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 assumptions underlying the task. Finally, also the researchers themselves may be held responsible for noise if they employ a poor questionnaire design, failing to control for biases and adverse effects, and lacking sufficient means to identify inapt participants.

We describe a stepwise process to reach elicited data at the highest attainable level in web-based Acceptability Judgment Tasks (AJTs). In doing so, 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. For the third step, the identification of inapt participants, we will present reflections on how to employ analyses of general *response times* as well as of *responses* to the specialized items so that a possible exclusion of participants is based on latency- as well as on 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 of participants to go to a laboratory and of researchers to instruct and supervise them in person. Instead, a very large and diverse population had become accessible with minimal recruitment time and costs. On the downside, conducting web experiments requires skills researchers typically lack (see Schnoebelen & Kuperman, 2010). In comparison to classic recruitment strategies like obliging one's own students, recruitment with web-based methods has been cumbersome (and slow) back in the days of early online studies (see Birnbaum, 2004).

However, it did not take long until extensive pools of potential participants were provided: With crowdsourcing platforms such as Amazon Mechanical Turk (AMT)[1] and clickworker[2] (both founded in 2005), researchers have been exempted from the task of coming up with a recruitment strategy. They provide a large pool of demographically and geologically diverse participants willing to take part in studies (to earn money) and are therefore time- and cost-effective (Gibson et al., 2011; Sprouse, 2011b). 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. Many researchers have compared online to laboratory results (e.g. Dandurand et al., 2008; Sprouse et al., 2013), and especially evaluated the reliability of results elicited with AMT (see Schnoebelen & Kuperman, 2010; Gibson et al., 2011; Sprouse, 2011b, for linguistic research). Thereby, laboratory experiments have been replicated with web-based methods 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 (Dandurand et al., 2008). Although distractions of mobile phones as well as non-cooperative behavior can occur in a lab as well, it is less likely than in remote elicitation due to Hawthorne effects (Phillips, 2009; Juzek & Häussler, 2017). Detrimental effects on data quality, leading to decreased statistical power (Oppenheimer et al., 2009), in online studies have been observed due to participants' behavior: online participants provide less accurate data and fail more frequently on attention checks, possibly due to adverse effects of distraction as well as for being less intrinsically motivated and unsupervised at the same time (Dandurand et al., 2008; Oppenheimer et al., 2009). An extensive amount of research has focused on identifying fraudulent participants who do not comply with the task (see Oppenheimer et al., 2009; Downs et al., 2010; Buchholz & Latorre, 2011; Eickhoff & de Vries, 2013; Gadiraju et al., 2015; Häussler & Juzek, 2016; Chandler & Paolacci, 2017). Indeed, it has been observed that malicious behavior in crowdsourcing is virulent (see Downs et al., 2010; Kazai et al., 2011; Eickhoff & de Vries, 2013). Eickhoff & de Vries (2013) further observe that simple tasks that can be easily automated attract money-driven workers (as opposed to entertainment-driven), which might be especially true for AJTs.

The remaining paper is structured as follows: Section 2 discusses AJTs in general, and

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

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

in particular its underlying assumptions and prerequisites. The remaining sections focus on different aspects of controlling web-based elicitation. The impact of the questionnaire design and different trial types is discussed in section 3.1, section 3.2 discusses control items, and section 3.3 introduces a new concept: attention items. Section 4 discusses control by choice of an appropriate participant pool, in particular regarding eligibility. 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* has been introduced in Chomsky (1965), where it 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 such as processing difficulties, semantic anomaly, or conformity to prescriptive norms (Dąbrowska, 2010). It is thus possible that *grammatical* sentences are judged as unacceptable. 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 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. Schütze & Sprouse (2014) – who define judging as *"report[ing] their spontaneous reaction"* – also highlight that a reported perception is different from concepts such as *introspection* or *intuition*, which would assume direct access to cognitive systems. 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 Chomsky, 1965; Schütze, 1996; Sprouse & Almeida, 2013, among others). 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, without reasoning, but with relative temporal immediacy (Devitt, 2006; Bader & Häussler, 2010; Maynes & Gross, 2013; Häussler & Juzek, 2020).

When dealing with AJs, we must also consider that acceptability can be gradient, as first observed in Bolinger (1961). Here again, acceptability differs from grammaticality, which is assumed to be dichotomous (see also Keller, 2000b). While gradience has been reflected in the amplification of diacritics used by syntacticans (such as * for unacceptable, and *?, ??, and ?** for decreasing states of acceptability), the meaning of the diacritics has never been formally defined, so it is unclear how to record and interpret them, particular so if authors use them inconsistently (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 (Keller, 2000b; Featherston, 2005; Sorace & Keller, 2005; Wasow & Arnold, 2005; Phillips, 2009). While there has been some debate in experimental linguistics whether linguistically naïve participants provide the same kind of judgment as informed linguists (see Culbertson & Gross, 2009; Dąbrowska, 2010), linguists now generally agree to collect judgments from naïve speakers using AJTs in a controlled experimental setting, e.g. by employing a factorial design, where each experimental condition is examined by multiple lexicalizations (see Gibson & Fedorenko, 2013).

To elicit AJs from naïve subjects (non-linguists), 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.

In a YN, participants indicate whether they find a stimulus acceptable (Yes) or not (No). Although participants cannot express (subtle) differences in perceived acceptability between conditions individually, aggregating the judgments provides gradient acceptability in terms of the proportions of the different judgments (Bader & Häussler, 2010).

In an LS, participants are asked to judge stimuli using a numerical (ordinal) scale. Participants can thus directly and intuitively express gradient perceptions. The scale employed is ordered and bipolar, ranging from unacceptable to acceptable, and is mostly chosen to be uneven by allowing five or seven points on the scale (5-pt-LS, 7-pt-LS). Odd scales of course provide a neutral option, 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).

ME (Bard et al., 1996; Featherston, 2005) is a task for eliciting relative judgments: participants ought to indicate how many times more acceptable a stimulus is compared to another one (’modulus’) by entering an arbitrary number. Thus, ME allows us to treat linguistic acceptability as a continuum and directly measures acceptability differences between stimuli (Bard et al., 1996; Sorace & Keller, 2005). 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).

A more intuitive form of eliciting relative judgments is provided by FC tasks. When used for AJTs, usually only two options are provided, hence this task is often called Two-Alternative Forced-Choice (2AFC). From this (minimal) pair of alternatives, participants simply must choose the one which they consider more acceptable. By this, they neither have to state the perceived strength of difference between these acceptabilities, nor how acceptable the presented stimuli are by themselves. Nonetheless, similar to YN tasks, it is possible to deduce the size of the difference between a pair from the proportions of chosen answers (Myers, 2009; Schütze & Sprouse, 2014).

