# 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: Such effects may emerge due to the loss of control in now widely favored unsupervised web-based elicitation. But noise may also be task-related, namely if participants lack understanding and do not satisfy the underlying assumptions. Finally, also the researchers themselves may be held responsible for noise if they employ a poor questionnaire design that fails to control for biases and adverse effects, and lacks 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 on careful constructions of appropriate filler and control items, and induce 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 present reflections on how to employ analyses of general *response times* as well as of *responses* to the specialized items so that potential inapt participants are *reliably* detected by latency- and response-based methods.

**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 as well as accurate measurements of response times (see Musch & 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 Gibson et al., 2011; Sprouse, 2011b), 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 Schnoebelen & Kuperman, 2010; Gibson et al., 2011; Sprouse, 2011b, for linguistic research) and have proven to only differ in terms of higher dropout-rates (Dandurand et al., 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 due to participants' behavior in online studies (see Dandurand et al., 2008; Oppenheimer et al., 2009). Many researchers have been concerned with identifying fraudulent participants who do not comply with the task (see, among others, Buchholz & Latorre, 2011; Gadiraju et al., 2015; Häussler & Juzek, 2016; Chandler & Paolacci, 2017) as it has indeed been observed that malicious behavior in crowdsourcing is virulent (see Downs et al., 2010; Eickhoff & de Vries, 2013).

The remaining paper is structured as follows: Section 2 discusses AJTs in general as well as its underlying assumptions and prerequisites in particular, and how these can be reflected in the instructions. The impact of the 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 as well as 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 in the absence of disturbing factors), acceptability takes factors of performance into account (Dąbrowska, 2010). 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 in an informal or an experimental setting, to make sure that all

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

confounding factors are excluded. If this is the case, we can assume that a sentence will be judged as acceptable if it is grammatical.

Acceptability itself is assumed to be a percept which is, like other percepts (e.g. pain), 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 & 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; Sprouse & Almeida, 2013). 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 Devitt, 2006; Bader & Häussler, 2010; Maynes & Gross, 2013; Schütze & Sprouse, 2014; Häussler & Juzek, 2020). A further characteristic of acceptability is its gradience (Bolinger, 1961), which has been reflected in the amplification of diacritics used by syntacticans (such as * for unacceptable, and ?, ??, and ?* for decreasing states of acceptability). But the meaning of the diacritics has never been formally defined, leading to inconsistent usage (see Sorace & Keller, 2005; Bader & Häussler, 2010).

Arguably, complex patterns of acceptability cannot reliably be accessed with traditional linguistic methods, as e.g. by micro-experiments where linguists ask themselves (or a few colleagues) whether a sentence is acceptable or not (see Keller, 2000b; Featherston, 2005; Sorace & Keller, 2005; Wasow & Arnold, 2005; Phillips, 2009). As current research heavily relies on gradient, i.e. subtle and potentially controversial judgments (see Schütze, 1996), linguists now generally agree to collect judgments from naïve speakers using AJTs in a controlled experimental setting, i.e. by employing a factorial design, where each experimental condition is examined by multiple lexicalizations (see Gibson & Fedorenko, 2013).

To elicit AJs from naïve subjects, well-known procedures used for sensory or social stimuli can be utilized (Bard et al., 1996; Cowart, 1997; Sorace & Keller, 2005). The four broadly accepted types of AJTs are yes-no task (YN), Likert-Scale (LS), Magnitude Estimation (ME), and Forced-Choice (FC). YN and LS elicit *absolute* statements because individual stimuli are judged in isolation. One can thus assume that they are assessing acceptability directly, whereas ME and FC take an indirect, or *relative*, approach and compare one stimulus to another. Sprouse & Almeida (2017) found that the FC task is (among the four AJTs presented above) the most sensitive when it comes detecting differences between two conditions, whereas the YN task proved to be the least sensitive. LS and ME tasks are about equally sensitive.[3] In general, AJTs have been proven to be reliable and robust, producing consistent judgments (Cowart, 1997; Sprouse, 2011b; Sprouse & Almeida, 2012; Sprouse et al., 2013; Sprouse & Almeida, 2017; Leivada & Westergaard, 2020).

In an LS, participants are asked to judge stimuli using 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

[3] However, previous research suggests that participants treat ME as a modified LS task due to their incapability to carry out the ME task properly when faced with linguistic material (Sprouse, 2011a; Weskott & Fanselow, 2011; Sprouse & Almeida, 2017).

be uneven by allowing five or seven points on the scale (5-pt-LS, 7-pt-LS). Odd scales of course provide a neutral point, to which participants can retreat if they have no clear opinion. Hence, forced random choices can be avoided. Sometimes, however, the scale is chosen to be even, e.g. by using a 4-pt-LS, to coerce participants into resolving their indecision (Schütze & Sprouse, 2014).

FC tasks provide an intuitive form of eliciting relative judgments. When used for AJTs, usually only two options are provided, and the task is hence often called Two-Alternative Forced-Choice (2AFC). From this (minimal) pair of alternatives, participants simply choose the one which they consider more acceptable. By this, 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 & Sprouse, 2014).

It holds for all types of AJTs that when conducting experiments, one needs to make sure that participants indeed base their decision on the *acceptability* of stimuli despite not being familiar with this linguistic concept. As one cannot assume that naïve speakers grasp the intended meaning (see Schütze, 1996), the concept of naturalness is often used to avoid misinterpretations of the term acceptability, e.g. as judgments based on prescriptive rules (see Featherston, 2021).[4] The following instruction (taken from Wasow & Arnold, 2005) is illustrative in preventing prescriptive judgments.

(1) Rely on your own intuitions of what sounds good, not on what you think is correct according to the experts.

As was already pointed out, experimental linguistics assumes that stimuli are assessed spontaneously by participants, and hence, this should be made clear to the participants as well. By this, they won't have time to apply norms such that these instructions help to avoid judgments based on prescriptive rules (see Phillips, 2009). It is remarkable that, although various suggestions can be found in the literature to achieve spontaneity, such as refraining from changing any given response or from rereading sentences already judged (Schütze, 1996), the ban on reflective (conscious) judgments, where temporal immediacy is not given, has not always been imposed in linguistic practice (as pointed out by Maynes & 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.

