additional qualification step. In other words, the extra task diminishes cost- and time-efficiency, which are a major motivation for using a crowdsourcing platform in the first place.

In sum, filtering can be used to fend off non-cooperative participants. However, effective filtering has its downside. Very strict participant selection criteria are costly and increase the risk of a sampling bias. We suggest to use rather lenient criteria despite the risk of including a considerable proportion of non-cooperative participants and to counter this risk by applying some of the discouraging techniques discussed in turn.

3.3.3 Discouraging Non-Cooperative Behaviour

Most detection strategies can be applied as discouraging strategies using feedback loops. In that case, however, we need absolute criteria, e.g. a minimum rating or minimum response time. For response times, we have shown that a warning popping-up when the participant's response times repeatedly fall below a predefined threshold, effectively discourages non-cooperative behaviour (Juzek 2016; Häussler & Juzek 2016).

---

screening. We are not sure this is right strategy, though, as we see two problems with this: 1) The number of completed HITs is not necessarily a robust predictor of worker type, i.e. "casual worker" vs "full-time worker". 2) Being a "full-time worker" does not necessarily mean that such a participant is (highly) familiar with linguistic tasks, i.e. it does not necessarily mean that such workers are biased.

[11] Some platforms, e.g. Clickworker, require workers to pass assessment tests before offering real jobs.

This warning mechanism produced an alerting pop-up window when a participant's response times were extremely short – less than 400 ms for binary rating and ratings on a 7-point Likert scale, and 800 ms for magnitude estimation and thermometer judgements. To estimate reading times, we turned to the psycholinguistic literature. Outliers are typically identified based on the overall distribution using a cut-off value that is derived by subtracting 2.5 standard deviations from the mean (e.g. Baayen & Milin 2010). Only few studies report absolute cutoffs. Luce (1986) argues that response times below 100 ms are physically impossible. For self-paced reading paradigms 200 ms is considered the minimum per word (Jegerski 2014). For comparison: 200 ms corresponds to a single eye fixation (Rayner 1978).

Since the shortest sentence in our materials has two words, we set a threshold at 400 ms. Most of the sentences are much longer, hence 400 ms is a rather conservative threshold. In fact, response times only rarely fell below this threshold. But if they did so repeatedly for a given participant, an alert popped up – giving a friendly and moderate feedback ("Ooopsie, you're going a bit too fast. Please, do not just 'click your way through'.") and a serious warning announcing consequences on the next occasion ("Sorry, you're going too fast and you might not get approved. If you are getting this message although you're doing the task properly, please continue as before."). Including this warning mechanism brought down the proportion of non-cooperative participants from 11% to 3–4% (for details see Häussler & Juzek 2016).

Just like response times, performance on control items can be monitored to give a warning when participants provide inappropriate responses. In that case, however, we need an absolute criterion, e.g. a minimum or maximum rating. Likewise, comprehension questions can be used for this purpose. This seems appropriate since there is a correct response, at least in most cases. Comprehension questions can help to discourage non-cooperative behaviour for two more reasons. First of all, it ruptures a clicking routine, especially if it requires to type an answer, even it is just N for "No" and Y for "Yes". Secondly, it reminds participants to pay attention.

In summary, data quality can be improved by giving participants a warning when their performance (in terms of responses to specific items or in terms of response times) indicates non-cooperative behaviour. So it's worth the effort to monitor participants' responses in the course of the experiment. However, this is hard to achieve on most crowdsourcing platforms. We recommend to use crowdsourcing platforms as a recruitment tool but to run the actual experiment on a separate website, e.g. using WebExp (http://groups.inf.ed.ac.uk/webexp/).

3.4 Choice of Scale

Weskott & Fanselow (2011) found no difference in informativity between a gradient Likert scale, magnitude estimation, and a binary Likert scale. However, when it comes to comparing results obtained with a binary Likert scale vs a gradient Likert scale, differences in data quality are well established for psychological research. Ghiselli (1939) reports that a binary Likert scale leads to more favourable results (i.e. participants tend to give higher ratings). Cox (1980) discusses potential information loss when using a binary Likert scale, in contrast to a gradient Likert

scale. Further, Weijters et al. (2010) report that aggregated data from binary Likert scales have a tendency towards the endpoints. Data from Häussler & Juzek (ms) suggests that Weijters et al.'s findings (2010) also hold for syntactic data, as illustrated in Fig. 2. That is, the quality of the results depends on the scale that is used, which can be problematic, as linguists choose their own scale for their introspective enquiries, but experimental subjects typically have to use the predefined scale that was given to them. If subjects are free to choose their own scale, most subjects opt for five degrees or more (cf. Bard et al. 1996: 45).

