Identifying Non-Cooperative Participation in Web-Based Elicitation of Acceptability Judgments – How to Get Rid of Noise in your Data

Jutta Pieper, Alicia Katharina Börner, Tibor Kiss
Linguistic Data Science Lab
Ruhr-University Bochum
{jutta.pieper, alicia.boerner, tibor.kiss}@rub.de

Abstract

In this paper, we discuss different sources of noise or other detrimental effects in the elicitation of experimental data: These effects may emerge due to the loss of control in unsupervised web-based elicitation, may be task-related, or finally emerge due to poor questionnaire designs lacking sufficient means to identify inapt participants.
We describe a stepwise process to reach elicited data at the highest attainable level in web-based Acceptability Judgment Tasks (AJTs). In the first step, the questionnaire design, we focus especially on a careful construction of appropriate filler and control items and introduce an alternative to instructional manipulation checks appropriate for AJTs, namely attention items. The second step is to choose the right platform to elicit experimental data from. Lastly, we will show how to employ latency- and response-based methods – analyses of response times as well as of responses to the specialized items – to reliably detect inapt participants.

Keywords: Acceptability Judgments, Questionnaire Design, Eligibility Screening, Online Participant Pools, Attention Checks, Control Items

1 Advantages and Challenges of Web-based Elicitation

After the time-consuming and error-prone procedure of paper & pencil studies could be simplified in the 1970ies, providing standardized and controlled presentation of stimuli and accurate measurements of response times (see Musch and Reips, 2000), a second revolution began in the 1990ies: internet-based technologies now removed the necessity to instruct and supervise participants in person. Instead, a very large and diverse population had become accessible with minimal recruitment time and costs (see e.g., Gibson, Piantadosi and Fedorenko, 2011), in particular by crowdsourcing platforms such as Amazon Mechanical Turk (AMT)[1] and clickworker[2], which have both been founded in 2005. Although these early platforms are geared towards micro-tasks (Human Intelligence Tasks (HITs)), such as labeling images, they have proven versatile for conducting experimental research, since laboratory experiments have been replicated with web-based methods (see e.g., Schnoebelen and Kuperman, 2010 for linguistic research) and have proven to only differ in terms of higher dropout-rates (Dandurand, Shultz and Onishi, 2008) and higher rates of uncooperative participants (Sprouse, 2011b).
However, abandoning control of the experimental setting comes at a price regarding possible distractions, completion speed and task understanding (Sprouse, 2011b). So does the loss of

personal acquaintance with participants as it weakens their engagement, and consequently, detrimental effects on data quality, leading to decreased statistical power, have been observed (see Dandurand et al., 2008; Oppenheimer, Meyvis and Davidenko, 2009). Many researchers have been concerned with identifying fraudulent participants who do not comply with the task (see, among others Gadiraju, Kawase, Dietze and Demartini, 2015; Chandler and Paolacci, 2017) as it has indeed been observed that malicious behavior in crowdsourcing is virulent (see e.g., Häussler and Juzek, 2016).

The remaining paper is structured as follows: Section 2 discusses acceptability judgment tasks (AJTs) in general as well as its underlying assumptions, and how these can be reflected in the instructions. The impact of questionnaire design and the purposes of different trial types are discussed in section 3, especially shedding light onto the nature of control items, and introducing a new trial type, namely attention items. Section 4 discusses why the choice of an appropriate participant pool is a relevant step in avoiding uncooperative behavior and ineligible participants. Section 5.1 discusses latency-based methods to identify participants who do not commit themselves sufficiently to the task at hand. Lastly, section 5.2 discusses how responses to control and attention trials can be used to identify incompetent and uncooperative participants. In particular, this concerns calculating the necessary amounts of control and attention trials in a questionnaire and determining appropriate thresholds. Section 6 concludes the paper.

2 Setting up the Experimental Frame: Acceptability and AJTs

The concept acceptability is distinguished from grammaticality. While grammaticality can be characterized as abstract rule conformity (reflecting competence), acceptability takes factors of performance into account (Juzek, 2016). It is thus possible that grammatical sentences are judged as unacceptable, and vice versa. Grammaticality, however, is only quantifiable by means of acceptability, and it is hence the task of the linguist, be it informal or experimentally, to exclude all confounding factors. If this is the case, we can assume that a sentence will be judged as acceptable if it is grammatical.

Acceptability itself is as a percept by nature not directly measurable. Moreover, a reported perception is different from concepts such as introspection or intuition, which would assume direct access to cognitive systems (Schütze and Sprouse, 2014). As such an access has never been assumed in modern linguistics, we can characterize an acceptability judgment (AJ) as a reported perception of acceptability (following, among others, Schütze, 1996). AJTs provide an indirect way to measure acceptability by asking participants to report their perceptions, for example, by using a scale. In such a setting, judgments are expected to occur spontaneous, automatic, unconscious and without reasoning but with relative temporal immediacy (see Maynes and Gross, 2013; Schütze and Sprouse, 2014). A further characteristic of acceptability is its gradience, which is reflected in the amplification of diacritics, the meaning of which has never been formally defined, leading to inconsistent usage (see e.g., Sorace and Keller, 2005).

Arguably, complex patterns of acceptability cannot reliably be accessed by traditional 'armchair linguists' (see e.g., Phillips, 2009). As current research heavily relies on gradient, i.e. subtle and potentially controversial judgments, linguists now generally agree to collect judgments from naïve speakers using AJTs in a controlled experimental setting (see Gibson and Fedorenko, 2013).

To elicit AJs from naïve subjects, well-known procedures used for sensory or social stimuli can be utilized (e.g. Sorace and Keller, 2005). In general, AJTs have been proven to be reliable and robust (Sprouse, Schütze and Almeida, 2013; Sprouse and Almeida, 2017;

Leivada and Westergaard, 2020). In the following, we will restrict ourselves to two broadly accepted task types, Likert Scales (LS) and Forced Choice (FC).
LS tasks elicit absolute judgements of isolated stimuli on a numerical (ordinal) scale, thus allowing for direct and intuitive expression of gradient perceptions. The scale employed is ordered and bipolar, ranging from unacceptable to acceptable, and is mostly chosen to be uneven (avoiding forced random choices by means of a neutral point).
FC tasks provide an intuitive form of eliciting relative judgments. When used for AJTs, usually only two options are provided (Two-Alternative Forced-Choice (2AFC)) from which participants simply choose the one which they consider more acceptable. They neither state the perceived strength of difference between these acceptabilities nor how acceptable the presented stimuli are by themselves. Nonetheless, it is possible to deduce the size of the difference between a pair from the proportions of chosen answers (Myers, 2009; Schütze and Sprouse, 2014).
It holds for all types of AJTs that when conducting experiments, one needs to make sure that naïve participants indeed base their decision on the acceptability of stimuli despite not being familiar with this linguistic concept. To this end, the concept of naturalness is often used in instructions to avoid misinterpretations of the term acceptability, e.g., as judgments based on prescriptive rules (see Featherston, 2021).[3]
As was already pointed out, experimental linguistics assumes that stimuli are assessed spontaneously by participants, and hence, this as well should be made clear to the participants by carefully worded instructions, thus helping avoid judgments based on prescriptive rules (see Phillips, 2009). Remarkably, although various suggestions can be found in the literature to achieve spontaneity (see e.g., Schütze, 1996), the ban on reflective (conscious) judgments has not always been imposed in linguistic practice (see e.g., Maynes and Gross, 2013).

3 Control by Architecture: The Structure of Questionnaires

Taking precautions in designing questionnaires is the first step to guarantee high-quality data. Apart from avoiding certain biases and unintentional effects, the detection of participants that are unable or unwilling to cope with the task needs to be facilitated.

