Amsterdam University Press

NEDERLANDSE TAALKUNDE 29.1 (2024) 49-75
HTTPS://DOI.ORG/10.5117/NEDTAA2024.1.005.SCHO

DISCUSSIE 2: SYNTAXIS

Consistency and variability in acceptability judgments from naive native speakers

Gert-Jan Schoenmakers
Utrecht University
g.t.schoenmakers@uu.nl

Roeland van Hout
Radboud University
roeland.vanhout@ru.nl

Abstract
Syntactic theories are typically construed based on acceptability judgments. These judgments are increasingly often collected experimentally, testing larger sets of linguistically naive participants. An important assumption is that participants have a very clear understanding of what it is they are asked to do, which can be assessed by establishing their internal consistency. The question we address in this paper is whether 'human measuring instruments' are consistent in their judgments. To this end, we re-examined the judgment data from Schoenmakers (2023), where three types of violations of the prescriptive norm and object scrambling sentences were evaluated. We used Generalizability Theory to investigate the degree of covariation in the judgments and found that the internal consistency was poor in the norm violation item sets, but excellent in the scrambling item set. A difference between the data patterns is that the former item sets led to 'sledgehammer' effects between the stigmatized and non-stigmatized variants, which left little room for participant variation. Our analyses show that judgments from naive native speakers can adequately serve linguistic theorizing, both in the case of stigmatized and non-stigmatized variation. Furthermore, we performed cluster analyses to identify subgroups of participants to get a better grasp on the variation in the data set. We conclude that specific statistical analyses can help understand data and advance linguistic theory building.

Keywords: reliability, Generalizability Theory, cluster analysis, scrambling, prescriptive norm violations

1. Introduction

Intuitive judgments of linguistic acceptability are fundamental in syntactic research. These judgments have traditionally been provided by the language researcher themself, and occasionally a handful of their direct colleagues, presumably all linguistic experts. However, there has been a proliferation of acceptability judgment experiments in the syntactic literature at least since Schütze (1996) and Cowart (1997), that is, experiments in which the acceptability of larger item sets is systematically rated by larger numbers of native speakers who are typically naive to linguistic theory. One advantage of this method of data collection is that, once enough data are collected in a properly controlled design, it permits statistical analysis to test specific hypotheses (see also Gibson & Fedorenko 2010, 2013; Gibson, Piantadosi & Fedorenko 2013).[1] The observed distinctions or patterns in the data can then be linked to linguistic theory via a 'linking hypothesis', which translates linguistic processes or operations into concrete experimental outcomes (Phillips et al. 2021, Francis 2022). This is not a trivial step, given that 'grammaticality' has now become a psychometric construct. That is to say, human beings do not have direct conscious access to the syntactic structure building mechanism (cf. Schütze 1996: § 2.4), and so the grammatical status of a stimulus sentence can only be inferred based on the acceptability judgments from (naive)[2] listeners (see e.g. Birdsong 1989 for discussion of the terms *grammaticality* and *acceptability*).

In judgment experiments, humans essentially function as 'measuring instruments' (Rietveld & van Hout 1993, Rietveld 2021) that are employed to reveal underlying cognitive or emotional abilities, concepts, or structures. This approach of data collection may lead to questionable results, however, because humans, or language users, may provide judgments seemingly at random. They may apply different criteria for judgment, or unclear criteria,

1 Gibson and Fedorenko (2010, 2013; Gibson, Piantadosi & Fedorenko 2013) claim that formally collected quantitative data are inherently superior to nonquantitative researcher intuitions (see also Featherston 2020). We disagree with this claim, because it is not supported by empirical evidence (see Sprouse & Almeida 2012, Sprouse, Schütze & Almeida 2013, Mahowald et al. 2016, Chen, Xu & Xie 2020). Instead, we support the view that the two data sources can be used for different purposes, i.e. for theory building and for hypothesis testing (cf. Phillips 2010, Newmeyer 2020, Sprouse 2020). But as Edelman and Christiansen (2003: 60) put it, "putting forward a theory is like taking out a loan, to be [repaid] by gleaning an empirical basis for it; theories that fail to do so (or their successors that may have bought their debts) are declared bankrupt."

2 Judgments provided by the linguistic expert are arguably also experimental in nature, although they come from a fully uncontrolled experiment with only one participant or a small number of (potentially subconsciously biased) participants (Gibson & Fedorenko 2013).

or no criteria at all, and so they may not even be capable of systematically or consistently evaluating the precise meaning or acceptability of utterances at all (cf. Schütze 1996: § 6.3.2). The grammaticality status of a sentence may thus be obscured in various ways when acceptability judgments are collected, which is due to the nature of the human measuring instrument. Discussing such individual variation, Birdsong (1989: 69) suggests that "[m]etalinguistic data are like 25-cent hot dogs: they contain meat, but a lot of other ingredients, too. Some of these ingredients resist ready identification." Yet, the internal consistency of judgments collected from naive participants can be established, by computing the *reliability* of the judgments. Computing reliability scores is common in the social sciences and in phonetic research, but completely absent in the evaluation of acceptability scores in grammar. It is crucial to establish the reliability of judgments before addressing their validity, however, i.e. whether the collected judgments in fact measure what they intend to measure. Any question of validity on grammatical evaluations cannot be answered properly when the question of reliability or internal consistency is not addressed first.

We stress that distinguishing between data patterns and theoretical claims is crucial. Certain researchers claim that acceptability judgments from linguistically naive participants should not and cannot be used to inform linguistic theory, because such participants are not trained in detecting grammatical operations and they cannot successfully suppress performance effects, i.e. effects that are not due to the language system proper but are instead caused by factors such as memory limitations, semantic or pragmatic interpretation, and/or prosody (cf. Francis 2022). Note, however, that this does not mean that linguistically naive participants do not have intuitions about sentences. These intuitions are immediate and unreflective, and it is up to the researcher to reflect on them and to link their observations to linguistic theory (i.e. to apply *grammatical reasoning*, Häussler & Juzek 2021). With regard to the nature of intuitions collected from naive participants, reliability analysis can provide valuable insights: based on the rhetoric used to justify the claim that these data should not and cannot be used in linguistic theorizing, the collected judgments should be inconsistent, as indicated by low reliabilities.

High reliability, by contrast, can serve as an indication of the success of the chosen method, as it is a reflection of the extent to which participants understood the assignment and could in fact perform the task in a similar and consistent manner. Various procedures and techniques can be applied to assess the internal consistency of human acceptability judgments, which we explain in Section 2. We introduce the terminology that is common in reliability

research, but rather unknown in the field of experimental syntax (Langsford et al. 2018 is a notable exception, as are studies on the language proficiency of second language learners; e.g. Hartshorne, Tenenbaum & Pinker 2018, who report the reliability of their collected judgments that they use to measure the language proficiency of L1 and L2 speakers of English). We discuss the methods and data of an acceptability judgment experiment from Schoenmakers (2023) in Section 3, which we use as a case study for the present paper.

In Section 4, we compute the reliabilities of three data sets from Schoenmakers (2023) and investigate the patterns of item and individual variation in them. Schoenmakers investigated the acceptability of three different violations of the prescriptive norm (stigmatized variation, viz. subject *hun*, comparative *als*, and auxiliary *doen*, see (1)). Participants rated these items on 100-point scales.

(1) a. *Anna hoopt dat **hun** de finale bereiken.* [subject *hun*]
'Anna hopes that them will reach the finals.'
b. *Leon drinkt minder koffie **als** zijn collega.* [comparative *als*]
'Leon drinks less coffee as his colleague.'
c. *Johan **doet** dan normaal gesproken altijd sporten.* [auxiliary *doen*]
'Johan usually always does sports then.'