Indeed, 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. 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).

It holds for all types of AJTs that when conducting experiments, one needs to make sure that the participants indeed judge the acceptability of stimuli despite not being familiar with this linguistic concept. It is thus problematic to retreat to concepts such as grammaticality or acceptability since one cannot assume that naïve speakers grasp their intended meanings (as was also pointed out by Schütze (1996)). So, participants may provide judgments based on interpretability, frequency, or plausibility (Cornips & Poletto, 2005). It is thus crucial that participants understand exactly what they are supposed to do. As a case at hand, consider the direction of an LS. 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). It must be further ensured that participants understand that they should judge the well-formedness and thus 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). The scale labels might thus read as follows: *entirely unnatural – rather unnatural – no tendency – rather natural – perfectly natural*. In addition to that, it must be conveyed to participants that they should stick to their spontaneous reactions so that they won't have time to instead of apply norms (see Phillips, 2009). All this must be achieved by providing explicit instructions, which should not be too brief. The instruction should illustrate triggers for unacceptability, as well as point out to reasons why a sentence cannot be judged as unacceptable (Schütze, 1996). 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.

It should also be explicitly stated that there are no right and wrong answers (see Juzek, 2015). 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 asking for immediate reactions without conscious reasoning. It is remarkable that 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)). 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). Sometimes, instructions to this end can be a two-pronged sword (in web-based elicitation in particular): asking participants to work as quickly as possibly (without being careless) requires the participants to determine the cut-off point between speed and diligence.

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. We will discuss how to design questionnaires to avoid biases and unintentional effects in section 3.1, and preparations against unable or unwilling participants in sections 3.2 and 3.3. To do so, we will adduce exemplary stimuli taken from the experimental studies presented in Kiss et al. (2021) that employ an LS and a 2AFC task (see section 2) to examine constraints on word order. The actual identification of malicious behavior will be discussed in section 5.

3.1. General properties of questionnaires

The starting point of each experiment 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 the lexical content or 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 widely known to be necessary to include so-called filler trials. These differ from test trials in the fact that they are not related to the phenomenon under investigation (and the factorial design). Their importance stems from the need to exclude certain effects and biases, in particular, confirmation-, and order-biases, 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 (see Wasow & Arnold, 2005; Gibson & Fedorenko, 2013; Schütze & Sprouse, 2014). By this, confirmation bias, emerging from participants not being naïve and oblivious, can be avoided. They further serve to impede linguistically naïve participants from familiarizing themselves with the test stimuli, as this might 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 lower the proportion of test trials in a questionnaire is (Schütze & Sprouse, 2014). Only if enough filler trials are present, the aim of veiling test trials can be achieved. It is thus somewhat surprising that the ratio of fillers to test trials is rarely discussed, nor reported in experimental studies in linguistics. Weskott & Fanselow (2009) form an exception in referring to Cowart (1997) who proposes that the ratio of test to filler trials (test-filler-ratio (TFR)) should at least be 1:2.[3]

The length of the questionnaire hence becomes an area of multiple conflicts: between the necessity to avoid confirmation bias 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. Whereas Schütze & Sprouse (2014)

[3] More recently, Goodall (2021) proposes a minimal proportion of 1:1. He also states that having more fillers is better, but fears that a larger proportion of fillers may lead to lists of unreasonable length. As will become clear below, we assume that this problem can be resolved by choosing the appropriate type of AJT.

recommend eight or more lexicalizations for each experimental condition, Goodall (2021) reasons that it is often (only) possible to employ at least four or five lexicalizations due to restrictions of questionnaires length. We infer that using at least six lexicalizations should be a minimal standard. We propose that the best strategy to avoid overly lengthy questionnaires but to retain enough power (i.e. six LPC) as well as enough filler trials (i.e. 1:2 TFR) is the reflected choice of the AJT, and thus the factorial design of the study.[4]

It is the factorial design in tandem with these minimum standards which 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. Consider for example Experiment 1 in Kiss et al. (2021) that examined anaphoricity constraints on word order of event-internal adverbials realized as prepositional phrases (PPs). In this study, object-oriented internal locative (ILOC(O)), object-oriented comitative (COM(O)), and instrumental (INSTR) adverbials were tested in two positions: above the object (PP>OBJ) or below the object (OBJ>PP). Kiss et al. (2021) employ an LS-task with a $3{\times}2$ design, as presented in Ex. (2).

(2) Rating (1-5) ~ AdvType (INSTR, ILOC(O), COM(O)) × Pos (PP>OBJ, OBJ>PP)

Lexical variation requires that each of the six experimental conditions is represented by at least six different lexical instances, yielding a total of 36 test trials, randomized with 72 filler trials (employing a ratio of 1:2), yielding a minimum questionnaire length of 108 trials in this case. Compare this to Experiment 2 (Kiss et al., 2021), in which we wanted to explore an additional factor, namely the presence or absence of thematic integration. Thematic integration was introduced by prepositions with affirmative readings (i.e. by the preposition *with*, *via*), but not by preposition with private ones (*without*). Despite reducing the amount of adverbial types from three to two (i.e. subject-oriented comitative (COM(S)) and instrumental (INSTR)), we were constrained to use an FC task in order to obtain a feasible questionnaire (length) due to this additional factor: In an FC task the positioning becomes the dependent factor and we obtain again a factorial design with only two factors ($2{\times}2$), as represented in Ex. (3).

(3) Pos (PP>OBJ, OBJ>PP) ~ AdvType (INSTR, COM(S)) × Integration (YES,NO)

Thus, only four experimental conditions need to be presented in six different lexical instances each, amounting to a total of 24 test trials, randomized with 48 filler trials.[5] Using instead a LS would have yielded a $2{\times}2{\times}2$ factorial design, resulting in eight experimental conditions, yielding 48 test trials interspersed with 96 filler trials.[6] Thus, changing the AJT

[4] We agree with Goodall (2021) that fatigue effects impose restrictions on the factorial design of an experimental study. However, we do not agree with Goodall's further conclusion that more complex designs than $2{\times}2$ should be avoided if possible. As the following discussion will make clear, there are alternative methods to take an appropriate experimental design and the length of a questionnaire into account.

[5] Here, the test filler ratio could even be increased: Sprouse & Almeida (2017) state that a length of about 100 trials comes without risking fatigue (for LS, FC, YN and ME tasks).