3.1 General Properties of Questionnaires

The starting point of each study is the formulation of experimental hypotheses, which are reflected in the factorial design: The possible combinations of factor levels constitute the experimental conditions. The test trials in a questionnaire comprise various lexicalizations per experimental condition (LPC) so that conclusions are not falsely drawn from peculiarities of individual lexical stimuli (see e.g., Gibson and Fedorenko, 2013). Moreover, multiple of lexicalizations (and multiple participants) increase statistical power (see e.g., Goodall, 2021).
Naturally, test trials build the core of each questionnaire. However, it is necessary to include so-called *filler trials*, which differ from test trials in not being related to the phenomenon under investigation. Their importance stems from the need to exclude certain effects and biases, confirmation- and order-biases in particular, but also scale-biases. A high test-filler ratio (TFR) serves to impede linguistically naïve participants from familiarizing themselves with the structure of test stimuli, as this may result in the unwarranted formation of conscious response strategies (Schütze and Sprouse, 2014).
The length of questionnaires (about 100 trials are seen as an upper boundary, see e.g. Sprouse et al., 2013) becomes an area of multiple conflicts: between the necessity to avoid confirmation biases and strategic responding (high TFR) and the necessity to attain reasonable power (many LPC) on the one hand and the necessity to avoid fatigue or boredom

effects (limited survey length) on the other hand. However, as for test and filler trials a minimum amount is required to serve their purposes, their share should not be arbitrarily lowered. Regarding the minimum requirement for LPC, Schütze and Sprouse (2014) recommend at least eight ones, whereas Goodall (2021) reasons that it is often (only) possible to employ at least four or five lexicalizations due to length restrictions. Admitting possible challenges in creating many test stimuli (see also Goodall, 2021), we propose at least six LCP as an absolute minimal standard. Concerning the TFR, we argue that at least two fillers for each test trial (1:2) should be employed (see e.g., Weskott and Fanselow, 2009) in order to prevent By adjacent test trials, thus impeding strategy building as well as priming resulting from adjacent stimuli of the same experimental condition (see Häussler and Juzek, 2021). Employing a 1:1 TFR (see Goodall, 2021) would thus result in noticeable patterns as every second trial was a test trial.

We propose that the best strategy to avoid overly lengthy questionnaires while retaining reasonable power (i.e. six LPC) and impeding strategy building (i.e. 1:2 TFR) is the reflected choice of the AJT, and thus the factorial design of the study. It is the factorial design in tandem with these minimum standards that lays the foundation for computing the required questionnaire length. It determines the number of experimental conditions, and thus the minimum required number of test trials (and subsequently of fillers). Changing the AJT from LS to an 2AFC reduces the length of the questionnaire by half, as in an FC task, trials consist of variants of an experimental conditions to pick from, one independent factor (e.g. in a LS task) becomes the dependent factor.[4]

On top of practice and fatigue effects, judgments are relative to each other and may thus be influenced by the presentation order of stimuli, also known as carryover effects (e.g., Weskott and Fanselow, 2009). It been demonstrated experimentally that the repeated exposure of ungrammatical or borderline structures lead to improved respective judgments in the course of the questionnaire (see e.g., Cornips and Poletto, 2005) and to participants becoming more accepting in general (Snyder, 2021). To prevent major influence of these order biases, it is mandatory to present the questionnaires in a randomized order, up to individual randomizations for each participant. For 2AFC tasks, the presentation order of the conditions of a minimal pair also needs to be randomized (see Sprouse and Almeida, 2017).

Moreover, in a within-items design, no participant is allowed to be subjected to the same test item more than once (neither in different experimental conditions) as to avoid carryover effects see Goodall, 2021). It is thus imperative to distribute the experimental conditions of the same test item across counter-balanced lists (e.g. Latin Square design, see e.g., Schütze and Sprouse, 2014).

Of course, not only test items but also filler items need to be carefully constructed. To avoid scale biases (scale compression[5] and scale skew[6]), filler stimuli need to cover different degrees of acceptability, ranging from fully unacceptable to fully acceptable (Schütze, 1996).[7] It is further strongly recommended to match acceptability degrees of filler stimuli to those of the test stimuli in order to avoid blinding participants towards potentially subtle differences in test stimuli (see Phillips, 2009).

It is thus clear that the choice of appropriate filler items depends on the nature of the test items. Thus, the nature of filler items can only be exemplified with regard to specific test items. For illustration purposes, we refer to the experiments presented in Kiss, Pieper and Börner (2022) that examined constraints on word order of event-internal adverbial prepositional phrases (PPs) in German. As an illustration, consider as the item in (1), taken from the FC experiment, where participants needed to select their preferred serialization of a comitative (*zusammen mit einem General* 'together with a counsellor') and the direct object (PP>OBJ vs. OBJ>PP).

(1) Test item

(a) *Ich habe gehört, dass ein Mininster zusammen mit einem General was*
I have heard that a secretary together with a counsellor something
unterzeichnet hat. Was es war, weiß ich aber nicht. (PP>OBJ)
signed has what it was know I but not

(b) *Ich habe gehört, dass ein Minister was zusammen mit einem General unterzeichnet hat. Was es war, weiß ich aber nicht.* (OBJ>PP)
'I have heard that a secretary decided something in tandem with a counsellor. But I don't know what it was.'

The test environment consisted of verb final embedded clauses. The objects were realized as *wh*-indefinites (*was* 'something'), which are assumed to be syntactically fixed if they receive an existential interpretation (see e.g., Haider, 2010). The required interpretation was enforced by including a disambiguating addendum. An essential point to grasp in creating test stimuli is that these should only differ structurally with respect to the experimental factors manipulated and should otherwise be constructed in such a way that they exclude (extra-grammatical) confounding factors (see e.g., Schütze, 1996).

However, this strict compliance to specific characteristics may render test trials quite catchy, especially in case of prominent aspects like the disambiguating addenda in (1). Thus, as *filler trials* are not related to phenomenon under investigation but are needed in order to obfuscate test trials, the filler stimuli in Kiss et al. (2022) were superficially similar to test stimuli (by sticking to word order manipulations), blurring the distinction between those groups (see also Goodall, 2021).

Although their main objective is the obfuscation of test trials, filler items can carry different functions. Besides ordinary fillers, we advise the integration of calibration trials and pseudo fillers, as well as (two types of) control items, and attention items. The latter two trial types are essential in identifying uncooperative participants and discussed in extension in the next two subsections.

Calibration trials aim at accustoming participants to the scale (implicit training). We recommend using six calibration trials (also in FC tasks), covering three degrees of well-formedness in two lexicalizations each (see supplementary materials).[8]

So-called *pseudo-fillers* (Patterson, 2014) pick up central aspects of test items in different context in order to prevent a dop in processing capacity due to over familiarization.

3.2 Including Control Trials in Questionnaires

Even the best efforts to account for biases and effects through questionnaire design will of course not block participants who are unwilling or unable to cope with the task at hand. Addressing this issue, we will discuss two types of filler items, related and unrelated *control items*, that can be used to identify these uncooperative participants in order to discard responses that harm the quality of the elicited data. The essential property of these control stimuli is their well-established status: They must be clearly (un)acceptable to facilitate the identification of participants failing to give correct responses. Of course, this raises the question how the status of control stimuli can be established. It should be evident that relying on the intuition of the designers of an experimental study may lead to vicious circles and we hence propose to verify the acceptability status of control stimuli in separate scale-based studies (as proposed for standard items in Gerbrich et al., 2019). We propose that it is

sufficient to carry out a study of unrelated control stimuli only once, but since related control stimuli bear a similarity to the test stimuli, they have to be carried out for each experimental study.
As an illustration for an unrelated unacceptable control stimulus, consider the German verb final (VF) clause in (2)[9] in which the order of the syntactic arguments violates up to three word order constraints: [+nom]<[+acc], [+pro]<[-pro], and [+animate]<[-animate]:

(2) *Max hat erzählt, dass Schwarzweißfilme er liebt.*
Max has told that black.and.white.films he loves
'Max said that he loved black-and-white films.'

