Gert-Jan Schoenmakers*

Linguistic judgments in 3D: the aesthetic quality, linguistic acceptability, and surface probability of stigmatized and non-stigmatized variation

https://doi.org/10.1515/ling-2021-0179
Received October 4, 2021; accepted October 11, 2022; published online January 11, 2023

Abstract: Linguistic judgment experiments typically elicit judgments in terms of the acceptability or surface probability of a sentence. There is evidence that the dimension of the scale on which sentences are judged influences the outcome of the experiment, but to date this evidence is only limited. This is not a trivial matter, as the elicited judgment data are increasingly considered the basis for inferences about linguistic representation. The present study investigates whether the dimension of the scale influences judgments. Sentences are judged in one of three dimensions: *acceptability*, *probability*, or *aesthetics*. Two distinct sets of experimental items are tested; one with cases of stigmatized variation (violations of the prescriptive norm) and another with cases of non-stigmatized variation (middle-field scrambling) in Dutch. The results show that participants take into account the scale dimension, both in stigmatized and in non-stigmatized variation. The results for stigmatized variation reflect a certain degree of conscious reflection based on the judgment scale; the effects in non-stigmatized variation, by contrast, are only main effects of instruction without changes in the relative pattern of judgments between conditions. These findings corroborate the idea that linguistic judgments of non-stigmatized variation are not the result of introspection in the technical sense, but automatic, multi-dimensional responses to a stimulus.

Keywords: Dutch; instructions; judgment tasks; prescriptive grammar; scrambling

1 Introduction

A common way in which language researchers with a penchant for experimental research collect data is through pooling sentence judgments from a large sample of linguistically naïve participants (see Cowart 1997; Goodall 2021; Schindler et al. 2020;

***Corresponding author: Gert-Jan Schoenmakers**, Radboud University, Postbus 9103, 6500 HD Nijmegen, The Netherlands, E-mail: gert-jan.schoenmakers@ru.nl. https://orcid.org/0000-0002-0666-6001

Schütze 1996). The question how the data thus collected should be interpreted, however, is a pesky one, as it is not entirely clear what the reported judgments quantify. On one interpretation, linguistic judgments are used as a window into "grammaticalness" (Chomsky 1965); that is, they are taken to reflect the syntactic status of a given structure and thereby inform linguistic theory directly. However, as Chomsky himself also acknowledges, there are many more underlying factors that contribute to a linguistic judgment than the grammatical status of an utterance alone. Thus, while the grammatical status of a linguistic string is certainly an important component of its acceptability, it is not the sole contributor to the reported reaction. It widely assumed, for example, that linguistic judgment data are prone to confounding factors, or "performance effects", which do not belong to the language system proper (e.g., Bever 1970; Fanselow and Frisch 2006; Hofmeister et al. 2014). And while speakers may have certain intuitions about a hypothetical utterance, as human beings they should not have conscious access to the cognitive mechanisms that regulate syntactic structure building or judgment processes (Schütze 1996: §2.4).[1] Schütze (1996) argues that linguistic judgments are therefore not introspective in the technical sense; rather, they are conscious reports of automatic reactions to a stimulus sentence (see also Sprouse 2020).

Featherston (2021) points out that what experimental researchers should ask their participants in linguistic judgment experiments is consequently not entirely clear. Most experiments prompt responses in a particular dimension through the instructions of the task, such as the *grammaticality* or *acceptability* of a set of stimuli.[2] However, it cannot simply be assumed that linguistically naïve participants will recognize and understand the intended meaning of these two notions (see Schütze 1996: §6.3.2). Some researchers therefore recommend the use of an alternative scale dimension which instead measures the naturalness of a stimulus sentence, asking whether someone would sound like a native speaker of a language when they would utter it (e.g., Featherston 2008, 2021; Schütze and Sprouse 2014). Although there is no guarantee that judgments given in terms of naturalness are any different from those given in terms of acceptability or grammaticality, Featherston

1 Note that participants in linguistic judgment experiments usually do not know the grounds for their judgments (Botha 1981), and asking them to justify these judgments will likely result in them inventing post-hoc rationalizations for their responses in an attempt to explain the causal relation between the stimulus and their judgment (cf. Nisbett and Wilson 1977).

2 The terms *grammaticality* and *acceptability* are sometimes used interchangeably, but refer to distinct concepts (e.g., Bard et al. 1996; Chomsky 1965; Häussler and Juzek 2020; Schütze 1996; Sprouse 2007). One pertinent difference also alluded to in the previous paragraph is that the grammaticality of a sentence is evaluated against a (subconscious) knowledge state (i.e., it relates to linguistic competence), whereas the acceptability of a sentence is impacted by many different (behavioral) factors other than grammaticality alone (i.e., it relates to linguistic performance).

(2021) maintains that such a difference would not necessarily be problematic, because the reported judgments will nonetheless represent all relevant features of the stimulus. And on the view that linguistic judgments are automatic responses to a stimulus sentence, one would not expect differences that are particularly relevant to linguistic theory (cf. Cowart 1997); that is, only main effects of the scale manipulation would be expected without interaction effects between the scale manipulation and item-internal factors. Yet, the question is relatively understudied to what extent naïve participants in linguistic judgment experiments take into consideration the dimension of the scale and, furthermore, to what extent the dimension of the scale can contribute to the acceptability core of linguistic judgments.

Evidence that a manipulation of the scale dimension affects the output of linguistic judgment experiments is restricted to cases with grammatical illusions or stigmatized variation (Langsford et al. 2019; Vogel 2019; see also Bennis and Hinskens 2014). The impact of the task instructions in cases of non-stigmatized morpho-syntactic variation has not remained uninvestigated, but the results of these studies are inconclusive (Cowart 1997; Langsford et al. 2019). The question remains whether experimental researchers can guide participants towards a particular dimension in linguistic judgment experiments and, furthermore, whether putative effects of the judgment dimension will interact with their hypothesized effects. To that end, the present study investigates to what extent naïve participants take into account the intended scale dimension in a judgment experiment, using stimulus sentences that contain cases of stigmatized and non-stigmatized variation. Different participant groups are instructed to judge the same sentences on three different scales: *acceptability*, *probability*, and *aesthetics* (following Vogel 2019). These dimensions are chosen to elicit judgments that are conceptually distinct from one another, namely, judgments based on a language system,[3] judgments about the linguistic reality, and judgments in terms of the linguistic feeling (referred to collectively as the "grammatical trinity" in Coppen 2011). The experiment thus explicitly tests tasks that are intended to go *beyond* the perception of acceptability (as discussed in e.g., Schütze 1996), and in so doing seeks to unearth effects of additional processes or factors influencing the judgments of aesthetic quality and surface probability.

Following Vogel (2019), for the stigmatized variation the experiment tests three different violations of the Dutch prescriptive norm: comparative *als* (Hubers and de Hoop 2013; Hubers et al. 2020; van der Meulen 2018), subject *hun* (van Bergen et al. 2011; van Bree 2012), and auxiliary *doen* (Cornips 1994, 1998; Giesbers 1983/1984; Sert et al. in prep.). These norm violations are expected to receive low judgment scores in

3 Note that judgments are interpreted here as perceptions à la Schütze (1996). Judgments of acceptability are not taken to directly reflect the linguistic competence and they may be influenced by prescriptive norms of a language.

the dimension of acceptability, because they are subject to strong sociolinguistic stigmatization. Such cultural pressure likely affects judgments in the aesthetic dimension as well. But native speakers of Dutch are aware that other speakers (or they themselves) occasionally violate the prescriptive norm (cf. Bennis and Hinskens 2014), and such violations have been acknowledged in prescriptive grammars for a long time (e.g., van der Meulen 2018, 2020). Judgments in the dimension of probability are therefore expected to be more lenient (cf. Vogel 2019 for German).

The stimulus material also includes cases of Dutch (A-)scrambling to investigate the effect of the scale dimension on sentence appreciation in non-stigmatized variation. Scrambling is a type of word order variation that is claimed to be motivated by information structural considerations in the theoretical literature (Broekhuis 2008; Broekhuis and Corver 2016: Ch. 13; Neeleman and van de Koot 2008; Neeleman and Reinhart 1998; Schaeffer 1997, 2000; Verhagen 1986). No explicit sociolinguistic conventions exist around this type of word order variation. However, the judgments reported in the literature are matter of debate (see Broekhuis 2016; de Hoop 2016), and experimental approaches to the phenomenon do not fully corroborate the claims made by theoretical linguists (Schoenmakers 2020; Schoenmakers et al. 2022; de Swart and van Bergen 2011). The discourse structure of the stimulus sentences in the experiment presented here are manipulated in such a way that the results are not only potentially informative about effects of the different scale dimensions; they may also contribute to an on-going discussion about a controversial type of word order variation.

The paper proceeds as follows. Section 2 discusses the cognitive nature of linguistic judgments and presents a discussion on the three dimensions of linguistic judgments under investigation in this paper in Section 2.1. Section 2.2 reviews previous studies which report on judgment experiments in which the dimension of the scale is manipulated, and Section 2.3 presents a description of (A-)scrambling in the Dutch middle-field. The sentence judgment experiment is presented in Section 3. Sections 4 and 5 contain general discussion and conclusions.

2 Theoretical background

Before discussing the putative influence of the scale dimension in linguistic judgment experiments, let me briefly digress to give some background on the cognitive nature of linguistic judgments. Schütze (1996) notes, following Ringen (1977), that Chomsky's (1965) (implicit) suggestion that linguistic judgments should emulate research methods from introspective psychology is not borne out. Introspection in the technical (Wundtian) sense entails careful examination of physical objects and/or experiences, and scrutinous discernment of the impressions surrounding them

(Wundt 1896). Wilhelm Wundt asked his psychology students to report their elementary impressions in personal, reductionist terms, such as what the object or experience makes them think of or how it makes them feel. Wundt's idea was that in order to understand the human psyche one must first make sense of the processes involved in experiencing.

This endeavor received an abundance of criticism, however, since it is not possible to discern e.g., the elements of water (hydrogen, oxygen) just by looking at it (Dellarosa 1988). Moreover, Dellarosa (1988) notes, Wundt's (1896) students received special training to be able to perform the task – and still there was no consensus on the perceptions they reported. A related problem is that conflicting reports can never be reconciled in scientific fashion because the impressions are internal (see also Santana 2020). Schütze (1996) argues that this does not preclude scientific analysis per se and suggests to look for ways in which judgment data can inform linguistic theory. However, (linguistic) judgments should then not be considered as directly reflective of cognitive mechanisms, because human beings do not have conscious access to such systems; rather, they are the conscious reports of automatic reactions to a stimulus, such as pain or the brightness of light – or the acceptability of a sentence (cf. Pateman 1987; Ringen 1977; Sprouse 2020). To illustrate the difference, Schütze distinguishes between two formulations of a research question a linguist collecting judgment data may have:

Introspection: What must be in the minds of participants for the sentence to have the [syntactic] status that they claim it has?
Perception: What must be in the minds of participants in order for them to react this way to the sentence?

(reproduced from Sprouse 2020: 219; cf. Schütze 1996: §2.4)

The perception-based definition of linguistic judgments seems to come closest to what linguists are working with. That is, judgment data are not commonly taken to reflect the linguistic competence directly; rather, they are subject to many interfering factors that arise during sentence processing as well as social judgments of linguistic acceptability. Crucially, however, the initial reaction to a stimulus sentence cannot be actively disengaged (Sprouse 2020). A manipulation of the judgment scale is therefore expected to only change one component that contributes to the automatic reaction, imbuing its acceptability core with contributions from other dimensions, without affecting the patterns between conditions (which would be possible on the introspective conception of judgments), at least in cases of non-stigmatized variation. Violations of the prescriptive norm are exceptional in that there is a social stigma against these forms despite their being frequent in colloquial speech. The question, then, is not only whether a manipulation of the judgment scale affects judgments, but also how it does.

2.1 Linguistic judgments in three dimensions

Recent linguistic judgment experiments often ask participants to rate how natural stimulus sentences sound to them with respect to their grammaticality, a formulation explicitly recommended by e.g., Featherston (2008, 2021) and Schütze and Sprouse (2014). This formulation elicits responses in terms of "naturalness" and hence prompts judgments about the linguistic system, although the reported judgments arguably reflect intuitions about the surface probability of a string as well (on the assumption that linguistic judgments can be multi-dimensional). Featherston (2008: 74) argues that this formulation "highlights the receptive aspect and the speaking mode and, crucially, avoids confusing or leading informants with association-laden terms such as 'grammatical'." It moreover excludes associations with status and prestige (Featherston 2021). Schütze and Sprouse (2014) similarly recommend a judgment scale in terms of native speaker ability rather than frequency or plausibility, in an effort to elicit judgments that come closer to judgments of acceptability (linguistic system) than judgments of surface probability (linguistic reality).

Sometimes, however, participants are asked how likely they think it is that a stimulus sentence is produced by a native speaker of a language (see e.g., Trotzke et al. 2015 for an elaborate version of this). This formulation seems comparable to Schütze and Sprouse's (2014) native speaker ability at first, except it crucially takes into account the surface probability of the string. This is an important difference conceptually, because participants are now asked to provide a judgment about the linguistic reality, and not (only) about the linguistic system. The linguistic reality sometimes meshes with (judgments of) the grammar linguists attempt to describe only poorly. Many researchers have reported considerable differences between elicited acceptability judgments and frequency data extracted from a corpus (Adli 2015; Arppe and Järvikivi 2007; Bader and Häussler 2010; Bermel and Knittl 2012; Divjak 2008; Featherston 2005; Kempen and Harbusch 2005, 2008). A common finding is that low-frequency forms are judged as acceptable. In other cases, participants give low judgment scores to a particular structure because of the experimental setting, or they claim to never use it, while in reality they use it regularly in colloquial speech (Labov 1975; Schmidt and McCreary 1977). This is especially problematic in cases of stigmatized variation when the prescriptive norm is violated. As a case in point, van Bergen et al. (2011) describe that a speaker of Dutch uses the pronoun *hun* 'them' as a subject, in a conversation where they claim to be aware of the prescriptive rule that rejects this construction. When this error is pointed out to them, the speaker continues to express their disgust with their own error and the fact that the conversation is being recorded. Participants in linguistic judgment experiments in which

prescriptive norm violations are tested may consequently feel inclined to demonstrate their knowledge of the language rules which they learned in school, even when they are not instructed to do so. Judgments about the surface probability of a string may thus not be completely independent from the corresponding acceptability judgments at least in the cases of stigmatized variation. Furthermore, it is not clear whether participants are able to accurately reflect on the frequency of linguistic constructions, or how these estimations relate to the corresponding judgments of acceptability.