[4] The intuitive understanding of an LS can be supported by using labels instead of or additionally to numbers and by highlighting these in different colors (e.g. from red over gray to blue). The scale labels could thus read as follows: 'entirely unnatural' – 'rather unnatural' – 'no tendency' – 'rather natural' – 'perfectly natural'.

### 3.1. General properties of questionnaires

The starting point of each study is marked by 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 (Featherston, 2007; Gibson & Fedorenko, 2013). Moreover, employing a higher amount of lexicalizations (as well as of participants) increases statistical power, i.e. the ability to detect real differences (see Sprouse & Almeida, 2017; 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 (and the factorial design). Their importance stems from the need to exclude certain effects and biases, confirmation- and order-biases in particular but also scale-biases.

The objective of filler trials is to complement a questionnaire so that participants cannot identify test trials and thereby the research subject, which participants should be kept unaware of to avoid confirmation biases (see Wasow & Arnold, 2005; Gibson & Fedorenko, 2013; Schütze & Sprouse, 2014). They further serve 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. The risk of strategic responding is particularly high if participants are subjected only to test trials but is reduced the higher the test-filler-ratio (TFR) is (Schütze & Sprouse, 2014).

As it has been argued that questionnaires should be limited to about 100 trials to avoid risking fatigue (Sprouse et al., 2013; Sprouse & Almeida, 2017), the length of the questionnaire hence becomes an area of multiple conflicts: between the necessity to avoid confirmation biases and strategic responding (*high TFR*) as well as 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 due to this conflict. Regarding the minimum requirement for LPC, Schütze & 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 restrictions of questionnaires length. Admitting possible challenges in creating many test stimuli (see also Goodall, 2021; Häussler & Juzek, 2021), we infer that using at least six lexicalizations should be an absolute minimal standard. Concerning the TFR, we argue that at least two fillers for each test trial (1:2) should be employed (see Cowart, 1997; Weskott & Fanselow, 2009) because it is advisable to make sure that filler trials intersperse with test trials in questionnaires such that no adjacent test trials are found, thus impeding strategy building on the part of participants as well as immediate repetition priming resulting from adjacent stimuli of the same experimental condition (see Häussler & 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 can be determined as the sixfold of this number. The number of required fillers is in turn determined as the twofold amount of test trials. As in an FC task trials consist of multiple experimental conditions to pick from, one independent factor becomes the dependent factor such that the factorial design is reduced by this independent variable. Thus, changing the AJT from LS to an 2AFC reduces the length of the questionnaire by half.[5]

One more point to consider is the order of individual stimuli in a questionnaire due to the necessity to counterbalance the influence of *order effects*. On top of practice and fatigue effects, judgments to (or choices of) individual stimuli are influenced by judgments to other stimuli, i.e. judgments are relative to each other and may thus be influenced by the presentation order of stimuli, also known as *carryover effects* (Bard et al., 1996; Weskott & Fanselow, 2009). Not only has it been demonstrated experimentally that the repeated exposure of (some) ungrammatical or borderline structures can make them sound more acceptable such that judgments improve in the course of the questionnaire (see Cornips & Poletto, 2005; Dąbrowska, 2010), but these can also coexist with 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, where (minimal) pairs of two experimental conditions are presented, the presentation order of the conditions also needs to be randomized to avoid response biases to this effect, possibly leading to false effects (see Sprouse & Almeida, 2017).

Moreover, in a within-items design, a test item will consist of stimuli for each experimental condition, which only differ with regard to the factors manipulated. This design is desirable to 'ensure that judgment differences between conditions are as much as possible due to the independent variables under investigation' (Häussler & Juzek, 2021, p. 100). Therefore, 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 due to lexical material (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 Gibson & Fedorenko, 2013; Schütze & Sprouse, 2014).[6]

Of course, not only test items but also filler items need to be carefully constructed. To avoid *scale biases*, acceptability degrees of filler stimuli need to be carefully selected such that in LS-studies the whole scale will be used (Schütze, 1996). This is done by including

[5] An alternative strategy to reduce questionnaire lengths is to employ a mixed design with a between-subject variable: Thereby, one would employ different groups of subjects (one for each level) and each group only is shown one level. However, between-groups designs are less sensitive and only recommended if they are without alternative as potential incidental systematic differences between groups will be confounded with the experimental factor (Myers, 2009).

[6] However, creating (plausible) lexicalizations for each experimental condition can be cumbersome such that within-items designs are often hard to achieve, especially for designs beyond $2\times2$ (see also Goodall, 2021; Häussler & Juzek, 2021). Alternatively, one factor can be designed as between-items variable such that each item corresponds to only one type, whereas as a within-items variable, the item exists in all levels.

stimuli with different degrees of acceptability, ranging from fully unacceptable to fully acceptable, and thus to prevent participants to use only a limited range (*compression*) or one end of the scale (*skew*).[7] It is further strongly recommended to match acceptability degrees of filler stimuli to those of the test stimuli in order to avoid that ratings of test stimuli will be comprised in that only a very limited range of the scale will be used, and further to avoid blinding participants towards potentially subtle differences in test stimuli (see Phillips, 2009).

Due to the need to properly blend test trials in with filler trials, 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 et al. (2022) that examined anaphoricity constraints on word order of event-internal adverbials realized as prepositional phrases (PPs). Consider as an illustration Ex. (2), an item taken from the FC experiment, where participants needed to select their preferred serialization: the comitative ('with a counsellor') is either realized above the object (PP>OBJ) or below (OBJ>PP).