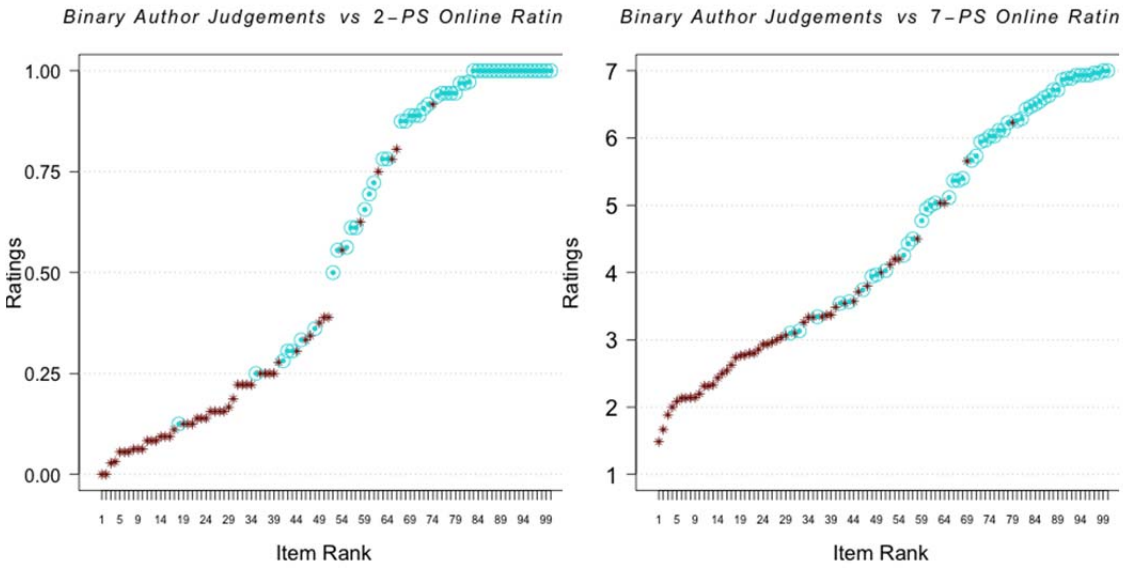

Fig. 2: From Häussler & Juzek (ms), illustrating how data obtained with binary Likert scales (left) have a tendency towards the endpoints, particularly when contrasted to data from gradient Likert scales (right). Items are lined up on the x-axis, ordered by their online ratings (y-axis). Red items were marked by authors in their original *Linguistic Inquiry* papers, from which items were sampled. Blue items were unmarked

4 Gradience and Endpoints

We decided to include the issue of gradience (a term coined by Bolinger 1961), as it has been repeatedly observed in experimental work and as it is an issue that keeps coming up in discussions of such work. At the same time, there is no consensus on how to deal with the observed gradience. Often, it is dismissed as a product of performance factors (e.g. Newmeyer 2003); but such explanations are often vague, without specifying the exact mechanisms and their workings. As we will see below, accounting for the observed gradience is difficult, which adds to the relevance of this section. However, an exhaustive discussion of the gradience controversy is beyond the scope of this paper. Instead, we present some general thoughts and while our aim is not to solve this issue, we hope that we can still contribute to the debate by pointing into the direction in which we think the discussion should be moving.

That gradience occurs in experimental data is widely accepted – in fact, we do not know of a single syntactician who does not accept this. It is also widely accepted

that the observed gradience points to acceptability as a gradient phenomenon. However, it is contested whether or not this gradience also points to grammaticality as a gradient phenomenon (for discussion, e.g. see the contributions in Fanselow et al. 2006; Wasow 2009).

To understand this better, let us briefly review the distinction between grammaticality and acceptability. In the dichotomy of competence vs performance (cf. Chomsky 1965), one could view grammaticality reflecting competence and acceptability reflecting performance. In this view, acceptability can be observed experimentally; notably through acceptability judgements, which are the product of grammaticality and performance factors, including general effects on decision making. Grammaticality, on the other hand, is impossible to observe experimentally and it is, arguably, derived by conjecture.