This raises the question to what extent the contribution of the judgment's dimension (e.g., acceptability vs. surface probability) shines through in the reported judgment. Recall that Schütze's (1996) conception of linguistic judgments is that they are reactions which cannot be suppressed (nor verified or falsified by mere observation). Schütze therefore suggests that they are better described as *grammaticality sensations* (cf. Pateman 1987; Ringen 1977). This definition of linguistic judgments is reminiscent of what is known in the literature as the "linguistic feeling" (or *Sprachgefühl*, or *sentiment de la langue*). This term is commonly used in various guises in philosophy of language, but although it is generally understood what it refers to in abstract terms, it is notoriously difficult to define in scientific terms (see e.g., Schulte 1988). Romand (forthcoming) demonstrates that the term has historically been used with different definitions in different disciplinary fields, by different authors, and even in different publications by the same authors. Siouffi (2018) suggests that the term in its current use is in some sort of *entre-deux* state between a technical and a "lay" sense.[4] Essentially, it refers to subconscious and conceptually pluralistic opinions about the aesthetic value of an expression (Fortis 2019) or the affective state that emerges from the interaction between a structural representation and the actual expression uttered (Romand 2019, forthcoming).

Romand (2019, forthcoming) discusses the difference between the two common definitions of the "linguistic feeling" in detail and crucially distinguishes between *form feeling* and *formal feeling*. The term *form feeling* was coined by aestheticians and art historians, and principally refers to subconscious feelings about the patterning harmony within the art that is language (Fortis 2019); the term *formal feeling* hails from the field of affective psychology and refers to "more abstract organizational dimensions of conscious experience" (Romand forthcoming: 22). Especially the latter definition is reminiscent of Schütze's (1996) definition of linguistic acceptability

4 Samuel Jay Keyser (p.c.) suggests that abstract or "lay" terminology may help to better understand art, in this case the linguistic feeling, because when we talk about such matters in scientific terms, the art itself may get lost. For example, the use of metaphor and impressionistic description, rather than standard technical vocabulary, is very common in music instruction and has been shown to be pedagogically effective (e.g., Barten 1992, 1998).

judgments. To mark the contrast with the dimension of linguistic acceptability, in what follows I will interpret the term "linguistic feeling" as pertaining to psycho-aesthetic feelings of speakers towards the form and meaning of a linguistic sequence, without much conscious linguistic reflection, and which is "tainted with a suspicion of subjectivism" (Siouffi 2018: 98, my translation). The psycho-aesthetic and subjective nature of the linguistic feeling is also evident from one of Schulte's (1988: 137) uses of the term, as he claims that "[…] there are people who simply are naturally clever at using words in surprisingly suitable or subtle ways; it does not cost them any effort to find the most adequate turn of phrase in the right situation; that is, they display a certain form of *Mutterwitz*, a gift or talent which we may envy but cannot acquire or imitate" (Schulte, however, also makes a "rough and ready" distinction between judgments of acceptability and judgments of aesthetic quality). The dimension of aesthetic quality is therefore substantially different from the dimensions of linguistic acceptability and surface probability in that it does not necessarily entail considerations of a grammatical system (as much as acceptability judgments) or hypothetical encounters with a language (as much as probability judgments). Rather, participants are encouraged to use their personal criteria when reporting judgments about their linguistic feeling, that is, the subjective component is one of the aspects that sets it apart from more specific (quasi-)scientific dimensions. Note, however, that the relation between the linguistic feeling and the other two dimensions is relatively unclear, as it is not common in the field of linguistics to ask participants for judgments of aesthetic quality.

Having identified three distinct dimensions of linguistic judgment, which are taken to reflect considerations of the linguistic system, linguistic reality, and linguistic feeling, the experiment in Section 3 will elicit the corresponding types of judgments: judgments of the linguistic acceptability, the surface probability, and the aesthetic quality of a set of stimulus sentences. The question is whether, and how, a manipulation of the judgment scale influences the outcome of the experiment.[5]

5 It has been reported that participants perform rather inconsistently when asked to judge sentences without explicit instructions (Bley-Vroman et al. 1988). But even when instructed to consider a specific dimension, it is conceivable that participants spend as little effort as possible on processing the instructions (cf. Noordman and Vonk 1987), especially those more familiar with linguistic judgment experiments. Participant pools often consist for the most part of students, whose main reason to participate in the experiment is that they receive a gift voucher or course credit upon completion of the task. One possible consequence is that participants do not feel motivated to pay close attention to the instructions, since their "reward" does not depend on the quality of their answers. They may then attempt to use simple rules of thumb to check whether they can proceed with the experiment by relying on their previous experience with similar experiments. If this is the case, a manipulation of the judgment scale will not have an effect on the outcome of the experiment and the question is reinforced what exactly it is linguistic judgments quantify (cf. Featherston 2021). There is evidence,

2.2 Previously reported effects of the judgment scale

Evidence that the dimension of the scale in linguistic judgment experiments impacts the outcome is limited. Cowart (1997: Ch. 4) maintains that the influence of the instructions can never be assumed, and presents an experiment in which two participant groups took part in the same task, but under different instruction sets. He distinguishes "intuitive instructions", which instruct the participants to use any grounds available (apart from prescriptive grammar rules) in judging the stimulus sentences, from "prescriptive instructions", which were designed to invoke careful examination of the structural well-formedness of the sentences. However, this distinction did not yield any differences in the pattern of results that are relevant to linguistic theory; that is, the contrasts between conditions remained intact. Cowart (1997: 57) concludes with the general impression that participants are poor at intentionally adjusting their judgment criteria, noting that "[a]pparently, the same factors govern informant responses under both types of instructions." The researcher therefore appears to have little control over these criteria (modulo matters of non-interest to the experiment, see e.g., Schütze 1996; Schütze and Sprouse 2014). Schütze (1996) suggests that Cowart's results could be interpreted as an indication that the instructions in a linguistic judgment experiment are ineffective. He maintains that participants may simply not have the cognitive ability to run different judgment processes in linguistic judgment experiments, and argues that the instructions will consequently not have an impact on judgment patterns between conditions. Cowart recognizes, however, that his results do not provide direct evidence against the possibility that the instructions affect linguistic judgment processes and suggests that future research might demonstrate systematic differences between judgment patterns.

Bennis and Hinskens (2014) investigate linguistic judgments about different kinds of prescriptive norm violations in Dutch (including subject *hun*) using a large scale questionnaire with 1,515 self-selected participants. Participants rated ten constructions with non-standard morphosyntactic inflection on four different scales: "good–bad Dutch", "ugly–beautiful", "sloppy–diligent", and "dialect–standard language". The judgment scores were remarkably similar across the board and each correlation between judgment scales was highly significant. Bennis and Hinskens conclude based on factor analyses of the judgment scales that only the "dialect–standard language" scale elicited judgments in a truly different dimension than the other three judgment scales. Notably, 283 participants in Bennis and Hinskens's

however, that the instructions of linguistic judgment experiments affect the elicited responses, see Section 2.2.

questionnaire signed up specifically for the purpose of their study on prescriptive norm violations, whereas the others were already registered as part of the survey panel. The "new" participants were more critical of four prescriptive norm violations in terms of their aesthetic quality. Bennis and Hinskens infer from this that these panelists are purists when it comes to grammar. Their findings indicate that participants are sufficiently invested to engage with the instructions at least in cases of stigmatized variation, although they were each given all four scales and may have had different motivations to use them. Yet, the high correlations corroborate Schütze's (1996) suspicion that the experimental instructions may be relatively trivial at least when it comes to linguistic theorizing (i.e., there was no difference in judgment patterns).

Langsford et al. (2019) investigate the difference between two instruction types, measuring in the dimensions of *acceptability* and (confidence of) *grammaticality*. Their choice for these two dimensions is based on the existence of so-called "grammatical illusions", i.e., sentences which are fleetingly accepted by most participants in judgment experiments, but turn out to be completely nonsensical on closer scrutiny (Bock and Miller 1991; Drenhaus et al. 2005; Leivada and Westergaard 2020; Parker and Phillips 2016; Phillips et al. 2011; Vasishth et al. 2008; Wagers et al. 2009; Wellwood et al. 2018). Langsford et al. (2019) include three grammatical illusions in their experiment, exemplified in (1).

(1) a. *More people have been to Russia than I have.*
b. *A man who had no beard was ever thrifty.*
c. *The key to the cabinets are on the table.*

The experiment moreover includes sentences with multiple center embeddings (e.g., *The rat the cat the dog chased killed ate the malt*; see Chomsky and Miller 1963). Such sentences are grammatical, but typically receive low acceptability scores in judgment experiments because they are unparseable due to resource limitations of the comprehension system. The constructions in (1) and the center-embedding sentences have in common that their acceptability and grammaticality statuses diverge.

In addition to these constructions, Langsford et al.'s (2019) experiment contains a subset of the stimulus sentences used in Sprouse et al. (2013). Sprouse et al. accumulated samples of judgment contrasts from papers published in *Linguistic Inquiry* (2001–2010) and collected the corresponding nonlinguist judgments in a series of judgment experiments. The convergence rate between the two data sources was extremely high.[6] Langsford et al. specifically include a subset of sentences in their

6 This finding implies that neither method is empirically superior to another, a position defended in e.g., Gibson and Fedorenko (2013). Empirical evidence from massive replication studies indicates that linguist and nonlinguist judgments are highly similar (e.g., Chen et al. 2020; Häussler and Juzek 2017;

experiment for which the judgments from linguists differed from those from nonlinguists, on the hypothesis that a different construct may have been measured in these cases (cf. Juzek and Häussler 2020). Langsford et al. hypothesize that the judgment scores for the grammatical illusions in (1), the sentences with multiple center embeddings, and the subset drawn from Sprouse et al. (2013) will differ depending on the specific instructions participants receive.

The results of Langsford et al.'s (2019) experiment indicate that the acceptability scores are more uniformly distributed over the six-point scale than the grammaticality scores; however, there is no clear discrepancy between the mean judgment scores for the two instruction types, in all four construction types. The largest differences are found in the agreement attraction sentences (1c), which receive higher judgment scores on the acceptability scale than on the grammaticality scale, and in the multiple center-embedding sentences, which show the reverse pattern. These findings are in accordance with the linguistic literature. Crucially, the sentences drawn from Sprouse et al. (2013) receive similar judgment scores under both instruction types. Langsford et al. (2019) perform a State Trace Analysis (Kalish et al. 2016) on their data, but do not find evidence that distinct dimensions were contemplated in the judgment process under the two instruction types (i.e., the state trace plot is one-dimensional). However, they do find evidence that the instructions in a judgment experiment impact the outcome; that is, the instructions can be used to guide the decision-making process of the participants to a certain extent, but again not in a way that is relevant to linguistic theory.

Vogel (2019) reports on a linguistic judgment experiment which takes German cases of stigmatized variation, or "grammatical taboos", as its stimuli. The experiment consists of three subexperiments, which differ only in the dimension of the scale. The experiment elicits judgments in terms of *normativity*, *possibility*, and *aesthetics*. Vogel hypothesizes that prescriptive norm violations are a specific type of morphosyntactic variation, since their markedness is caused by extra-grammatical (sociolinguistic) factors. Based on the historical development of the stigmatized auxiliary use of *tun* 'do' in German (see Davies and Langer 2006; Langer 2001), Vogel suggests that the relation between the three scale dimensions in question might be as illustrated in (2). The reason for this is that auxiliary *tun* 'do' was first rejected in

Langsford et al. 2018; Linzen and Oseki 2018; Mahowald et al. 2016; Munro et al. 2010; Spencer 1973; Sprouse and Almeida 2012; Sprouse et al. 2013). Note that these findings support Schütze's (1996) original take on linguistic judgments as conscious reports of reactions to (or perceptions of) stimuli which cannot be suppressed. An anonymous reviewer points out that speakers should not be able to readily provide judgments of structural well-formedness (or grammaticality) separately from all of the other factors that contribute to acceptability (as suggested in e.g., Juzek and Häussler 2020), because linguists and nonlinguists alike are human beings and therefore have the same automatic processes (see also Schütze 1996: §4.4.1).

poetic registers; the more general grammars of (written) German only followed several decades later. Aesthetic judgments about the auxiliary use of *tun* 'do' are therefore potentially less compromising than judgments in terms of the linguistic norm. Judgments about whether an expression is possible in German at all are considered the most liberal.

(2) *beautiful language* < *norm-compliant language* < *informal language*

Vogel's (2019) experiment crosses a taboo and a non-taboo variant of 32 experimental items (of four different taboo phenomena) with a grammatical and an ungrammatical variant, in which the ungrammaticality was due to an agreement error on the finite verb. A sample item is given in (3), adapted from Vogel (2019: 56, his (15)), with the different types of markedness indicated in boldface. The grammaticality manipulation was added to the design on the hypothesis that the between-participant variation in judgments of taboo phenomena is larger than in judgments of grammar-internal markedness, because the former are subject to sociolinguistic conventions.

(3) a. *Damals hat Hans gut gelesen.* [+gramm., –taboo]
then have.3SG Hans well read
'In those days, Hans was a good reader.'

b. *Damals* ***tat*** *Hans gut lesen.* [+gramm., +taboo]
then **do**.PST3SG Hans well read

c. *Damals hab**en** Hans gut gelesen.* [–gramm., –taboo]
then have.3PL Hans well read

d. *Damals* ***taten*** *Hans gut lesen.* [–gramm., +taboo]
then **do**.PST3PL Hans well read

Vogel (2019) reports that, across all sentence types (including filler items), the aesthetic scores are only slightly lower than the normativity scores. These two judgment types show a comparable pattern overall, which Vogel takes to suggest that the dimensions of aesthetics and normativity may not be completely independent from each other. Yet, the difference between the two is relatively large in the +taboo conditions (aesthetics: 18.7%, normativity: 24.5%), implying an additional aesthetic disadvantage for the cases of stigmatized variation. The elicited possibility scores are considerably higher than the aesthetic and normativity scores, providing evidence for the prediction in (2). When compared to unmarked grammatical sentences, however, the +taboo conditions receive a relatively low average possibility score of 36.1%.