(2) *Test item*

a. *Ich habe gehört, dass ein Minister zusammen mit einem Berater was*
I have heard that a secretary together with a counsellor what
*entschieden hat. Was es war, weiß ich aber nicht.*
decided has what it was know I but not

'I have heard that a secretary decided something in tandem with a counsellor. But I don't know what it was.' (PP>OBJ)

b. Ich habe gehört, dass ein Minister was zusammen mit einem Berater entschieden hat. Was es war, weiß ich aber nicht. (OBJ>PP)

The test environment in Ex. (2) consisted of verb final (VF) embedded clauses where the object was realized as the *wh*-indefinite *was* ('what', 'something'). Subject–object–verb (SOV) structures are considered to reflect the basic order in German clauses (see Keller, 2000a), and the *wh*-indefinites are assumed to be fixed if they receive an existential interpretation (see e.g. Haider, 2010). The required interpretation was enforced by including a disambiguating addendum in the test stimuli that renders the *wh*-indefinite incompatible with a specific interpretation. 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 ought to exclude (extragrammatical) confounding factors as well as possible such that, at best, no other factors than the experimental conditions may influence ratings (see Schütze, 1996; Cowart, 1997; Featherston, 2005; Dąbrowska, 2010).[8]

[7] It is therefore advisable that randomization should not only consider the distribution of item functions (i.e. test vs. filler) but also of acceptability degrees (Cornips & Poletto, 2005). To achieve this, creating blocks of mixed item functions and conditions helps: The order of sentences within blocks is randomized, and the order of blocks within a questionnaire is randomized (see Zanuttini et al., 2018).

[8] All VF sentences were subcategorized by a propositional argument selecting matrix verb in perfect tense with the subject realized as first person pronoun or named entity. The disambiguating addendum remained the same over all test stimuli. They all showed roughly the same length, and all arguments were realized as indefinite, nominal phrases (NPs) thus avoiding alternating patterns of indefinites and definites, which are

However, this strict compliance to specific characteristics may render test trials quite catchy, in particular by such prominent ones like the disambiguating addendum. 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) employed a different core (VF clause) in a similar frame (e.g. matrix clause and disambiguating addendum), which exhibited some variance, as illustrated in Table 1.[9] By this, filler stimuli were superficially similar to test stimuli, blurring the distinction between those groups (see also Goodall, 2021).

| matrix | C | VF | disambiguating addendum |
|---|---|---|---|
| Ich hoffe, | | | Ob das stimmt, weiß ich aber nicht. |
| 'I hope' | | | 'But I don't know whether this is true.' |
| Isa hat versprochen, | dass | [...] | Welchen genau, weiß ich aber nicht. |
| 'Isa has promised' | 'that' | | 'But I don't know which one exactly (it was).' |
| Tim hat erzählt, | | | Warum er das getan hat, weiß ich aber nicht. |
| 'Tim has recounted' | | | 'But I don't know why he did that.' |
| Weißt du vielleicht, | ob | | Ich hoffe, dass [...]. |
| 'Do you know by any chance' | 'whether' | | 'I hope that' |
| | | | Ich bin froh, dass [...]. |
| | | | 'I am glad that' |

Table 1: Exemplative frames of filler stimuli

Although their main objective is the obfuscation of test trials, filler items can carry different functions. Despite not being fillers in the strict sense, such special fillers count as fillers in terms of the TFR. A questionnaire typically starts with some calibration trials. As their purpose is to accustom participants to the scale without telling them, they can also be thought of as implicit training trials. We recommend using six implicit training trials in FC tasks as well so that participants become familiar with the task.[10] These stimuli should cover the full range of acceptability, including the mid-range (see Häussler & Juzek, 2021), such that participants ideally have used every point of the scale right at the start of the survey. Kiss et al. (2022) have for example used the calibration stimuli listed under Ex. (3).

(3) *Calibration stimuli*

known to be sensitive to word order constraints (see Lenerz, 1977). All NPs were also unmodified as syntactic weight may influence word order (Hawkins, 1992). Certain PP-modifiers had a disambiguating function, and the modifier *zusammen* ('together') was employed to restrict the PP to a comitative interpretation. This method also ensured that an unwarranted syntactic attachment of the PP to the *wh*-indefinite, which would arguably prevent an existential interpretation, is blocked in Ex. (2). Furthermore, all objects of VF, i.e. *wh*-indefinites, were inanimate because animacy has been claimed to influence word order (Keller, 2000a).

[9] In the following, we will restrict the presentation of stimuli to the VF clause.

[10] Goodall (2021) estimates that already three to five trials are sufficient, whereas Schütze & Sprouse (2014) recommend five up to ten trials.

a. * *Jan hat versprochen, dass sich er beeilt. [...]*
Jan has promised that REFL he hurries
'Jan has promised to hurry up.'

b. ? *dass Pia für ihre Verhaltensweise sich geschämt hat.*
that Pia for her behavior REFL feel.ashamed has
'I have heard that Pia was ashamed of her behavior.'

c. *dass sein Sohn einigen Nachbarn auf die Nerven gegangen ist.*
that his son some neighbors on the nerves walked has
'Ben related that his son annoyed some neighbors.'

By sticking to using word order constraints violations, i.e. soft constraints violations leading to mild unacceptability (as opposed to hard constraints violations leading to strong unacceptability, see Sorace & Keller, 2005), the acceptability degrees were more comparable to those of the test stimuli. Thereby, scale compression and blinding participants towards subtle effects of test items has been avoided; this applied to all filler stimuli (see Phillips, 2009). Likewise, all ordinary fillers made use of ungrammatical, grammatical and marked conditions. They consisted of several groups employing different syntactic structures in different ranges of length (Patterson, 2014).

Unlike ordinary fillers, so-called pseudo-fillers are partially related to test items. By this, they arguably support obfuscating the research subject by picking up conspicuous characteristics of test items. For the test items in Kiss et al. (2022), these characteristics comprised on the one hand the *wh*-indefinites, which resurfaced in pseudo-fillers as interrogative pronouns, as e.g. in Ex. (4a). On the other hand, the adverbials themselves, which are realized as PPs, constituted striking features of the test items. These can resurface in pseudo-fillers in the shape of (additional) adverbials of another type that are not currently under investigation, as e.g. a manner adverbial in Ex. (4b), or in identical wording but employing a different function, e.g. comitatives can resurface as complements.

(4) *Pseudo-filler*

a. ? *Weißt du vielleicht, wen Max eingeladen hat zur Party?*
know you maybe whom Max invited has to.the party
'Do you know by any chance whom Max has invited to the party?'

b. *dass Tim zusammen mit Ben sich auf verdächtige Art weggeschlichen hat*
that Tim together with Ben himself on suspicious way slipped.out has
'[...] that Tim has slipped out in a suspicious manner in tandem with Ben.'