Early discussions concerning the dichotomy of grammaticality vs acceptability can be found in Chomsky (1955/75) and Bolinger (1961). A few decades later, methodological advances sparked a renewed interest in issues surrounding gradience, when data from experiments and corpora repeatedly exhibit gradience (e.g. Featherston 2007). The same applies to the large-scale study in Sprouse et al. (2013); although not intended this way, combining the graphs in their Fig. 1 gives a new graph that covers the entire scale and not just the endpoints. Results in a recent study of our own point into the same direction, as illustrated in Fig. 3. In this study, we collected experimental rating for sentences we sampled from *Linguistic Inquiry* articles (for details see Section 2 and Häussler & Juzek ms). N.B.: the items shown were rated as unmarked and *-marked by the authors from which they were sampled.

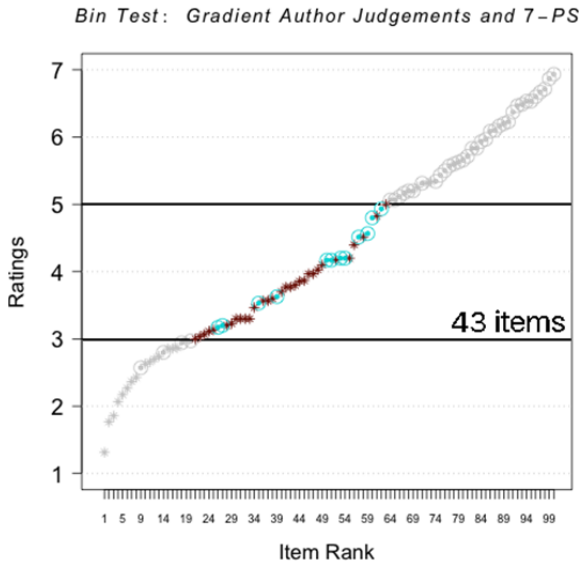

Fig. 3: From Häussler & Juzek (ms). Items are lined up on the x-axis, ordered by their experimental ratings (y-axis). Items with a mean rating between 3 and 5 are highlighted in red and blue; the remaining items are in grey. Red items were marked with an asterisk in the corresponding Linguistic Inquiry paper. Blue items were unmarked in the paper

There is the question of how much one can conclude from an intermediate degree of acceptability about a possible intermediate degree of grammaticality. Various factors may ameliorate or deteriorate acceptability ratings, while the grammaticality of the item(s) in question is unchanged. Two factors could be relevant in particular: 1) The choice of critical items and fillers and 2) intelligibility considerations. As to 1), the same item might receive higher or lower ratings, depending on the other items in the experiment. If the other items are mainly extremely bad, the ratings for a certain intermediate item might be "pushed up" to some extent; on the other side, if other items in an experiment are mainly good, a certain item might receive lower ratings. Hence, it is the relative difference between experimental conditions which is meaningful in an acceptability rating study; and critics might argue that not much can be interpreted into the observed intermediate acceptability. However, the importance of a well-balanced test set and well-balanced fillers is known to most experimentalists (certainly since Cowart 1997), which is why we don't think that 1) should be used to brush aside the issue of gradience. As to 2), intelligibility might influence the acceptability of an item. Consider two items, one syntactically bad but meaningful / intelligible, the other syntactically equally bad and also semantically nonsensical / unintelligible. One might expect the meaningful / intelligible item to receive considerably higher ratings than the semantically nonsensical / unintelligible item. One might even expect the meaningful / intelligible item to get as high as into the upper mid-range of the experimental scale. In our view, this effect could very well be real. And a possible explanation for such an effect could be this: just like humans, primates, and other animals seem to have a *Mitteilungsbedürfnis* (Fitch 2011; loosely: the need to forward information), it would be reasonable to hypothesise that there is also a *Bedürfnis-zu-Verstehen* (loosely: the need to make sense of an input, particularly of language input). By-and-large, however, (2) is understudied and requires further research.