The +taboo conditions nonetheless receive higher judgment scores than the cases of intra-grammatical markedness, with the largest difference in the dimension of possibility. This finding is not surprising: despite their high degree of

sociolinguistic stigmatization, taboo sentences do exist in the linguistic reality. Vogel (2019) refers to this as the *paradox of grammatical taboos*. His data furthermore show larger between-participant variation in the taboo sentences than in grammar-internal markedness.[7] The degree of variation is the largest in the aesthetic scores (which highlights the subjective nature of the dimension) and decreases via the normativity scores to the possibility scores. This same pattern was found for the grammatical taboo phenomena and the grammar-internally marked sentences, except there was no decrease in variation from the normativity scores to the possibility scores in the grammatical taboos. Vogel concludes that the difference in scale dimensions cannot neutralize the high level of sociolinguistic controversy.

The findings in Cowart (1997), Bennis and Hinskens (2014), and Langsford et al. (2019) indicate that the instructions in linguistic judgment experiments can impact the outcome of the experiment, although they only impinge on the relative judgment scores between conditions in sentences with a prescriptive norm violation (Vogel 2019). These constructions must therefore be accounted for somehow in Schütze's (1996) theory of linguistic judgments; I will return to this in Section 4.1. The present study investigates Dutch cases of stigmatized variation and non-stigmatized variation in a novel judgment experiment, eliciting judgments in the dimensions of *acceptability*, *probability*, and *aesthetics*. The stimuli contain three Dutch prescriptive norm violations, in an attempt to replicate Vogel's (2019) findings for German, as well as an item set with middle-field scrambling sentences, which are not subjected to sociolinguistic stigmatization.

2.3 Scrambling in the Dutch middle-field

In Dutch, definite objects may occupy various positions in the middle-field of the clause, i.e., the typological region between the finite verb, or the complementizer in embedded clauses, and the clause-final main verb. An example is given in (4), where the object *het boek* 'the book' may appear on the left or right side of the clause adverb *waarschijnlijk* 'probably'.

(4)

Jan heeft	*(het boek)*	*waarschijnlijk*	*(het boek)*	*gelezen.*
Jan has	the book	probably	the book	read

'Jan probably read the book.'

7 It is not clear whether this comparison was with the grammar-internally marked test items or with separate filler items with grammar-internal markedness, since the experimental lists contain 21 ungrammatical filler items as well as nine filler items which are "syntactically marked according to the standard criteria" (Vogel 2019: 57). Vogel (2019) refers to the "marked filler items" in this discussion, and no reference is made to the ungrammatical test items until much later in the paper.

There is a consensus in most of the theoretical literature that scrambling is regulated by discourse packaging conditions: objects which appear to the left of clause adverbs (in "scrambled" position) are claimed to be presuppositional (i.e., topical and/or anaphoric), and objects which appear to their right (in "unscrambled" position) non-presuppositional (i.e., focused and/or non-anaphoric) (e.g., Broekhuis 2008; Broekhuis and Corver 2016: Ch. 13; Neeleman and van de Koot 2008; Neeleman and Reinhart 1998; Schaeffer 1997, 2000; Verhagen 1986). Deviations from this discourse template are considered highly awkward (see Schoenmakers 2020), yet some researchers argue that scrambling is much more optional in this regard than is generally assumed (van Bergen and de Swart 2009, 2010; van der Does and de Hoop 1998; de Hoop 2000, 2003, 2016; Schoenmakers 2020; Schoenmakers et al. 2022; de Swart and van Bergen 2011). De Hoop (2016), for instance, expresses her disagreement with some of the judgments reported in Broekhuis and Corver (2016: Ch. 13). Consider the brief dialogue in (5). Broekhuis and Corver mark Speaker B's utterance as an unacceptable answer to Speaker A's question, because the direct object *de verkeerde* 'the wrong one' is new to the discourse and should thus not be able to surface on the left side of the clause adverb. Not agreeing with this judgment, de Hoop calls for efforts to collect empirical data to test such theoretical assumptions.

(5) ***Speaker A:***
Ik heb het aan Peter verteld.
I have it to Peter told
'I told it to Peter.'
Speaker B:
Dan heb je de verkeerde waarschijnlijk ingelicht.
then have you the wrong.one probably informed
'Then you probably informed the wrong person.'

Such empirical data are reported in van Bergen and de Swart (2009, 2010), who conduct a large-scale corpus study in order to investigate the scrambling behavior of different types of direct objects in spontaneous speech. One of the factors they investigate is the anaphoricity of direct objects. They find 7% of the non-anaphoric definite objects and, strikingly, only 22% of the anaphoric definite objects in scrambled position. This finding indicates that anaphoricity might influence object placement in the expected direction, but also that there is a general preference for the unscrambled word order. De Swart and van Bergen (2011) do not find support for the claims that non-anaphoric definite objects obligatorily surface in unscrambled position, or vice versa, in a follow-up study either: in a sentence completion task, as much as 8.2–32.7% of the non-anaphoric definite objects (depending on the order of presentation of constituents) and only 13.7–34.1% of the anaphoric definite objects were produced in scrambled position.

Schoenmakers et al. (2022) conduct a similar sentence completion experiment in which the anaphoricity and topicality of definite objects in scrambling structures is manipulated. They report that anaphoric definite objects were produced in scrambled position more frequently than non-anaphoric objects, and topics more frequently than foci (in line with the above-mentioned discourse template). However, the scrambling proportions are nowhere near categorical: 34% of non-anaphoric foci were produced in scrambled position, and 43% of anaphoric topics were left unscrambled. The authors conclude that scrambling is relatively optional; it may be influenced, but it is not determined, by information structure. So far, experimental studies on Dutch scrambling which report on a sentence judgment experiment have not investigated the influence of information structure, but they do show that scrambled and unscrambled sentences receive similarly high judgments scores when presented free of context (Schoenmakers and de Swart 2019; de Swart and van Bergen 2011). These studies asked participants to rate the likelihood of someone saying the sentence (de Swart and van Bergen 2011) and how native-like a friend would sound when they would utter the sentence (Schoenmakers and de Swart 2019). The question remains whether judgments change when the discourse status of the object is manipulated, as predicted in most of the theoretical literature, and to what extent this finding hinges on the scale dimension.

Recall that Schütze's (1996) account of linguistic judgments asserts that they are the conscious report of automatic reactions to a stimulus sentence and not the result of Wundtian introspection. The findings from the experimental studies discussed above indicate that scrambling is at least a grammatical option in Dutch that in constrained language production may be sensitive to, but does not depend on, information structure. Regarding the different scale dimensions investigated in the experiment, one may expect one of two judgment patterns depending on whether one sides with the Wundtian or the Schützean conception of judgment data. If participants engage in Wundtian introspection, one might expect differences between the dimensions of acceptability and probability, in that these dimensions are conceptually at least somewhat comparable to the notions of competence and performance (cf. Section 2.1). Technical introspection, when properly executed, entails prolonged reflection and careful consideration (*grammatical reasoning* in Juzek and Häussler 2020), and should to some extent permit suppression of performance effects,[8] which in turn enhances the making of inferences about the language system

8 It is matter of debate whether linguistically naïve participants are even able to successfully engage in grammatical reasoning, since they are not aware of the theoretical assumptions that trained linguists are aware of. I will leave this matter aside, noting that there is no empirical evidence for the claim that linguist and non-linguist judgments represent distinct cognitive processes (see also footnote 6).

proper. The acceptability judgment scores should then show a stronger contrast between configurations which do and do not deviate from the discourse template (based on the theoretical literature); a contrast stronger than in the dimension of probability at least (based on the production data reported in Schoenmakers et al. 2022). Here, however, I will follow the argumentation in Schütze (1996) instead, and so only main effects of the use of the scale are expected. That is, the elicited judgment scores may move up or down the scale due to processes influencing the aesthetic and probability judgments *in addition to* the core acceptability judgment, but effects are crucially not expected to impinge on the relative acceptability between conditions. More specifically, regardless of the judgment scale it is expected that the elicited judgments of scrambling structures will not reflect the strong information structural contrasts assumed in most of the theoretical literature; that is, topical and focused definite objects should be (more or less) acceptable, aesthetically pleasing, and probable, regardless of their position vis-à-vis the adverb (although the experiment might reveal slight preferences in the direction of the discourse template, cf. Schoenmakers et al. 2022).

3 A sentence judgment experiment

This section presents a novel sentence judgment experiment, inspired by the experiment reported in Vogel (2019). The stimulus material contains an item set with prescriptive norm violations and an item set with scrambling sentences. Different participant groups are instructed to judge sentences on one of three scales, designed to elicit judgments in terms of the linguistic system (*acceptability*), the linguistic reality (*probability*), or their linguistic feeling (*aesthetics*). Based on Vogel's (2019) findings, the acceptability and aesthetic scores are expected to be lower than the probability scores across the board, but especially in the sentences with a prescriptive norm violation, because they are subject to strong sociolinguistic stigmatization. A second expectation is that the difference between the acceptability and aesthetic scores is larger in items with a prescriptive norm violation than in items without a prescriptive norm violation (cf. Vogel 2019). Only main effects of the judgment dimension are expected in the scrambling item set, on the hypothesis that the contributions of the alternative scales can only add to the acceptability core of a linguistic judgment of non-stigmatized variation (cf. Schütze 1996).

Further, the experiment manipulates the information structure of the scrambling sentences to test the assumption of a strict discourse template. Topical definites should then show a preference for the scrambled position, and focused definites for the unscrambled position, although all combinations are predicted to receive judgment scores at the high end of the scale (independently of the dimension of the scale).

3.1 Participants

Two hundred and four native speakers of Dutch volunteered to take part in the online questionnaire, the vast majority of whom was recruited via an e-mail chain with the aim of sampling from a heterogeneous population. Data from 36 participants were discarded, because they did not complete the full survey. Data from thirteen participants were removed because they gave more than a quarter of the fillers an unexpected judgment score. Unexpected scores were defined prior to statistical analysis as scores under 40% for the grammatical fillers and scores over 60% for the ungrammatical fillers. Data from one participant were removed, because they mentioned the term *scrambling* in a follow-up question about the purpose of the experiment. Data from one participant were removed, because they did not show variance in their responses (the overall standard deviation in their responses was 1.75). In the end, data from 153 participants (90 female, 62 male, 1 "other"; mean age = 48.51, age range = 18–91, SD = 20.89) were entered into statistical analysis.

3.2 Materials

The questionnaire closely followed the design of the experiment in Vogel (2019). There were three versions of the experiment, which only differed in the experimental instructions. The experimental items in each version were identical. The different versions of the experiment were designed to elicit judgments in the dimensions of *aesthetics*, *acceptability*, and *probability*; see (6) for the corresponding questions and scale labels. These were repeated for each experimental item. Note that judgments are given on a semi-continuous scale (0–100) and the labels presented here correspond to the endpoints.

(6) a. ***Aesthetic judgment:***
Hoe mooi vind je de formulering van de bovenstaande zin?
How nice do you find the wording of the above sentence?
niet mooi Nederlands—heel mooi Nederlands
not nice Dutch—very nice Dutch

b. ***Acceptability judgment:***
Hoe goed vind je de bovenstaande zin als Nederlandse constructie?
How good do you find the above sentence as a construction of Dutch?
niet goed Nederlands—heel goed Nederlands
not good Dutch—very good Dutch

c. ***Probability judgment:***
Hoe waarschijnlijk vind je het dat de bovenstaande zin is uitgesproken door een moedertaalspreker van het Nederlands?
How likely do you think it is that the above sentence has been uttered by a native speaker of Dutch?
niet waarschijnlijk—heel waarschijnlijk
not likely—very likely

The experiment contained 108 experimental items in total, 36 of which contained a prescriptive norm violation (stigmatized variation) and 24 a scrambling structure (non-stigmatized variation). The target sentences of the stigmatized items were constructed in a similar way as in Vogel (2019) and contained a prescriptive norm violation or its prescriptively correct equivalent. Each target sentence was then paired with ungrammatical variants containing an agreement error on the finite verb. Thus, the factors in this item set were ±*norm violation* and ±*grammaticality*.

Each target sentence was preceded by a short preamble which served no independent function other than establishing a degree of similarity with the scrambling items. The stigmatized items contained three violations of the prescriptive norm: errors in the use of the comparative particles *als* 'as' and *dan* 'than', the use of *hun* 'them' as a subject, and the use of auxiliary *doen* 'do' in habitual or intentional contexts. A sample item from each norm violation is given in (7), (8), and (9), with the causes of markedness indicated in boldface (±*gramm.* stands for *grammatical*, ±*viol.* stands for *(norm) violation*). Items from the stigmatized item set never contained a clause adverb or scrambling structure, so as to avoid priming effects with the scrambling sentences.