The three different prescriptive norm violations (subject *hun*, comparative *als*, and auxiliary *doen*) each have their own history of stigmatization. Comparative *als* is the most well-known among the three and can already be found in prescriptive grammars from the 16th century (van der Meulen 2018). Van der Meulen (2018) shows that the comparative *als* variation is being rejected in prescriptive grammars more and more in the 20th century, and the public opinion is equally if not more dismissive. Subject *hun*, by contrast, had not been described until the early 20th century (van Bree 2012), but is nonetheless highly salient in contemporary discussions, where it is considered extraordinarily bothersome (van Hout 2003, 2006). This is different for auxiliary *doen*, which can be found in certain dialectal regions (Giesbers 1983/1984) and might not even be ubiquitously recognized as a norm violation (Sert et al. 2023). If participants are sensitive to the prescriptive norm, there may be differences between the types of norm violations, but there will also be a sharp distinction in their acceptability judgments between non-violations and violations. This distinction may in fact overshadow all other sources of variation. Our hypothesis is that naive listeners are highly sensitive to violations, particularly because the task they carry out solicits apparent grammatical evaluations (see also Schoenmakers 2023). If this is true, the main distinction in the data concerns violation

versus non-violation, without much room for variability between items and/or participants. And if there is no systematic variation between items and participants, reliability scores are low.

To further investigate the balance between the three prescriptive norms, we applied a cluster analysis to compare how the judgments of the three norm violation types may vary between individuals. More specifically, we investigate whether we can distinguish subgroups of participants on the basis of patterns in the judgments, given that participants may vary in the extent to which they reject the three different violation types.

The outcomes of the item set with prescriptive norm violations form the backdrop against which the outcomes of a fourth item set from Schoenmakers (2023) will be interpreted. This set is related to scrambling. Scrambling is a type of word order variation in which the direct object can surface on the right or left side of an adverb, see (2). Crucially, scrambling constructions are not involved in any discussions on the prescriptive norm, that is to say, there are no stigmatized variants, nor is there generally known language advice about the appropriate word order. Both the scrambled and unscrambled variant are, *ceteris paribus*, acceptable constructions of Dutch (see Schoenmakers & de Swart 2019).

(2) a. Milan gaat absoluut **het boek** lezen. [unscrambled]
b. Milan gaat **het boek** absoluut lezen. [scrambled]
'Milan will absolutely read the book.'

The theoretical literature, however, claims that scrambling is determined by information structure. Specifically, the claim is that focused definite objects must be located in unscrambled position, to the right of an adverb (2a), and topical definite objects in scrambled position, to the left of the adverb (2a) (e.g. Schaeffer 2000, Broekhuis 2008, Neeleman & van de Koot 2008). Experimental data do not corroborate this claim. Schoenmakers, Poortvliet & Schaeffer (2022) find that, while information structure affects word order in constrained language production, it does not determine it. Furthermore, Schoenmakers (2023) reports on a follow-up acceptability judgment experiment and shows that the two word orders are both fully acceptable for both topics and foci. Unlike in the case of the norm violations, there are no wide gaps between the judgments of the sentences in the different conditions.

How should we interpret this outcome? One may conclude that naive speakers are unable to handle such subtle distinctions, associated with complicated concepts such as information structure. Linguistic experts, on

the other hand, are trained to be sensitive to such fine-grained distinctions. If naive speakers are unable to give systematic judgments, their judgments should be random with low reliabilities as indicators of unsystematic evaluation. Thus, the discrepancy between scrambling judgments reported in theoretical work and judgments collected from naive native speakers highlights the importance of determining the reliability of a data set of judgments. When the judgments are consistent, investigating the potential subgroups of individual participants is a next step. High consistency on the overall level does not preclude the existence of smaller subgroups of naive speakers who are somehow deviant from the majority. These steps are crucial in determining how we can use judgment data from naive listeners in linguistic theory building. In Section 5, we apply reliability and cluster analyses to the scrambling item set and show that, while the overall reliability of the data set is high, smaller subgroups of naive speakers do differ in their evaluations, so that we can distinguish subgroups. This is an important conclusion. Perhaps the scrambling process we investigate is part of processes of language variation and change, with the consequence that naive speakers are not homogeneous in their judgments. Speakers may be different in the extent to which they accept or reject syntactic constructions and the existence of subgroups may reflect the different stages in processes of variation and change.

2. Reliability and reliability measures

The *reliability* of our data pertains to the internal consistency of the judgments provided by participants, that is, to the extent to which participants provided judgments that are correlated (Rietveld & van Hout 1993, Rietveld 2021). This definition crucially does not refer to absolute agreement between participants, the extent to which participants return identical values, as measured by e.g. Krippendorff's alpha (Krippendorff 2013: Ch. 11), but to the degree of covariation in their judgments, i.e. the extent to which the relative patterns of their responses 'move' in the same direction. More technically, in reliability analysis acceptability ratings are partitioned into two components. One component represents the true score, the other contains 'distortions' or the 'error', i.e. by-participant deviations from these true scores. Note that the error variance is inherent to the procedure of human measurement, as participants may be prone to learning and fatigue effects, they may use the judgment scale in different ways, and/or they may use different criteria to base their judgments on. The assumption

in experimental research is that, at least in experiments that test large enough samples,[1] the true variance comes to light through averaging of the judgments for all items (within a condition) by all participants, because their individual errors will cancel each other out. The reliability of the data set is then determined by the relationship between the two variance components, that is, it is the proportion of true variance in the sum of all variances. Thus, data are most reliable with a relatively small error component, in that the mean judgments are most representative of the true scores; the reliability of a data set is low, by contrast, when the error component is substantial relative to the contribution of the true component. Rietveld (2021) illustrates that reliability is by no means a trivial concept, because low reliability will lead to drastic and unpredictable effects on correlations between the (acceptability) judgments and other variables in the equation.

The statistical measure that represents reliability on the definition outlined above is the *intraclass correlation coefficient* (ICC; Shrout & Fleiss 1979). The ICC is a suitable metric for interval data and permits generalization to participants not included in a data set. However, an important condition for the calculation of the ICC is that *all participants* must have rated *all items*, that is, it is unable to handle nested structures. The ICC is therefore not an appropriate reliability measure when experimental lists are used with the items distributed according to a Latin Square configuration, i.e. where different participant groups rate items in different conditions (as was the case in Schoenmakers 2023).

2.1 Generalizability theory

A reliability metric that does allow for alternative research designs is the *generalizability coefficient* (G-coefficient, ρ^2) employed in Generalizability Theory (GT; Cronbach et al. 1972, Shavelson & Webb 1991, Brennan 2001, see also Briesch et al. 2014). An advantage of GT is that it can accommodate multiple sources of variance, or *facets*, at once (e.g. participant, item, occasion). In GT, the error component is further decomposed into a systematic and a random or 'residual' error component. The systematic error component represents the error attributable to the different facets; the residual error component contains the error attributable to the interactions between facets as well as unidentified and/or random variability (these components cannot be disentangled from one another, Shavelson & Webb 1991). Like the ICC,

1 Sprouse and Almeida (2017) reassuringly show that different types of acceptability experiments with smaller sample sizes are relatively well-powered to detect common syntactic phenomena.

the G-coefficient is calculated on the basis of the relative contribution of the true score component and an error component, but it takes into account variances that can be attributed to multiple predefined facets. The focus of GT is therefore "the average score that would be expected across all possible variations in the measurement procedure (e.g., different raters, forms, or items)" (Briesch et al. 2014: 15) and the G-coefficient can be considered as a "stepped-up intraclass correlation coefficient" (Brennan 2001: 35). The G-coefficient permits drawing relative inferences about the performance of participants, i.e. inferences about their performance with respect to the performance of the other participants. GT can also be used to draw absolute inferences about their performance, by calculating the *index of dependability* (D-coefficient, Φ). With the D-coefficient, inferences can be drawn about the absolute performance of participants, i.e. inferences about their performance with respect to a fixed point (e.g. 60% in a school test represents a passing grade) rather than to the rank-order of items. Lacking such an anchor on the acceptability judgment scales used in Schoenmakers (2023), here we will be concerned with the G-coefficient to assess the proportion of replicable variance in the data set.