Related to this is the question concerning the burden of proof. Should proponents of a categorical view show why gradience in acceptability does not necessitate gradience in grammaticality? Or should proponents of gradient views show how gradience in acceptability is best explained with gradience in grammaticality? We cannot give a full answer to the question concerning the burden of proof here, but the following analysis discusses what each side would have to do if the burden of proof was in their court. The Items (2) to (5)[12] are useful for our analysis, as each side would have to explain how the ratings for these items came about. Importantly, individual ratings are gradient as well, as indicated by the median and mean ratings attached to the items below.

(2) **John pleaded to take care of oneself.*
(Item 46c in Culicover & Jackendoff 2001; median rating in our study: 3.5, mean: 3.3)

(3) **Boags was very tall a basketball player.*
(Item 5a in Borroff 2006; median rating in our study: 2.5, mean: 3.3)

(4) **Rose saw some taller man than my father.*
(Item 45b in Larson & Marusic 2004; median rating in our study: 3, mean: 3.3)

(5) **Mary discovered the book about himself yesterday that Bob wrote.*

[12] The items, given with their original author judgement, correspond to item ranks 30-33 when ordered by mean rating.

(Item 46b in Takahashi & Hulsey 2009; median rating in our study: 3, mean: 3.3)

Proponents of a categorical model have to explain why most online participants rated Items (2) to (5) as being "in-between". This might be easier with in-between items that are purportedly unmarked, because then one could argue that their ratings are being dragged down by performance-noise. However, the difficulty is that with Items (2) to (5), we observe that their ratings are "dragged up". Though performance factors can have the effect of creating the illusion of well-formedness despite a grammatical violation (Frazier 2008; Phillips et al. 2011), this is likely not the case for (2) to (5). None of the four sentences is an instance of any known grammatical illusion.

Categorical grammars do not necessarily have to blame performance factors to account for gradience. As an alternative, one can assume competition between multiple grammars in the head of a single speaker. Cornips (2006), for instance, argues that internal multilingualism may drive gradience. Speakers speaking multiple varieties of a language may fail to differentiate between these varieties. However, for such an explanation to work, one would need to show which varieties and which constructions compete in our examples and how the overall rating come about.

In principle, accounting for gradient data in gradient grammars is straightforward – but can be tricky when it comes to the details. First of all, one would have to come up with principal properties of the grammar architecture that produce gradience. In particular, a gradient grammar has to have some quantitative component. For example, Linear Optimality Theory assigns weights to constraints (Keller 2000). As result, violation costs vary and perceived ill-formedness is a matter of degree. Another example is Stochastic Optimality Theory (Boersma & Hayes 2001). In this model, constraint rankings are not stable but subject to a probability distribution. This design feature explains gradient data by constant re-ranking. The difficulty for proponents of gradient grammars really is the following: they need to define concrete constraints in accordance with their principles so that those constraints can then be used to correctly predict the status of sentences like those in Items (2) to (5).

Related to the question of gradience is the point that it is hard to experimentally observe the endpoints in their absoluteness. For instance, on a 7-point scale, it becomes increasingly unlikely that the aggregated rating for any given item comes out as a perfect 1 or a perfect 7 as the number of participants increases (to some extent this might be an experimental artefact caused by the chosen scale and the chosen materials). Again, one could seek to explain this observation with the distinction between acceptability and grammaticality. Items and the endpoints of the scale might be equally (un)grammatical but differ along various other dimensions. Under this view, acceptability would be the joint product of factors of which grammaticality is only one. This gives room for variation under the assumption that at least some of the other factors differ across participants.

However, such an answer gives rise to another question: what is the empirical/experimental evidence for assuming the existence of absolute endpoints in the first place? Would the opposite assumption not be equally plausible, i.e. no absolute endpoints exist, they are just theoretical constructs (i.e. platonic concepts in a theorised space)? Those assuming the existence of absolute endpoints need to

(i) either point to a way how one can test and validate their existence or (ii) give strong theoretical reasons why their assumption is necessary.

5 Concluding Remarks

We have looked at various hot topics in experimental syntax, starting with the question whether or not experimental work is needed at all. We argued that based on the literature, one cannot conclude that researcher introspection is sufficiently reliable as the only basis of syntactic theory. In fact, the opposite is the case: our own work shows that there are reliability issues with researcher introspection. We conclude that researcher introspection is a valuable source but needs to be accompanied by sound experimental data.