(7) ***Comparative als***
Vincent heeft aan een hardloopwedstrijd meegedaan. In zijn categorie deden 50 mannen mee. Vincent is als 48^{e} 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 ALS 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 ALS the most men

(8) ***Subject hun***

Arthur is een cadeau aan het bedenken voor het jubileum van zijn ouders. Opeens heeft hij een geweldige ingeving: hij gaat ze een weekendje weg aanbieden.
'Arthur is thinking of a present for his parents' anniversary. Suddenly, he has a great idea: he is going to offer them a weekend away.'

a. *Arthur weet dat zij naar Parijs willen.* [+gramm., –viol.]
Arthur know.SG that they to Paris want
'Arthur knows that they want to go to Paris.'

b. *Arthur weet dat **hun** naar Parijs willen.* [+gramm., +viol.]
Arthur know.SG that HUN to Paris want

c. *Arthur wet**en** dat zij naar Parijs willen.* [–gramm., –viol.]
Arthur know.PL that they to Paris want

d. *Arthur wet**en** dat **hun** naar Parijs willen.* [–gramm., +viol.]
Arthur know.PL that HUN to Paris want

(9) ***Auxiliary doen***

Rosa heeft een belangrijke brief ontvangen van haar makelaar. Ze weet niet goed hoe ze erop moet antwoorden. Na veel wikken en wegen vraagt ze haar moeder om hulp.
'Rosa received an important letter from her real estate agent. She doesn't quite know how to answer it. After much deliberation, she asks her mother for help.'

a. *Rosa gaat vanavond op de brief reageren.* [+gramm., –viol.]
Rosa go.SG tonight on the letter respond
'Rosa will respond to the letter tonight.'

b. *Rosa **doet** vanavond op de brief reageren.* [+gramm., +viol.]
Rosa **do**.SG tonight on the letter respond

c. *Rosa **gaan** vanavond op de brief reageren.* [–gramm., –viol.]
Rosa go.PL tonight on the letter respond

d. *Rosa **doen** vanavond op de brief reageren.* [–gramm., +viol.]
Rosa **do**.PL tonight on the letter respond

The scrambling sentences were constructed in four variants, crossing the factors *object position* (scrambled, unscrambled) and *context type* (topic, focus). Each target sentence was preceded by a preamble which served to identify the object in the target sentence as the topic or focus. The object of the target sentence was always introduced in the first sentence of the preamble;[9] the remainder of the preamble

9 This entails that the target objects were always anaphoric. It has been claimed that it is the object's anaphoricity, and not its topicality, that determines its position relative to adverbs in scrambling structures (e.g., Erteschik-Shir 2007; Schaeffer 1997, 2000). However, the distinction between anaphoric topics and anaphoric foci yielded a significant effect in the expected direction in

either revolved around the object, licensing it as the topic in the target sentence, or around the subject, licensing the object as the focus in the target sentence.

In the topic condition, the preamble explicitly mentioned the target object a second time and continued to provide more information about it. In the focus condition, the target object was not mentioned again or referred to in any other way. Moreover, the subject of the target sentence in this condition took the form of a pronoun to mark it as the topic (cf. Givón 1988). Using a pronominal subject made the target sentences sound more natural and reinforced the focused status of the object. Care was taken that the preambles did not contain scrambling clauses. Each target sentence had a S-Aux-O-V structure, with the auxiliary *gaan* 'will' (lit. 'go'), and a clause adverb on the left or right side of the object. All objects were referential nouns preceded by a definite article, placed in scrambled (OBJ – ADV) or in unscrambled (ADV – OBJ) position. The objects and adverbs were matched for length in syllables to avoid effects of grammatical weight. A sample item is given in (10) and (11).

(10) ***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.'

(11) ***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.'

The stigmatized items and the scrambling items were distributed over four experimental lists according to a Latin Square design. Fourty eight unrelated filler items were

Schoenmakers et al.'s (2022) sentence completion experiment. It is therefore anticipated that a manipulation of topicality will yield similar effects in the experiment presented here.

added to each experimental list, which were identical across lists. Twelve filler items were unmarked grammatical sentences. Twelve filler items were ungrammatical sentences and contained a violation of V2, a violation of verb-final in complex main clauses, or an error in gender agreement (e.g., **het papegaai* 'the parrot'). Twenty four filler items were "marked", in the sense that they contained an error that is not generally considered a serious (grammatical) error. These items contained anglicisms (Zenner et al. 2012), violations of the *Animate First* principle (Lamers and de Hoop 2014), or past participles in first sentence position (Schoenmakers and Foolen 2022). Each experimental list contained 108 items. These lists were pseudorandomized in two distinct orders, using the software Mix (van Casteren and Davis 2006). Each list started with at least three filler items (one from each category). Scrambling items were at least four items apart, as were prescriptive norm violations of the same type (comparative *als*, subject *hun*, auxiliary *doen*) and, within the item set with prescriptive norm violations, items in the same condition of either factor (±*norm violation* and ±*grammaticality*). The experimental lists and data are available in a repository available at https://doi.org/10.34973/k01s-ez17. The experiment was conducted using Qualtrics software (Qualtrics 2021).

3.3 Procedure

The experiment took the form of an online questionnaire in which participants were randomly assigned to one of the three versions. After reading an information document and providing their consent, participants were asked to carefully read the general instructions of the experiment. These general instructions explained what the experiment would look like and what was expected from the participant, i.e., to carefully read the stories and to rate a sentence in a given dimension (depending on the version of the experiment) on a scale from 0 to 100%, and to follow their first intuition. The general instructions did not include an example nor training items and were followed by questions about the participant's age and gender. Participants were then moved into the main part of the experiment and were asked to rate sentences in the instructed dimension (cf. (6)). After the experiment started, participants would read brief preambles to a sentence and were asked to judge this sentence on the basis of the preceding information, using a slider bar on a scale from 0 to 100%. The slider bar was initially set to 50% for each trial and participants were forced to move it to continue to the next trial, which was presented on a new page. Participants saw percentage labels on the scale in 10-point increments, in addition to the verbal labels of the scale, but they did not see the precise percentage corresponding to their responses. After the main part of the experiment, participants were asked for qualitative comments about the questionnaire and their ideas about the purpose of the experiment.

3.4 Results

The results for the filler items across the elicited dimensions are visually represented in Figure 1 (the same information can also be found in table form for the fillers and for both item sets in Appendix A). The data confirm that the filler item categories were judged in the intended manner. Grammatical filler items received the most positive judgment scores at the high ends of the scales, and ungrammatical filler items received judgment scores at the very low ends of the scales. Marked filler items received judgment scores near 50%, which indicates that participants were indecisive about their aesthetic quality, acceptability, and surface probability. Thus, participants were invested enough to follow the instructions given to them and carefully read the sentences. Notice finally that the judgment scores for the (unmarked) grammatical fillers are numerically much lower in the aesthetic dimension than in the two other dimensions.

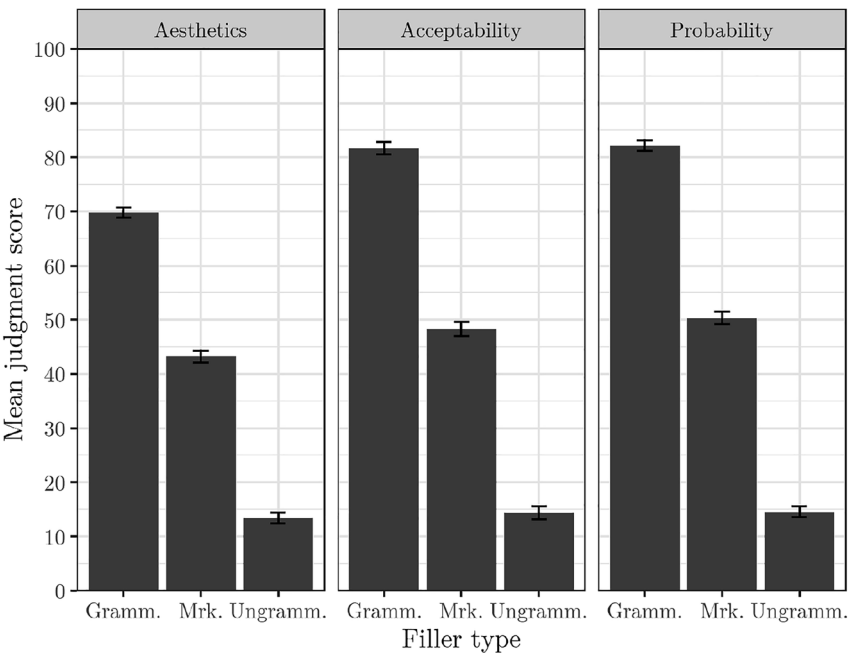

Figure 1: Mean judgment scores per filler item category in three dimensions (error bars indicate within-subject standard errors from the mean).

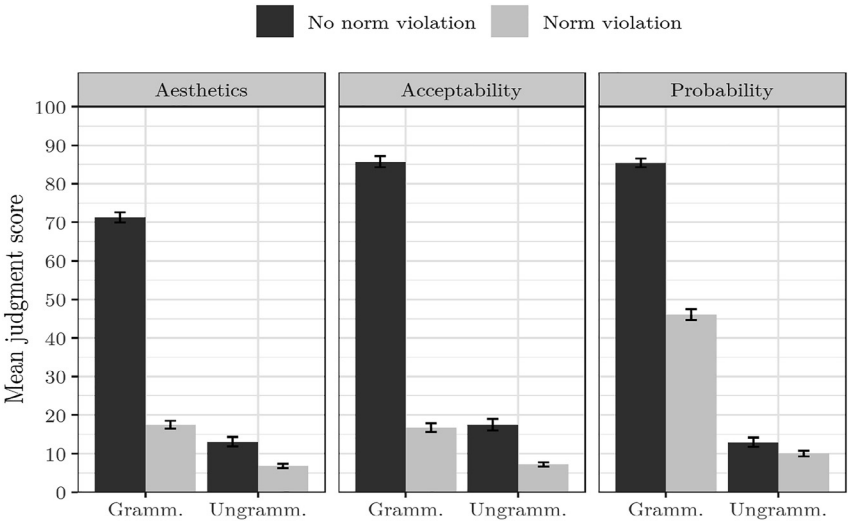

Figure 2: Mean judgment scores per condition for the stigmatized item set in three dimensions (error bars indicate within-subject standard errors from the mean).

3.4.1 Prescriptive norm violations

The results for the stigmatized item set are visually represented in Figure 2. The judgment patterns are generally similar in all dimensions, with the grammatical variants receiving the highest scores by far and the ungrammatical variants receiving judgment scores at the very low ends of the scales. There are two deviations from the general pattern, which are related to the dimension of the scale. First, while the items with a prescriptive norm violation received similarly low judgment scores in the dimensions of aesthetics (17.49%) and acceptability (16.71%), they were much better appreciated in the dimension of probability (46.08%). Second, the unmarked variants (i.e., grammatical items without a prescriptive norm violation) were judged considerably lower on the aesthetics scale (71.25%) than on the other scales (acceptability: 85.72%; probability: 85.40%). Notice finally that the two types of markedness seem to have triggered an additive effect: variants that contained both a violation of the prescriptive norm and an agreement error received judgment scores visibly lower than variants with only one type of markedness. This effect emerges in all three dimensions, but is smallest in the dimension of probability (if it exists there in the first place).

The results per type of prescriptive norm violation are displayed in Figure 3. The general pattern is the same for each norm violation: they receive judgment scores at

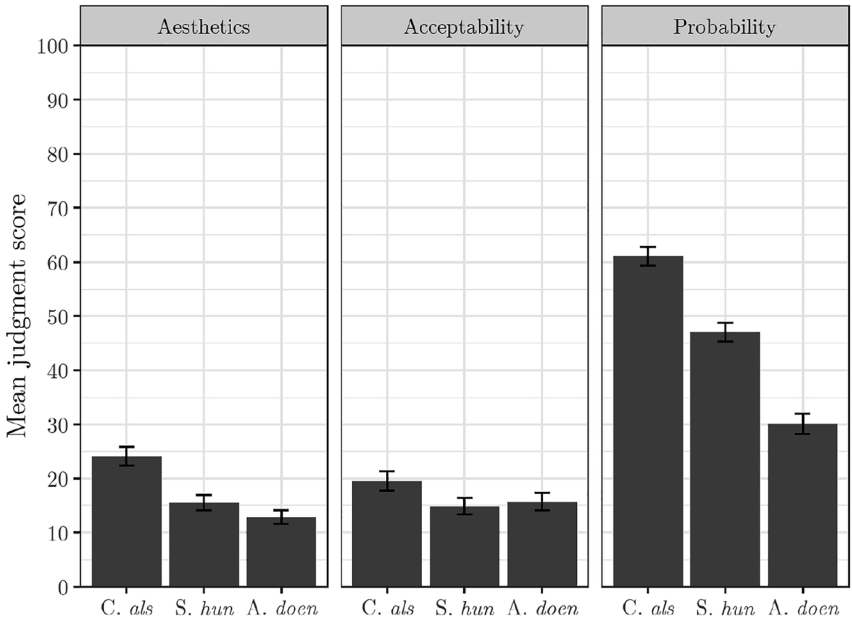

Figure 3: Mean judgment scores per prescriptive norm violation (comparative *als*, subject *hun*, auxiliary *doen*) in three dimensions (error bars indicate within-subject standard errors from the mean).

the low end of the aesthetics and acceptability scales, and much higher judgment scores on the probability scale. The cases with comparative *als* received the highest mean scores in each dimension, with a considerable margin, especially in the dimension of probability. This discrepancy may be related to the fact that the *als/dan* variation is one of the most well-known cases of stigmatized variation, and made its first appearance in prescriptive grammars as early as in the 16th century (van der Meulen 2018). By contrast, subject *hun* was not described until 1911 (van Bree 2012), and auxiliary *doen* is currently still mostly restricted to specific dialectal regions of the Netherlands (Giesbers 1983/1984) and may not even have been recognized as a violation of the prescriptive norm by some participants (as also suggested in Sert et al. in prep.). It must therefore be noted that, while judgments of prescriptive norm violations are clearly influenced by the dimension of the scale, their overall appreciation differs per violation type.

A linear mixed-effects model was performed on the *z*-transformed judgment data using the software R (version 4.0.5, R Core Team 2020) and the *lme4* package (Bates et al. 2015). The variables *judgment dimension*, *±norm violation*, and *±grammaticality* were entered into the model as fixed effects. The dimension of acceptability was set as the reference category for the variable *judgment dimension*. Both

two-level factors were coded using deviation contrasts (−0.5, 0.5). The random structure of the initial version of the model contained by-participant as well as by-item intercepts and slopes for the effects of the two two-level fixed factors and their interaction. When the model failed to converge, its random structure was simplified by removal of the by-item interaction term. The reported *p*-values were calculated with the normal approximation to the *t*-value.