It is important to note that *reliability* is not the same as *validity*, i.e. whether the judgments represent what they were intended to measure. Brennan (2005) shares an old adage to illustrate the difference between the two terms: 'A man with one watch can tell the time; a man with two watches can never be sure.' The G-coefficient reflects the extent to which the two watches tell the same time, not whether they accurately reflect the international atomic time. Thus, as Brennan (2005: 10), puts it, "[i]nvestigators searching for platonic truth in G theory are doomed to disappointment." But in order to assess the validity of a set of judgments, it is necessary to first establish its reliability.

2.2 Cluster analysis

When participants vary in their judgments, with regard to the extent in which they reject or accept specific syntactic constructions, it can be helpful to investigate whether subgroups or 'clusters' of individual participants can be distinguished. A known technique to do so is cluster analysis. In cluster analysis, individual speakers are compared based on their judgments. Given these judgments (measurements), their (dis)similarity can be computed. By comparing the (dis)similarities between all participants, an algorithm can be applied to investigate whether subgroups can be distinguished, as participants can respond to the stimulus materials in a similar or in a completely dissimilar fashion. There are many variants of cluster analysis to allocate participants

to specific subgroups. We applied the best-known variant, so-called *hierarchical clustering*, with squared Euclidean distances between participants, in combination with Ward's method (see e.g. Hennig et al. 2016).

3. Data set composition

In this paper, we first compute the reliability of the judgments of the three prescriptive norm violations collected in Schoenmakers (2023). The experiments followed a 2×2 design, crossing the factors ± *norm violation* and ± *grammatical*; see (3) for a sample item with a comparative *als* violation, with the source(s) of markedness indicated in boldface, reproduced from Schoenmakers (2023). The experiment also included items with subject *hun* and auxiliary *doen* items, cf. (1). Each target sentence was preceded by a short preamble. Twelve items per norm violation were distributed over four experimental lists according to a Latin Square design, so that each participant saw each condition of all three norm violations three times, and 36 items in total. For the purposes of the analyses performed for the present paper, we discarded the ungrammatical variants of these items (see (3c) and (3d)). We report on our reliability and cluster analyses for these data sets in Section 4.

(3) *Vincent heeft aan een hardloopwedstrijd meegedaan. In zijn categorie deden 50 mannen mee. Vincent is als 48e geëindigd.*
'Vincent participated in a running race. 50 men took part in his category. Vincent finished 48th.'
   a. *Vincent is langzamer dan de meeste mannen.* [+gramm., –viol.]
      Vincent is slower than the most men
      'Vincent is slower than most men.'
   b. *Vincent is langzamer* ***als*** *de meeste mannen.* [+gramm., +viol.]
      Vincent is slower as the most men
   c. *Vincent* ***zijn*** *langzamer dan de meeste mannen.* [–gramm., –viol.]
      Vincent are slower than the most men
   d. *Vincent* ***zijn*** *langzamer* ***als*** *de meeste mannen.* [–gramm., +viol.]
      Vincent are slower as the most men

In Section 5, we investigate the reliability and variation pattern of the judgments for 24 scrambling items from the same experiment in Schoenmakers (2023). This item set similarly crossed two factors in a 2×2 design, viz. *object position* (scrambled vs. unscrambled) and *object discourse status* (topic vs. focus). The

scrambled object position is on the left side of the (clause) adverb and the unscrambled position on the right side. The other factor, *object discourse status*, was manipulated using the preamble, which marked the object in the target sentence as topical (4) or focused (5). Participants saw each condition six times.

(4) **Topic condition**
Nora heeft een interessant museum ontdekt. Het is een wetenschappelijk museum met een uitgebreide collectie. Binnenkort wordt een nieuwe expositie geopend.
'Nora discovered an interesting museum. It is a science museum with an extensive collection. A new exposition will be opened soon.'
Target sentence:
Nora gaat (*het museum*) *absoluut* (*het museum*) *bezoeken.*
Nora goes the museum absolutely the museum visit
'Nora will absolutely visit the museum.'

(5) **Focus condition**
Nora heeft een interessant museum ontdekt. Ze wil zich al een tijd meer verdiepen in de archeologie. Binnenkort heeft ze een weekendje vrij.
'Nora discovered an interesting museum. She has been wanting to indulge more in archeology for a while. She has a weekend off soon.'
Target sentence:
Ze gaat (*het museum*) *absoluut* (*het museum*) *bezoeken.*
she goes the museum absolutely the museum visit
'She will absolutely visit the museum.'

Finally, each list contained (the same) 48 filler items, twelve of which were grammatically unmarked sentences, twelve ungrammatical sentences, and 24 'marked' in some way. This last category contained anglicisms, fronted participles, and violations of the Animate First principle.

The experiment was an online questionnaire conducted in Qualtrics. Participants were recruited through network and snowball sampling. Upon starting the experiment, participants were presented with one of three versions of the questionnaire and instructed to rate sentences in terms of their linguistic acceptability, surface probability, or aesthetic quality (see Schoenmakers 2023). For the purposes of the present paper, however, we discarded the participants who filled out the surface probability and aesthetic quality questionnaire variants and included only those participants who filled out the acceptability questionnaire in our analyses. These analyses were performed on data from 46 participants (mean age = 48.91, SD = 21.17,

age range 18–91). Participants were instructed to rate the sentences using a slider bar on a scale from 0% to 100%. The slider bar was initially set at 50% and the actual values were not visible to the participants.[2]

In the following, we report on our analyses, for which we used the software R (version 4.2.3, R Core Team 2022).

4. Results: Norm violations

First, we performed a linear mixed effect model on the acceptability scores (using the option **lmer** from the R-package *lme4*, Bates et al. 2015), with ± *norm violation* entered as a fixed effect, and with by-participant and by-item random intercepts. We visualized the random structure using the R-package *sjPlot* (Lüdecke et al. 2023) to determine whether particular items were judged in an unexpected deviant manner, see Figure 1.

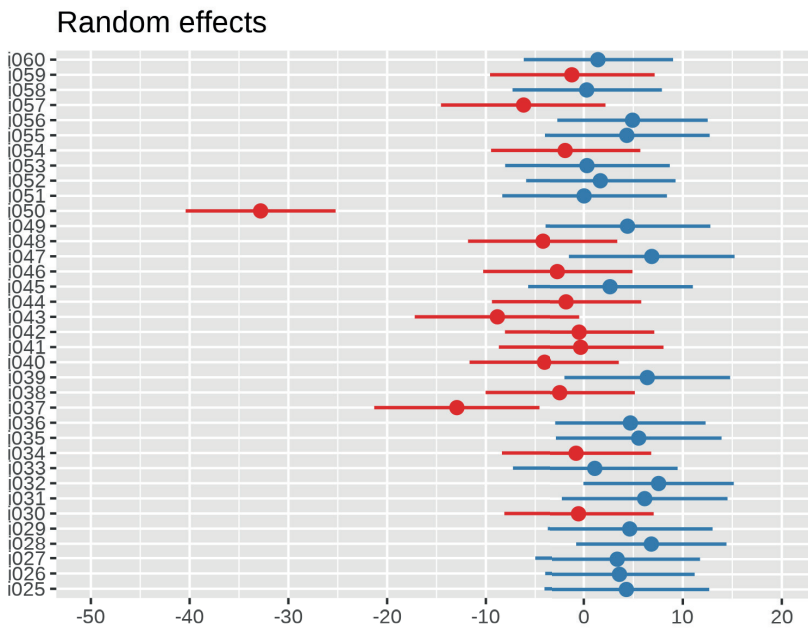

Figure 1 By-item random effects structure of the norm violation item set, with red items representing negative effects and blue elements representing positive effects.