We then moved on to practical questions concerning experimental work. There are reliability issues, too: we briefly discussed sampling issues and then devoted a major part of the subsection to reliability concerns related to running experiments over the internet. Furthermore, we raised issues concerning the choice of scale.

In the last section, we looked into issues surrounding gradience. Here, our message is that one cannot just dismiss gradience as a mere side-phenomenon. It has to be taken seriously and be accounted for by syntacticians. Arguably, the attempt to do so will lead to progress in the field. However, there is the possibility that syntacticians might not wish to include gradience in their models after all – but then one has to put convincing arguments forward why this should be the case. Such arguments would have to be more convincing than existing arguments.

References

Baayen, H. & P. Milin (2010) Analyzing reaction times. *International Journal of Psychological Research*, 3(2): 12-28.

Bard, E. G., D. Robertson, & A. Sorace (1996) Magnitude Estimation of linguistic acceptability. *Language*, 72(1): 32-68.

Berinsky, A. J., G. A. Huber, & G. S. Lenz (2012) Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, 20: 351-368.

Boersma, P. & B. Hayes (2001) Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry*, 32(1): 45-86.

Bolinger, D. L. (1961) *Generality, gradience, and the all-or-none*. Mouton, The Hague.

Bornkessel-Schlesewsky, I. & M. Schlesewsky (2007) The wolf in sheep's clothing: against a new judgment-driven imperialism. *Theoretical Linguistics*, 33: 319-333.

Borroff, M. L. (2006) Degree phrase inversion in the scope of negation. *Linguistic Inquiry*, 37(3): 514-521.

Buchholz, S. & J. Latorre (2011) Crowdsourcing preference tests, and how to detect cheating. In *12th Annual Conference of the International Speech Communication Association (INTERSPEECH) (27-31 August 2011, Florence, Italy)*. International Speech Communication Association (ISCA), Florence: 3053-3056.

Chomsky, N. (1955/1975) *The Logical Structure of Linguistic Theory*. The University of Chicago Press, Chicago.

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

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences*. Erlbaum, Hillsdale, N.J.

Cornips, L. (2006) Intermediate Syntactic Variants in a Dialect-Standard Speech Repertoire and Relative Acceptability. In G. Fanselow, C. Féry, M. Schlesewsky, & R. Vogel, eds., *Gradience in grammar: Generative perspectives*. Oxford University Press, Oxford: 85-105.

Cowart, W. (1997) *Experimental syntax: applying objective methods to sentence judgments*. Sage Publications, London, UK.

Cox, E. P. (1980) The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research*, 17(4): 407-422.

Culbertson, J. & S. Gross (2009) Are linguists better subjects? *British Journal for the Philosophy of Science*, 60: 721-736.

Culicover, P. W. & R. Jackendoff (2010) Quantitative methods alone are not enough: response to Gibson and Fedorenko. *Trends in Cognitive Sciences*, 14: 234-235.

Culicover, P. W. & R. Jackendoff (2001) Control is not movement. *Linguistic Inquiry*, 32(3): 493-512.

Dabrowska, E., (2010) Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, 27: 1-23.

Dandurand, F., T. R. Shultz, & K. H. Onishi (2008) Comparing online and lab methods in a problem-solving experiment. *Behavioral Research Methods*, 40(2): 428-434.

den Dikken, M., J. Bernstein, C. Tortora, & R. Zanuttini (2007) Data and grammar: means and individuals. *Theoretical Linguistics*, 33: 335-352.

Downs, J. S., M. B. Holbrook, S. Sheng, & L. F. Cranor (2010) Are Your Participants Gaming the System? Screening Mechanical Turk Workers. In E. Mynatt, ed., *CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Atlanta, GA: 2399-2402.

Edelman, S. & M. Christiansen (2003) How seriously should we take minimalist syntax? *Trends in Cognitive Sciences*, 7: 60-61.

Eickhoff, C. & A. P. de Vries (2013) Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2): 121-137.

Erlewine, M. Y. & H. Kotek (2016) A streamlined approach to online linguistic surveys. *Natural Language & Linguistic Theory*, 34(2): 481-495.

Fanselow, G., C. Féry, M. Schlesewsky, & R. Vogel, eds., (2006) *Gradience in grammar: Generative perspectives*. Oxford University Press, Oxford.

Featherston, S. (2007) Data in generative grammar: the stick and the carrot. *Theoretical Linguistics*, 33: 269-318.