The estimates of the final model for the stigmatized item set are presented in Table 1. The model yielded a significant effect of ±*norm violation* (β = −1.00, SE = 0.06, t = −18.25, p < 0.001) and ±*grammaticality* (β = 0.99, SE = 0.07, t = 14.03, p < 0.001). This confirms that, across all dimensions, the variants without a norm violation were better appreciated than the variants with a norm violation, and the grammatical variants were better appreciated than the ungrammatical variants. The model moreover yielded a significant interaction effect between the two factors (β = −1.47, SE = 0.07, t = −19.75, p < 0.001); this effect is likely driven by the high judgment scores of the grammatical sentences without a norm violation. Further, items were judged significantly higher in the dimension of probability than in the dimension of acceptability (β = 0.09, SE = 0.02, t = 3.72, p < 0.001). This effect appears to be driven by the relatively high judgment scores of items containing a prescriptive norm violation

Table 1: Model specifications of the linear mixed-effects model for the stigmatized item set (number of observations: 5,506, groups: participant, 153; item, 36).

	Fixed effects				Random effects (SDs)	
Parameters	***β***	**Std. error**	***t*-value**	***p***	***by-participant***	***by-item***
(Intercept)	−0.475	0.023	−20.452	<0.001	0.082	0.095
Norm violation	−0.996	0.055	−18.248	<0.001	0.210	0.230
Grammaticality	0.994	0.071	14.029	<0.001	0.220	0.351
Dimension (aesthetics)	−0.011	0.023	−0.492	0.623	–	–
Dimension (probability)	0.086	0.023	3.719	<0.001	–	–
Dimension (aesthetics) * norm violation	0.097	0.053	1.823	0.068	–	–
Dimension (probability) * norm violation	0.439	0.053	8.328	<0.001	–	–
Dimension (aesthetics) * grammaticality	0.066	0.055	1.203	0.229	–	–
Dimension (probability) * grammaticality	0.440	0.054	8.077	<0.001	–	–
Norm violation * grammaticality	−1.467	0.074	−19.752	<0.001	0.388	–
Dimension (aesthetics) * norm violation * grammaticality	0.009	0.102	0.086	0.931	–	–
Dimension (probability) * norm violation * grammaticality	0.544	0.101	5.392	<0.001	–	–

in the probability dimension (see Figure 2). This is supported by the significant two-way interaction effects between *judgment dimension* (probability) and ±*norm violation* (β = 0.44, SE = 0.05, t = 8.33, p < 0.001) and between *judgment dimension* (probability) and ±*grammaticality* (β = 0.44, SE = 0.05, t = 8.08, p < 0.001), and the significant three-way interaction between *judgment dimension* (probability), ±*norm violation*, and ±*grammaticality* (β = 0.54, SE = 0.10, t = 5.39, p < 0.001). The difference between judgment scores in the dimensions of aesthetics and acceptability was not significant (β = −0.01, SE = 0.02, t = −0.49, p = 0.623), nor were the two-way interaction effects between *judgment dimension* (aesthetics) and ±*norm violation* (β = 0.10, SE = 0.05, t = 1.82, p = 0.068) and between *judgment dimension* (aesthetics) and ±*grammaticality* (β = 0.07, SE = 0.06, t = 1.20, p = 0.229), or the three-way interaction effect between *judgment dimension* (aesthetics), ±*norm violation*, and ±*grammaticality* (β = 0.01, SE = 0.10, t = 0.09, p = 0.931). Thus, the judgment patterns, in terms of the variables ±*norm violation* and ±*grammaticality*, did not differ significantly between the acceptability and aesthetics scales, but they did differ significantly between the acceptability and probability scales.

3.4.2 Scrambling

The results for the scrambling item set are visually represented in Figure 4. The judgment patterns are similar across dimensions, with lower judgment scores on the aesthetics scale than on the acceptability and probability scales. The scrambled variants received higher judgment scores than the unscrambled variants on each scale, yet all four conditions received a judgment score at the high end of the scale. Definite objects are thus better appreciated in scrambled than in unscrambled position, regardless of their topicality. But when compared to the judgment scores of the filler items, all scrambling clauses are judged as acceptable, likely to be produced by a native speaker, and to a lesser extent aesthetically pleasing (scrambled or unscrambled, topical or focused). The manipulation of the discourse context did not have an effect that is clearly visible from Figure 4.

A linear mixed-effects model was performed on the *z*-transformed judgment data using the software R (version 4.0.5, R Core Team 2020) and the *lme4* package (Bates et al. 2015). The variables *judgment dimension*, *context type* (topic, focus), and *object position* (unscrambled, scrambled) were entered into the model as fixed effects. The dimension of acceptability was set as the reference category for the variable *judgment dimension*. Both two-level factors were coded using deviation contrasts (−0.5, 0.5). The random structure of the initial version of the model contained by-participant as well as by-item intercepts and slopes for the effects of the two two-level fixed factors and their interaction. When the model failed to converge,

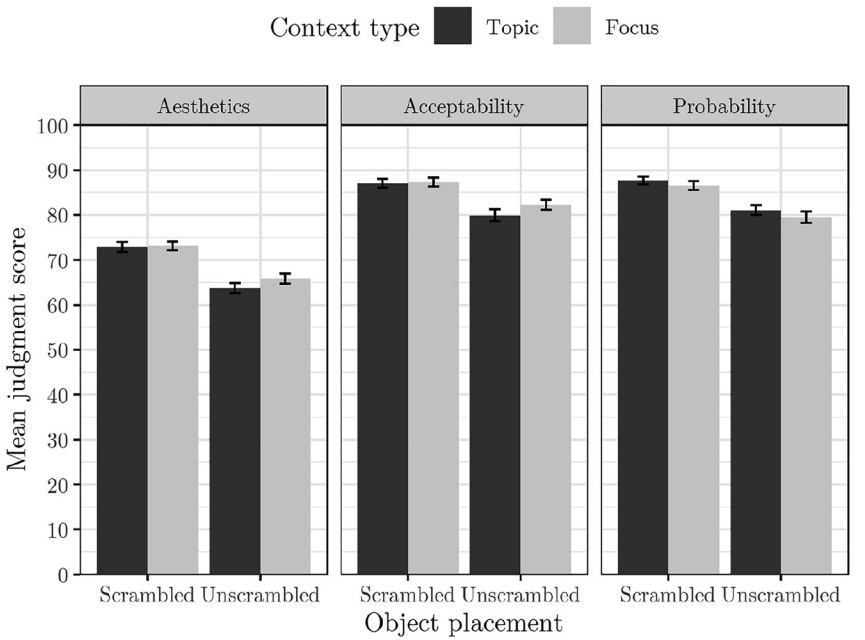

Figure 4: Mean judgment scores per condition for the scrambling item set in three dimensions (error bars indicate within-subject standard errors from the mean).

its random structure was simplified by removing the by-participant interaction term. The reported *p*-values were calculated with the normal approximation to the *t*-value.

The estimates of the final model for the stigmatized item set are presented in Table 2. The model yielded a significant effect of *object position* (β = 0.16, SE = 0.05, t = 3.43, p < 0.001), confirming that scrambled variants were rated higher than unscrambled items in general. The effect of *context type* was not significant (β = 0.03, SE = 0.03, t = 0.89, p = 0.374), nor was the interaction effect between *context type* and *object position* (β = –0.04, SE = 0.06, t = –0.62, p = 0.533). This means that the present experiment did not provide evidence for a discourse template of sorts. Regarding the scale dimensions, the difference between the acceptability and probability scores was not significant (β = –0.05, SE = 0.03, t = –1.52, p = 0.129), but the aesthetic scores were significantly lower than the acceptability scores (β = –0.08, SE = 0.03, t = –2.56, p = 0.010). The model did not yield a significant interaction effect between *judgment dimension* (aesthetics) and *context type* (β = –0.02, SE = 0.04, t = –0.42, p = 0.677) or between *judgment dimension* (aesthetics) and *object position* (β = 0.09, SE = 0.06, t = 1.55, p = 0.121). The interaction effects between *judgment dimension* (probability) and *context type* (β = –0.05, SE = 0.04, t = –1.33, p = 0.185) and between *judgment*

Table 2: Model specifications of the linear mixed-effects model for the scrambling item set (number of observations: 3,671, groups: participant, 153; item, 24).

	Fixed effects				Random effects (SDs)	
Parameters	***β***	**Std. error**	***t*-value**	***p***	***by-participant***	***by-item***
(Intercept)	0.842	0.036	23.281	<0.001	0.126	0.139
Context type	0.026	0.029	0.888	0.374	0.054	0.058
Object position	0.164	0.048	3.429	<0.001	0.227	0.113
Dimension (aesthetics)	−0.079	0.031	−2.564	0.010	–	–
Dimension (probability)	−0.046	0.030	−1.519	0.129	–	–
Dimension (aesthetics) * context type	−0.015	0.036	−0.416	0.677	–	–
Dimension (probability) * context type	−0.048	0.036	−1.327	0.185	–	–
Dimension (aesthetics) * object position	0.089	0.058	1.549	0.121	–	–
Dimension (probability) * object position	0.004	0.057	0.064	0.949	–	–
Context type * object position	−0.038	0.061	−0.624	0.533	–	0.168
Dimension (aesthetics) * context type * object position	−0.009	0.069	−0.132	0.895	–	–
Dimension (probability) * context type * object position	0.027	0.068	0.393	0.694	–	–

dimension (probability) and *object position* (β = 0.00, SE = 0.06, t = 0.06, p = 0.949) were not significant. Finally, neither of the three-way interactions was significant (aesthetics: β = −0.01, SE = 0.07, t = −0.13, p = 0.895; probability: β = 0.03, SE = 0.07, t = 0.39, p = 0.694).

3.5 Discussion

The results of the judgment experiment demonstrate that the manipulation of the scale dimension had an effect on the outcome, although what the effect looks like is dependent on the type of variation in the stimulus sentences. In the following, I will discuss the results for the stigmatized and non-stigmatized variation separately.

3.5.1 The (non-)appreciation of prescriptive norm violations

The results of the present study replicate the results in Vogel (2019) for cases of stigmatized variation in Dutch: items with a prescriptive norm violation received higher judgment scores on the scale of probability than on the scale of acceptability.

The results presented here differ from Vogel's results in one crucial respect. Vogel reports that the aesthetic judgment scores were only slightly lower than the normative judgment scores especially in the variants without a norm violation, yet the difference between the aesthetic and acceptability judgment scores presented here was only small for items with a prescriptive norm violation—not for the unmarked variants. Instead, the acceptability scores for these sentences were much higher, closer to the probability scores. One possible reason for this discrepancy is that the stimulus sentences in the experiment presented here were accompanied by linguistic context. The preambles may have influenced the aesthetic judgments in a negative way, through lack of poetic flourish. Moreover, the "P-judgments" in Vogel's experiment were judgments in terms of mere *possibility* in all varieties of German, whereas the experiment presented here elicited judgments in terms of *probability* (i.e., likelihood of a sentence having been pronounced by a native speaker). Judgment in terms of possibility might be more lenient than judgment in terms of probability. Another possibility, finally, is that the difference can be partially or additionally attributed to cultural differences between the Netherlands and Germany regarding the prescriptive norm, that is, German speakers might simply draw a closer connection between prescriptive norms and aesthetics than Dutch speakers do.

As noted, judgment scores in the stigmatized item set of the present experiment were significantly higher in the dimension of probability than in the dimension of acceptability. Moreover, the two-way interactions as well as the three-way interaction between *judgment dimension* (acceptability vs. probability), *norm violation*, and *grammaticality* were also significant. These effects are driven by the relatively high probability scores for grammar-internally unmarked items that contained a violation of the prescriptive norm: the constructions exist in the linguistic reality despite the prescriptive norms that reject them, and native speakers of Dutch are aware of this. The results show that one group of participants did not deem such constructions particularly good in terms of the language system (based on prescriptive rules); another group of participants judged their surface probability much higher. This indicates that participants in a linguistic judgment experiment take into consideration the dimension of the judgment scale and are therefore able to consciously control the judgment routine they invoke based on the instructions – at least in case of stigmatized variation. Therefore, Schütze's (1996) perception-based theory of linguistic judgments needs to allow for a degree of variability around prescriptive norms (cf. Labov 1996). The question which cognitive processes play a role in judging prescriptive norm violations and how these processes interact, as well as the question how these processes relate to those operational in the judgment of non-stigmatized variation, I will mostly leave for future research (although I will briefly return to this discussion in Section 4.1).

Considering the prevalence (or salience) of prescriptive norm violations in colloquial language use, however, one could argue that the mean probability score is still rather low (46.08%). This suggests that participants in the probability condition did not judge the sentences completely independently of the dimension of acceptability; that is, the reported judgments may represent considerations from a composite of multiple dimensions. Participants may have wanted to demonstrate their knowledge of the prescriptive rules despite not being instructed to do so,[10] which would mean that linguistic judgments can in fact be multi-dimensional. Moreover, note that the status of prescriptive norm violations in a grammatical system is not at all clear. In the experiment presented here, for example, the acceptability and the aesthetic scores for items with a prescriptive norm violation were comparable to the corresponding scores for ungrammatical items. This replicates earlier findings from a sentence-matching experiment reported in Hubers et al. (2020), who study comparative *als* 'as' and find that the prescriptive norm violations slow down the sentence-matching process as much as ungrammatical sentences do. However, Hubers et al. also conduct an eye-tracking experiment in which they observe that prescriptive norm violations and ungrammatical sentences do not in fact pattern alike in the earliest stages of the parse. Thus, prescriptive norm violations are crucially different from their grammatical and ungrammatical counterparts and in a sense form some sort of an in-between category (see also Hubers et al. 2016 for an fMRI study with the same conclusion).

Finally, in the present study I collected judgment data of three violations of the Dutch prescriptive norm, each with a vastly different history. The disparate (non-) appreciation of these norm violations can serve as an important consideration for much needed future research on the topic, which is relatively understudied (Bennis and Hinskens 2014 is a notable exception).