[6] Rating (1-5) ~ AdvType (INSTR, COM(S)) × Integration (YES,NO) × Pos (PP>OBJ, OBJ>PP)

from LS to FC reduced the length of the questionnaire by half, from 144 to 72 trials (meeting minimum standards of six lexicalizations per condition and a filler ratio of 1:2).[7]

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 might 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, also known as syntactic satiation (see Dąbrowska, 2010) or **habituation effect** (Cornips & Poletto, 2005), but these can also coexist with across-the-board effects causing participants to become more accepting in general (Snyder, 2021). To prevent major influence of these **order biases**, i.e. to counterbalance their effects across stimuli, it is mandatory to present the questionnaires in a randomized order, up to individual randomizations for each participant. In doing so, it is advisable to make sure that filler trials intersperse with test trials, 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). 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. After all, order might affect the choice, e.g. in always choosing the first option (Sprouse & Almeida, 2017) or in favoring the lower option (as it is nearer to the continue button), in particular, if participants are indecisive. Failing to randomize the presentation order might thus lead to differences in proportions only due to the presentation order, which might then be mistaken as a (true) effect.

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). Therefore, no participant is allowed to be subjected to see 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 Dean & Voss, 1999; Gibson & Fedorenko, 2013; Schütze & Sprouse, 2014). Consider for example the experiment described in Ex. (2), exhibiting a $3{\times}2$ design: here, we would need one list for each of the six experimental conditions to allocate participants to such that each participant only is subjected to a single condition of a test item, i.e. each test item is presented exactly once on each list.[8] For the study described in Ex. (3) we would thus need four lists as a $2{\times}2$ design

[7] An alternative would be 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).

[8] 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 2x2 (see also Häussler &

results in four experimental conditions.

Of course, not only the sheer amount of filler trials and their positioning in a questionnaire is highly relevant, but also their nature: not only test items, but also filler items need to be carefully constructed. To avoid the last type of biases, **scale bias**, 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 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*). 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).[9] It is further strongly recommended to match acceptability degrees of filler stimuli to those of the test stimuli in order to avoid that ratings for 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).

The choice of appropriate filler items depends on the nature of the test items, not only due to the need to create lists balanced for acceptability (see Goodall, 2021), but also to properly blend test trials in with filler trials. Thus, the nature of filler items can only be exemplified with regard to specific test items. For this illustration purposes, we refer to the experiments presented in Kiss et al. (2021). Consider as an illustration Ex. (4), an item taken from the FC experiment illustrated in Ex. (3) where participants needed to select their preferred serialization.

(4) Test Item: affirmative COM(S)

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. (4) 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

Juzek, 2021, p. 100; Goodall, 2021, pp. 11f). Alternatively, for example, the factor AdvType 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. Employing AdvType as a between-item variable reduces the number of required lists to two.

[9] 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).

a disambiguating addendum in the test stimuli, which 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 might influence ratings (see Schütze, 1996; Cowart, 1997; Featherston, 2005; Dąbrowska, 2010).[10]

However, this strict compliance to specific characteristics might render test trials quite catchy, in particular by such prominent characteristics like the disambiguating addendum. Thus, as filler trials are not related to phenomenon under investigation, but are needed in order to obfuscate test trials, in Kiss et al. (2021), the filler stimuli 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.[11] By this, filler stimuli were superficially similar to test stimuli, blurring the distinction between filler and test trials (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 function is the obfuscation of test trials, filler items can carry different functions. Despite these different functions, they count as fillers in terms of the test filler ratio, even if we cannot consider them fillers in the strict sense. This applies to calibration items

[10] 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 known to be sensitive to word order constraints (see Lenerz, 1977). All NPs were also unmodified as syntactic weight might 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. (4). Furthermore, all objects of VF, i.e. *wh*-indefinites, were inanimate, as animacy has been claimed to influence word order (Keller, 2000a).

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

as well as for fillers used for control purposes (see sections 3.2 and 3.3). A questionnaire typically (at least in LS tasks) starts with some calibration trials, which can also be thought of as implicit training trials, as their purpose is to accustom participants to the scale, but they are not explicitly labeled as training trials. We therefore recommend using implicit training trials in FC tasks as well so that participants become familiar with the task. Here again, the amount of required trials to achieve this purpose is controversial: whereas Schütze & Sprouse (2014) recommend five up to ten trials, Goodall (2021) estimates that already three to five trials are sufficient. 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. (2021) have for example used the calibration stimuli listed under Ex. (5).

(5) exemplary calibration stimuli

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.'

In 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, we avoided scale compression and blinding participants towards subtle effects of test items; this applied to all filler stimuli (see Phillips, 2009).

Also, all other ordinary fillers (with no other function than obfuscating the study goal), made use of ungrammatical, grammatical and marked conditions. They consist 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 research subject by picking up conspicuous characteristics of test items. For the test items in Kiss et al. (2021), these characteristics comprised on the one hand the *wh*-indefinites, which resurfaced in pseudo-fillers as interrogative pronouns, as e.g. in Ex. (6a). On the other hand, the adverbials themselves that 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. (6b).

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

3.2. Including control trials in questionnaires

Even the best efforts to account for the biases and effects through questionnaire design will not necessarily block participants, who are unwilling or unable to cope with the task at hand, and thus may harm the quality of the elicited data.

Addressing this issue, we will discuss two types of filler items that fulfill special purposes: (related and unrelated) control items, and attention items. Control items are defined as triggering a reaction that is known to the designers of the experimental study. In the case of AJTs, this means that the acceptability status of their conditions is uncontroversial (see below). With a known answer, they can be used to identify participants who are unable or unwilling. While control items have been known in AJTs for a while, their exact properties, and their required distribution in a questionnaire have remained dubious (nor are these figures reported often). We will thus not only assess control items as such, but also discuss the required number of control trials per questionnaire. We also see the need to employ different types of control items: unrelated control items, and related control items that bear some resemblance to the structure of the test items. While control trials are known in questionnaires, we have added attention trials as a further extension, with a particular focus on sloppy participants, whose attention possibly falls short of the task.

The essential property of control stimuli is their well-established status: They must be clearly acceptable, or unacceptable, to facilitate the identification of participants failing on these trials. Of course, this raises the question how the status of control stimuli can be established – an issue that has received surprisingly little attention in experimental linguistics. It should be evident that relying on the intuition of the designers of an experimental study may lead to vicious circles and 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) acceptable or (fully) unacceptable without much variation in such a study (see e.g. Gerbrich et al., 2019). Given the nature of related and unrelated control stimuli (which will be discussed below), 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 clauses in which the order of the syntactic arguments violates up to three word order constraints: [+NOM]≺[+ACC], [+PRO]≺[-PRO], and [+ANIMATE]≺[-ANIMATE], as exemplified in Ex. (7) (see also Keller, 2000a).