### 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 participants so that their responses can be discarded as they are harming the quality of the elicited data. The essential property of these control stimuli is thus 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. Appropriate control stimuli are those which have been rated (fully) natural or (fully) unnatural without much variation in such a study (Gerbrich et al., 2019). 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 German verb final (VF) clauses in which the order of the syntactic arguments violates up to three word order constraints: [*+nom*]$\prec$[*+acc*], [*+pro*]$\prec$[*-pro*], and [*+animate*]$\prec$[*-animate*], as exemplified in Ex. (5) (see also Keller, 2000a).

(5) * *dass Schwarzweißfilme er liebt.*
that black-and-white.films.ACC he.NOM loves
'[...] that he loves black-and-white films.' (OSV)

What counts as related or unrelated of course depends on the study. Ex. (5) can be considered unrelated in the context of the studies reported in Kiss et al. (2022), which investigate adverbials by means of *wh*-indefinites – both of which are missing in Ex. (5). 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 adverbials (in the shape of PPs) are present, Ex. (6) is a related control stimulus in the context of the studies in Kiss et al. (2022).

(6) * *dass was der Student zum Verkauf angeboten hat.*
that what the student to.the sale offered has.
'[...] that the student offered something for sale.' (OSV)

Ex. (6) bears some resemblance to test items – see Ex. (2) – but differs from them in that the uncontroversially offending element is the order between the subject and the *wh*-indefinite, and that the adverbial 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 (Ex. (4)), and thus help preventing confirmation biases as well as conscious response strategies. Secondly, they are presumably less salient than unrelated control trials (see Ex. (5)) and thus counterbalance the possibility that trained participants identify control trials as outstanding, 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, consider word order variation of the negation in the addendum of the stimulus in Ex. (7).

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

Here, the violating negation ('*nicht*') occurs late in the example, is a short element, which is surrounded by other short elements. 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, whereby inattentive participants are recognized by their disregard of exact orders given in the (trial-related) instructions given directly above one question (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 attention checks must preclude any misinterpretations. As a fair example, the question repeated in Ex. (8) is provided:

(8) Please indicate your agreement to the statements below:
It is important that you pay attention to this study. Please tick 'Strongly Disagree'.

We see various reasons for not using IMCs in AJTs. To begin with, trials in linguistic studies typically deviate from those in e.g. psychology or sociology in that instructions throughout an AJT remain the same such that changing them must be considered unfair. What is more, the requested clarity can never be reached in an AJT as participants may be confused whether they are supposed to rate the naturalness of a stimulus or to follow the instructions it gives. On top of that, there are various concerns regarding attention checks: Oppenheimer et al. (2009) point out that attention checks may harm the data quality as diligent participants may feel insulted whereas careless participants may feel embarrassed. In both cases, the behavior of the participants is affected. Vannette (2016) assumes that exclusions due to attention checks may lead to demographic biases, and also concludes that attention checks (negatively) affect the performance of the participants. Finally, attention checks may simply be too noticeable, particularly for participants that are experienced in answering questionnaires on crowd-sourcing platforms (Gadiraju et al., 2015; Juzek & Häussler, 2017).

None of these problems emerge for the attention stimuli presented above. The feeling of being tried to be trapped is prevented as the checks are not that obvious. It is hence unlikely that *attention stimuli* (as opposed to IMCs) will change participants' behavior. However, we must consider that the level of attention may vary throughout the experiment even within diligent participants (Paolacci & Chandler, 2014), which implies that attention as well as

control trials should be spread across the whole survey, which impacts the number of control and attention trials required. It should be clear that using control as well as attention trials only at the beginning of a study will not help detecting participants who 'doze off' during a study (see Juzek, 2016) but placing them only in the last part will not help detecting participants who have been distracted at the beginning of a survey.

## 4. Control by choosing an adequate participant pool

Although the questionnaire is now equipped with the means to identify inapt participants, it is of course desirable to employ a high-quality participant pool to elicit data from 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: this does not only concern the structure of the task but especially the demands they pose upon *participants*, which already becomes apparent by the fact that they are typically called *workers* in HITs.

First of all, HITs require much smaller assignment numbers as quality control can be determined by majority votes (see Mason & Suri, 2012). Thus, AMT charges an additional 20 % fee (on top of its 20 % fee) if more than nine participants processed the trials (= HIT) in a batch. AJTs require – as a rule of thumb – 24 to 36 participants to yield useful results (in terms of statistical power, see Sprouse & Almeida, 2017; Goodall, 2021), since variability in the data elicited is expected. Consequently, special trials with verifiable answers are needed to vet participants. Thus, a major difference consists in the need of AJTs to be embedded in a carefully designed questionnaire (as has been discussed in section 3), whereas HITs are typically independent from each other and presented in small batches. Accordingly, employing smaller batch sizes in order to ensure attentiveness and to deter spammers is clearly not an option for AJTs (see Schnoebelen & Kuperman, 2010; Eickhoff & de Vries, 2013).

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 (as we have discussed in section 2 and section 3). In AJTs we need judgments of *native speakers* that are *naïve*, i.e. had no prior exposure to the test stimuli, and are preferably non-linguists. As for HITs eligibility is a minor issue, few possibilities exist to constrain it in crowdsourcing platforms like AMT (see Mason & Suri, 2012), leading to the need to reject *workers* that should not have participated in the study in the first place (see Schnoebelen & Kuperman, 2010).[11]

This issue has been addressed by new platforms that are geared towards the needs of academic researchers in the meantime: e.g. Prolific (founded in 2014),[12] CloudResearch

[11] 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 (Hunt, 2015; Robinson et al., 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).