3.5.2 The (non-)appreciation of middle-field scrambling

Based on the results from Schoenmakers et al. (2022), the prediction for the scrambling sentences was that all variants would receive high judgment scores, with discourse conditions motivating object placement. Specifically, topics were expected to be (slightly) better appreciated in scrambled position, and foci in unscrambled position. This is not what we find. Instead, scrambled objects received higher

10 An alternative interpretation of the findings is that the prescriptive norms are effective to the extent that the actual surface probability of the prescriptive norm violations is in fact low (see van Bergen et al. (2011) and Hubers and de Hoop (2013) for corpus data of subject *hun* and comparative *als*). However, the high general salience of prescriptive norm violations suggests that surface probability estimations are not likely to be affected in this way.

judgment scores than unscrambled objects regardless of their topicality, in each dimension of judgment. This finding is unexpected, as most of the theoretical literature postulates a strict discourse template which reserves the scrambled position for topics and the unscrambled position for foci (Broekhuis 2008; Broekhuis and Corver 2016: Ch. 13; Neeleman and van de Koot 2008; Neeleman and Reinhart 1998; Schaeffer 1997, 2000; Verhagen 1986). Deviations from the discourse template (i.e., unscrambled topics and scrambled foci) were therefore expected to receive lower judgment scores than their information structurally better-behaved counterparts, although they were not expected to receive judgment scores at the very low end of the scale either (based on previous experimental results).

The general preference for scrambled definite objects is unexpected as well, since van Bergen and de Swart (2009, 2010) find a preference for the unscrambled position in their corpus data, and Schoenmakers and de Swart (2019) moreover find in a sentence judgment experiment that scrambled and unscrambled sentences with a clause adverb receive similar judgment scores at the high end of the (naturalness) scale when presented free of context. Note, however, that the differences in judgment scores for scrambled and unscrambled items in the experiment presented here are smaller than 10%, and that the unscrambled sentences still received judgment scores at the very high end of the scale.

A possible explanation for the two observations reported above is that the objects in the present experiment were part of the common ground in both the topic and focus conditions, because they were always anaphoric (see also footnote 9). The possibility that presuppositionality (or anaphoricity) is the most important determinant in scrambling, and not discourse prominence (topicality), could explain the general preference for scrambled objects as well as the non-significant difference between context types. However, a manipulation of topicality was enough to elicit a significant effect in Schoenmakers et al.'s (2022) experiment. The discrepancy with the findings presented here can then be understood as a task difference. Recall that two sentences can be equally acceptable while there is a distinct preference for one of them in language production (e.g., Bader and Häussler 2010; Kempen and Harbusch 2008). Such a discrepancy has in fact been reported for Dutch scrambling sentences before (Schoenmakers and de Swart 2019; de Swart and van Bergen 2011).

Another possible explanation for the unexpected findings is that (some) participants may not have been fully engaged in the context because of its non-evident role in (and the overall length of) the experiment. If this was the case, the manipulation of context type was unsuccessful, and participants may have used the definite article as a proxy for information structure instead (Givón 1988, cf. also Coussé 2009). That is, participants may have interpreted the target objects as presuppositional

regardless of the manipulation of context type. Note, however, that on either possibility, the relatively high judgment scores for the unscrambled sentences are unexpected on the assumption that this position is reserved for non-presuppositional (discourse-new) information.

3.5.3 A comparison of raw scores and *z*-scores[11]

Recall that, following Schütze's (1996) seminal work, Sprouse (2020) defines linguistic judgments as the conscious reports of automatic responses to a stimulus sentence. The critical issue for this theory of judgments is to what extent the effects are simply main effects of the use of the scale and to what extent they impinge on the relative acceptability between conditions. Now, the *z*-score transformation is explicitly designed to "capture all of the available information about the relative acceptability of the tested sentence types, while abstracting away from the specifics of the scale used to collect the data" (Cowart 1997: 14), i.e., judgment scores are standardized across scales, thereby eliminating the main effects of the use of the scale. For that reason, the *z*-score plots of the data should look identical in the three dimensions, on Schütze's account, despite the clear numerical differences depicted in Figures 2 and 4 (viz. between the dimensions of aesthetic quality and linguistic acceptability). If the effects would involve the judgment pattern between conditions, those should become more pronounced after the *z*-score transformation, because the main effect of the scale would not distract us from it.

The mean *z*-scores and distributional information of the two data sets are given in Figures 5 and 6 (see Figure B1 in Appendix B for a violin/boxplot of the filler items). On visual inspection, the differences in the judgment patterns between dimensions are much less pronounced in the *z*-scores than in the raw data (see Figures 2 and 4), except in the case of (grammatical) prescriptive norm violations. These items are rated considerably higher in the dimension of probability than in the dimensions of acceptability and aesthetic quality. That is, the use of the scale has a clear effect on the judgment patterns *between conditions*. This finding does not readily corroborate Schütze's (1996) account of linguistic judgments, an issue I will address in Section 4.1. But in the cases of non-stigmatized variation, in the scrambling set, the judgment patterns between dimensions in the *z*-score plots look nearly identical, unlike those in the raw data plots, where we can see that the judgment scores of aesthetic quality are much lower. This is exactly what Schütze's account predicts; I will elaborate on this finding in Section 4.2.

11 I would like to thank an anonymous reviewer for recommending the addition of this subsection. In retrospect, the manuscript would not have been complete without it.

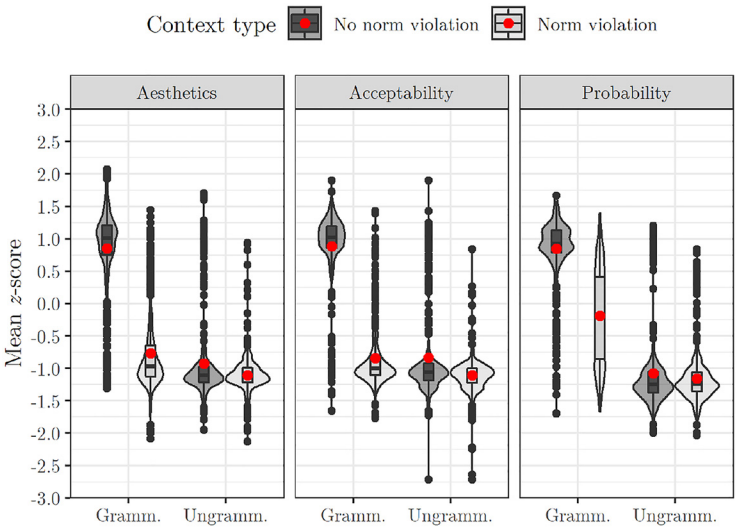

Figure 5: Distribution of *z*-scores per condition (stigmatized item set) in three dimensions (the red dot encodes the mean, the horizontal black line the median, and the edges of the boxplot the interquartile range; the violins provide a density plot).

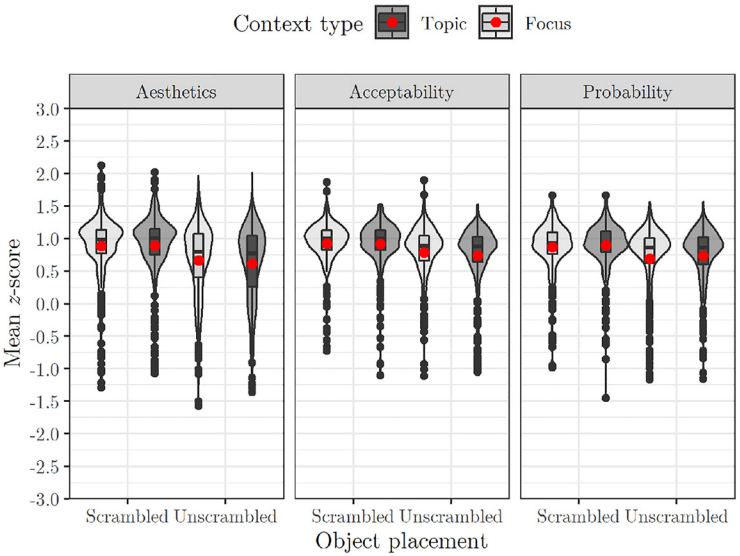

Figure 6: Distribution of *z*-scores per condition (scrambling item set) in three dimensions (the red dot encodes the mean, the horizontal black line the median, and the edges of the boxplot the interquartile range; the violins provide a density plot).

Note, however, that the statistical analysis reported in Section 3.4.2 was performed on the *z*-transformed data. This analysis reveals a significant main effect between the dimensions of aesthetic quality and acceptability. The significance of this main effect is likely due to bottom effects of the ungrammatical non-experimental conditions (i.e., fillers and stigmatized items). This should not distract us from the discrepancy between the judgment patterns visualized in the raw data plots and the *z*-score plots. This discrepancy is striking and speaks against the claim that participants in judgment experiments engage in Wundtian introspection to arrive at their reported judgment, and in favor of Schütze's (1996) perception-based theory. More specifically, it indicates that participants take into account the dimension of the judgment scale, but this does not affect the relative judgments between conditions (in non-stigmatized variation).

4 General discussion

The main research question of this study was whether the dimension of the scale in linguistic judgment experiments can impact their outcome. Earlier studies on the topic report inconclusive results (Cowart 1997; Langsford et al. 2019) and/or examined specific constructions with a grammatical illusion or prescriptive norm violation only (Bennis and Hinskens 2014; Langsford et al. 2019; Vogel 2019). The present study provides evidence for the impact of the experimental instructions: stimulus sentences were rated differently in terms of the aesthetic quality, linguistic acceptability, and surface probability both in cases of stigmatized (prescriptive norm violations) and non-stigmatized (middle-field scrambling) variation. What exactly these differences between judgment scales look like depends on the type of variation.

4.1 Reasoning about and processing of sentences

The manipulation of the scale dimension triggered differences in the judgment pattern in the stigmatized item set only, i.e., there were interactions between the scale dimension and the two other factors under investigation. The crucial difference between this item set and the non-stigmatized item set is that participants have conscious access to the prescriptive rules of their language. As Schütze (1996: 83) puts it, "we could imagine that expected judgment causes people to revert to conscious reasoning *about* sentences, rather than processing *of* them." This warning applies

especially to the acceptability scores, and likely also the aesthetic scores, which may very well reflect the participants' knowledge of the prescriptive rules rather than the extent to which they really accept the sentences or find them beautiful. The difference with the probability scores reported in the previous section can thus be considered a difference between, on the one hand, a conscious mapping between a stimulus sentence and a prescriptively ideal grammar, and on the other hand, an estimation of the frequency with which the stimulus sentence occurs – where the latter is the more automatic reaction. Notice that the judgments of linguistic acceptability and aesthetic quality then reflect a binary opposition, in that the stimulus sentence either does or does not match a prescriptively correct form, whereas the frequency estimations are much more open-ended. The idea that there are different types of judgment in itself is not new, although it has been presented in different guises. Featherston (2005), for example, argues that categorical and relative judgments tap into different cognitive processes,[12] and Bader and Häussler (2010) present a model which can be used to derive binary judgments from gradient judgments (i.e., automatic responses), implementing a so-called *decision criterion* (cf. Bader and Häussler 2020). The stigmatized item set may thus not be particularly relevant for Schütze's (1996) perception-based theory of linguistic judgments; however, the results do indicate that, when rating sentences with a prescriptive norm violation, participants are able to toggle between judgment strategies based on the instructions provided to them.

The different strategies may also explain the processing effects reported in Hubers et al. (2020). In an eye-tracking experiment, Hubers et al. find no reading time differences between prescriptive norm violations and grammatical sentences in the earliest stages of the parse. But then later on in processing, the reading times for the norm violations start to increase compared to the grammatical items, and they start to resemble the ungrammatical items more (at least for some participants; the authors report considerable individual variation here). Thus, participants may have had an automatic reaction to the stimulus sentence and initially accepted the prescriptive norm violation (note that forms that are frequent in the linguistic reality are considered grammatical in theories of grammar, cf. Hubers et al. 2016). But when the participant recognizes the structure as a prescriptive norm violation, they revert to conscious reasoning about it and, in so doing, they slow down. The reading times consequently increase and start to resemble those due to processing difficulties associated with ungrammaticality. The unique way in which violations of the

12 For Featherston (2005: 201), categorical judgments are actually expressions of "the likelihood that a structure *is good enough to occur in practice*," yielding a binary opposition reflecting occurrence (yes or no). Relative judgments, on the other hand, reflect the processing cost of a structure's form, meaning, and the mapping between the two.

prescriptive norm are processed may thus be due to a tension between automatic and more conscious processing strategies, which also affect the corresponding judgment data.

4.2 An acceptability core with aesthetic add-ons

Regarding the grammatical sentences that do not contain a prescriptive norm violation, Schütze's (1996) account of linguistic judgments predicts that there should only be main effects of the scale dimension, without changes in the relative judgment pattern between conditions (cf. Cowart 1997). The reason for this is that the core component of the judgments is driven by an automatic process (cf. Sprouse 2020). The results of the experiment presented here show exactly this pattern: the mean judgment scores only move up or down the scale without affecting the relative patterns between conditions. Specifically, the judgment scores for the grammatical fillers were considerably lower in the dimension of aesthetics than in the other two dimensions (at least numerically, see Appendix A), and this same penalty also emerged in the grammatical sentences without a norm violation from the stigmatized item set as well as in the scrambling sentences. In line with the prediction, the interactions between the scale dimensions and the other two factors in the scrambling item set were not significant, and the judgment patterns look remarkably alike when the effect of the judgment scale has been neutralized (see Section 3.5.3). So, participants were more critical when instructed to judge sentences in terms of their linguistic feeling (aesthetics) than in terms of the linguistic system (acceptability), but the experiment presented here did not yield evidence that the scale dimension influenced the pattern of judgments between conditions. The results can thus be interpreted as supporting and extending Schütze's (1996) account: linguistic judgments are the conscious reports of automatic reactions to stimulus sentences, consisting of an acceptability core and additional effects from other cognitive processes that influence judgments of aesthetic quality (and presumably also judgments of surface probability, or other types of judgment, although the results reported here do not provide evidence for this).