(7) * *Schwarzweißfilme er liebt.*
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. (7) can be considered unrelated in the context of the studies reported in Kiss et al. (2021), which investigate *wh*-indefinites and adverbials – both of which are missing in Ex. (7). Related control stimuli expedite a thorough reading of the stimuli as they deviate only slightly from the test stimuli (Patterson, 2014, see). In the context of the studies in Kiss et al. (2021), Ex. (8) is a related control stimulus, as both *wh*-indefinites and adverbials are present.

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

Ex. (8) bears some resemblance to a test item – see Ex. (4) – but differs from the test stimuli 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 further advantages on top of the identification of problematic participants: First, they arguably support further disguising the test trials, particularly in combination with pseudo fillers (Ex. (6)). Thus, they help preventing confirmation biases as well as conscious response strategies. Secondly, we suppose that they are less salient than unrelated control trials (see Ex. (8)). They thus counterbalance the possibility that trained participants identify control trials as outstanding, and thus enhance the possibility of identifying non-cooperative participants, particularly in combination with a proper distribution of related and unrelated control trials, to be discussed in section 5.2.

3.3. Including attention trials in questionnaires

While the primary purpose of control trials is to identify participants that are unable or unwilling to cope with the task, the purpose of attention trials is to identify participants who do not pay enough attention. Attention stimuli provide a known (gold) answer, just like control items, but are always clearly unacceptable.[12] They also differ from control items in that the relevant manipulation is comparatively obscure, introduced at a late point of the stimuli and should also consist of rather small elements, and hence can be overlooked. As an illustration, consider word order variation of the negation in the addendum to the stimulus in Ex. (9).

(9) * *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.'

[12] In FC tasks, the unacceptable attention condition is paired with its perfectly acceptable alternative, so that an attention stimulus showing word order violations is paired with a stimulus in which all these violations are reversed, or an attention stimulus showing agreement violations is paired with a stimulus in which the violations are not present.

Here, the violating negation ('*nicht*') occurs late in the example, is a short element, which is surrounded by other short elements. Attention items 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). Oppenheimer et al. (2009) argue that participants failing on an attention check will also fail to follow the instructions of other trials, resulting in less reliable data. However, Prolific Team (2021) define a fair attention check as a test whether a participant has paid attention to the question, not so much to the instructions above it. Prolific Team (2021) further suggest that attention checks must not *"leav[e] room for mis-interpretations"*. As an example, they provide the question repeated in Ex. (10):

(10) 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. What is more, attention checks are simply not feasible in an AJT, where a participant is asked to rate the naturalness of a stimulus since the requested clarity can never be reached even if instructions are changed – which must be considered unfair anyway. 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. It is even possible that attention checks lead to Hawthorne (observer) effects as socially desirable responses might be triggered (see Clifford & Jerit, 2015). Finally, attention checks may simply be too noticeable, particularly for participants that have been trained in answering questionnaires on crowd-sourcing platforms (Gadiraju et al., 2015; Juzek & Häussler, 2017).

None of these problems emerge for attention trials. The feeling of being tried to be trapped is prevented as the checks are not that obvious. It is hence unlikely that **attention trials** (as opposed to attention checks) will change participants' behavior.

3.4. And how many?

It should be clear that the effectiveness of control and attention trials depends on the proportion of such trials included in a questionnaire. We have noticed that even if authors report on the number of fillers included in a questionnaire, they frequently do not disclose how many fillers were used as control trials (of course, attention trials cannot be reported by prior studies, as we have introduced the concept). We will discuss the required proportion of control and attention trials in detail in section 5.2, where we will discuss appropriate measures for FC and LS studies, but the following points need to be considered: First and foremost, the number of control (and attention) trials included in the questionnaire must be such that participants should not be able to answer enough control (and attention) trials by chance alone. Furthermore, 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 (after calibration trials have been presented), which again impacts the number of control and attention trials. It should thus be clear that using control trials only at the beginning of a study will not help detecting participants who "doze off" during a study; but again using control trials only in the last part – as proposed by Juzek (2015) – will not help detecting participants who have been distracted at the beginning of a survey. Considering randomization, should also be clear that subsequent control trials should be avoided, so that response strategies can be inhibited. While these maxims should be followed, we will show that precise measures can be given to determine the amount of control and attention items per questionnaire in section 5.2.

4. Control by choosing an adequate participant pool

Control and attention trials provided the questionnaires with the necessary means to identify inapt participants. As a further step, let us consider the careful choice of an adequate online platform to recruit participants from. The studies concerning fraudulent behavior in section 1 (see Oppenheimer et al., 2009; Downs et al., 2010; Buchholz & Latorre, 2011; Eickhoff & de Vries, 2013; Gadiraju et al., 2015; Häussler & Juzek, 2016; Chandler & Paolacci, 2017) relate to crowdsourcing platforms geared towards HITs. By now it should be clear how AJTs differ from HITs; 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.

A first difference manifests itself in the combination of fees paid by AMT and the number of workers needed (e.g. to attain reliable labels). When compared to AJTs, both figures are smaller. While 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), HITs require much smaller batch sizes. 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.

The small numbers of workers in HITs can be justified by the assumption that quality control is determined by majority votes (see Mason & Suri, 2012). On the contrary, in AJTs, variability in the data elicited is expected, such that trials with verifiable answers are needed to vet participants. We therefore introduced control and attention trials in sections 3.2 and 3.3. In section 5 we will discuss how many of them are needed to be considered a valid control measure. Thus, a major difference consist 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. Although it is in principle possible to replicate full questionnaires as one single batch, workers on AMT generally expect shorter projects. Schnoebelen & Kuperman (2010) recommend small batches to ensure attentiveness and to prevent dropouts of participants. Therefore, they adopted their Cloze completion test to the usual format of AMT which renders the process entirely different from that in the laboratory. As a consequence, elicitation stretches over several days, where the exposure of the same stimuli presented the day before to the same participant cannot be excluded. For these reasons and the issues discussed in section 3, like repeated exposure, we argue that Schnoebelen & Kuperman's (2010) and Eickhoff & de Vries's (2013) recommendation to employ smaller batch sizes in order to ensure attentiveness and to deter spammers is clearly not an option for AJTs.