2 See Schoenmakers (2023) for a more detailed description of the experimental materials and procedure.

There are two ill-behaving items, viz. item 37 and item 50; the corresponding sentences are given in (6) and (7). Item 37 may have been considered pragmatically awkward. That is, participants may have expected a reflexive or some other sort of direct object in the target sentence (6b). Regarding item 50, we realized upon closer scrutiny of the stimulus materials that there was an error in the grammatical variant of the target sentence (7b): it contained a participle rather than an infinitival verb form, i.e. **Lorna gaat hem expres niet verteld dat …* 'Lorna goes him deliberately not told that…'. We consequently removed the two items from further analysis.

(6) **Item 37**

a. *Brian had zich aangemeld als coach voor het nieuwe basketbalteam. De meeste teamleden waren al erg goed toen hij begon, maar sommigen moest hij nog iets meer sturing geven.*
'Brian had signed on to coach the new basketball team. Most of the team members were already good players when he started, but he to some of them he still had to give some more guidance.'

b. *Brian denkt dat zij/hun nog wel zullen verbeteren.*
'Brian thinks that they/them will probably improve.'

(7) **Item 50**

a. *Lorna heeft haar vader al lang niet gezien. Ze woont namelijk sinds een jaar of twee in het buitenland. Morgen komt ze naar Nederland om hem te verrassen.*
'Lorna hasn't seen her father in a long time. She has been living abroad for about two years now. Tomorrow she is coming back to the Netherlands to surprise him.'

b. *Lorna gaat/doet hem expres niet vertellen dat ze komt.*
'Lorna will/does not tell him that she's coming.'

These outcomes demonstrate that the application of a regression analysis that is able to deal with the nested structure of the data and with random effects supports the interpretation of the outcomes, including detecting unintended, deviant items. The confidence intervals (the lines in Figure 1) of all other items include the value of zero (no effect), with only a slight exception of item 43.

Next, having excluded the two deviant items, we analyzed the data using the **gstudy()** and **dstudy()** functions of the *gtheory* package (Moore 2016) to evaluate the reliabilities and the contribution of the different sources

of variation. This package make use of the *lmer* analysis. *Participant* was entered as the object of measurement, since we are interested in the differentiation among participants.

The (proportions of) estimated variance attributable to the different facets, i.e. the sources of variation, are given in Table 1. The participant component, or the estimated universe score variance, shows how much participants differ in their judgment scores; the item component shows how different the mean score (over participants) of a given item is expected to be from the mean over all items in the universe (specifically, it reflects the squared distance of the data point to the mean; see Shavelson & Webb 1991). Table 1 shows that judgments are much more variable between participants than between items. That items did not influence the judgment scores much could already be seen in Figure 1: after we removed the outliers, the remaining items (almost) crossed the zero value. Yet, the largest component is the residual component, which encompasses the interaction between participants and items as well as random or unidentified noise. The large values indicate that the relative standing of participants differs from one item to the next, and/or that other facets that were not included in the G-study design played a considerable role as well. The values in Table 1 indicate that a test with one participant would not be provide a good estimate of acceptability.

Table 1 G-study components for the norm violation item sets.

	Participant	Item	Residuals
Norm violations	81.70 (16.7%)	9.37 (1.9%)	397.85 (81.4%)
Comparative *als*	96.19 (23.4%)	1.12 (0.3%)	314.64 (76.4%)
Subject *hun*	101.22 (16.7%)	9.21 (1.5%)	496.38 (81.8%)
Auxiliary *doen*	50.67 (11.6%)	0 (0%)	385.24 (88.4%)

However, these reported components are based on a model that takes into account the effect of violation vs. non-violation, because a model without this factor was not supported by the data. The factor has a drastic effect on the judgments, as shown in Figure 2. Figure 2 shows that all participants make a clearcut distinction between the two categories for all three violations, with only marginal overlap. The judgments represent the high and low ends of the scale, and are extremely different to such an extent that analysis of the separate violations with no fixed effect of violation does not work (with systematic warning of so-called singularity problems).

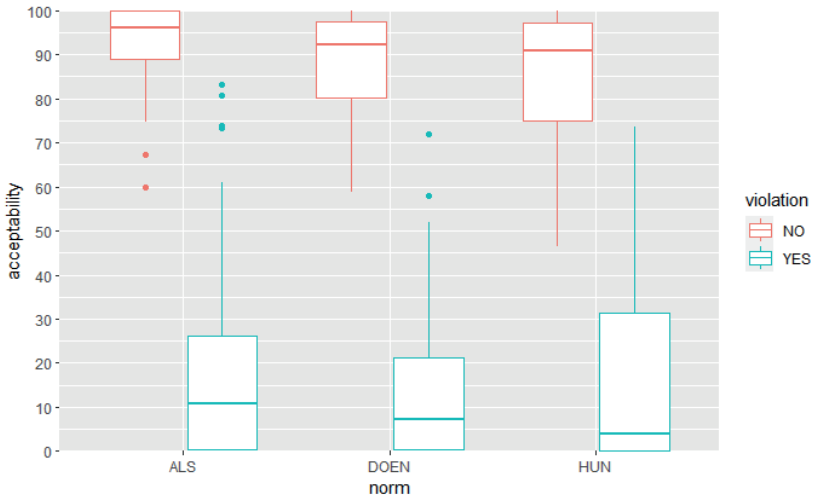

Figure 2 Boxplots of the mean acceptability scores in the violation and non-violation conditions for each norm.

Our reliability analysis indicates that the reliability of the full data set (all three types of norm violations taken together), with 46 participants and 34 items, was good (ρ^2 = .78).[3] Suspecting that the grouping together of the different norm violations may have increased overall reliability, we ran separate analyses for the different norm violations. The reliability levels were moderate for comparative *als* items (k = 12, ρ^2 = .65) and subject *hun* items (k = 11, ρ^2 = .55), and poor for auxiliary *doen* (k = 11, ρ^2 = .40).[4] These findings may seem surprising at first, because the reliability scores are rather low. However, the inclusion of the fixed factor ± *norm violation* in the model produces an important imbalance between the components of the variation (cf. Shavelson & Webb: Ch. 5).

There are two possible factors that may underlie the low reliability coefficients (see Rietveld 2021: § 4.6.1). The first is that participants did not fully understand the task and/or applied idiosyncratic criteria when judging the

3 In what follows, the index k is used for the number of items and the index n for the number of participants. Reliability scores below 0.5 are considered poor, scores between 0.5 and 0.75 moderate, scores between 0.75 and 0.9 good, and scores above 0.9 excellent (Koon & Li 2016).

4 The auxiliary *doen* analysis ran into singularity issues, that is, the model could not estimate the variance component attributable to the item facet (no variation between the items), and so the random effects structure was inconclusive. The model is therefore not supported by the data, as we can rule out the possibility that the item variance component is truly zero (cf. Matuschek et al. 2017).

items, that is, they judge sentences seemingly at random. Their judgments are consequently not highly correlated and the internal consistency is low. This conclusion must be rejected for our data, however, because participants systematically separated the violations from the non-violations (see Figure 2). This distinction appears to override all other facets or variation components, for all three types of violations. This leads us to the second possible factor, that the judgments do not vary *sufficiently*. Participants are near-unanimous in their judgments of the experimental conditions. They completely agree in rejecting the stigmatized variants and accepting the non-stigmatized variants: there is no grey area. With so little variation between participants, the relative contribution of random or unidentified variation is 'inflated', as the G-coefficient equation divides the variance between participants by the sum of the participant variance and the residual variance (Cronbach et al. 1972), and the reliability coefficient is low. This was the case in our norm violation item sets.