The results indicate that the aesthetic judgment scores were decreased across the board when compared to the acceptability judgment scores. One possible explanation is that the two judgment scales reflect Romand's (2019, forthcoming) distinction between *form feeling* and *formal feeling*. The main difference between the two terms is that the form feeling pertains to the psycho-aesthetic, in part socially determined, impression of a string (cf. Fortis 2019), whereas the formal feeling describes an affective state towards the mapping of the mental representation of a

sentence and its structure onto the actual utterance (here, terms such as *expectation*, *satisfaction*, and *deception* are used). Note that the latter definition is reminiscent of Schütze's (1996) take on linguistic judgments as automatic perceptions of a sentence. Although aesthetic judgments are not commonly elicited in linguistic research, here they indicate that participants in linguistic judgment experiments do take into account the dimension of the scale (while still representing all relevant features of the stimulus, cf. Featherston 2021). The instructions in the aesthetics condition possibly motivated participants to take into account considerations of their form feeling in addition to their normal judgment criteria (which constitute their formal feeling and are arguably elicited by the acceptability instructions, cf. Schütze 1996).

This is an important conclusion in light of the mapping between experimental data and linguistic theory. When interpreting the results of their experiments, experimental linguists must disentangle the grammatical and extra-grammatical factors that weigh in on the judgment process, and then make inferences about the grammaticality status of a construction based on explicit theoretical assumptions (Juzek and Häussler 2020). This process is very intricate if the results reflect an amalgamation of (partially non-instructed) considerations on the part of the participant, especially since these considerations can be vastly different conceptually. However, the findings presented here imply that the experimental instructions did not influence the outcome of the judgment experiment in a way that impinges on the relative acceptability between conditions, and so the observed differences between judgment scales are not necessarily problematic for linguistic theory-building (cf. Featherston 2021). The experimental researcher is nonetheless advised to carefully consider the dimension of the scale prior to testing (Chaudron 1983; see also Featherston 2021; Marty et al. 2020; Schütze 1996; Schütze and Sprouse 2014 for other considerations), and to be explicit about whether the reported judgments are primarily judgments about the linguistic system, the linguistic reality, or the linguistic feeling (or any other dimension).

5 Conclusion

The dimension of the scale in judgment experiments impacts the results, whether the stimuli contain cases of stigmatized or non-stigmatized variation. Judgments about the surface probability of prescriptive norm violations in Dutch are higher than judgments about their aesthetic quality or linguistic acceptability, which reflects the prevalence (or salience) of such constructions in the linguistic reality. However, the judgments of linguistic acceptability (and aesthetic quality) are likely informed by conscious reflection of prescriptive language rules. The probability scores, by contrast, are likely automatic responses to the stimulus sentence, but they are still

only at approximately 50%. Given the prevalence of violations of the prescriptive norm in the linguistic reality, this finding suggests that participants take into account considerations from multiple dimensions.

Regarding the scrambling item set, the experimental results did not reveal judgment patterns that follow the discourse template assumed in most theoretical literature. Instead, definite objects in scrambled position were better appreciated than those in unscrambled position (by a small margin), regardless of the manipulation of information structure. This finding corroborates theoretical accounts of Dutch middle-field scrambling that allow for a degree of optionality (e.g., de Hoop 2000, 2003; Struik and Schoenmakers 2022), as well as Schütze's (1996) perception-based theory of linguistic judgments, in that participants do not engage in conscious Wundtian introspection, but rather report their automatic responses to the stimulus.

The probability and acceptability scores for all grammatical sentences in the experiment presented here were considerably higher than the aesthetic scores for the same sentences. Thus, participants were more critical when asked to judge sentences in terms of their linguistic feeling than in terms of the linguistic system or the linguistic reality, but the judgment scale manipulation did not impinge on the relative acceptability between conditions. The different scale dimensions may, however, trigger additional cognitive processes that add to the acceptability core of a linguistic judgment. What it is exactly that linguistic judgments quantify is matter of debate, although definitions from the field of philosophy of language highlight its subjective nature, relating it to the broader perspective of affective psychology (Romand 2019, forthcoming) and art appreciation (Fortis 2019). The present study disentangled judgments in terms of linguistic acceptability and aesthetic quality (in cases of non-stigmatized variation). Linguistic research has not been concerned with this question as much; instead, judgments have for the most part been elicited in terms of acceptability and surface probability. What led participants to their eventual judgment scores can never be assumed, but the dimension of the scale can serve as a first indication and should receive special attention in the experimental design stage.

Data availability statement

The data generated and analyzed during this study are available in the Radboud University repository: https://doi.org/10.34973/k01s-ez17.

Acknowledgments: I would like to thank the anonymous reviewers, Stacie Chadwick, Cas Coopmans, Ad Foolen, Helen de Hoop, and Peter de Swart for their

helpful comments on an earlier version of this paper. Their feedback really improved the paper.

Appendix A

Table A1: Mean judgment scores and standard deviations (between brackets) for the different categories of filler items in three dimensions.

	Aesthetics	Acceptability	Probability
Grammatical fillers	69.77 (21.29)	81.69 (22.45)	82.14 (20.51)
Marked fillers	43.19 (31.84)	48.26 (37.30)	50.36 (35.66)
Ungrammatical fillers	13.39 (19.92)	14.38 (24.19)	14.53 (22.06)

Table A2: Judgment scores and standard deviations (between brackets) for the different conditions in the stigmatized item set in three dimensions.

	Aesthetics	Acceptability	Probability
Grammatical, norm not violated	71.25 (24.04)	85.72 (24.56)	85.40 (20.59)
Grammatical, norm violated	17.49 (22.58)	16.71 (24.02)	46.08 (29.82)
Ungrammatical, norm not violated	13.09 (23.78)	17.44 (29.47)	13.00 (24.44)
Ungrammatical, norm violated	6.85 (12.39)	7.23 (13.44)	10.01 (16.32)

Table A3: Judgment scores and standard deviations (between brackets) per prescriptive norm violation in three dimensions.

	Aesthetics	Acceptability	Probability
Comparative *als*	24.08 (26.43)	19.55 (26.48)	61.08 (26.58)
Subject *hun*	15.53 (21.19)	14.89 (23.01)	47.07 (27.89)
Auxiliary *doen*	12.85 (17.88)	15.70 (22.29)	30.08 (26.63)

Table A4: Judgment scores and standard deviations (between brackets) for the different conditions in the scrambling item set in three dimensions.

	Aesthetics	Acceptability	Probability
Scrambled, topic	72.88 (21.19)	87.04 (17.59)	87.66 (14.92)
Scrambled, focus	73.13 (19.31)	87.30 (16.79)	86.54 (17.27)
Unscrambled, topic	63.74 (23.21)	79.88 (22.94)	81.04 (20.44)
Unscrambled, focus	65.85 (22.34)	82.20 (19.47)	79.46 (23.31)

Appendix B

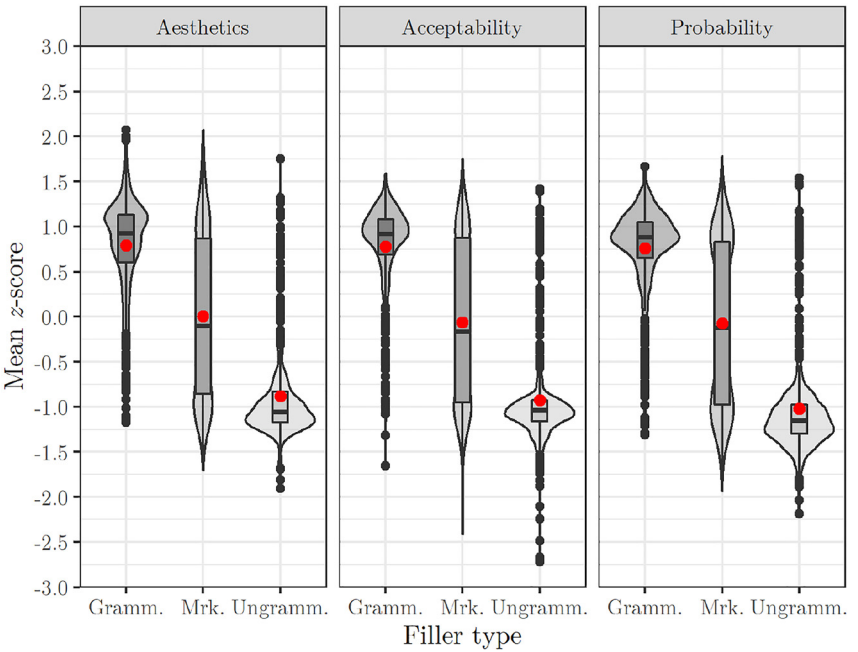

Figure B1: Distribution of *z*-scores per filler item category in three dimensions (the red dot encodes the mean, the horizontal black line the median, and the edges of the boxplot the interquartile range; the violins provide a density plot).

References

Adli, Aria. 2015. What you like is not what you do: Acceptability and frequency in syntactic variation. In Aria Adli, Marco García García & Göz Kaufmann (eds.), *Variation in language: System- and usage-based approaches*, 173–199. Berlin & Boston: de Gruyter.

Arppe, Antti & Juhani Järvikivi. 2007. Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3(1). 99–109.

Bader, Markus & Jana Häussler. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46(2). 273–330.

Bader, Markus & Jana Häussler. 2020. How to get from graded intuitions to binary decisions. In Sam Featherston, Robin Hörnig, Sophie von Wietersheim & Susanne Winkler (eds.), *Experiments in focus: Information structure and semantic processing*, 183–207. Berlin & Boston: de Gruyter.

Bard, Ellen, Dan Robertson & Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72(1). 32–68.

Barten, Sybil. 1992. The language of musical instruction. *Journal of Aesthetic Education* 26(2). 53–61.

Barten, Sybil. 1998. Speaking of music: The use of motor-affective metaphors in music instruction. *Journal of Aesthetic Education* 32(2). 89–97.
Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.
Bennis, Hans & Frans Hinskens. 2014. Goed of fout: Niet-standaard inflectie in het hedendaags Standaardnederlands [Right or wrong: Non-standard inflection in Present-day Standard Dutch]. *Nederlandse Taalkunde* 19(2). 131–184.
Bermel, Neil & Luděk Knittl. 2012. Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech. *Corpus Linguistics and Linguistic Theory* 8(2). 241–275.
Bever, Thomas. 1970. The cognitive basis for linguistic structures. In John Hayes (ed.), *Cognition and language development*, 277–360. New York: Wiley.
Bley-Vroman, Robert, Sascha Felix & Georgette Loup. 1988. The accessibility of universal grammar in adult language learning. *Second Language Research* 4(1). 1–32.
Bock, Kathryn & Carol Miller. 1991. Broken agreement. *Cognitive Psychology* 23(1). 45–93.
Botha, Rudolf. 1981. *The conduct of linguistic inquiry: A systematic introduction to the methodology of generative grammar*. The Hague: Mouton.
Broekhuis, Hans. 2008. *Derivations and evaluations: Object shift in the Germanic languages*. Berlin: de Gruyter.
Broekhuis, Hans. 2016. Syntax of Dutch: The data set. *Nederlandse Taalkunde* 21(2). 297–327.
Broekhuis, Hans & Norbert Corver. 2016. *Syntax of Dutch: Verbs and verb phrases*, vol. 3. Amsterdam: Amsterdam University Press.
Chaudron, Craig. 1983. Research on metalinguistic judgments: A review of theory, methods, and results. *Language Learning* 33(3). 343–377.
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.
Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge: MIT Press.
Chomsky, Noam & George Miller. 1963. Introduction to the formal analysis of natural languages. In Robert Bush, Robert Luce & Eugene Galanter (eds.), *Handbook of mathematical psychology*, 269–321. New York: Wiley.
Coppen, Peter-Arno. 2011. Grammatica is een werkwoord [Grammar is a verb]. In Steven Vanhooren & André Mottart (eds.), *Vijfentwintigste conferentie Het Schoolvak Nederlands [Twenty-fifth conference 'The School Subject Dutch']*, 222–228. Ghent: Academia Press.
Cornips, Leonie. 1994. De hardnekkige vooroordelen over de regionale doen+infinitiefconstructie [The persistent prejudices about the regional doen+infinitive construction]. *Forum der Letteren* 35(4). 282–294.
Cornips, Leonie. 1998. Habitual doen in Heerlen Dutch. In Ingrid Tieken-Boon van Ostade, Marijke van der Wal & Arjan van Leuvensteijn (eds.), *Do in English, Dutch and German: History and present-day variation*, 83–101. Amsterdam & Münster: Stichting Neerlandistiek/Nodus Publikationen.
Coussé, Evie. 2009. Focus, complexiteit en extrapositie: Over de veranderende woordvolgorde in het Nederlands [Focus, complexity and extraposition: About the changing word order in Dutch]. *Neerlandistiek* 9(4). 1–32.
Cowart, Wayne. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks: SAGE.
Davies, Winifred & Nils Langer. 2006. *The making of bad language*. Frankfurt am Main: Peter Lang.
de Hoop, Helen. 2000. Optional scrambling and interpretation. In Hans Bennis, Martin Everaert & Eric Reuland (eds.), *Interface strategies*, 153–168. Amsterdam: Royal Netherlands Academy of Arts & Sciences.

de Hoop, Helen. 2003. Scrambling in Dutch: Optionality and optimality. In Simin Karimi (ed.), *Word order and scrambling*, 201–216. Oxford: Blackwell.

de Hoop, Helen. 2016. Woordvolgordevariatie: Theorie versus empirie? [Word order variation: Theory versus empiricism?]. *Nederlandse Taalkunde* 21(2). 275–284.

Dellarosa, Denise. 1988. A history of thinking. In Robert Stenberg & Edward Smith (eds.), *The psychology of human thought*, 1–18. Cambridge: Cambridge University Press.

de Swart, Peter & Geertje van Bergen. 2011. Definiteness and adverb–object order in Dutch. Unpublished manuscript.

Divjak, Dagmar. 2008. On (in)frequency and (un)acceptability. In Barbara Lewandowska-Tomaszczyk (ed.), *Corpus linguistics, computer tools, and applications: State of the art*, 213–234. Frankfurt am Main: Peter Lang.

Drenhaus, Heiner, Stefan Frisch & Douglas Saddy. 2005. Processing negative polarity items: When negation comes through the backdoor. In Stephan Kepser & Marga Reis (eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 145–165. Berlin: de Gruyter.

Erteschik-Shir, Nomi. 2007. *Information structure: The syntax–discourse interface*. Oxford: Oxford University Press.

Fanselow, Gisbert & Stefan Frisch. 2006. Effects of processing difficulty on judgments of acceptability. In Gisbert Fanselow, Caroline Féry, Matthias Schlesewsky & Ralf Vogel (eds.), *Gradience in grammar: Generative perspectives*, 291–316. Oxford: Oxford University Press.