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). Consequently, Schnoebelen & Kuperman (2010) need to reject *workers* that should not have participated in the study in the first place (e.g. people who they consider ineligible as they grew up multilingually or abroad the US).[13]

This issue has been addressed by new platforms which are geared towards the needs of academic researchers in the meantime: e.g. Prolific[14] (founded in 2014), CloudResearch[15] (founded 2016), and Positly[16] (founded in 2017). 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 those participants that match these prescreening criteria.[17]

Prescreening for AJTs mainly concerns native language but might 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. This might not be surprising after all since fluency is in many cases absolutely sufficient to be eligible, i.e. for studies in other areas than linguistics. For linguists, Prolific provides the most elaborate screening options and is thus by far the most suitable platform for AJTs.[18]

[13] 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).

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

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

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

[17] 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.

[18] 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.

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 (or repeating a study).

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.

Not only do you want to screen eligible participants, but also high-quality participants. To provide optimal support by impeding fraudulent behavior, 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 combination of the methods discussed in section 3 with the careful selection of an appropriate participant pool is an important step in ensuring data quality. Yet, the choice of an appropriate platform as well as the inclusion of control and attention items only provide the means to determine whether participants in a study were inapt or unwilling. In the following section, we will show how this means can be employed to determine problematic participants based on both latency- and response-based methods.

5. Control by analyses: latency- and response-based identification of non-cooperative participants

To remove noise emerging from inapt participants, we need to inspect the elicited data for signs of fraudulence, incompetence, or inattention. In doing so, we employ latency-based methods, 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 by RTs, which can thus be used to identify adverse behavior on the part of participants: short outliers (on the left tail) might indicate guessing, whereas long outliers (on the right tail) might 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), 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 as *"frequent distractions would arguably reduce the quality of responses"*. Distractions could indicate that participants are rereading the stimulus multiple times or that they are consulting with a third person. It follows from both assumptions that participants do not report their immediate perceptions as required (see section 2).

In contrast, short outlying RTs have been commonly used in web-based experiments to identify 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 tend to exhibit long intermissions, resulting in single extreme long RTs, affecting their mean RT so that it becomes inconspicuous. Despite fears that participants become 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**, since participants might also switch back and forth between rushing and dawdling around. The identification of outliers, however, is not a 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, 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 (SD), 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 is due to participants and different (groups of) items (attention and control trials, in particular) exhibiting different RT distributions (see Figure 1), differing in just the characteristics determining the exclusion criterion.

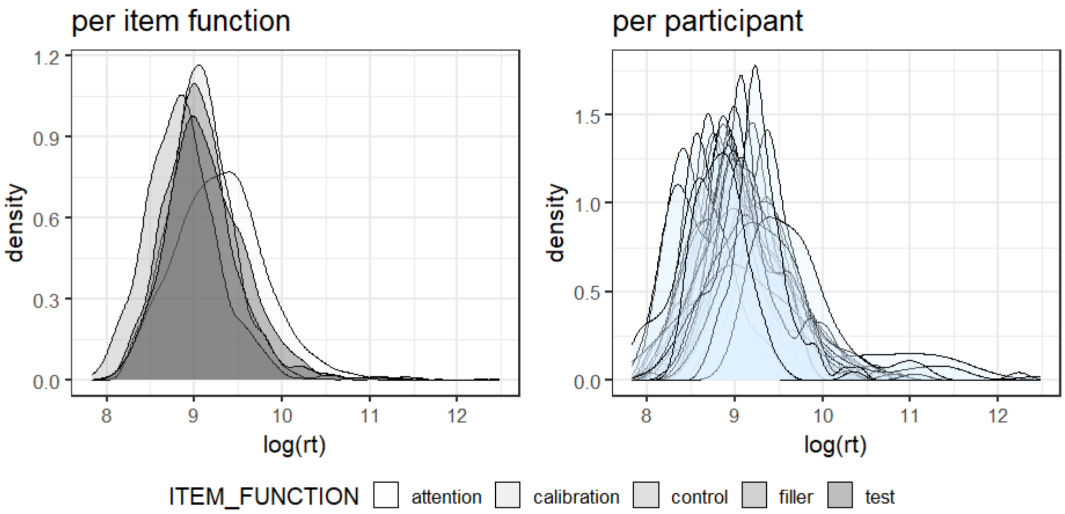

Figure 1: Density of RTs (Experiment 2 from Kiss et al., 2021)

Naturally, participants exhibit different reading and judging times (see Dandurand et al., 2008; Häussler & Juzek, 2016). However, 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 might be a normal RT in case of attention trials, might 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, if distributions are indeed dissimilar, determining only one general cutoff point would arguably fail in separating spurious from genuine outliers. And, even worse, it might truncate the distribution of one or several groups (see Lachaud & Renaud, 2011). It is not unlikely that excluding slow but genuinely responding participants might induce biases, like IMCs do (see section 3.3, Vannette (2016)).

Taking these considerations into account, we are proposing a *recursive multi-factorial outlier detection* (ReMFOD) for capturing genuine intermissions and rushes. ReMFOD aims at identifying individual trials with outlying RTs. In doing so, we are accounting for different RT distributions of different participants and item functions, as well as swamping and masking effects. Underpinned by these individual trials identified as intermissions or rushes as outlier counts, we will detect underperforming participants by means of proportion of trials not responded to wholeheartedly.

For identifying individual trials with outlying RTs, we compare the RT of each trial to two cutoff points: The first one 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 **both** cutoff points, it will be designated as a **genuine intermission** (see Equation (1)). In the same vein, we define a lower cutoff point for capturing trials whose RT fall below both cutoff points as **genuine rushes** (see Equation (2)).

$$\begin{aligned}\text{cutoff\_intermission} = \max\{ & \\ & \text{median(RTs:participant)} + 2.5 \times \text{mad(RTs:participant)}, \\ & \text{median(RTs:item\_function)} + 2.5 \times \text{mad(RTs:item\_function)}\} \quad (1) \\ \text{cutoff\_rushes} = \min\{ & \\ & \text{median(RTs:participant)} - 1.5 \times \text{mad(RTs:participant)}, \\ & \text{median(RTs:item\_function)} - 1.5 \times \text{mad(RTs:item\_function)}\} \quad (2)\end{aligned}$$

To account for swamping and masking effects (Ben-Gal, 2005), 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. The iteration number, i.e. *outlier degree*, can be seen as describing the severity of the outlier.

Consider Figure 2 for an illustration of the different cutoff points. Different outlier types are marked by different shapes: solid circle shaped trials do not exceed any cutoff points, i.e. these are normal or valid RTs. Plus shaped trials are considered outlying with respect to all trials of the same participant regardless of the item function, but not with respect to all trials with the same item function regardless of the participant. The opposite is true for cross shaped trials i.e. these are outlying with respect to their item function but not with respect to the responding participant. Star (up and down triangle) shaped trials are the only RTs we consider as genuine intermissions or rather genuine rushes: they exceed the cutoff points of both groups.[19] The upper plots are grouped by item function, displaying RTs of all participants, the lower plots are grouped by exemplative participants, showing RTs in response to all item functions. The red dotted lines indicate singular absolute cutoff points computed with respect to all trials of all participants.