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

(founded 2016),[13] and Positly (founded in 2017).[14] These platforms provide methods to address a specific subset of their pool (*prescreening*). Hence, the researchers define the target population, and the platform only provides access to their study to participants who match these prescreening criteria.[15]

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. While 'Current Country of Residence' (next to 'Age', 'Gender' and 'Education') seems to be a standard screening option, 'native' language is not. For linguists, Prolific provides the most elaborate screening options and is thus by far the most suitable platform for AJTs.[16]

Another eligibility criteria for AJTs is naïveté (as opposed to HITs where experience probably leads to more accurate results): i.e. participants should not be familiar with test stimuli or with the research subject (see section 3). Accordingly, these new platforms provide the exclusion of participants that haven taken part in previous studies and are by this again geared towards researchers. 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. To ensure naïveté, still, eligible participants need to be prevented from conducting the exact same study repeatedly. Many platforms integrate with other software for survey creation, like e.g. Jatos (Lange et al., 2015), such that researchers provide a link to the website hosting their study. Consequently, the platforms can certainly guarantee that participants can have only one submission per study on that platform (and can only be paid once), but they cannot guarantee that participants repeatedly visit the website hosting the study. Thus, all platforms can do to prevent *multiple submissions* (and payments) is to inhibit people from creating various accounts on the platform.

To ensure the quality of their participant pools, some platforms do not only protect against bots but are *vetting participants* such that they provide evaluations of participants' trustworthiness regarding attention, engagement, English comprehension, and data quality beforehand (Chandler et al., 2019; Litman et al., 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 section, we will show how to determine problematic participants based on latency- and

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

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

[15] Nonetheless, Prolific Team (2018b) draws attention to the fact that prescreening information is self-reported and cannot be verified (except for maybe their current country of residence). Thus, Prolific Team (2018a) suggests to validate the screening questions by repeating them in your survey and by checking whether the answers are in accordance with the prescreening criteria as profiles might eventually become outdated. Moreover, incentives to lie cannot be fully excluded as they supposedly provide access to a wider range of studies.

[16] They differentiate between 'First Language' and 'Fluent Languages'. Furthermore, the screening questions 'Were you raised monolingual?', 'Apart from your native language, do you speak any other languages fluently?' and 'Ethnicity' exist. Next to 'Current Country of Residence', also 'Country of Birth' is available. On top of that, for the question 'What subject do you study?' the option 'Languages' is available and may be excluded.

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, i.e. the inspection of response times (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 & Juzek (2016) aim at excluding participants who are frequently distracted.[17]

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 & 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 Downs et al., 2010; Buchholz & Latorre, 2011; 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 a-priori knowledge about general RTs in a specific task (Berger & Kiefer, 2021). On the other hand,

[17] Distractions could indicate that participants are rereading the stimulus multiple times or that they are consulting with a third person.

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 & Juzek, 2016). To ensure capturing all trials where participants where 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 (and thus exhibit a strong difference in FC tasks), whereas the test stimuli probably do not (see Ratcliff et al., 2018; 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.

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* (i.e. attention trials only, control trials only, etc.) 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

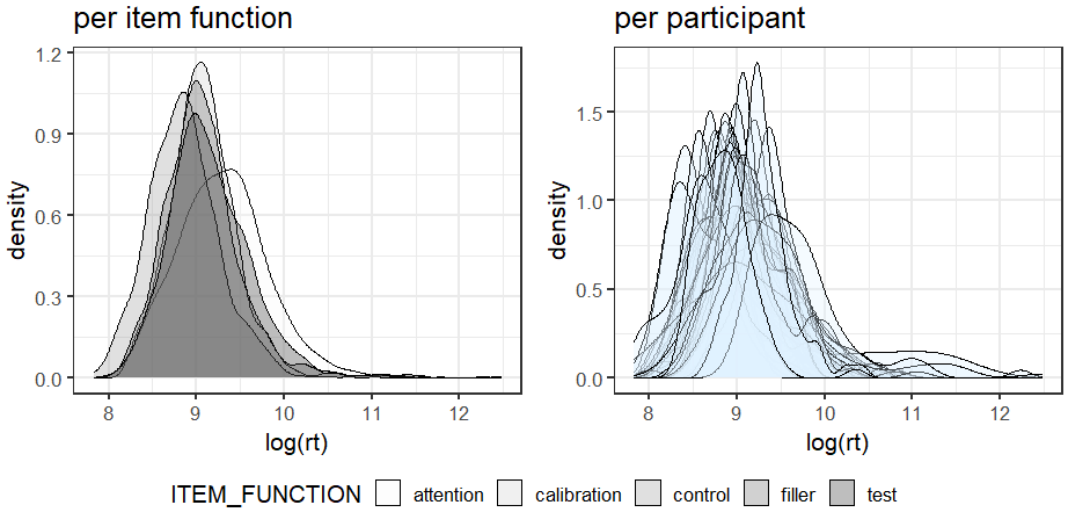


Figure 1: Density of RTs (Experiment 2 from Kiss et al., 2022)

as a *genuine intermission* (see Equation (1)) or as a *genuine rush* (see Equation (2)).[18]

$$\begin{aligned} \textit{cutoff\_intermission} = \max\{ & \\ & \text{median}(\textit{RTs:participant}) + 2.5 \times \text{MAD}(\textit{RTs:participant}), \\ & \text{median}(\textit{RTs:item\_function}) + 2.5 \times \text{MAD}(\textit{RTs:item\_function}) \} \end{aligned} \tag{1}$$

$$\begin{aligned} \textit{cutoff\_rush} = \min\{ & \\ & \text{median}(\textit{RTs:participant}) - 1.5 \times \text{MAD}(\textit{RTs:participant}), \\ & \text{median}(\textit{RTs:item\_function}) - 1.5 \times \text{MAD}(\textit{RTs:item\_function}) \} \end{aligned} \tag{2}$$

To account for swamping and masking effects, the process described above will be repeated on a reduced data set (i.e. excluding already detected outliers) 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.[19]