What counts as related or unrelated of course depends on the study. The item in (2) can be considered unrelated in the context of the studies reported in Kiss et al. (2022), as they neither contain adverbial PPs nor *wh*-indefinites. Related control stimuli deviate only slightly from the test stimuli such that they expedite a thorough reading of the stimuli (see Patterson, 2014). As both *wh*-indefinites and PPs are present in the unacceptable item in (3), it is a related control stimulus in the context of the studies in Kiss et al. (2022).

(3) *Lea hat erzählt, dass was der Student zum Verkauf angeboten hat.*
Lea has told that something the student for sale offered has
'Lea said that the student offered something for sale.'

This item differs from the test items in that the uncontroversially offending element is the order between the subject and the *wh*-indefinite object, and that the PP employed does not belong to the class of adverbials under investigation. We assume that the similarity between related control stimuli and test stimuli provides two advantages: First, they arguably support further disguising the test trials, particularly in combination with pseudo-fillers. Secondly, as they are less salient than unrelated control trials, they minimize the possibility that trained participants identify control trials as such, hence enhancing the chances to identify non-cooperative participants.

3.3 Including Attention Trials in Questionnaires

While the primary purpose of control trials is to identify participants that are unable or unwilling to comply with the task, the purpose of *attention trials* is to identify participants who do cooperate but do not commit themselves sufficiently. Attention stimuli provide a known answer but differ from unacceptable control stimuli in that the relevant manipulation is comparatively obscure, introduced at a late point of the stimulus that is usually not affected and consisting of rather small elements, and hence can be easily overlooked. As an illustration for an unacceptable attention trial, consider the following word order variation of the negation (*nicht* ('not')) in the addendum in example (4):

(4) *Kannst du mir sagen, wen Uwe gefeuert hat? Ich hoffe, es nicht war Ben.*
can you me say whom Uwe fired has I hope it not was Ben
'Can you tell me whom Uwe has fired? I hope it wasn't Ben.'

Attention stimuli have not been used in prior experimental studies in linguistics. If attention is treated as an issue at all, attention checks are considered instead, which aim at identifying inattentive participants by their disregard of exact orders given in (trial-related) instructions

(instructional manipulation check (IMC), Oppenheimer et al., 2009). However, Prolific Team (2021) defines a fair attention check as a test whether a participant has paid attention to the question (i.e., stimulus), not so much to the instructions above it, and further requires that it must preclude any misinterpretations.

We see various reasons for not using IMCs in AJTs. First, trials in linguistic studies typically deviate from those in e.g., psychology in that instructions throughout an AJT remain the same. We deem a change of instructions unfair, as participants may be confused whether they are supposed to rate the naturalness of a stimulus or to follow the instructions it gives. Secondly, there are various concerns regarding attention checks: Oppenheimer et al. (2009) point out that they may harm the data quality as participants may feel insulted or embarrassed, affecting performances. Finally, attention checks may simply be too noticeable, particularly for experienced participants on crowd-sourcing platforms (e.g., Gadiraju et al., 2015).

None of these problems emerge for the attention stimuli presented above. As these checks are not obvious as such, it is unlikely that they (as opposed to IMCs) will influence participants' behavior. However, as we the level of attention may vary throughout the experiment even within diligent participants (Paolacci and Chandler, 2014), attention as well as control trials should be spread across the whole survey (see e.g., Juzek, 2016), which impacts the number of control and attention trials required.

4 Control by Choosing an Adequate Participant Pool

Next to employing specific item types as means to identify inapt participants, it is desirable to employ a high-quality participant pool in order to reduce the amount of potential uncooperative or ineligible participants beforehand. Early crowdsourcing platforms, like AMT, are geared towards HITs. But these differ from AJTs, in the structure of the task and in the demands posed upon participants.

First of all, HITs require much smaller assignment numbers as quality control can be determined by majority votes (see Mason and Suri, 2012). AJTs require – as a rule of thumb – 24 to 36 participants to yield useful results (see e.g., Goodall, 2021), since variability in the data elicited is expected. Thus, a major difference consists in the need of AJTs to be embedded in a carefully designed questionnaire (see section 3). Accordingly, employing smaller batch sizes as in HITs in order to ensure attentiveness and to deter spammers is clearly not an option for AJTs (see e.g., Schnoebelen and Kuperman, 2010).

The most marked difference between HITs and AJTs consists in the eligibility of participants: Whereas in the former (almost) no restrictions apply, multiple restrictions apply to participants of experiments. In AJTs, we need judgments of native speakers that are naïve, i.e. non-linguists. As for HITs eligibility is a minor issue, few possibilities exist to constrain it in crowdsourcing platforms like AMT (see Mason and Suri, 2012), leading to the need to reject workers that should not have participated in the study in the first place (see Schnoebelen and Kuperman, 2010).[10]

This issue has been addressed by new platforms that are geared towards the needs of academic researchers: e.g. Prolific (founded in 2014),[11] CloudResearch (founded 2016),[12] and Positly (founded in 2017).[13] They provide methods to address a specific subset of their pool (prescreening) such that the researchers can define a concrete target population.[14]

Prescreening for AJTs mainly concerns native language but may be extended to the country of residence (where your target language is spoken), monolinguals, and arguably non-linguists. For linguists, Prolific provides the most elaborate language-related screening options and is thus by far the most suitable platform for AJTs.

Another eligibility criterion for AJTs is naïveté, i.e., participants should not be familiar with test stimuli or with linguistic structures as the research subject (see section 3). Accordingly, these new platforms provide the exclusion of participants that haven taken part in previous studies. As syntactic satiation, i.e., the loss of strong native-speaker intuitions due to repeated exposure, persists over time and can be observed weeks later (Snyder, 2021), this feature is especially useful when running multiple similar studies (especially in addition to a ban of multiple accounts of one and the same person on part of the platforms).
To ensure the quality of their participant pools, some platforms do not only protect against bots but are vetting participants' trustworthiness beforehand (Litman, Rosenzweig and Moss, 2021). The careful selection of an appropriate participant pool is thus an important step in ensuring data quality by providing access to the study only to eligible and trustworthy participants. Yet, the choice of an appropriate platform only reduces the proportion of inapt participants but cannot fully exclude them. In the following, we will show how to determine problematic participants based on latency- and response-based methods.

5 Control by Analyses: Latency- and Response-Based Identification of Non-Cooperative Participants

To remove noise emerging from inapt participants, we need to inspect the elicited data for signs of fraudulence, incompetence, or inattention. In doing so, we employ latency-based methods targeting RTs, as well as response-based methods, i.e., the analysis of responses to control and attention trials. Whereas latency-based methods provide insights into behavioral patterns of participants, response-based methods provide insights into the data quality provided. Taken together, these provide a powerful tool to detect inapt participants.

5.1 Latency-Based Identification

In general, there are two ways in which inapt behavior manifests itself in RTs: short outliers may indicate guessing, whereas long outliers may indicate distraction or underperformance (Ratcliff, 1993).
Although there is concern that the loss of control in web-based elicitation leads to participants being frequently distracted (see section 1), and although intermissions could indicate that participants do not report their immediate perceptions as required (see section 2), little has been said about exceedingly long RTs: Whereas Dandurand et al. (2008) propose to exclude participants with single excessively long RTs, Häussler and Juzek (2016) aim at excluding participants who are frequently distracted.[15]
In contrast, short outlying RTs have been commonly used in web-based experiments to identify fraudulent participants who are not complying with the task but are guessing or randomly selecting responses. Following the research tradition, these participants can be called spammers, and can be further subdivided into simple and clever spammers (Häussler and Juzek, 2016). The latter group tends to exhibit long intermissions, resulting in single extreme long RTs, affecting their mean RT such that it becomes inconspicuous. Despite fears that participants became more advanced cheaters with the advent of crowdsourcing platforms (see e.g. Gadiraju et al., 2015), we argue that the identification of spammers is a minor issue due to the increase of control imposed by the choice of an appropriate online platform (see section 4).
Taken together, we propose to combine slow and long outliers into one measure of underperformance as being distracted and being in a rush both are signs of uncooperative behavior, and participants might also switch back and forth between rushing and dawdling around. The identification of outliers, however, is no straightforward issue.