As a next step, we performed cluster analyses to see whether we could identify subgroups of participants with similar properties. We found four clusters or subgroups of participants. Figure 3 displays the difference scores between the non-violation and the violation mean scores for each of the three norms. All difference scores are positive for all participants: they all judge the prescriptively acceptable counterparts more positively. The first subgroup of participants (n = 6) was relatively mild in their judgment of comparative *als* and auxiliary *doen* items, and rejected subject *hun* especially (i.e. the difference between the conditions was relatively large). The second subgroup (n = 26) strongly rejected all three norm violations. This was the largest cluster, encompassing those participants who view violations of the prescriptive norm as unacceptable constructions of Dutch. The third subgroup (n = 10) rejected comparative *als* and auxiliary *doen* items, but was relatively mild with regards to subject *hun*. Interestingly, this subgroup evaluated the different norm violations in the exact opposite pattern of the first subgroup. The final subgroup (n = 4) was small: these participants were relatively tolerant of all three violations, and in particular of comparative *als*. Interestingly, subject *hun* shows a clear difference between clusters 1 and 2 versus clusters 3 and 4.[5]

5 The original experiment in Schoenmakers (2023) did not include questions about demographical information other than age and gender. This type of information could have been insightful with regard to the identified participant subgroups, so as to get a better grasp on the variation patterns. We thus recommend future researchers to include more questions about the social status of their participants.

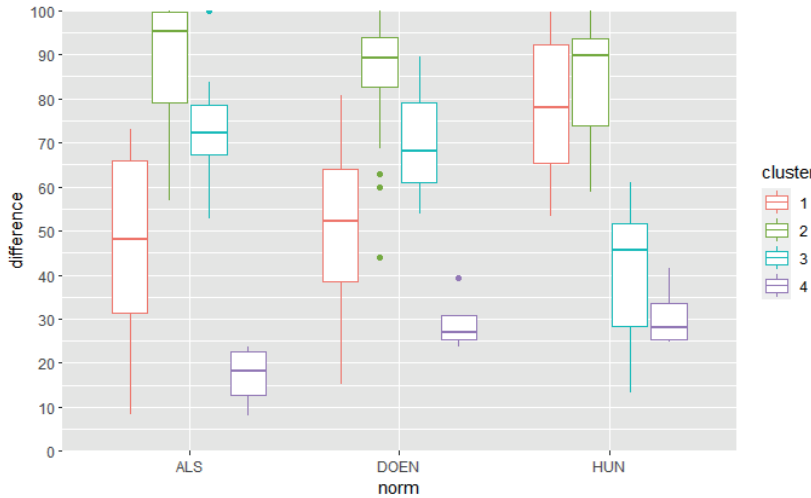

Figure 3 Boxplots of the difference scores in acceptability of the three norms for four participant clusters. Cluster 1 (n = 6), moderate rejection, but strong rejection of subject *hun*; cluster 2 (n = 26): extreme rejection overall; cluster 3 (n = 10): strong rejection, but moderate rejection of subject *hun*; cluster 4 (n = 4): low rejection overall.

5. Results: Scrambling

We applied the same analyses to the scrambling item set, starting with a linear mixed effect model, with *object position* (scrambled vs. unscrambled) as a fixed effect, and with by-participant and by-item random intercepts. Addition of *object information status* (topic vs. focus) did not significantly improve the model fit; this factor was therefore not retained. We plotted the by-item random structure, see Figure 4. Visual inspection reveals two deviating items (12 and 21). Upon closer scrutiny, these items both contained the adverb *kennelijk* 'apparently', which may have been considered as archaic by (some) participants. These items were thus excluded from further analysis.

The (proportions of) estimated variance attributable to the different facets are given in Table 2, again indicating that a one-participant experiment would not yield good estimates of acceptability. The variance attributable to the facet of item was very small, which indicates that the judgments for the different items did not vary substantially overall. Participants judgments, by contrast, varied to some extent. The residual variance component represented the largest chunk of the total variance, however, showing that individual

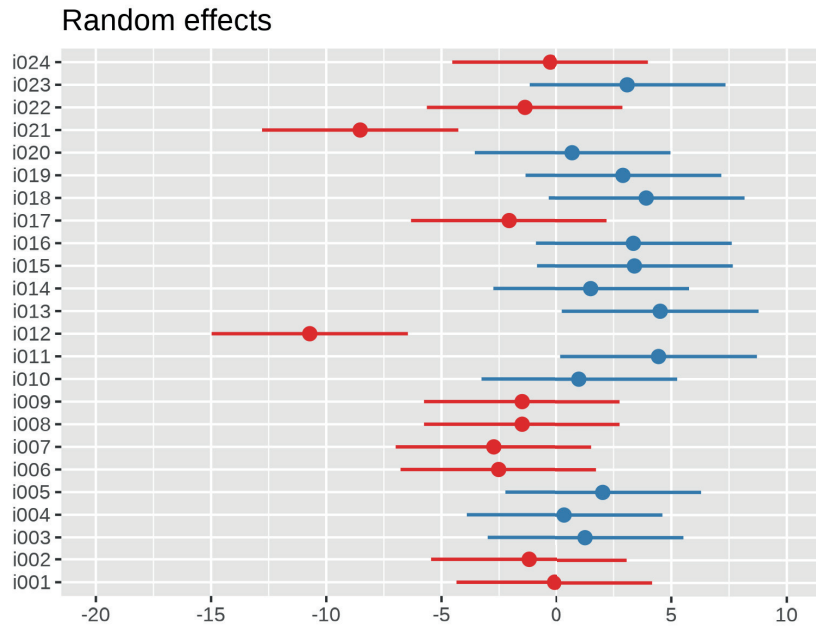

Figure 4 By-item random effects structure of the scrambling item set, with red items representing negative effects and blue elements representing positive effects.

participants' assessment of the particular items varied considerably, or that there was much random noise.

Table 2 G-study components for the scrambling item set.

	Participant	Item	Residuals
Scrambling	89.65 (27.9%)	4.48 (1.4%)	227.10 (70.7%)

The reliability of the full scrambling item set, with 46 participants and 24 items, was excellent (ρ^2 = .90). The internal consistency of the judgments is high: participants were consistent in providing judgments of a non-stigmatized type of variation, and so the reliability score unequivocally speaks in favor of the use of acceptability judgments from naive native speakers as a data source in linguistic theorizing.

The factor *object position* (scrambling) was found to have a significant effect on judgments (see Schoenmakers 2023). Figure 5 shows the density plot of this effect. It visualizes how the scrambling effect is distributed over participants.

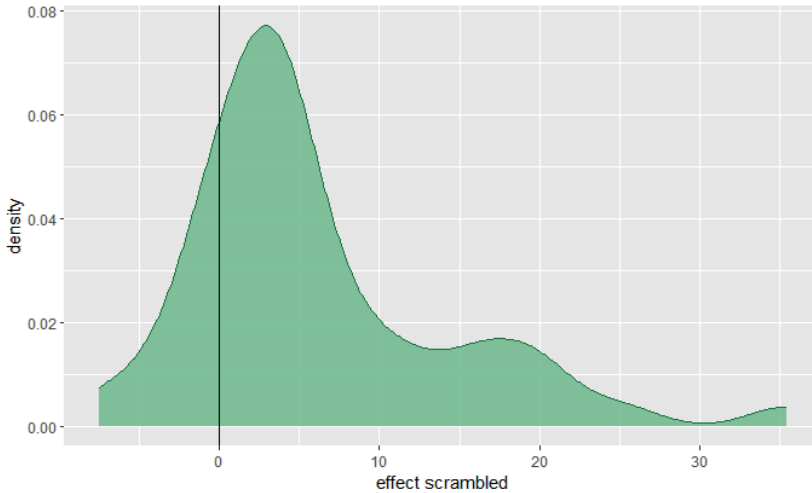

Figure 5 Density plot visualizing the scrambling effect.