Fitch, W. T. (2011) "Deep Homology" in the Biology & Evolution of Language. In A. M. Di Sciullo & C. Boeckx, eds., *The Biolinguistic Enterprise: New Perspectives on the Evolution and Nature of the Human Language Faculty*. Oxford University Press, Oxford: 135-166.

Frazier, L. (2008) Processing Ellipsis: A Processing Solution to the Undergeneration Problem? In C. B. Chang & H. J. Haynie, eds., *Proceedings of the 26th West Coast Conference on Formal Linguistics*. Cascadilla Proceedings Project, Somerville, MA: 21-32.

Ghiselli, E. E. (1939) All or none versus graded response questionnaires. *Journal of Applied Psychology*, 23: 405-415.

Gibson, E. & E. Fedorenko (2013) The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1-2): 88-124.

Gibson, E. & E. Fedorenko (2010) Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences*, 14: 233-234.

Gibson, E., S. Piantadosi, & E. Fedorenko (2013) Quantitative methods in syntax / semantics research: a response to Sprouse and Almeida (in press). *Language and Cognitive Processes*, 28(3): 229-240.

Gibson, E., S. Piantadosi, & K. Fedorenko (2011) Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass*, 5(8): 509-524.

Grewendorf, G. (2007) Empirical evidence and theoretical reasoning in generative grammar. *Theoretical Linguistics*, 33: 369-381.

Häussler, J. & T. S. Juzek (ms). Issues surrounding the reliability and structure of syntactic data.

Häussler, J. & T. S. Juzek (2016) Detecting and discouraging non-cooperative behavior in online experiments using an acceptability judgement task. In H. Christ, D. Klenovsak, L. Sönning & V. Werner, eds., *Methods and Linguistic Theories*. Bamberg: University of Bamberg Press: 73-99.

Huber, P. (1981) *Robust statistics*. John Wiley, New York.

Ipeirotis, P. (2010a) Demographics of Mechanical Turk. *NYU Working Paper No. CEDER-10-01*. New York: New York University. Available at: https://archive.nyu.edu/bitstream/2451/29585/2/CeDER-10-01.pdf (last access July 7, 2017).

Ipeirotis, P. (2010b) Be a top Mechanical Turk worker: You need $5 and 5 minutes. http://behind-the-enemy-lines.blogspot.com/2010/10/be-top-mechanical-turk-worker-you-need.html (last access July 14, 2017).

Jegerski, J. (2014) Self-paced reading. In J. Jegerski & B. VanPatten, eds., *Research Methods for Second Language Psycholinguistics*. Routledge, New York: 20-49.

Juzek, T. S. (2016) *Acceptability judgement tasks and grammatical theory*. PhD thesis, University of Oxford.

Kazai, G. (2011) In search of quality in crowdsourcing for search engine evaluation. In P. Clough, C. Foley, C. Gurrin, G. Jones, W. Kraaij, H. Lee, & V. Murdock, eds., *Advances in Information Retrieval*. Springer, Heidelberg: 165-176.

Kazai, G., J. Kamps, & N. Milic-Frayling (2011) Worker Types and Personality Traits in Crowdsourcing Relevance Labels. In B. Berendt, A. de Vries, W. Fan, C. Macdonald, I. Ounis, & I. Ruthven, eds., *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM'11)*. ACM, New York: 1941-1944.

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

Keller, F., M. Corley, S. Corley, L. Konieczny, & A. Todirascu (1998) *Web-Exp: A Java toolbox for web-based psychological experiments*. Technical Report HCRC/TR-99, Human Communication Research Centre, University of Edinburgh.

Keller, F., S. Gunasekharan, N. Mayo, & M. Corley (2009) Timing accuracy of Web experiments: A case study using the WebExp software package. *Behavior Research Methods*, 41: 1-12.

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

Larson, R. K. & F. Marusic (2004) On indefinite pronoun structures with APs: reply to Kishimoto. *Linguistic Inquiry*, 35(2): 268-287.

Leys, C., C. Ley, O. Klein, P. Bernard, & L. Licata (2013) Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4): 764-766.

Lin, M., H. C. Jr. Lucas, & G. Shmueli (2013) Too Big to Fail: Large Samples and the p-Value Problem. *Information Systems Research*, 24(4): 906-917.