On the one hand, defining absolute cutoff points is far from trivial as it requires *apriori* knowledge about general RTs in a specific task (Berger and Kiefer, 2021). On the other hand, standardized approaches to determine relative cutoff points suffer from being sensitive to outliers. Hence, a single extreme outlier can have an immense effect on measures of average and shape (such as skewness), resulting in the outliers themselves influencing the exclusion criterion (see Miller, 1991; Ratcliff, 1993). Thus, if choosing such a procedure, median and median absolute deviation (MAD) should be preferred above mean and standard deviation as these are more robust against outliers.

For both approaches, it cannot be known whether indeed all and only invalid RTs have been identified as outliers (Berger & Kiefer, 2021), i.e., whether all outliers indeed have been generated by uncooperative behavior. Swamping and masking effects illustrate this problem: the former term applies to situations in which a group of outlying instances skews the mean, making other non-outlying instances look like outliers, the latter term describes a situation in which an outlier is only classified as such as soon as another outlier is removed (Ben-Gal, 2005). What is more, it is known that usually long spurious outliers overlap with long genuine ones (see Ratcliff, 1993): The former are in truth generated by proper behavior on part of the participant, i.e. attentive complying with the task, but they do not differ in magnitude from the latter, which are indeed generated by adverse behavior. In the context of AJTs, this may be caused by the fact that participants exhibit different reading and judging times (see Dandurand et al., 2008; Häussler and Juzek, 2016). To ensure capturing all trials where participants were rushing or dawdling, we need to consider all trials of the entire questionnaire, which contains different types of items (see section 3.1). But what may be a normal RT in case of attention trials may already be considered outlying in case of control trials as the manipulations in the former are designed to be hard to grasp (in particular in case of FC tasks; see section 3.2 and 3.3). Moreover, short RTs in response to control stimuli can be expected as compared to test stimuli as the former must have clear acceptability status (see e.g. Featherston, 2021). Thus, as participants and different (groups of) items exhibit different RT distributions (see Figure 1), differing in just the characteristics determining the exclusion criterion, determining only one general cutoff point would arguably fail in separating spurious from genuine outliers.

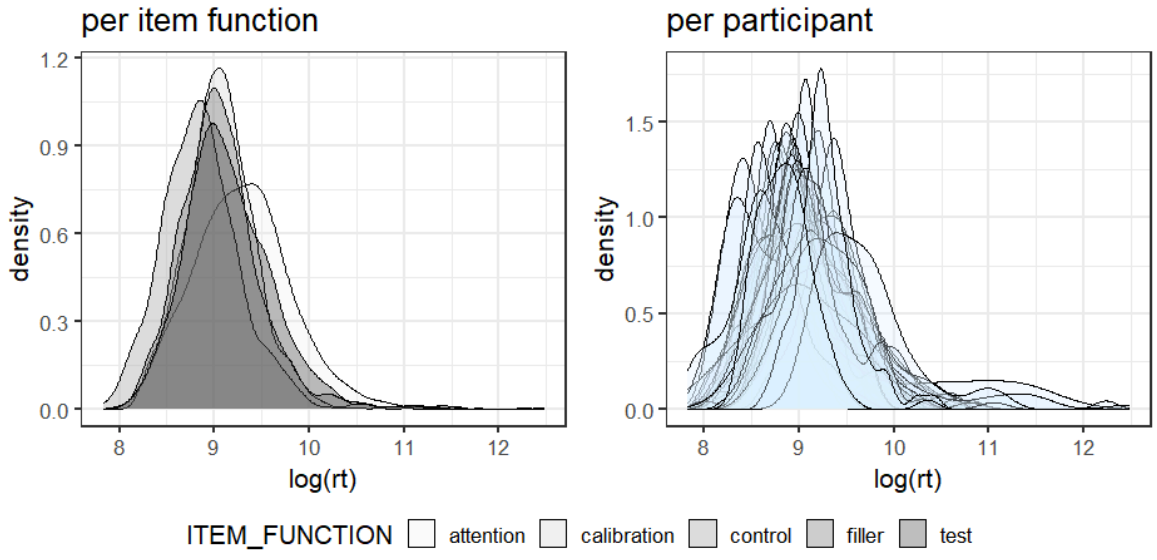

Figure 1: *Density of RTs (Experiment 2 from Kiss et al. 2022)*

Taking these considerations into account, we propose a recursive multi-factorial outlier detection (ReMFOD) for capturing outlying RTs. ReMFOD aims at identifying individual trials as genuine intermissions and rushes. In doing so, ReMFOD accounts for different RT distributions of different participants and item functions, as well as swamping and masking effects. Underpinned by these suspicious individual trials, underperforming participants can be determined by means of proportion of trials not responded to wholeheartedly. We propose to discard participants who have responded genuinely to less than 90 % of trials because they supposedly did not meet the task with the necessary seriousness.

To account for different RT distributions, ReMFOD compares the RT of each trial to a lower and an upper cutoff point, which each consider two cutoff criteria, respectively: The first criterion is computed with respect to the group of trials with the same item function regardless of the participant responding, the second one is computed with respect to all trials of the corresponding participant (regardless of the item function). Only if an RT surmounts or falls below both criteria, it will be designated as a *genuine intermission* (see Equation 1 in (5)) or as a *genuine rush* (see Equation 2 in (5)).[16]

(5) Equation1: cutoff_intermission = max {
median(RTs:participant) + 2.5 x MAD(RTs:participant),
median(RTs:item_function) + 2.5 x MAD(RTs:item_function)}

Equation 2: cutoff_rush = min {
median(RTs:participant) – 1.5 x MAD(RTs:participant),
median(RTs:item_function) – 1.5 x MAD(RTs:item_function)}

To account for swamping and masking effects, the process described above will be repeated on a reduced data set until no more outliers can be found. Therefore, in each iteration step, the cutoff points must be computed afresh. Consider Figure 2 for an illustration of the different cutoff points. Different outlier types, computed with respect to different groups, are marked by different shapes: Box-shaped trials are the only RTs we consider as genuine intermissions or rushes.[17]