Thus, the overlap of genuine and spurious RTs is illustrated in Figure 2 by shaped points inside the cluster of solid circle points. Note that the RTs in the plot are log-transformed to facilitate their presentation, resulting in a condensed representation (in the upper part), so that they appear to be much closer than they really are.

After having identified individual trials as genuine intermissions or rushes using ReMFOD, we will now turn to identifying participants as underperforming in general. As being distracted and being in a rush both are signs of uncooperative behavior, we define a threshold of how many trials participants are required to respond to genuinely. We propose that participants who have responded genuinely to less than 90 % of trials, will be discarded because

[19] Note that the shapes might overlap as these outliers have been computed by different ReMFOD procedures differing in the groups it employed to identify outliers, and also include outliers of higher degrees.

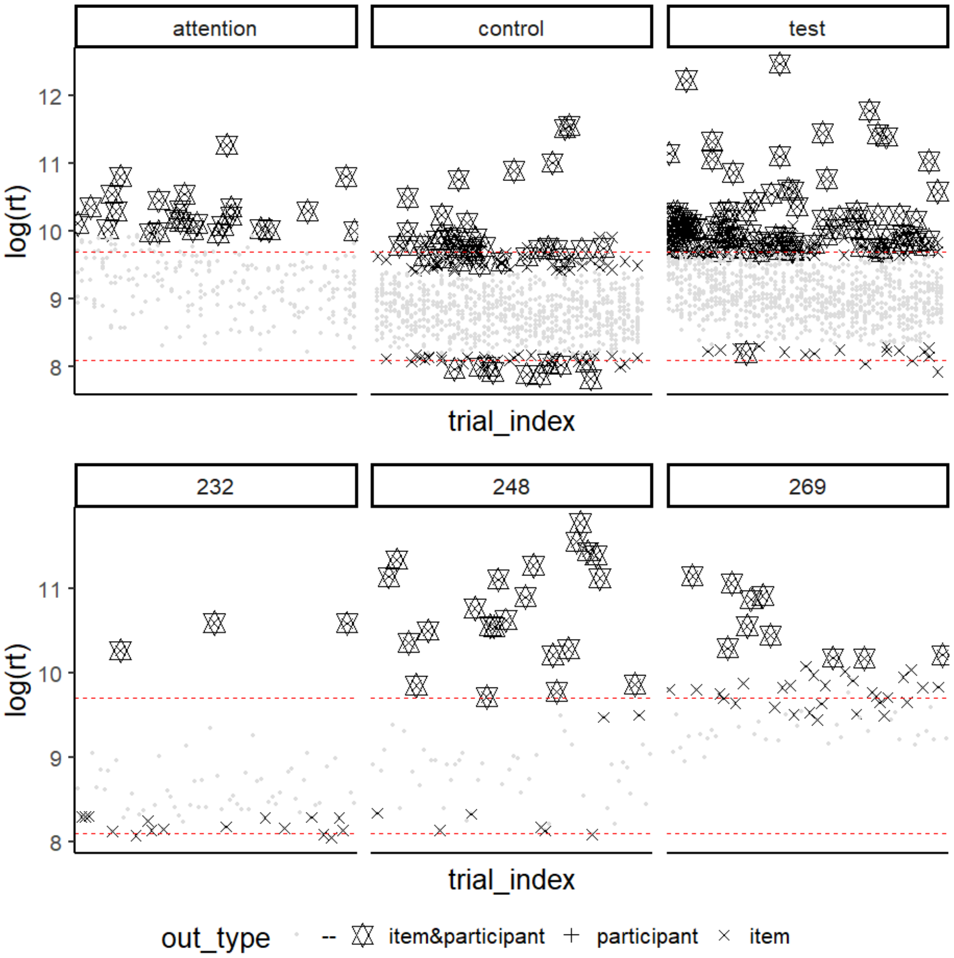

Figure 2: *Recursive multi-factorial outlier detection* (ReMFOD)

we suppose that they do not meet the task with the necessary seriousness.

5.2. Response-based identification

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. It should be noticed that approaching FC studies is easier than LS studies, due to the nature of the responses in the different types of studies.

5.2.1. Forced-Choice

For the sake of simplicity, we will start with the analysis of responses to control and attention trials in FC studies. Contrary to test trials, "correct responses" are known in advance, as a grammatical stimulus is directly compared to its ungrammatical version. Thus, 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 % (see 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):

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

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. For example, if a 75 out of 100 trials are answered correctly, it can be assumed that 50 (i.e. 75 - 25) trials have been *solved*, while 25 have been *guessed* correctly (and another 25 wrongly) (Frederick & Speed, 2007). We assume that this assumption can be transferred to AJs if native speakers are employed, and in particular to attention trials, which can seduce participants to guess if they lose one's temper searching for the manipulation. Thus, this differentiation between correctly *solved* and *guessed* answers shows that an assumed threshold of 50 % correctly *guessed* responses is insufficient. Frederick & Speed (2007) have further outlined that scores reached by guessing, i.e. by chance, lie in fact within a broader (symmetrical) range around a chance threshold. Hence, a higher threshold is needed. They report that 90 % correct responses are required in an FC task.

N					k				
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**
6	.66	.34	.11	**.02**					
8	.86	.64	.36	.15	**.04**	**< .00**			
10	.95	.83	.62	.38	.17	.05	**.01**	**< .00**	
12	.98	.93	.81	.61	.39	.19	.073	**.02**	**< .00**
14	.99	.97	.91	.79	.61	.40	.21	.09	**.01**
16	1.00	.99	.96	.90	.77	.60	.40	.23	**.04**

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

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 passing 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 only 12, i.e. 75 %, correct responses are sufficient.

Following this lead, we strongly argue against a decision criterion based on a 50 % proportion (as is illustrated by the probabilities in the diagonal in Table 2), as well as against using less than eight attention/control trials. Instead, a threshold of about 90 % seems reasonable for AJs, too.

When employing related and unrelated control items, as proposed in section 3.2, and if only one of the groups fail, we need to consider combined chances. If each group is evaluated by itself, it should comprise of at least eight members. If a joint evaluation of unrelated and related control trials is carried out, then four trials in each group will suffice. When employing eight trials per group, then the separate evaluation of each group will result in lower chances to pass both tests by guessing than the chances to pass one single but more extensive test, i.e. the joint evaluation of both groups, by chance. Thus, whenever possible, we recommend the separate analysis of the groups.