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

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

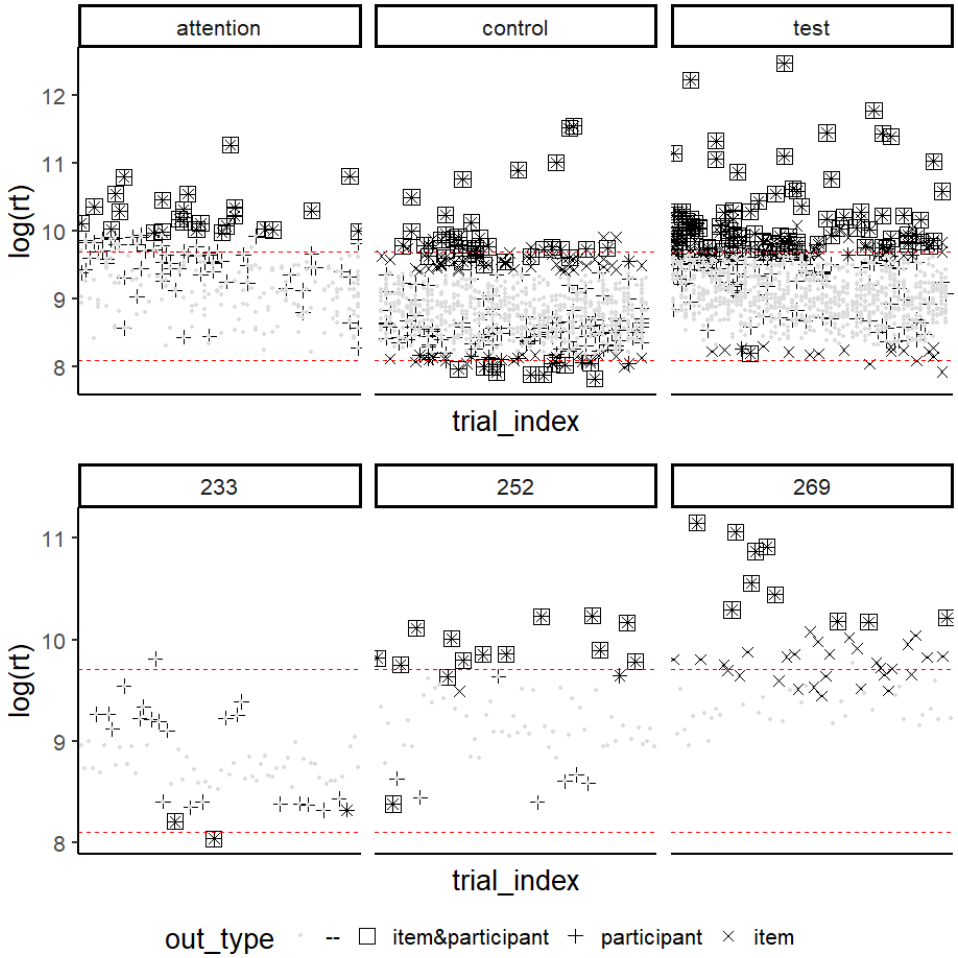


Figure 2: *Recursive multi-factorial outlier detection* (ReMFOD)

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 having examined 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 determining which participants should be considered malicious due to their responses to these trial types. 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 stimulus is directly compared to its ungrammatical version in FC tasks, correct responses to control and attention trials 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.

Although it should be self-evident that a threshold should be set in a way that participants perform (well) above chance, the practice in experimental linguistics often shows negligence of this issue. This holds 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 does not consider that the sum of trials correctly answered consists of the sum of trials correctly *solved* and correctly *guessed* (Frederick & Speed, 2007). A simple method to estimate the number of trials correctly solved is depicted in Equation (3):

$$\textit{Number of trials correctly solved} = \textit{Rights} - \textit{Wrongs} \tag{3}$$

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, which can seduce participants to guess if they lose their temper searching for the manipulation. 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 & 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 2), then the proportion of correct responses minimally required (slowly) 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 and resulting slips in a comparatively long questionnaire. Following this lead, we strongly argue against using less than six but for employing at least eight attention/control trials in 2AFC tasks.

### 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. This is so 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

| $N$ | $k$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **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** |

Table 2: Probabilities of responding to $\geq k$ out of $N$ (2AFC) trials correctly by chance (see Equation (6) in Appendix A)

trials. This raises the issue, however, whether the neutral point in an LS 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 3a). If we take the neutral point not to be acceptable, this number is reduced to five correct responses out of six trials (see Table 3b). As the neutral point arguably does not reliably indicate that the participant has indeed spotted the manipulation, we recommend excluding the neutral point from the acceptable responses.

**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 in Ex. (9) for a *5-pt*-LS.

(9) *Different accounts for control trials on 5-pt (bipolar) LSs*

a. **positional:** all stimuli should receive expected ratings, i.e.
$< 3$ *for ungrammatical stimuli*
$\geq 3$ *for grammatical stimuli*

b. **relational:** grammatical stimuli should receive higher ratings 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, however, the positioning on the scale does not matter, 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, see section 3.1). 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.

| $N$ | $k$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **12** |
| **4** | .821 | .475 | .130 | | | | | | | |
| **6** | .959 | .821 | .544 | .233 | **.047** | | | | | |
| **8** | .991 | .950 | .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** | .954 | .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** |

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

Table 3: Probabilities of responding to $\geq k$ out of $N$ unacceptable (5-pt-LS) trials correctly by chance (see Equation (5) in Appendix A). 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).

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. As has become apparent by the discussion of attention trials, one point of discussion of the positional account is 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 (3) 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 4.[20] 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 $\geq 3$ 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. $1\times2$ or $2\times2$), the questionnaire should include at least 24 or 48 filler trials, when employing the minimum TFR of 1:2 (see section 3.1). As the related control items also support obfuscating the research subject, this goal can also be reliably obtained with the remaining ordinary fillers.

| $N$ | grammatical ($\geq 3$) | | 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 |

Table 4: Joint probabilities of passing grammatical as well as ungrammatical controls on a positional account

[20] 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.

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 et al., 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 declaration of the neutral option as either acceptable or unacceptable, Zanuttini et al. (2018) hence employ mean values, as is done in the relational account, the problems of which will be discussed immediately.

The more prominent relational account, as e.g. advocated for in Häussler & 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 & Perla (2008, p. 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 & Juzek (2016) did, a treatment of ratings as interval instead of ordinal 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 (although their status is well-established). 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 that should be removed. Relational accounts thus fail to control for scale biases, especially, scale compression.