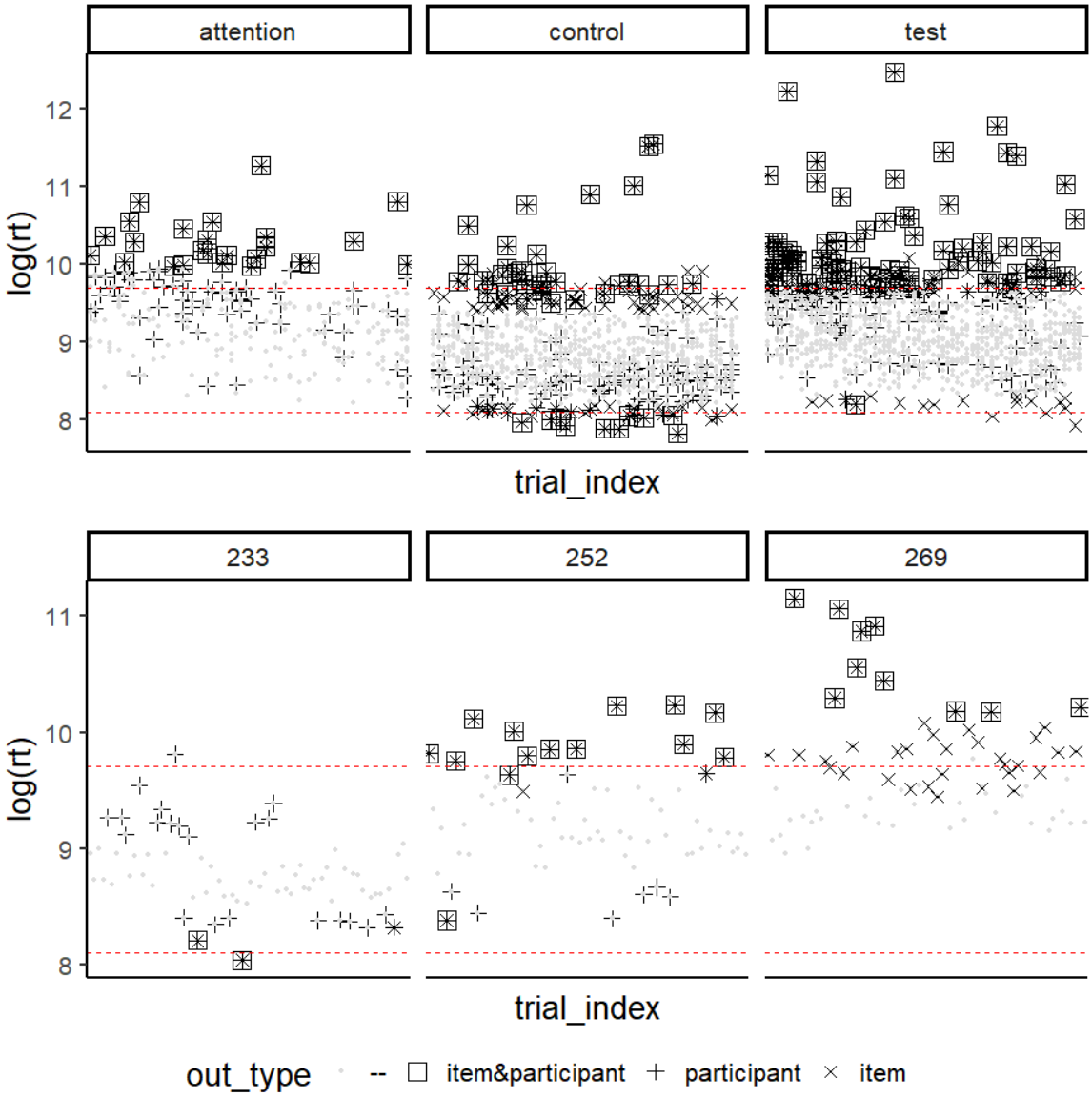

Figure 2: *Recursive Multi-factorial Outlier Detection (ReMFOD)*

The red dotted lines indicate singular absolute cutoff points computed with respect to all trials of all participants: They falsely suspect long genuine RTs of attention trials or slow participants, but are missing out (short and) long outliers in case of (test and) control trials, and are hence indicative of the superiority of ReMFOD.

5.2 Response-Based Identification

After examining the behavior of participants by means of RTs, let us now take a look at the quality of data provided: Section 3.2 and 3.3 have characterized control and attention trials. Now the objective is to define decision criteria for identifying maliciously responding participants. We will provide concrete suggestions on how many trials are needed for a reliable exclusion process, discussing FC and LS studies in turn.

5.2.1 Forced-Choice

As a grammatical control and attention stimulus is directly compared to its ungrammatical version in FC tasks, correct responses are known in advance such that the analysis simply amounts to counting correct choices. The identification of malicious participants is determined by applying a threshold for the proportion of choices required to be correct. It should be self-evident that a threshold should be set in a way that participants perform (well) above chance. But this issue is often neglected in experimental practice, e.g., if researchers assume that the proportion of correct responses should correspond to responding to a single trial correctly by chance, which would amount to a proportion of 50 % (cf. Dandurand et al., 2008). Such an assumption ignores that the sum of trials correctly answered consists of the sum of trials correctly solved and correctly guessed (Frederick and Speed, 2007). A simple method to estimate the number of trials correctly solved is depicted in Equation 3 in (6).

(6) Equation 3: Number of trials correctly solved = Rights – Wrongs

Here, the underlying assumption is that stimuli are so simple to respond to correctly that wrong responses must be wrong guesses. Guessing, then, results in correct responses with equal probability. This assumption holds for AJs if native speakers are employed, and in particular for attention trials, in which the manipulation is not easy to detect. Since scores reached by guessing thus lie in fact within a broader (symmetrical) range around a chance threshold, an assumed threshold of 50 % correctly guessed responses is insufficient. Hence, the required threshold even increases to 90 % correct responses in an FC task (Frederick and Speed, 2007). It should be noted that the decision for a threshold depends on the amount of trials used (see Table 2). If we set – as a rule of thumb – that the probability to pass control trials by guessing should not exceed 5 % (indicated by bold numbers in Table 1), then the proportion of correct responses minimally required decreases as the number of trials increases: When employing six trials, all of them must be responded to correctly, whereas when employing 16 trials, 12 correct responses are sufficient. In our view, it is too strict to assume that all trials need to be answered correctly as even diligent participants may become subject to fatigue in a comparatively long questionnaire. Thus, we strongly argue against using less than six but for employing at least eight attention/control trials in 2AFC tasks.

Table 1
Probabilities of responding to ≥ k out of N (2AFC) trials correctly by chance (see Equation 6)

N	k									
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**
4	.688	.312	.062							
6	.891	.656	.344	.109	**.016**					
8	.965	.855	.637	.363	.145	**.035**	**.004**			
10	.989	.945	.828	.623	.377	.172	.055	**.011**	**.001**	
12	.997	.981	.927	.806	.613	.387	.194	.073	**.019**	**.000**
14	.999	.994	.971	.910	.788	.605	.395	.212	.090	**.006**
16	1.000	.998	.989	.962	.895	.773	.598	.402	.227	**.038**

5.2.2 Likert-Scale

Attention Trials. Attention trials in LS studies differ from control trials in that only ungrammatical conditions will be presented to participants because experimental conditions are presented in isolation in LS, and grammatical stimuli will fail to indicate attention. Thus, like for FC trials, the analysis amounts to counting correct responses for attention trials. This raises the issue, however, whether the neutral point of an uneven scale should count as correct response or not.

Consider that the neutral point should qualify as acceptable first. If we assume a probability of less than 5 % to pass a test by chance and combine it with the qualification that not all trials need to be answered correctly, it follows that nine out of ten trials need to be responded to correctly (see Table 2a). If we take the neutral point not to be acceptable, this number is reduced to five correct responses out of six trials (see Table 2b). As the neutral point arguably does not reliably indicate that the participant has indeed spotted the manipulation, we recommend excluding it from the acceptable responses.

Table 2

Probabilities of responding to ≥ k out of N unacceptable (5-pt-LS) trials correctly by chance (see Equation 5 below). We test different options for p (chances to respond correctly) and q (chances to respond incorrectly) as these depend on whether the neutral point (3) is considered correct or not (for unacceptable stimuli).

N	*k*									
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**
4	.821	.475	.130							
6	.959	.821	.544	.233	**.047**					
8	.911	.955	.826	.594	.315	.106	**.017**			
10	.998	.988	.945	.834	.633	.382	.167	**.046**	**.006**	
12	1.000	.997	.985	.943	.842	.665	.438	.225	.083	**.002**
14	1.000	.999	.996	.982	.942	.850	.692	.486	.279	**.040**
16	1.000	1.000	.999	.995	.981	.942	.858	.716	.527	.167

a. chances are at 0.6 (ratings 1,2,3) to respond correctly and at 0.4 (ratings 4,5) to respond wrongly

N	*k*									
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**
4	.525	.179	**.026**							
6	.767	.456	.179	**.041**	**.004**					
8	.894	.685	.406	.174	.050	**.009**	**.001**			
10	.9554	.833	.618	.367	.166	.055	**.012**	**.002**	**.000**	

12	.980	.917	.775	.562	.335	.158	.057	**.015**	**.003**	**.000**
14	.992	.960	.876	.721	.514	.308	.150	.058	**.018**	**.001**
16	.997	.982	.935	.833	.671	.473	.284	.142	.058	**.005**

a. chances are at 0.4 (ratings 1,2) to respond correctly and at 0.6 (ratings 3,4,5) to respond wrongly

Control Trials. As the grammatical and ungrammatical conditions of a control item are shown in isolation in an LS, the analysis is more complicated than in FC tasks. In general, there are two different practices to define cooperative behavior on LS control trials, as depicted for a 5-pt-LS in (7).