The size of the scrambling effect is between -7 and 10 numerical points on the judgment scale for the majority of participants, but the density plot is trimodal: for a second group of participants, the size of the scrambling effect is between 15 and 20 numerical points, and for a small minority of participants (presumably only one) the effect size is around 35 numerical points. Thus, despite the high reliability of the data set, there is individual variation.

Insights into the subgrouping of participants can once again shed light on individual patterns. Hierachical cluster analysis with Ward's method returned three clusters of participants. The boxplots of these three groups are given in Figure 6 for all four experimental conditions.

Figure 6 shows that the subgroups of participants are marked by separate patterns. Cluster 1, by far the largest subgroup (n = 36), makes no distinction at all between the four conditions. Participants in cluster 2 (n = 4) reject topics in unscrambled position, in line with theoretical analyses of scrambling (e.g. Broekhuis 2008, Neeleman & van de Koot 2008). Note, however, that this is only a small subgroup that happily accepts scrambled foci, an unexpected outcome according to the same theories (cf. Schoenmakers 2023). Participants in cluster 3 (n = 6) disfavor unscrambled sentences, although they do not reject them altogether. The median scores for the two conditions were at approximately 70, close to 10 numerical points lower than in the scrambled conditions, and the interquartile ranges are rather small. The pattern we observe in the scrambling item set is not quite as extreme as in the sentences that violate a prescriptive norm, but the distinctions are systematic.

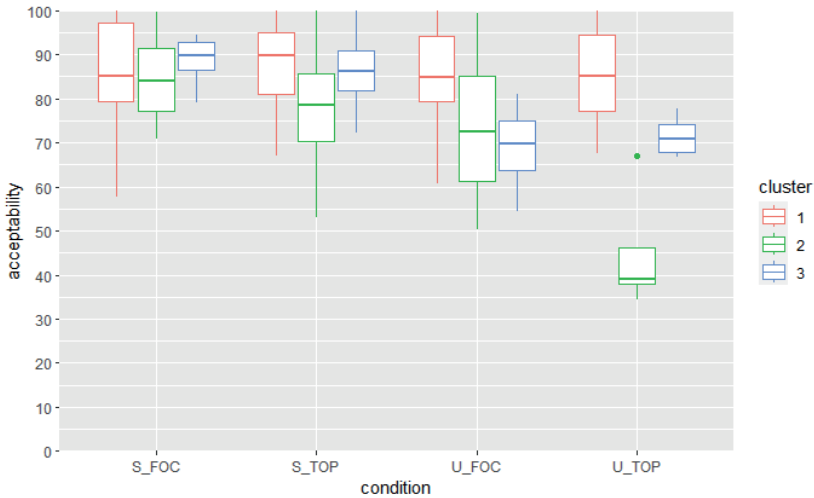

Figure 6 Boxplots of the scrambling conditions sets with three participant clusters (S = scrambled, U = unscrambled; FOC = focus, TOP = topic). Cluster 1 (n = 36): high acceptance of all utterances; cluster 2 (n = 4): acceptance of all utterances except the unscrambled sentences with topical objects; cluster 3 (n = 6): high overall acceptance of scrambled utterances, lower acceptance of unscrambled sentences.

6. Discussion

6.1 Summary of the research findings

In this paper we further explored the data sets reported in Schoenmakers (2023). Specifically, we were interested in the reliability of judgments of cases of stigmatized and non-stigmatized variation collected from linguistically naive participants. Low reliability scores can be a sign that participants did not understand or could not properly execute the task they were given, whereas high reliability scores indicate that 'human measurement' from naive participants can be a viable source of acceptability data that can be used in linguistic theorizing. We thus used Generalizability Theory (GT) to establish the degree of internal consistency in the data sets, by calculating the G-coefficient for each of the data sets.

The item sets with violations of the prescriptive norm yielded low reliability scores, although we were forced to include ± *norm violation* as a fixed factor in the model to avoid running into issues of singularity. This factor led to a 'sledgehammer effect', in that all participants made a strong distinction in their judgments of violations and non-violations, an effect that overshadowed all other sources of variation. Given that the G-coefficient

is calculated based on the variance components, reliabilities were low. The G-coefficient of the scrambling item set, by contrast, was high: the data set has excellent reliability. There were no sledgehammer effects in this item set, the mean judgment scores for each of the conditions (*object position* × *object discourse status*) were rather similar in the acceptable region of the scale. Our results show that participants were highly consistent in their judgment of non-stigmatized, or 'regular morpho-syntactic', variation. We conclude that these data can properly serve linguistic theory because of their consistency.

Individual variation was nonetheless a clear feature of the variability in each of the data sets. We performed two cluster analyses to identify subgroups of participants. Regarding the prescriptive norm violations, we find that most participants dismiss all three types: comparative *als*, subject *hun*, and auxiliary *doen*. A small second subgroup did not seem to be fazed by any of the prescriptive norm violations all too much, in particular with regard to comparative *als* (the oldest and likely the most 'socially integrated' of the three types). Interestingly, a third subgroup of participants was rather lenient towards subject *hun*, while a fourth subgroup was particularly mild towards comparative *als* and auxiliary *doen*. Thus, while most participants do not accept norm violations as acceptable constructions of Dutch, there is a degree of individual variation in the data patterns. Cluster analysis can be a useful tool to bring such types of individual variation to the surface and as such can serve as a basis for more in-depth interpretations of the data.

In the scrambling item set, we identify three subgroups of participants. The largest subgroup did not differentiate between the four experimental conditions at all. A small number of participants reject topics in the unscrambled position, in line with the theoretical literature, but at the same time they accept scrambled foci, contra this same literature. This finding speaks in favor of an analysis of Dutch scrambling that allows for a degree of optionality (see also Schoenmakers, Poortvliet & Schaeffer 2022). Finally, six participants slightly disfavored the unscrambled word order, regardless of the information structural manipulation. It is suggested in Schoenmakers's (2023) discussion that this factor may not have been picked up by some participants, as they may not have properly read the preambles in which the discourse status of the target sentence object was manipulated. These participants may then assume that the object in the target sentence must be part of the common ground and, according to linguistic theories, it should therefore be located in scrambled position. We stress that this suggestion does not explain the rather high acceptability scores for the unscrambled order, which according to most theoretical linguists should be

rejected categorically, if the object is presuppositional. Still, this subgroup of participants may entertain a grammatical system in which scrambling adheres to the discourse template to the extent that the template *influences* but does not *determine* word order preferences (cf. Schoenmakers, Poortvliet & Schaeffer 2022). Yet, we reiterate that 36 of the participants (78,3%) did not show any distinct preference in this regard. Moreover, in the grand scheme of things our findings run counter to Broekhuis's (2023: 149) claim that "the experiment [in Schoenmakers (2023)] may have been unsuccessful in manipulating the context type, which means that we still have to await successful experiments […]"; our data show that this claim is erroneous in that it does not apply to the data set at large.