Confronted with this result, we could consider an extension of the 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):

$$\frac{\text{mean}(grammatical) - \text{mean}(ungrammatical)}{\max(LSrating) - \min(LSrating)} \geq 0.5 \tag{4}$$

Thus, for a 5-pt-LS the minimum distance of 2 ($0.5 * (5 - 1)$) between both conditions is required. By this, both conditions (grammatical and ungrammatical) 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.[21]

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. Determining these probabilities is analogous to estimating the FC guessing probabilities: We determine

[21] 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. By not allowing for any slips, the account of Zanuttini et al. (2018) is stricter, and thus rejecting a good part of the outcomes accepted by the extended relational account.

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 where each pair consist 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 Ex. (9b), i.e. $> 0$ and Equation (4), i.e. $\geq 0.5$.

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 5). 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 %.[22] The relational account in Ex. (9b) 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 from conceptional problems – the acceptance of scale compression – but hence has also small chances to detect guessing participants (see Table 5). 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).

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, which also manifests itself in the discussion of whether the scale should include one or not in the first place (see section 2), 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

[22] 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.

| pairs | possible outcomes | pure: $s > 0$ | extended: $s \geq 0.5$ |
|---|---|---|---|
| 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 | 152,587,890,625 | 0.4651 | 0.0026 |

Table 5: Acceptance rates of guessed outcomes, i.e. all possible response combinations, according to relational accounts

be clearly and intuitively understandable. Another difference between the present account and Zanuttini et al. (2018) consist 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 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, i.e. that they should not apply prescriptive rules or judge plausibility but should stick to their immediate reactions. 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, 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 platform to recruit from needs to be carefully selected as common platforms differ in being geared towards HITs or being more suitable for research tasks such as AJTs: Differences concern not only the structure of the task but the eligibility of participants in particular. Modern platforms especially geared towards researchers do not only alleviate some of the problems resulting from the loss of supervision but more importantly allow targeting the population. 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 participant platforms. Thus, we propose to additionally combine slow and long outliers

into one measure of underperformance. However, detecting outliers is far from trivial: This is due to masking and swamping effects as well as the 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. Relying on these, underperforming participants can be identified. 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 should 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 as the former come with major drawbacks, as the need to justify an interpretation as interval data and the 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. Taken all of these measures together, researchers can be confident that the data entering analysis reaches the highest attainable level in web-based elicitation of AJTs.

## A. Guessing probabilities

Frederick & 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) gives the probability of (exactly) $k$ correct answers out of $N$ trials:

$$\frac{N!}{k!(N-k)!}p^k q^{N-k}, \text{ where} \tag{5}$$

$$p = \text{ probability of a correct response}$$
$$q = \text{ probability of an incorrect response}$$

When $p = q$, which occurs when there are only two answer choices as in 2AFC, this formula can be simplified (see Equation (6)).

$$\frac{N!}{k!(N-k)!}p^N \tag{6}$$

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).

## B. Data Availability

Supplementary materials is available at
https://github.com/Linguistic-Data-Science-Lab/AJTs-eligibility-screening

## Abbreviations

| | |
|---|---|
| **2AFC** | Two-Alternative Forced-Choice |
| **AJ** | acceptability judgment |
| **AJT** | Acceptability Judgment Task |
| **AMT** | Amazon Mechanical Turk |
| **HIT** | Human Intelligence Task |
| **FC** | Forced-Choice |
| **IMC** | instructional manipulation check |
| **LPC** | lexicalizations per experimental condition |
| **LS** | Likert-Scale |
| **MAD** | median absolute deviation |
| **ME** | Magnitude Estimation |
| **PP** | prepositional phrase |
| **ReMFOD** | recursive multi-factorial outlier detection |
| **RT** | response time |
| **TFR** | test-filler-ratio |
| **VF** | verb final |
| **YN** | yes-no task |

## References


Bader, M., & Häussler, J. (2010). Toward a model of grammaticality judgments. *Journal of Linguistics*, *46*(2), 273–330. https://doi.org/10.1017/S0022226709990260

Bard, E., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, *72*. https://doi.org/10.2307/416793

Ben-Gal, I. (2005). Outlier detection. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 131–146). Springer US. https://doi.org/10.1007/0-387-25465-X_7

Berger, A., & Kiefer, M. (2021). Comparison of different response time outlier exclusion methods: A simulation study. *Frontiers in Psychology*, *12*, 2194. https://doi.org/10.3389/fpsyg.2021.675558

Bolinger, D. L. (1961). Syntactic blends and other matters. *Language*, *37*(3), 366–381. https://doi.org/10.2307/411078

Buchholz, S., & Latorre, J. (2011). Crowdsourcing preference tests, and how to detect cheating. *12th Annual Conference of the International Speech Communication Association*, 3053–3056.

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

Chandler, J., & 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

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, *51*(5), 2022–2038. https://doi.org/10.3758/s13428-019-01273-7

Cornips, L., & 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

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments*. SAGE Publications.

Dąbrowska, E. (2010). Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, *27*(1), 1–23. https://doi.org/10.1515/tlir.2010.001

Dandurand, F., Shultz, T., & 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

Devitt, M. (2006). *Ignorance of language*. Oxford: Clarendon Press. https://doi.org/10.1093/0199250960.001.0001

Downs, J., Lanyon, M., Sheng, S., & Cranor, L. (2010). Are your participants gaming the system? Screening Mechanical Turk workers. *Conference on Human Factors in Computing Systems - Proceedings*, *4*, 2399–2402. https://doi.org/10.1145/1753326.1753688

Eickhoff, C., & de Vries, A. (2013). Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval - IR*. https://doi.org/10.1007/s10791-011-9181-9

Featherston, S. (2005). Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua*, *115*, 1525–1550. https://doi.org/10.1016/j.lingua.2004.07.003

Featherston, S. (2007). Data in generative grammar: The stick and the carrot. *Theoretical Linguistics*, *33*(3), 269–318. https://doi.org/10.1515/TL.2007.020

Featherston, S. (2021). Response methods in acceptability experiments. In G. Goodall (Ed.), *The Cambridge handbook of experimental syntax* (pp. 39–61). Cambridge University Press. https://doi.org/10.1017/9781108569620.003

Frederick, R. I., & 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. (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., & 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. https://doi.org/10.1145/2702123.2702443

Gerbrich, H., Schreier, V., & Featherston, S. (2019). Standard items for English judgment studies: Syntax and semantics. In S. Featherston, R. Hörnig, S. von Wietersheim, & S. Winkler (Eds.), *Experiments in focus: Information structure and semantic processing* (pp. 305–328). De Gruyter Mouton. https://doi.org/10.1515/9783110623093-012

Gibson, E., & 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., & 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* (pp. 7–38). Cambridge University Press. https://doi.org/10.1017/9781108569620.002

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

Häussler, J., & 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, & V. Werner (Eds.), *A blend of MaLT. Selected contributions from the methods and linguistic theory symposium 2015* (pp. 73–99). University of Bamberg Press.