(7) **positional**: all stimuli should receive expected ratings, i.e.
< 3 for ungrammatical stimuli
≥ 3 for grammatical stimuli

relational: grammatical stimuli should be rated higher than ungrammatical ones, i.e.
mean(grammatical) > mean(ungrammatical)

In a relational account the only concern is that grammatical control stimuli are judged in average better than ungrammatical ones, irrespective of the positioning on the scale, whereas the positional account assumes that both conditions need to be on the intended side of scale, since control stimuli are designed to be maximally unacceptable or maximally acceptable (in context of the study). Furthermore, the two groups – ungrammatical and grammatical controls – are dependent on each other as concerns the evaluation in a relational account but not in a positional account.
On a positional account, the assessment of participants' performance is carried out separately for grammatical and ungrammatical stimuli, focusing on the pertinent side of the scale in each case. One point of discussion of the positional account is again how to treat the neutral point. We recommend allowing the neutral point as legitimate response to grammatical stimuli but not to ungrammatical stimuli as it does neither reliably indicate the rejection of the grammatical version nor the rejection of the ungrammatical version. The chance to pass control trials is then the joint probability of passing the two groups individually. Consequently, less strict requirements can apply to the individual groups, i.e. (un)grammatical controls, by themselves. If we consider group sizes (N) from 2–8, and also allowing for different decision criteria (i.e. k out of N correct responses required), as well as considering the neutral point as correct response for grammatical but not for ungrammatical stimuli, we receive various joint probabilities of less than 5 % by multiple decision criteria, some of which are depicted in Table 3.[18] Thus, under the assumptions listed above, at least ten control trials are needed – namely five grammatical and five ungrammatical – of which participants are required to answer at least four per group correctly (i.e. < 3 for ungrammatical and ≥ 3 for grammatical trials) in order to be considered cooperative. Together with the minimum amount of six attention trials, we hence need at least 16 fillers for controlling functions. Although this number might appear extensive, it should be no cause of concern: Even in the smallest factorial designs (i.e. 1×2 or 2×2), the questionnaire should include at least 24 or 48 filler trials, when employing the minimum TRF of 1:2 (see Section 3.1).

Table 3
Joint probabilities of passing grammatical as well as ungrammatical controls on a positional account

N	grammatical (≥3)		ungrammatical (<3)		joint
	k	*p*	*k*	*p*	
5	4	.337	4	.087	.029
6	5	.233	4	.179	.042
6	4	.544	5	.041	.022
6	5	233	5	.041	.010
7	6	.159	4	.290	.046
7	5	.420	5	.096	.040
7	6	.159	5	.096	.015
7	5	.420	6	.019	.008
7	6	.159	6	.019.	.003
8	5	.594	6	.050	.030
8	6	.315	6	.050	.016

The positional account is thus intuitive and does not pose any problems. Surprisingly, we are only aware of a single advocate of the positional account (Zanuttini, Wood, Zentz and Horn, 2018), which however, differs from the present account in major respects. First, Zanuttini et al. (2018) do not allow for slips: They reject participants who used the wrong side of scale (excluding the neutral point) at least once. As mentioned above, we are afraid that this is overly strict when employing a sufficiently large number of control trials. Secondly, they do not tolerate participants with an average rating > 2 for ungrammatical and < 4 for grammatical control items. Thirdly, the status of the neutral point is fuzzy: to avoid a clear declaration of the neutral option, Zanuttini et al. (2018) employ mean values, as is done in the relational account,

The more prominent relational account, as e.g., advocated for in Häussler and Juzek (2016), is confronted with two major problems. First, an analysis of LS data based on mean values requires a specific number of trials to justify an interpretation of LSs as intervals. Carifio and Perla (2008:1150) report that 'at least eight reasonably related trials' produce interval data, while individual trials produce ordinal data. Thus, it should be noted that when using only four control trials, like Häussler and Juzek (2016) did, a treatment of ratings as interval data is not justified. Instead, using at least eight control trials (in each condition) is necessary to properly analyze these trials in a relational account.

Secondly, it allows the acceptance of participants not distinguishing ungrammatical from grammatical stimuli. This might happen if grammatical stimuli are rated as more natural than ungrammatical ones, but both are placed on the same side of the scale. In this case, participants do not use the scale as intended, and their judgments will introduce noise. Relational accounts thus fail to control for scale biases (especially compression).

Thus, we could consider an extended relational account requiring a minimum distance between both conditions, expressed as the proportion of the maximal possible distance between ratings for grammatical and ungrammatical control trials, and set a threshold of at least 50 % of the optimal distance, as defined in Equation 4 in (8).

(8) Equation 4: $$\frac{\text{mean}\left(grammatical\right) - \text{mean}\left(ungrammatical\right)}{\max\left(LSrating\right) - \min(LSrating)} \geq 0.5$$

Thus, for a 5-pt-LS the minimum distance of 2 ($0.5 * (5 - 1)$) between both conditions is required. By this, both conditions being on the same pole of the scale is forbidden, and the issue of scale compression is circumvented. However, the state of the neutral point is again fuzzy.[19]

The motivation of a threshold of 0.5 is purely conceptual, its major intention being to control for scale compression. The specific threshold could be justified by determining the chances of guessing participants to pass control trials on relational accounts, analogous to estimating the FC guessing probabilities: We determine the proportion of outcomes out of the number of possible outcomes under which the decision criterion is fulfilled. To allow for comparisons of mean values, let us consider pairs of responses each consisting of one grammatical and one ungrammatical trial. On a 5-pt-LS, there are 25 different ways to respond to two trials (i.e. 1-1, ..., 5-5). All combinations of these amount to 625 possible outcomes for two pairs of trials (e.g. <1-1, 1-1>, ..., <5-5, 5-5>), and so on. For each possible outcome, we compute mean values for grammatical and ungrammatical trials, and the proportion of the optimal distance (s) as in Equation 4, and decide whether the distance between conditions is acceptable under the relational account, i.e. > 0 and Equation 4, i.e. ≥ 0.5.

Under the assumption that each outcome is equally probable when responding randomly, we can estimate the probabilities to pass control trials by chance by determining the proportion of outcomes passing the decision criteria (see Table 4). To be effective against random responding, the criterion should hence aim at excluding most (i.e. 95 %) of scores reached by random responding, this is, again, the rule of thumb that guessing chances should be less than 5 %.[20] The relational account in fails on this rule of thumb: more than 40 % of possible outcomes will be accepted by means of the relational account in every case, with numbers even increasing the more pairs are considered. The relational account hence does not only suffer from unfulfilled requirements and conceptional problems (acceptance of scale compression) but has also small chances to detect guessing participants (see Table 4). As the extended relational account has solved the issue of scale compression, the chances to detect guessing participants are much better and even sufficient when employing three pairs, i.e., six control trials in total. That is, however, no advantage when compared to the positional account as eight control trials are needed per group (i.e., 16 in total) nonetheless to treat LS as interval data. The extended relational account thus fares better in detecting random responding than the original but provides no further advantages over the much simpler positional account. The same considerations apply to the proposal by Zanuttini et al. (2018).