5.2.2. Likert-Scale

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 FC trials, the analysis amounts to counting correct responses for attention trials. This raises the issue, however, whether the neutral option in an LS should count as correct response or not.

Consider that the neutral 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).[20] If we take the neutral option not to be acceptable, this number is reduced to five correct responses out of six trials (see Table 3b).

Arguably, as the neutral point does not reliably indicate that the participant has indeed spotted the manipulation, we recommend excluding the neutral point from the acceptable responses.

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 correct responses to the totality of all LS control trials, as depicted in Ex. (11) for a *5-pt*-LS.

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

a. **positional:** all stimuli should receive expected ratings, i.e.
< 3 for ungrammatical stimuli
>= 3 for grammatical stimuli

[20] In our view, it is too strict to assume that all trials need to be answered correctly. Even diligent participants may become subject to fatigue and resulting slips in a comparatively long questionnaire.

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). Of course, the relational account is thereby inherent in the positional account. Furthermore, the two groups – ungrammatical and grammatical controls – are dependent on each other as concerns the evaluation in a relational account but not in a positional account.

On a positional account, the assessment of participants' performance is carried out separately for grammatical and ungrammatical stimuli, focusing on the pertinent side of the scale in each case. 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 (if existent).

We recommend allowing the neutral option as legitimate response to grammatical stimuli, but not to ungrammatical stimuli (cf. Table 3) as it does neither reliably indicate the rejection

N	k								
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**
6	.82	.54	.23	**.05**					
8	.95	.83	.59	.32	.11	**.02**			
10	.99	.95	.83	.63	.38	.17	**.05**	**.01**	
12	1	.98	.94	.84	.67	.44	.23	.08	**< .00**
14	1	1	.98	.94	.85	.69	.49	.28	.04
16	1	1	1	.98	.94	.86	.72	.53	.17

(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								
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**
6	.46	.18	**.04**	**< .00**					
8	.68	.41	.17	.05	**.09**	**< .00**			
10	.83	.62	.37	.17	.06	**.01**	**< .00**	**< .00**	
12	.92	.77	.56	.33	.16	.06	**.02**	**< .00**	**< .00**
14	.96	.88	.72	.51	.31	.15	.06	**.02**	**< .00**
16	.98	.93	.83	.67	.47	.28	.14	.06	**< .00**

(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 option (3) is considered correct or not (for unacceptable stimuli).

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 (e.g. k out of N correct responses required), as well as considering the neutral option (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.[21]

N	grammatical (>= 3)		ungrammatical (< 3)		joint
	k	prob	k	prob	
5	4	.34	4	.087	.0293
6	5	.23	4	.179	.0418
6	4	.54	5	.041	.0223
6	5	.23	5	.041	.0096
7	6	.16	4	.290	.0460
7	5	.42	5	.096	.0404
7	6	.16	5	.096	.0153
7	5	.42	6	.019	.0079
7	6	.16	6	.019	.0030
8	5	.59	6	.050	.0296
8	6	.32	6	.050	.0157

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

Thus, researchers may pick any of the decision criteria listed in Table 4. Under the assumptions listed above, at least ten control trials are needed – namely five grammatical and five ungrammatical – of which participants are required to answer at least four per group correctly (i.e. < 3 for ungrammatical and >= 3 for grammatical trials) in order to be considered cooperative. Together with the minimum amount of six attention trials, we hence need at least 16 fillers for controlling functions. Although this number might appear extensive, it should be no cause of concern: Even in the smallest factorial designs (i.e. $1{\times}2$ or $2{\times}2$), the questionnaire should include at least 24 or 48 filler trials, when employing the minimum test-filler-ratio 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.

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

[21] 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 (see requirement above).

(excluding the neutral point) at least once. We are afraid that this is overly strict when employing a sufficiently large number of items. 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) 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.[22] 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 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 items, and set a threshold of at least 50 % of the optimal distance, as defined in Equation (4):

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

Thus, for a 5-pt-LS the minimum distance of 2 $(0.5 * (5 - 1))$ between both conditions is required. By this, the positional account is inherent (by criterion ≥ 0.5) as 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, this is, average judgments of 1-3, or 3-5 will be accepted.[23]

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 the proportion of outcomes out of the number of possible outcomes under which the decision criterion (Equation (4)) is fulfilled. To allow for comparisons of mean values, let us consider

[22] Typically, Zanuttini et al. (2018) used 15 control stimuli (out of 45).

[23] The extended relational account differs from Zanuttini et al. (2018), as the score is conditioned on the position of the scale. Thus, mean responses to grammatical trials of 3 and to ungrammatical trials of 1 are acceptable under the extended relational account, but not under Zanuttini et al. (2018). By not allowing for any slips, Zanuttini et al. (2018) is stricter than the extended relational account, and thus accepts only a subset of the outcomes accepted by the extended relational account.

pairs of responses where each pair consist of one one grammatical and one ungrammatical stimulus as one event (e.g. 4-3). On a 5-pt-LS, there are 25 different ways to respond to two stimuli (i.e. 1-1, ..., 5-5). All combinations of these amount to 625 possible outcomes for two pairs of stimuli (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 stimuli, and the proportion of optimal distance as in Equation (4), and decide whether the distance between conditions is acceptable under Ex. (11b), i.e. > 0 and Equation (4), i.e. >= 0.5.

Under the assumption that each outcome is equally probable when responding randomly, we can estimate the probabilities to pass control trials by chance by determining the proportion of outcomes passing the decision criteria (see Table 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 %.[24]

pairs	possible outcomes	pure: s > 0	extended: s >= 0.5
1	25	.40	.240
2	625	.43	.112
3	15,625	.44	**.045**
4	390,625	.45	**.030**
5	9,765,625	.46	**.014**

Table 5: Acceptance rates of guessed outcomes, i.e. all possible response combinations, mixed account compared to relational ones

The relational account in Ex. (11b) 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 – 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 but provides no further advantages over the much simpler positional account. The same considerations apply to the proposal by Zanuttini et al. (2018).

[24] Apart from estimating the guessing probabilities for the given relational accounts, the outcomes can hence serve a further purpose: namely to deduce a threshold from them 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.

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, there are no other requirements on the number of trials than to provide high chances to detect guessing participants. Hence, the smallest number of control trials 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 option is a point of discussion, which also manifests itself in the discussion of whether the scale should include a neutral point or not in the first place (see section 2), the positional account can be easily adapted to include different interpretations of the neutral point. This may also include allowing for a maximal proportion of trials answered with the neutral option. By this, the state of the neutral point would be less fuzzy, and the decision criterion by itself would be clearly and more intuitively understandable than Zanuttini et al.'s (2018) approach. Another difference between the present account and Zanuttini et al. (2018) consist in the tolerance to slips: The more control trials are required, 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 a study.