Häussler, J., & Juzek, T. S. (2020). Linguistic intuitions and the puzzle of gradience. In S. Schindler, A. Drożdżowicz, & K. Brøcker (Eds.), *Linguistic intuitions: Evidence and method* (pp. 233–254). Oxford University Press. https://doi.org/10.1093/oso/9780198840558.003.0013

Häussler, J., & Juzek, T. S. (2021). Variation in participants and stimuli in acceptability experiments. In G. Goodall (Ed.), *The Cambridge handbook of experimental syntax* (pp. 97–117). Cambridge University Press. https://doi.org/10.1017/9781108569620.005

Hawkins, J. (1992). Syntactic weight versus information structure in word order variation. In J. Jacobs (Ed.), *Informationsstruktur und Grammatik. Linguistische Berichte Sonderhefte*. VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-663-12176-3_7

Hunt, N. (2015). A guide to using MTurk to distribute a survey or experimental instrument. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2664339

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

Juzek, T. S., & Häussler, J. (2017). Hot topics surrounding acceptability judgement tasks. In S. Featherston, R. Hörnig, R. Steinberg, B. Umbreit, & J. Wallis (Eds.), *Proceedings of linguistic evidence 2016: Empirical, theoretical, and computational perspectives*. https://doi.org/10.15496/publikation-19039

Keller, F. (2000a). Evaluating competition-based models of word order. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd annual conference of the cognitive science society* (pp. 747–752). Lawrence Erlbaum Associates. https://www.semanticscholar.org/paper/Evaluating-Competition-based-Models-of-Word-Order-Keller/f403393ea39d050b4fc99c3a9dc913c8f9aba4d9

Keller, F. (2000b). *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality* (PhD thesis). University of Edinburgh. https://doi.org/10.7282/T3GQ6WMS

Kiss, T., Pieper, J., & Börner, A. K. (2022). Supplementary material for: "Word order constraints on event-internal modifiers". https://doi.org/10.5281/zenodo.5898464

Lange, K., Kühn, S., & Filevich, E. (2015). Correction: "just another tool for online studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLOS ONE*, *10*, e0130834. https://doi.org/10.1371/journal.pone.0130834

Leivada, E., & 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

Lenerz, J. (1977). Zur Abfolge nominaler Satzglieder im Deutschen [On the serialization of nominal clauses in German]. In W. Abraham, W. Boeder, U. Engel, J. Leriot, O. Leys, & H. Vater (Eds.), *Studien zur deutschen Grammatik [Studies on German grammar]*. Stauffenburg.

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

Mason, W., & 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., & 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., & Reips, U.-D. (2000). Chapter 3 - a brief history of web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 61–87). Academic Press. https://doi.org/10.1016/B978-012099980-4/50004-6

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., & 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., & Chandler, J. J. (2014). Inside the Turk. *Current Directions in Psychological Science*, *23*, 184–188.

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

Phillips, C. (2009). Should we impeach armchair linguists? *Japanese-Korean Linguistics*, *17*.

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

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

Ratcliff, R., Voskuilen, C., & Teodorescu, A. (2018). Modeling 2-alternative forced-choice tasks: Accounting for both magnitude and difference effects. *Cognitive Psychology*, *103*, 1–22. https://doi.org/10.1016/j.cogpsych.2018.02.002

Robinson, J., Rosenzweig, C., & Litman, L. (2020). The Mechanical Turk ecosystem. In L. Litman & J. Robinson (Eds.), *Conducting online research on Amazon Mechanical Turk and beyond* (pp. 27–47). Sage Academic Publishing.

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija*, *43236599*, 441–464. https://doi.org/10.2298/PSI1004441S

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

Schütze, C. T., & Sprouse, J. (2014). Judgment data. In R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 27–50). Cambridge University Press. https://doi.org/10.1017/CBO9781139013734.004

Snyder, W. (2021). Satiation. In G. Goodall (Ed.), *The Cambridge handbook of experimental syntax* (pp. 154–180). Cambridge University Press. https://doi.org/10.1017/9781108569620.007

Sorace, A., & 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., & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's core syntax1. *Journal of Linguistics*, *48*, 609–652. https://doi.org/10.1017/S0022226712000011

Sprouse, J., & Almeida, D. (2013). The empirical status of data in syntax: A reply to Gibson and Fedorenko. *Language and Cognitive Processes*, *28*, 222–228. https://doi.org/10.1080/01690965.2012.703782

Sprouse, J., & 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., & 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

Vannette, D. L. (2016). Testing the effects of different types of attention interventions on data quality in web surveys. Experimental evidence from a 14 country study. *Annual Conference of the American Association for Public Opinion Research*, *71*.

Wasow, T., & Arnold, J. (2005). Intuitions in linguistic argumentation. *Lingua*, *115*, 1481–1496. https://doi.org/10.1016/j.lingua.2004.07.001

Weskott, T., & Fanselow, G. (2009). Scaling issues in the measurement of linguistic acceptability. In S. Featherston & S. Winkler (Eds.), *Volume 1: Process* (pp. 229–246). De Gruyter Mouton. https://doi.org/10.1515/9783110216141.229

Weskott, T., & Fanselow, T. G. (2011). On the informativity of different measures of linguistic acceptability. *Language*, *87*, 249–273. https://doi.org/10.1353/LAN.2011.0041

Zanuttini, R., Wood, J., Zentz, J., & 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