Table 4

Acceptance rates of guessed outcomes, i.e. all possible response combinations, in relational accounts

pairs	possible outcomes	pure:	extended:
		$s > 0$	$s \geq 0.5$

1	25	0.4000	0.2400
2	625	0.4320	0.1120
3	15,625	0.4440	0.0449
4	390,625	0.4511	0.0296
5	9,765,625	0.4561	0.0140
6	244,140,625	0.4598	0.0076
7	6,103,515,625	0.4627	0.0045
8	152,587,890,623	0.4651	0.0026

In comparison, the positional account shows various advantages over its competitors: First, it controls for scale compression *per se*. Secondly, there are no underlying assumptions to fulfill an interpretation of an interval scale, required to determine mean values. Thirdly, the minimum number of control trials to reliably detect guessing participants under the positional account is ten, while the relational account as well as Zanuttini et al. (2018) require at least 16 control trials (eight grammatical, eight ungrammatical).
Although the treatment of the neutral point is a controversial affair, the positional account can be easily adapted to various interpretations of its status. This may also include allowing for a maximal proportion of trials answered with the neutral point. By this, its state would be less fuzzy, and the decision criterion by itself would be clearly and intuitively understandable. Another difference between the present account and Zanuttini et al. (2018) consists in the tolerance to slips: The more control trials are required, and the longer the questionnaire, the stronger becomes the argument of allowing slips.

6 Conclusion

We have discussed different sources of noise in experimental data, and provided various measures against them, which apply to all stages of an AJ study. This does concern the instructions, which need to ensure that participants grasp what they ought to judge. Moreover, the questionnaire design must manage to control certain biases and effects, in particular to prevent conscious response strategies, and provide sufficient means to enable the identification of inapt participants. To strengthen the concealment of test trials by fillers, we recommend employing so-called pseudo fillers, which are picking up conspicuous characteristics of test items, and related control items (in addition to unrelated ones), which are designed to be very similar in structure to test items such that they further enhance the robustness of the detection of malicious participants. To complement means to identify inattentive, underperforming participants, we introduced attention stimuli as an alternative to IMCs, which we consider infeasible for AJTs. However, to reduce the amount of malicious participants, the recruitment platform needs to be carefully selected as these differ in being geared towards HITs or towards research tasks such as AJTs (with respect to the structure of the task and eligibility of participants). Research oriented platforms do not only alleviate some of the problems resulting from the loss of supervision but also allow for prescreening. Still, researchers are not exempted from the identification of potential inapt participants via the inspection of response times as well as of responses. Regarding the former, we argue that the identification of spammers is a minor issue due to the increase of control on part of online

platforms. Thus, we propose to additionally combine slow and long outliers into one measure of underperformance. However, detecting outliers is far from trivial due to masking and swamping effects as well as different RT distributions of participants and different item functions. We thus proposed the ReMFOD methodology, accounting for different cutoff points repeatedly, to identify individual trials as genuine intermissions as well as rushes. Regarding the inspection of responses, participants are at least required to perform well above chance. To determine the amount of control trials needed, we investigated probabilities of passing control trials by guessing, and set a rule of thumb that these should neither exceed 5 % nor be based on requiring perfection (to respond to all trials correctly). The results indicate that at least eight trials per questionnaire are necessary for 2AFC tasks, whereby participants are required to respond to seven of them correctly. Regarding the analysis of LS control trials, we have reviewed relational and positional accounts and argued for the use of the latter due to major drawbacks of the former come (interpretation as interval data and failure to control for scale compression). A major point of discussion of the positional account is how to treat the neutral point: We propose to include it as an acceptable option for grammatical but not for ungrammatical stimuli. Considering joint guessing probabilities of both conditions evaluated individually, only ten control trials are needed – whereby participants are required to give correct responses to four out of five trials in each group. Taking all these measures together, researchers can be confident that the data entering analysis reaches the highest attainable level in web-based elicitation of AJTs.

Guessing Probabilities

Frederick and Speed (see appendix of 2007) refer to the standard binomial expansion as a standard way to compute the probability of a score for FC. Equation 5 in (9) gives the probability of (exactly) k correct answers out of N trials:

(9) Equation 5: $\frac{N!}{k!(N-k)!)} p^k q^{N-k}$ where
p = probability of a correct response
q = probability of an incorrect response

When p = q, as in case of two answer choices as in 2AFC, this formula can be simplified (see Equation 6 in (10)).

(10) Equation 6: $\frac{N!}{k!(n-k)!} p^N$

To obtain the probability of answering at least k out of N, e.g., half of the items, right (by chance), we need the sum of all probabilities from k to N (cumulative probabilities).

Open Practices Statement

The data and materials for the current study are available at
https://github.com/Linguistic-Data-Science-Lab/AJTs-eligibility-screening

References

Ben-Gal, I. (2005) Outlier detection. In O. Maimon and L. Rokach (eds.) *Data Mining and Knowledge Discovery Handbook* 131–146. Springer US. https://doi.org/10.1007/0-387-25465-X_7

Berger, A. and Kiefer, M. (2021) Comparison of different response time outlier exclusion methods: A simulation study. *Frontiers in Psychology* 12: 675558. https://doi.org/10.3389/fpsyg.2021.675558

Carifio, J. and Perla, R. (2008) Resolving the 50-year debate around using and misusing Likert scales. *Medical Education* 42:1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x

Chandler, J. and Paolacci, G. (2017) Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science* 8(5): 500–508. https://doi.org/10.1177/1948550617698203

Cornips, L. and Poletto, C. (2005) On standardising syntactic elicitation techniques (part 1). *Lingua* 115(7): 939–957. https://doi.org/10.1016/j.lingua.2003.11.004

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

Featherston, S. (2021) Response methods in acceptability experiments. In G. Goodall (ed.),*The Cambridge Handbook of Experimental Syntax* 39–61. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108569620.003

Frederick, R. I. and Speed, F. (2007) On the interpretation of below-chance responding in forced-choice tests. *Assessment* 14: 3–11. https://doi.org/10.1177/1073191106292009

Freeman, L. and Greenberg, S. (2019) How do I screen participants for my study? 6 steps to screening success. https://www.positly.com/screening-research-participants/

Gadiraju, U., Kawase, R., Dietze, S. and Demartini, G. (2015) Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* 1631–1640. New York: Association for Computing Machinery. https://doi.org/10.1145/2702123.2702443

Gerbrich, H., Schreier, V. and Featherston, S. (2019) Standard items for English judgment studies: Syntax and semantics. In S. Featherston, R. Hörnig, S. von Wietersheim and S. Winkler (eds.) *Experiments in Focus* 305–328. Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110623093-012

Gibson, E. and Fedorenko, E. (2013) The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes* 28(1–2): 88–124. https://doi.org/10.1080/01690965.2010.515080

Gibson, E., Piantadosi, S. and Fedorenko, K. (2011) Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass* 5: 509–524. https://doi.org/10.1111/j.1749-818X.2011.00295.x

Goodall, G. (2021) Sentence acceptability experiments: What, how, and why. In G. Goodall (ed.) *The Cambridge Handbook of Experimental Syntax* 7–38. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108569620.002

Haider, H. (2010) *The Syntax of German*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511845314

Häussler, J. and Juzek, T. S. (2016) Detecting and discouraging non-cooperative behavior in online experiments using an acceptability judgment task. In H. Christ, D. Klenovšak, L. Sönning and V. Werner (eds.) *A Blend of MaLT* 73–99. University of Bamberg Press.

Häussler, J. and Juzek, T. S. (2021) Variation in participants and stimuli in acceptability experiments. In G. Goodall (ed.), *The Cambridge Handbook of Experimental Syntax* 97–117. Cambridge: University Press. https://doi.org/10.1017/9781108569620.005

Juzek, T. S. (2016) Acceptability judgement tasks and grammatical theory (PhD thesis). Oxford: Jesus College.