6.2 Dealing with variability in acceptability judgment data

Van Hout and Muysken (2016) describe how linguistics has come to terms with the chaos of language variation. They observe that the language sciences have two solutions: researchers either largely ignore it (e.g. within the generative grammar framework, which emphasizes homogeneity) or they integrate it in linguistic theory (e.g. in variationist sociolinguistics). The variability in acceptability judgment data is large. An attractive solution is therefore to ignore it, in particular because it only relates to the perception of sentences. The alternative is to integrate it in linguistic theory, i.e. to link it to grammar. That is, a grammar must not only produce (generate) utterances, these utterances need to be perceived (processed) and understood properly (see also Schütze 1996: Ch. 6).

Misperception is an important topic in the study of sound segments as well as in explaining language variation and change. A listener may misperceive speech forms (e.g. Ohala 1981; Blevins 2004). Another relevant topic in these fields is the appreciation of speech forms, leading to the desire to sound like and imitate other speakers (Giles 1973; Gussenhoven 2000; Pierrehumbert 2001; Bybee 2002). Perception and evaluation are essential components in understanding the phenomenon of language variability. Speakers differ in their language abilities and verbosity, in their communicative skills and styles, in their accents, timbre, and voice quality, but also in the perceptual systems they have built up.

But what about the variability in linguistic judgments of sentences? Our analyses show that systematic analysis can pay a significant contribution to tame forms of variation that seem to be chaotic without systematic analysis. We applied mixed regression, generalizability/decision studies (Generalizability Theory), and cluster analysis, and found interpretable patterns and results. Does the systematic variation of linguistic acceptability judgments

reflect variation in individual grammars? Our results can be seen as support for an affirmative answer (cf. Kovač & Schoenmakers, submitted). That is, the variation observed in the judgments can be linked to syntactic sources and to individual speakers. Differences between individual grammars can subsequently be linked to grammatical processes involved in variation and perhaps change, with different speakers representing different stages in such processes. Specifically, certain subgroups of participants may be more or less lenient towards certain structures involved in the process of language change.

Furthermore, differences between individual grammars may offer new insights in phenomena at the margins of grammar. With this, we allude to cases where the experimental researcher collects mean acceptability ratings of approximately 40% (on a scale of 0–100). The structures under investigation seem to have an in-between status when it comes to grammar: they are judged as worse than 'regular' grammatical items, but at the same time as better than ungrammatical filler items. Such awkward average ratings may be due to large degrees of individual variation, to the extent that participants may entertain individual grammars (cf. Schütze 1996: Ch. 4). Moreover, with large enough data sets, meticulous investigation of the patterns between different items may help uncover additional factors of various kinds that play a role in judgment (on top of the grammatical status of the stimuli).

We therefore believe that in-depth investigation of consistency and variability patterns is a promising direction in the field of experimental syntax. New insights may moreover be informative with regard to grammaticality contrasts theoretical linguists have disagreed on ('questionable judgments', see Phillips 2010), with scrambling sentences as a case in point. A related relevant perspective is the field of sociosyntax. The sociolinguistic study of syntactic phenomena may profit from using and dissecting gradual acceptability judgments of sentences in this manner as well, to investigate how differences between speakers can be linked to grammatical processes, sociodemographic characteristics, and language change.

7. Conclusion

We have shown that performing specific statistical analyses (which are not common in the field of experimental syntax) can be a fruitful exercise that can help understand the consistency and variability in linguistic judgment data. We recommend experimental researchers i) to explore the random effects structure of their models, so as to identify and discard ill-behaving

items (and/or participants), ii) to perform reliability analyses to establish the degree of internal consistency of the collected judgments, and iii) to conduct cluster analyses to identify potential subgroups of participants (and/or items). Systematic behavior within subgroups of participants may shed light on the individual variation in the data set, and perhaps even bring to light cases of individual grammars.

References

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker (2015). Fitting linear mixed effects models using lme4. *Journal of Statistical Software* 67(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Birdsong, David (1989). *Metalinguistic performance and interlinguistic competence*. Berlin: Springer. https://doi.org/10.1007/978-3-642-74124-1

Blevins, Juliette (2004). *Evolutionary phonology: The emergence of sound patterns*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486357

Brennan, Robert (2001). *Generalizability Theory*. Berlin: Springer. https://doi.org/10.1007/978-1-4757-3456-0_6

van Bree, Cor (2012). Hun als subject in een grammaticaal en dialectologisch kader. *Nederlandse Taalkunde* 17(2), 229-249. https://doi.org/10.5117/nedtaa2012.2.hun_527

Briesch, Amy, Hariharan Swaminathan, Megan Welsh & Sandra Chafouleas (2014). Generalizability Theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology* 52(1), 13-35. https://doi.org/10.1016/j.jsp.2013.11.008

Broekhuis, Hans (2008). *Derivations and evaluations: Object shift in the Germanic languages*. Berlin: De Gruyter. https://doi.org/10.1515/9783110207200

Broekhuis, Hans (2023). Scrambling of definite object NPs in Dutch: Formal theories, corpus data and experimental research. *Nederlandse Taalkunde* 28(2), 145-179. https://doi.org/10.5117/NEDTAA2023.2.001.BROE

Bybee, Joan (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14(3), 261–290. https://doi.org/10.1017/s0954394502143018

Chen, Zhong, Yuhang Xu & Zhiguo Xie (2020). Assessing introspective linguistic judgments quantitatively: The case of *The Syntax of Chinese*. *Journal of East Asian Linguistics* 29(3), 311-336. https://doi.org/10.1007/s10831-020-09210-y

Cowart, Wayne (1997). *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks: SAGE.

Cronbach, Lee, Goldine Gleser, Harinder Nanda & Nageswari Rajaratnam (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

Edelman, Shimon & Morten Christiansen (2003). How seriously should we take Minimalist syntax? *Trends in Cognitive Sciences* 7(2), 60-61. https://doi.org/10.1016/s1364-6613(02)00045-1

Featherston, Sam (2020). Can we build a grammar on the basis of judgements? In: Samuel Schindler, Anna Drożdżowicz & Karen Brøcker (eds.), *Linguistic intuitions: Evidence and method.* Oxford: Oxford University, 165-188. https://doi.org/10.1093/oso/9780198840558.003.0010

Francis, Elaine (2022). *Gradient acceptability and linguistic theory*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780192898944.001.0001

Gibson, Edward & Ev Fedorenko (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14(6), 233-234. https://doi.org/10.1016/j.tics.2010.03.005

Gibson, Edward & Ev Fedorenko (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, Edward, Steven Piantadosi & Ev Fedorenko (2013). Quantitative methods in syntax/semantics research: A response to Sprouse & Almeida (2013). *Language and Cognitive Processes* 28(3), 229-240. https://doi.org/10.1080/01690965.2012.704385

Giesbers, Herman (1983/1984). Doe jij lief spelen? Notities over het perifrastisch doen. *Mededelingen van de Nijmeegse Centrale voor Dialect- en Naamkunde* 19, 57–64.