First, we noticed that while web-based elicitation provides many advantages, uncooperative behavior has been identified as a virulent issue. This is due to the loss of experimental supervision, and a faint of participants. Furthermore, participants might be less intrinsically motivated and driven by monetary incentives, leading them to provide less accurate data, resulting in decreased statistical power.

Secondly, we have reflected upon some of the most common tasks used in experimental syntax, namely AJTs. Thereby, we have elucidated that when eliciting AJs following precautions need to be addressed: First of all, it is important that participants grasp what they ought to judge, i.e. that they should not apply prescriptive rules or judge plausibility. Therefore, they should stick to their immediate reactions, i.e. report their perceptions *spontaneously*. This issue has only been addressed by giving instructions to that effect, but has not been explicitly controlled for.

In section 3 we have elucidated the importance of a good questionnaire design that manages controlling certain biases and effects and provides sufficient means to enable the identification of inapt participants. Most importantly, we have highlighted the importance of carefully constructed filler items as a guard against confirmation biases and conscious response strategies by preventing participants from spotting test items. To strengthen the concealment, we recommend employing so-called pseudo-fillers, that are picking up conspicuous characteristics of test items, and related control items, that are designed to be very similar in structure to test items (as opposed to ordinary control items), and that are thereby additionally enhancing careful consideration of items (see Patterson, 2014), and the robustness of the detection of malicious participants. To complement means to identify inapt participants, we introduced **attention items** as an alternative to IMCs (see Oppenheimer et al., 2009) which we consider infeasible for AJTs. They are suitable to test whether participants comply with the instructions, and avoid various concerns regarding IMCs. Only by using a sufficient amount of these items, we are well equipped to identify inapt participants.

In section 4, we have elucidated the importance of a careful selection of a platform, and thus the participant pool to recruit from. Besides, we have been contemplating differences between HITs, towards which common platforms are geared, and AJTs: These concern not only the structure of the task, but in particular, the eligibility of participants. Modern platforms especially geared towards researchers do not only alleviate some of the problems resulting from the loss of supervision mentioned above but more importantly, allow targeting the population, e.g. monolingual native speakers being naïve towards the research subject.

In the last chapter, we discussed identification of uncooperative participants via latency-based methods, i.e. by inspection of *response times*, as well as response-based methods, i.e. by inspection of *responses*.

When considering latency-based methods, we argue that the identification of spammers is a minor issue due to the increase of control on part of online participant platforms, such that slow and long outliers can be combined 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, we identified underperforming participants.

When discussing response-based methods, we noticed that in general, it is expected that participants perform well above chance. Therefore, we investigated of probabilities of passing control trials, i.e. of giving more than a specific proportion of correct responses, by chance, 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). We thus suggest for FC tasks that using a threshold of only 50 % as well as using less than eight control/attention trials is heavily advised against. Using eight trials implies a proportion of 90 % correctly answered trials needed, whereas using at least 16 control trials lowers the proportion to 75 % (in order to properly identify complying participants). Regarding the analysis of LS control trials, we have reviewed relational and positional accounts and argued for the use of a positional account (i.e. counting right and wrong answers) as the relational accounts come with major drawbacks, as they need to justify an interpretation as interval data and fail to control for scale compression. A major point of discussion of the positional is however, how to treat the neutral point. We propose to include it as an acceptable option for grammatical but not for ungrammatical stimuli. Considering the joint guessing probabilities of both groups evaluated individually, only ten control trials are needed – five grammatical and five ungrammatical out of which four are required to be responded to correctly.

Taken all of these measures together, we are 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:

$$\begin{aligned} &\frac{N!}{k!(N-k)!}p^k q^{N-k}, \\ &\text{where} \\ p = &\text{ probability of a correct response} \\ q = &\text{ probability of an incorrect response} \end{aligned} \tag{5}$$

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).

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
ReMFOD	recursive multi-factorial outlier detection
RT	response time
SD	standard deviation
TFR	test-filler-ratio
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

Birnbaum, M. (2004). Human research and data collection via the internet. *Annual review of psychology*, *55*, 803–32. https://doi.org/10.1146/annurev.psych.55.090902.141601

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

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

Clifford, S., & Jerit, J. (2015). Do Attempts to Improve Respondent Attention Increase Social Desirability Bias? *Public Opinion Quarterly*, *79*(3), 790–802. https://doi.org/10.1093/poq/nfv027

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

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

Culbertson, J., & Gross, S. (2009). Are linguists better subjects? *The British Journal for the Philosophy of Science*, *60*, 721–736. https://doi.org/10.1093/bjps/axp032

Dąbrowska, E. (2010). Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, *27*(1), 1–23. https://doi.org/doi: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*, 428–34. https://doi.org/10.3758/BRM.40.2.428

Dean, A., & Voss, D. (1999). Designs with two blocking factors. *Design and analysis of experiments* (pp. 387–420). Springer. https://doi.org/10.1007/978-3-319-52250-0_12

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/doi: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.

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* (pp. 73–99).

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). 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/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. (2015). *Acceptability judgement tasks and grammatical theory* (PhD thesis). Jesus College, Oxford.

Juzek, T. S., & Häussler, J. (2017). Hot topics surrounding acceptability judgement tasks. https://doi.org/10.15496/publikation-19039

Kazai, G., Kamps, J., & Milic-Frayling, N. (2011). Worker types and personality traits in crowdsourcing relevance labels. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 1941–1944. https://doi.org/10.1145/2063576.2063860

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. (2021). *Word order constraints on event-internal modifiers*. https://lingbuzz.net/lingbuzz/006319

Lachaud, C., & Renaud, O. (2011). A tutorial for analyzing human reaction times: How to filter data, manage missing values, and choose a statistical model. *Applied Psycholinguistics*, *32*, 389–416. https://doi.org/10.1017/S0142716410000457

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]*.

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. A, Human experimental psychology*, *43*(4), 907–912. https://doi.org/10.1080/14640749108400962

Musch, J., & Reips, U.-D. (2000). A brief history of web experimenting. *Psychological experiments on the internet*. 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/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/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/https://doi.org/10.1016/j.cogpsych.2018.02.002

Robinson, J., Rosenzweig, C., & Litman, L. (2020). Conducting online research on Amazon Mechanical Turk and beyond. In L. Litman & J. Robinson (Eds.). 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. 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/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/doi: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/doi:10.1515/lingvan-2016-0070