Kiss, T., Pieper, J. and Börner A. K. (2022) Supplementary material for Kiss et al. 2021*Word order constraints on event-internal modifiers* (v2.0). https://doi.org/10.5281/zenodo.5898464

Leivada, E. and Westergaard, M. (2020) Acceptable ungrammatical sentences, unacceptable grammatical sentences, and the role of the cognitive parser. *Frontiers in Psychology* 11. https://doi.org/10.3389/fpsyg.2020.00364

Litman, L., Rosenzweig, C. and Moss, A. (2021) New solutions dramatically improve research data quality on MTurk. https://www.cloudresearch.com/resources/blog/new-toolsimprove-research-data-quality-mturk/

Mason, W. and Suri, S. (2012) Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44(1): 1–23. https://doi.org/10.3758/s13428-011-0124-6

Maynes, J. and Gross, S. (2013) Linguistic intuitions. *Philosophy Compass* 8: 714–730. https://doi.org/10.1111/phc3.12052

Miller, J. (1991) Reaction time analysis with outlier exclusion: Bias varies with sample size. *The Quarterly Journal of Experimental Psychology. Section A, Human Experimental Psychology* 43(4): 907–912. https://doi.org/10.1080/14640749108400962

Musch, J. and Reips, U.-D. (2000) A brief history of web experimenting. In M. H. Birnbaum (ed.) *Psychological Experiments on the Internet* 61–87. New York, Boston: Academic Press. https://doi.org/10.1016/B978012099980-4/50004-6

Myers, J. (2009) The design and analysis of small-scale syntactic judgment experiments. *Lingua* 119(3): 425–444. https://doi.org/10.1016/j.lingua.2008.09.003

Oppenheimer, D. M., Meyvis, T. and Davidenko, N. (2009) Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45(4): 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Paolacci, G. and Chandler, J. J. (2014) Inside the Turk. *Current Directions in Psychological Science* 23: 184–188. https://doi.org/10.1177/0963721414531598

Patterson, C. (2014) The role of structural and discourse-level cues during pronoun resolution (PhD thesis). Potsdam: University of Potsdam.

Phillips, C. (2009) Should we impeach armchair linguists? In S. Iwasaki, H. Hoji, P. M. Clancy and S.-O. Sohn (eds.) *Japanese-Korean Linguistics 17* 49–64. Stanford: CSLI Publications.

Prolific Team. (2018a) Can I screen participants within my survey? https://researcher-help.prolific.co/hc/en-gb/articles/360010165173-Can-I-screen-participants-withinmy-survey-

Prolific Team. (2018b) Using our demographic filters to prescreen participants. https://researcher-help.prolific.co/hc/en-gb/articles/360009221093-Using-our-demographicfilters-to-prescreen-participants

Prolific Team. (2021) Using attention checks as a measure of data quality. Retrieved April 8, 2021, from https://researcher-help.prolific.co/hc/en-gb/articles/360009223553

Ratcliff, R. (1993) Methods for dealing with reaction time outliers. *Psychological Bulletin* 114(3): 510–532. https://doi.org/10.1037/0033-2909.114.3.510

Robinson, J., Rosenzweig, C. and Litman, L. (2020) The Mechanical Turk ecosystem. In L. Litman and J. Robinson (eds.) *Conducting Online Research on Amazon Mechanical Turk and Beyond* 27–47. Sage Academic Publishing.

Schnoebelen, T. and Kuperman, V. (2010) Using Amazon Mechanical Turk for linguistic research. *Psihologija* 43(4): 441–464. https://doi.org/10.2298/PSI1004441S

Schütze, C. T. (1996) *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.

Schütze, C. T. and Sprouse, J. (2014) Judgment data. In R. J. Podesva and D. Sharma (eds.) *Research Methods in Linguistics* 27–50. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139013734.004

Snyder, W. (2021) Satiation. In G. Goodall (ed.) *The Cambridge Handbook of Experimental Syntax* 154–180. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108569620.007

Sorace, A. and Keller, F. (2005) Gradience in linguistic data. *Lingua* 115: 1497–1524. https://doi.org/10.1016/j.lingua.2004.07.002

Sprouse, J. (2011a) A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language* 87: 274–288. https://doi.org/10.1353/lan.2011.0028

Sprouse, J. (2011b) A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43(1): 155–167. https://doi.org/10.3758/s13428-010-0039-7

Sprouse, J. and Almeida, D. (2017) Design sensitivity and statistical power in acceptability judgment experiments. *Glossa* 2: 1–32. https://doi.org/10.5334/gjgl.236

Sprouse, J., Schütze, C. T. and Almeida, D. (2013) A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua* 134: 219–248. https://doi.org/10.1016/j.lingua.2013.07.002

Weskott, T. and Fanselow, G. (2009) Scaling issues in the measurement of linguistic acceptability. In S. Featherston and S. Winkler (eds.) *The Fruits of Empirical Linguistics. Volume 1: Process* 229–246. Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110216141.229

Zanuttini, R., Wood, J., Zentz, J. and Horn, L. (2018) The Yale Grammatical Diversity Project: Morphosyntactic variation in North American English. *Linguistics Vanguard* 4(1): 20160070. https://doi.org/10.1515/lingvan-2016-0070

---

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

[2] https://www.clickworker.com/

[3] The intuitive understanding of an LS can be supported by using labels (from 'entirely unnatural' to 'perfectly natural') instead of or additionally to numbers and by using different colors.

[4] An alternative reduction strategy is to employ a mixed design with a between-subject variable. However, between-groups designs are less sensitive (Myers, 2009).

[5] use of only a limited range

[6] use of only one end

[7] It is advisable to randomize not only the distribution of item functions but also of acceptability degrees (Cornips and Poletto, 2005; Zanuttini et al., 2018).

[8] Goodall (2021) estimates that three to five trials are sufficient, whereas Schütze and Sprouse (2014) recommend five up to ten trials. Gerbrich, Schreier and Featherston (2009) even recommend a higher number of standard items.

[9] For reasons of space, the items will be presented without the addenda from now on. The interested reader is referred to the supplementary data.

[10] Nonetheless, the possibility exists to conduct screening questions or screening studies beforehand, and then to only allow those participants that qualified in the screening to your actual study (e.g., Robinson, Rosenzweig and Litman, 2020). However, designing screening tasks is not straightforward as the questions should disguise what population is targeted to prevent participants from cheating to gain access to studies (Freeman, 2019).

[11] https://www.prolific.co/

[12] https://www.cloudresearch.com/

[13] https://www.positly.com/

[14] Nonetheless, Prolific Team (2018b) notes that prescreening information is self-reported and cannot be verified (except for maybe their current country of residence). Thus, they suggest validating the screening questions by repeating them in your survey.

[15] Distractions could indicate that participants are re-reading the stimulus multiple times or that they are consulting with a third person.

[16] Miller (1991) proposes the values of 3 (very conservative), 2.5 (moderately conservative) or even 2 (poorly conservative). Häussler and Juzek (2016) suggest using an asymmetric criterion (using standard deviations) of -1.5 for the lower and +4 for the upper cutoff point.

[17] Note that the shapes may overlap as these outliers have been computed by various procedures differing in the groups they included to identify outliers.

[18] Due to space restrictions, the table only displays options fulfilling the following conditions: each probability (passing grammatical and ungrammatical trials) needs to be below 0.6 and k needs to be less than N in both conditions.

[19] The extended relational account differs from Zanuttini et al. (2018) as the score is not conditioned on the position of the scale. Thus, mean responses to grammatical trials of 3 and to ungrammatical trials of 1 are only acceptable under the extended relational account. As the stricter account of Zanuttini et al. (2018) does not allow for slips, it rejects a good part of the outcomes accepted by the extended relational account.

[20] Apart from estimating the guessing probabilities for the given relational accounts, the outcomes can hence serve a further purpose: namely to deduce a threshold instead of setting it conceptually. That is to say, the 95% quantile of the distances reached by the possible outcomes corresponds to the threshold on distances required to hold chances to pass control trials at 5 %. For one pair the 95 % quantile lies at 0.75, for three pairs at 0.50, and for five pairs at 0.35. Thus, when using more trials, the conceptual score of 0.5 becomes stricter than required with regard to random responding, but is still advisable to avoid scale compression.