Featherston, Sam. 2005. The decathlon model of empirical syntax. In Stephan Kepser & Marga Reis (eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 187–208. Berlin & New York: de Gruyter Mouton.

Featherston, Sam. 2008. Thermometer judgments as linguistic evidence. In Claudia Riehl & Astrid Rothe (eds.), *Was ist Linguistische Evidenz?*, 69–89. Aachen: Shaker.

Featherston, Sam. 2021. Response methods in acceptability experiments. In Grant Goodall (ed.), *The Cambridge handbook of experimental syntax*, 39–61. Cambridge: Cambridge University Press.

Fortis, Jean-Michel. 2019. On Sapir's notion of form/pattern and its aesthetic background. In James McElvenny (ed.), *Form and formalism in linguistics*, 59–88. Berlin: Language Science Press.

Gibson, Edward & Ev Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes* 28(1/2). 88–124.

Giesbers, Herman. 1983/1984. Doe jij lief spelen? Notities over het perifrastisch doen [Do you play nice? Notes on periphrastic doen]. *Mededelingen van de Nijmeegse Centrale voor Dialect- en Naamkunde [Announcements of the Nijmegen Center for Dialectics and Onomastics]* 19. 57–64.

Givón, Talmy. 1988. The pragmatics of word-order: Predictability, importance and attention. In Michael Hammond, Edith Moravcsik & Jessica Wirth (eds.), *Studies in syntactic typology*, 243–284. Amsterdam: John Benjamins.

Goodall, Grant. 2021. Sentence acceptability experiments: What, how, and why. In Grant Goodall (ed.), *The Cambridge handbook of experimental syntax*, 7–38. Cambridge: Cambridge University Press.

Häussler, Jana & Tom Juzek. 2017. Hot topics surrounding acceptability judgement tasks. In Sam Featherston, Robin Hörnig, Reinhild Steinberg, Birgit Umbreit & Jennifer Wallis (eds.), *Proceedings of Linguistic Evidence 2016: Empirical, theoretical, and computational perspectives*, 1–21. Tübingen: University of Tübingen.

Häussler, Jana & Tom Juzek. 2020. Linguistic intuitions and the puzzle of gradience. In Samuel Schindler, Anna Drożdżowicz & Karen Brøcker (eds.), *Linguistic intuitions: Evidence and method*, 233–254. Oxford: Oxford University Press.

Hofmeister, Philip, Laura Stum Casasanto & Ivan Sag. 2014. Processing effects in linguistic judgment data: (Super-)additivity and reading span scores. *Language and Cognition* 6(1). 111–145.

Hubers, Ferdy & Helen de Hoop. 2013. The effect of prescriptivism on comparative markers in spoken Dutch. In Suzanne Aalberse & Anita Auer (eds.), *Linguistics in The Netherlands*, vol. 30, 89–101. Amsterdam: John Benjamins.

Hubers, Ferdy, Theresa Redl, Hugo de Vos, Lucas Reinarz & Helen de Hoop. 2020. Processing prescriptively incorrect comparative particles: Evidence from sentence-matching and eye-tracking. *Frontiers in Psychology* 11. 186.

Hubers, Ferdy, Tineke Snijders & Helen de Hoop. 2016. How the brain processes violations of the grammatical norm: An fMRI study. *Brain and Language* 163. 22–31.

Juzek, Tom & Jana Häussler. 2020. Data convergence in syntactic theory and the role of sentence pairs. *Zeitschrift für Sprachwissenschaft* 39(2). 109–147.

Kalish, Michael, John Dunn, Oleg Burdakov & Oleg Sysoev. 2016. A statistical test of the equality of latent orders. *Journal of Mathematical Psychology* 70. 1–11.

Kempen, Gerard & Karin Harbusch. 2005. The relationship between grammaticality ratings and corpus frequencies: A case study into word order variability in the midfield of German clauses. In Stephan Kepser & Marga Reis (eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 329–349. Berlin: de Gruyter.

Kempen, Gerard & Karin Harbusch. 2008. Comparing linguistic judgments and corpus frequencies as windows on grammatical competence: A study of argument linearization in German clauses. In Anita Steube (ed.), *The discourse potential of underspecified structures*, 179–192. Berlin: de Gruyter.

Labov, William. 1975. Empirical foundations of linguistic theory. In Robert Austerlitz (ed.), *The scope of American linguistics*, 77–133. Lisse: Peter de Ridder.

Labov, William. 1996. When intuitions fail. In Lisa McNair, Kora Singer, Lise Dobrin & Michelle Aucon (eds.), *Papers from the parasession on theory and data in linguistics*, vol. 32, 77–106. Chicago: Chicago Linguistic Society.

Lamers, Monique & Helen de Hoop. 2014. Animate object fronting in Dutch: A production study. In Brian MacWhinney, Andrej Malchukov & Edith Moravcsik (eds.), *Competing motivations in grammar and usage*, 42–53. Oxford: Oxford University Press.

Langer, Nils. 2001. *Linguistic purism in action: How auxiliary tun was stigmatised in Early New High German*. Berlin: de Gruyter.

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.

Langsford, Steven, Rachel Stephens, John Dunn & Richard Lewis. 2019. In search of the factors behind naive sentence judgments: A State Trace Analysis of grammaticality and acceptability ratings. *Frontiers in Psychology* 10. 2886.

Leivada, Evelina & Marit Westergaard. 2020. Acceptable ungrammatical sentences, unacceptable grammatical sentences, and the role of the cognitive parser. *Frontiers in Psychology* 11. 364.

Linzen, Tal & Yohei Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa: A Journal of General Linguistics* 3(1). 100.

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.

Marty, Paul, Emmanuel Chemla & Jon Sprouse. 2020. The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Glossa: A Journal of General Linguistics* 5(1). 72.

Munro, Robert, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen & Harry Tily. 2010. Crowdsourcing and language studies: The new generation of linguistic data. In Chris Callison-Burch & Mark Dredze (eds.), *Proceedings of the NAACL HLT 2010*

workshop on creating speech and language data with Amazon's mechanical turk, 122–130. Stroudsburg: Association for Computational Linguistics.
Neeleman, Ad & Hans van de Koot. 2008. Dutch scrambling and the nature of discourse templates. *Journal of Comparative Germanic Linguistics* 11(2). 137–189.
Neeleman, Ad & Tanya Reinhart. 1998. Scrambling and the PF interface. In Miriam Butt & Wilhelm Geuder (eds.), *The projection of arguments*, 309–353. Stanford: CSLI Publications.
Nisbett, Richard & Timothy Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84(3). 231–259.
Noordman, Leo & Wietske Vonk. 1987. De selectieve verwerking van tekst [The selective processing of text]. *Tijdschrift voor Taal- en Tekstwetenschap* 7(1). 57–69.
Parker, Dan & Colin Phillips. 2016. Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition* 157. 321–339.
Pateman, Trevor. 1987. *Language in mind and language in society: Studies in linguistic reproduction*. Oxford: Clarendon Press.
Phillips, Colin, Matthew Wagers & Ellen Lau. 2011. Grammatical illusions and selective fallibility in real-time language comprehension. In Jeffrey Runner (ed.), *Experiments at the interfaces*, vol. 37, 147–180. Leiden: Brill.
Qualtrics. 2021. Qualtrics XM: The leading experience management software. Provo. Available at: https://www.qualtrics.com.
R Core Team. 2020. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
Ringen, Jon. 1977. On evaluating data concerning linguistic intuition. In Fred Eckman (ed.), *Current themes in linguistics: Bilingualism, experimental linguistics, and language typologies*, 145–160. Washington D.C.: Hemisphere.
Romand, David. 2019. More on formal feeling/form-feeling in language sciences: Heinrich Gomperz's concept of "formal logical feeling" (logisches Formalgefühl) revisited. *Histoire épistémologie langage* 41(1). 131–157.
Romand, David. forthcoming. 'Formal feeling' or 'form-feeling': Genealogical and typological analysis of a concept between psychology, theory of language, aesthetics, and art history. In Willi Reinecke & Serge Tchougounnikov (eds.), *Die psychologische Ästhetik in der Jahrhundertwende: Zwischen Psychologismus und Formalismus*, 1–24. Berlin & Münster: LIT.
Santana, Carlos. 2020. How we can make good use of linguistic intuitions, even if they are not good evidence. In Samuel Schindler, Anna Drożdżowicz & Karen Brøcker (eds.), *Linguistic intuitions: Evidence and method*, 129–148. Oxford: Oxford University Press.
Schaeffer, Jeanette. 1997. *Direct object scrambling in Dutch and Italian child language*. Los Angeles: University of California dissertation.
Schaeffer, Jeanette. 2000. *The acquisition of direct object scrambling and clitic placement: Syntax and pragmatics*. Amsterdam: John Benjamins.
Schindler, Samuel, Anna Drożdżowicz & Karen Brøcker. 2020. *Linguistic intuitions: Evidence and method*. Oxford: Oxford University Press.
Schmidt, Richard & Carol McCreary. 1977. Standard and super-standard English: Recognition and use of prescriptive rules by native and non-native speakers. *TESOL Quarterly* 11(4). 415–429.
Schoenmakers, Gert-Jan. 2020. Freedom in the Dutch middle-field: Deriving discourse structure at the syntax–pragmatics interface. *Glossa: A Journal of General Linguistics* 5(1). 114.
Schoenmakers, Gert-Jan & Ad Foolen. 2022. At the margins of grammar: Dutch and German verb particles in first sentence position. *Nederlandse Taalkunde* 27(3). 368–393.

Schoenmakers, Gert-Jan, Marjolein Poortvliet & Jeannette Schaeffer. 2022. Topicality and anaphoricity in Dutch scrambling. *Natural Language & Linguistic Theory* 40(2). 541–571.

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*, 124–145. Tübingen: University of Tübingen.

Schulte, Joachim. 1988. Remarks on Sprachgefühl. In Kristóf Nyíri & Barry Smith (eds.), *Practical knowledge: Outlines of a theory of tradition and skills*, 136–146. London: Croom Helm.

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, Berlin.

Schütze, Carson & Jon Sprouse. 2014. Judgment data. In Robert Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 27–50. Cambridge: Cambridge University Press.

Sert, Cansel, Theresa Redl & Helen de Hoop. in prep. On the acceptability of the not so dummy auxiliary 'do' in Dutch.

Siouffi, Gilles. 2018. La notion de sentiment linguistique et la philologie au tournant des XIXe et XXe siècles. *Romanica Cracoviensia* 2. 97–104.

Spencer, Nancy. 1973. Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research* 2(2). 83–98.

Sprouse, Jon. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics* 1. 123–134.

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*, 215–232. Oxford: Oxford University Press.

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.

Sprouse, Jon & Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's 'Core syntax'. *Journal of Linguistics* 48(3). 609–652.

Struik, Tara & Gert-Jan Schoenmakers. 2022. When information structure exploits syntax: The relation between the loss of VO and scrambling in Dutch. *Journal of Linguistics* 1–36. https://doi.org/10.1017/s0022226722000172.

Trotzke, Andreas, Stefano Quaglia & Eva Wittenberg. 2015. Topicalization in German particle verb constructions: The role of semantic transparency. *Linguistische Berichte* 244. 407–424.

van Bergen, Geertje & Peter de Swart. 2009. Definiteness and scrambling in Dutch: Where theory meets practice. In Anisa Schardl, Martin Walkow & Muhammad Abdurrahman (eds.), *Proceedings of the North East Linguistic Society (NELS)*, vol. 38, 89–100. Amherst: GLSA.

van Bergen, Geertje & Peter de Swart. 2010. Scrambling in spoken Dutch: Definiteness versus weight as determinants of word order variation. *Corpus Linguistics and Linguistic Theory* 6(2). 267–295.

van Bergen, Geertje, Wessel Stoop, Jorrig Vogels & Helen de Hoop. 2011. Leve hun! Waarom hun nog steeds hun zeggen [Long live hun! Why hun still say hun]. *Nederlandse Taalkunde* 16(1). 2–29.

van Bree, Cor. 2012. Hun als subject in een grammaticaal en dialectologisch kader [Hun as subject in a grammatical and dialectological frame]. *Nederlandse Taalkunde* 17(2). 229–249.

van Casteren, Maarten & Matthew Davis. 2006. Mix, a program for pseudorandomization. *Behavioral Research Methods* 38(4). 584–589.

van der Does, Jaap & Helen de Hoop. 1998. Type-shifting and scrambled definites. *Journal of Semantics* 15. 393–416.

van der Meulen, Marten. 2018. Do we want more or less variation? The comparative markers als and dan in Dutch prescriptivism since 1900. In Bert Le Bruyn & Janine Berns (eds.), *Linguistics in The Netherlands*, vol. 35, 79–96. Amsterdam: John Benjamins.

van der Meulen, Marten. 2020. Language should be pure and grammatical: Values in prescriptivism in The Netherlands 1917–2016. In Don Chapman & Jacob Rawlins (eds.), *Language prescription: Values, ideologies and identity*, 121–144. Bristol: Clevedon.

Vasishth, Shravan, Sven Brüssow, Richard Lewis & Heiner Drenhaus. 2008. Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science* 32(4). 685–712.

Verhagen, Arie. 1986. *Linguistic theory and the function of word order in Dutch: A study on interpretive aspects of the order of adverbials and noun phrases*. Amsterdam: Vrije Universiteit dissertation.

Vogel, Ralf. 2019. Grammatical taboos. *Zeitschrift für Sprachwissenschaft* 38(1). 37–79.

Wagers, Matthew, Ellen Lau & Colin Phillips. 2009. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language* 61(2). 206–237.

Wellwood, Alexis, Roumyana Pancheva, Valentine Hacquard & Colin Phillips. 2018. The anatomy of a comparative illusion. *Journal of Semantics* 35(3). 543–583.

Wundt, Wilhelm. 1896. *Grundriss der Psychologie*. Leipzig: Engelmann.

Zenner, Eline, Dirk Speelman & Dirk Geeraerts. 2012. Cognitive Sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. *Cognitive Linguistics* 23(4). 749–792.