Luce, R. D. (1986) *Response times: Their role in inferring elementary mental organization*. Oxford University Press, New York.

Mahowald, K., P. Graff, J. Hartman, & E. Gibson (2016) SNAP judgments: A Small N Acceptability Paradigm (SNAP) for linguistic acceptability judgments. *Language*, 92(3): 619-635.

Mason, W. & S. Suri (2012) Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1): 1-23.

Munro, R., S. Bethard, V. Kuperman, V. T. Lai, R. Melnick, C. Potts, T. Schnoebelen, & H. Tily (2010) Crowdsourcing and language studies: the new generation of linguistic data. In *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*: 122-130.

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

Newmeyer, F. (2003) Grammar is grammar and usage is usage. *Language*, 79: 682-707.

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

Phillips, C. (2010) Should we impeach armchair linguists? In A. Iwasaki, H. Hoji, P. Clancy, & S.-O. Sohn, eds., *Japanese-Korean Linguistics*. CSLI Publications, Stanford, CA: 49-64.

Phillips, C. & H. Lasnik (2003) Linguistics and empirical evidence: reply to Edelman and Christiansen. *Trends in Cognitive Sciences*, 7: 61-62.

Phillips, C., M. W. Wagers, & E. F. Lau (2011) Grammatical illusions and selective fallibility in real-time language comprehension. In J. Runner, ed., *Experiments at the interfaces (Syntax & Semantics 37)*. Emerald Publications, Bingley, UK: 153-186.

Rayner, K. (1978) Eye Movements in Reading and Information Processing. *Psychological Bulletin*, 85: 618-660.

Runkel, P. (2012) Large Samples: Too Much of a Good Thing? *The Minitab Blog*. Available at: http://blog.minitab.com/blog/statistics-and-quality-data-analysis/large-samples-too-much-of-a-good-thing (last access October 13, 2016).

Schnoebelen, T. & V. Kuperman (2010) Using Amazon Mechanical Turk for linguistic research. *Psihologija*, 43(4): 441-464.

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

Soleymani, M. & M. Larson (2010) Crowdsourcing for affective annotation of video: development of a viewer-reported boredom corpus. In *ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*: 4-8.

Sorokin, A. & D. Forsyth (2008) Utility data annotation with amazon mechanical turk. Computer Vision and Pattern Recognition Workshops. In *IEEE Computer Society Conference on IEEE (CVPRW'08)*: 1-8.

Sprouse, J. (2011) A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, 43(1): 155-167.

Sprouse, J. & D. Almeida (2013) The role of experimental syntax in an integrated cognitive science of language. In K. Grohmann & C. Boeckx, eds., *The Cambridge Handbook of Biolinguistics*. Cambridge University Press, Cambridge, UK: 181-202.

Sprouse, J. & D. Almeida (2012) Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*, 48: 609-652.

Sprouse, J., C. T. Schütze, & D. Almeida (2013) A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, 134: 219-248.

Stewart, N., C. Ungemach, A. J. L. Harris, D. M. Bartels, B. R. Newell, G. Paolacci, & J. Chandler (2015) The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making*, 10(5): 479-491.

Takahashi, S. & S. Hulsey (2009) Wholesale late merger: beyond the A/Ā distinction. *Linguistic Inquiry*, 40(3): 387-426.

Vakharia, D. & M. Lease (2015) Beyond Mechanical Turk: an analysis of paid crowdwork platforms. In *Proceedings of the iConference*. Newport Beach, CA. https://www. ischool.utexas.edu/~ml/papers/donna-iconf15.pdf (last access October 13, 2016).

Wasow, T. (2009) Gradient Data and Gradient Grammars. *Chicago Linguistics Society*, 43: 255-271.

Wasow, T. & J. Arnold (2005) Intuitions in linguistic argumentation. *Lingua*, 115(11): 1481-1496.

Weijters, B., E. Cabooter, & N. Schillewaert (2010) The effect of rating scale format on response styles: the number of response categories and response category labels. *International Journal of Research in Marketing*: 27: 236-247.

Weskott, T. & G. Fanselow (2011) On the informativity of different measures of linguistic acceptability. *Language*, 87(2): 249-273.

Zhu, D. & B. Carterette (2010) An Analysis of Assessor Behavior in Crowdsourced Preference Judgments. In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*: 17-20.