Giles, Howard (1973). Accent mobility: A model and some data. *Anthropological Linguistics* 15(2), 87-105

Gussenhoven, Carlos (2000). On the origin and development of the central Franconian tone contrast. In: Aditi Lahiri (ed.), *Analogy, levelling, markedness: Principles of change in phonology and morphology.* Berlin: Mouton de Gruyter, 215-260. https://doi.org/10.1515/9783110808933.215

Hartshorne, Joshua, Joshua Tenenbaum & Steven Pinker (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition* 177, 263-277. https://doi.org/10.1016/j.cognition.2018.04.007

Hennig, Christian, Marina Meila, Fionn Murtagh & Roberto Rocci (2016). *Handbook of cluster analysis.* New York: Chapman & Hall/CRC Press. https://doi.org/10.1201/b19706

van Hout, Roeland (2003). Hun zijn jongens: Ontstaan en verspreiding van het onderwerp 'hun'. In: Jan Stroop (ed.), *Waar gaat het Nederlands naartoe? Panorama van een taal*. Amsterdam: Uitgeverij Bert Bakker, 277-286.

van Hout, Roeland (2006). Onstuitbaar en onuitstaanbaar: de toekomst van een omstreden taalverandering. In: Nicoline van der Sijs, Jan Stroop & Fred Weerman (eds.), *Wat iedereen van het Nederlands moet weten en waarom*. Amsterdam: Uitgeverij Bert Bakker, 42-54.

van Hout, Roeland & Pieter Muysken (2016). Taming chaos: Change and variability in the language sciences. In: Klaas Landsman & Ellen van Wolde (eds.), *The challenge of chance: A multidisciplinary approach from science and the humanities.* Berlin: Springer, 249-266. https://doi.org/10.1007/978-3-319-26300-7_14

Häussler, Jana & Tom Juzek (2021). Data convergence in syntactic theory and the role of sentence pairs. In: Samuel Schindler, Anna Drożdżowicz & Karen Brøcker (eds.), *Linguistic intuitions: Evidence and method*. Oxford: Oxford University Press, 233-254. https://doi.org/10.1093/oso/9780198840558.003.0013

Koon, Terry & Mae Li (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kovač, Iva & Gert-Jan Schoenmakers (submitted). An experimental-syntactic take on long passive in Dutch: Unraveling the patterns underlying its (un) acceptability. Unpublished manuscript.

Krippendorff, Klaus (2013). *Content analysis: An introduction to its methodology* (3rd ed.). Thousand Oaks: SAGE. https://doi.org/10.4135/9781071878781

Langsford, Steven, Amy Perfors, Andrew Hendrickson, Lauren Kennedy & Danielle Navarro (2018). Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: A Journal of General Linguistics* 3(1), 37. https://doi.org/10.5334/gjgl.396

Lüdecke, Daniel, Alexander Bartel, Carsten Schwemmer, Chuck Powell, Amir Djalovski & Johannes Titz. (2023). sjPlot: Data visualization for statistics in social science. R package version 2.8.14. Retrieved from https://CRAN.R-project.org/package=sjPlot

Mahowald, Kyle, Peter Graff, Jeremy Hartman & Edward Gibson (2016). SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language* 92(3), 619-635. https://doi.org/10.1353/lan.2016.0052

Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen & Douglas Bates (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language* 94, 305-315. https://doi.org/10.1016/j.jml.2017.01.001

van der Meulen, Marten (2018). Do we want more or less variation? The comparative markers *als* and *dan* in Dutch prescriptivism since 1900. *Linguistics in the Netherlands 35*, 79-96. https://doi.org/10.1075/avt.00006.meu

Moore, Christopher (2016). gtheory: Apply Generalizability Theory with R. R package version 0.1.2. Retrieved from https://CRAN.R-project.org/package=gtheory

Neeleman, Ad & Hans van de Koot (2008). Dutch scrambling and the nature of discourse templates. *Journal of Comparative Germanic Linguistics* 11(2), 137-189. https://doi.org/10.1007/s10828-008-9018-0

Newmeyer, Frederick (2020). The relevance of introspective data. In: Samuel Schindler, Anna Drożdżowicz & Karen Brøcker (eds.), *Linguistic intuitions: Evidence and method*. Oxford: Oxford University Press, 149-164. https://doi.org/10.1093/oso/9780198840558.003.0009

Ohala, John (1981). The listener as a source of sound change. In: Carrie Masek, Roberta Hendrick & Mary Frances Miller (eds.), *Proceedings of the Chicago Linguistics Society 17: Papers from the parasession on language and behavior.* Chicago: Chicago Linguistics Society, 178-203. https://doi.org/10.1075/cilt.323.050ha

Pierrehumbert, Janet (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In: Joan Bybee & Paul Hopper (eds.), *Frequency effects and the emergence of lexical structure*. Amsterdam: John Benjamins, 137-157. https://doi.org/10.1075/tsl.45.08pie

Phillips, Colin (2010). Should we impeach armchair linguists? In: Shoishi Iwasaki, Hajime Hoji, Patricia Clancy & Sung-Ock Sohn (eds.), *Japanese/Korean linguistics 17*. Stanford: CSLI Publications, 49-64.

Phillips, Colin, Phoebe Gaston, Nick Huang & Hanna Muller (2021). Theories all the way down: Remarks on "theoretical" and "experimental" linguistics. In: Grant Goodall (ed.), *The Cambridge handbook of experimental syntax*. Cambridge: Cambridge University Press, 587-616. https://doi.org/10.1017/9781108569620.023

Preston, Alvin (2021). *Grammaticality judgements: A linguistic perspective*. New York: States Academic Press.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rietveld, Toni (2021). *Human measurement techniques in speech and language pathology*. New York: Routledge. https://doi.org/10.4324/9781003053118-2

Rietveld, Toni & Roeland van Hout (1993). *Statistical techniques for the study of language and language behaviour*. Berlin: Mouton de Gruyter.

Schaeffer, Jeanette (2000). *The acquisition of direct object scrambling and clitic placement: Syntax and pragmatics.* Amsterdam: John Benjamins. https://doi.org/10.1075/lald.22

Schoenmakers, Gert-Jan (2023). Linguistic judgments in 3D: The aesthetic quality, linguistic acceptability, and surface probability of stigmatized and non-stigmatized variation. *Linguistics* 61(3), 779-824. https://doi.org/10.1515/ling-2021-0179

Schoenmakers, Gert-Jan, Marjolein Poortvliet & Jeannette Schaeffer (2022). Topicality and anaphoricity in Dutch scrambling. *Natural Language & Linguistic Theory 40*(2), 541-571. https://doi.org/10.1007/s11049-021-09516-z.

Schoenmakers, Gert-Jan & Peter de Swart (2019). Adverbial hurdles in Dutch scrambling. In: Anja Gattnar, Robin Hörnig, Melanie Störzer & Sam Featherston (eds.), *Proceedings of Linguistic Evidence 2018: Experimental data drives linguistic theory.* Tübingen: University of Tübingen, 124-145. http://doi.org/10.15496/publikation-32627

Schütze, Carson (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology.* Chicago: University of Chicago Press. Reprinted in 2016 by Language Science Press. https://doi.org/10.26530/OAPEN_603356

Sert, Cansel, Ferdy Hubers, Theresa Redl & Helen de Hoop (2023). On the acceptability of the not so dummy auxiliary 'do' in Dutch. *Linguistics in the Netherlands* 40, 210-229.

Shavelson, Richard & Noreen Webb (1991). *Generalizability Theory: A primer.* Thousand Oaks: SAGE. https://doi.org/10.1016/0886-1633(93)90019-l

Shrout, Patrick & Joseph Fleiss (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* 86(2), 420-428. https://doi.org/10.1037/0033-2909.86.2.420

Sprouse, Jon (2020). A user's view of the validity of acceptability judgments as evidence for syntactic theories. In: Samuel Schindler, Anna Drożdżowicz & Karen Brøcker (eds.), *Linguistic intuitions: Evidence and method.* Oxford: Oxford University Press, 215-232. https://doi.org/10.1093/oso/9780198840558.003.0012

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

Sprouse, Jon & Diogo Almeida (2012). Assessing the reliability of textbook data in syntax: Adger's 'Core syntax'. *Journal of Linguistics* 48(3), 609-652. https://doi.org/10.1017/s0022226712000011

Sprouse, Jon & Diogo Almeida (2017). Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: A journal of general linguistics* 2(1), Article 14. https://doi.org/10.5334/gjgl.236