Acceptability Judgement Tasks and Grammatical Theory

Tom S Juzek

D.Phil.

Jesus College, Oxford

2016

Abstract

This thesis considers various questions about acceptability judgement tasks (AJTs).

In Chapter 1, we compare the prevalent informal method of syntactic enquiry, researcher introspection, to formal judgement tasks. We randomly sample 200 sentences from Linguistic Inquiry and then compare the original author judgements to online AJT ratings. Sprouse et al., 2013, provided a similar comparison, but they limited their analysis to the comparison of sentence pairs and to extreme cases. We think a comparison at large, i.e. involving all items, is more sensible. We find only a moderate match between informal author judgements and formal online ratings and argue that the formal judgements are more reliable than the informal judgements. Further, the fact that many syntactic theories rely on questionable informal data calls the adequacy of those theories into question.

In Chapter 2, we test whether ratings for constructions from spoken language and constructions from written language differ if presented as speech vs as text and if presented informally vs formally. We analyse the results with an LME model and find that neither mode of presentation nor formality are significant factors. Our results suggest that a speaker's grammatical intuition is fairly robust.

In Chapter 3, we quantitatively compare regular AJT data to their Z-scores and ranked data. For our analysis, we test resampled data for significant differences in statistical power. We find that Z-scores and ranked data are more powerful than raw data across most common measurement methods.

Chapter 4 examines issues surrounding a common similarity test, the TOST. It has long been unclear how to set its controlling parameter δ. Based on data simulations, we outline a way to objectively set δ. Further results suggest that our guidelines hold for any kind of data.

The thesis concludes with an appendix on non-cooperative participants in AJTs.

A Note to the Reader

The target audience for the present thesis are syntacticians, non-experimental and experimental alike. However, for non-experimental syntacticians, Chapter 1 will be most appealing. In Chapter 1, we make a case for acceptability judgement tasks (and thus for experimental methods), in contrast to researcher introspection. Chapter 1 should be of interest to experimental syntacticians, too; certainly as a reference justifying their experimental work.

The remainder of the thesis (i.e. Chapter 2 to Chapter 4 and Appendix 6) is aimed at experimentalists working with acceptability judgement tasks. The scope of Chapter 4 ("The TOST as a Method of Similarity Testing in Linguistics") is even more general; it started off as a project on judgement data, but we expanded it to cover any kind of data.

The thesis might also be of interest to linguists in other areas. This applies to semanticists, pragmaticists, and psycholinguists in particular, as they deal with questions similar to those posed in this thesis (e.g. formal vs informal experiments, procedures for analysing the results of acceptability judgment tasks, etc.). Non-linguists might also take some interest, as this thesis might give them an idea of the state of the field.

As mentioned in the abstract, the present thesis comprises four papers, mirrored by the first four chapters. Each chapter is a paper in its own right, i.e. each chapter has its own introduction, body, conclusion, and (where applicable) appendices. This way, the reader can *pick'n'mix* which papers he or she wishes to read. This level of accessibility causes some redundancies; a trade-off we hope the reader will understand and accept. Further, the work presented here is methodological and should be accessible irrespective of one's theoretical background.

All chapters involve a great deal of additional data (individual ratings, reaction

times, filler items, audio recordings, the source code of our analyses, the source code of our resampling procedures, further data sets, etc.). This is more than the appendices could carry, which is why we made many additional materials available online (tsjuzek.com/thesis.html).

Lastly, a technical note: We use "we" throughout the thesis. Chapter 1 and Chapter 4 are based on joint work, so "we" is adequate. In the remainder we decided to adhere to the "we" for stylistic reasons.

Acknowledgements

I am extremely grateful to my supervisors Mary Dalrymple and Greg Kochanski for their continuous guidance, support, and advice, which extend well beyond this thesis. My thanks also go to Google for allowing Greg Kochanski to supervise me in the manner he has done. Further, I am extremely grateful to Tom Wasow for his guidance during the academic year 2012/13 and his comments on parts of this thesis. I would also like to thank Ulrich Schade and his team for fostering my interest in computational linguistics. And special thanks go to my collaborators Jana Häussler of the University of Potsdam and Johannes Kizach of Aarhus University, both of whom it has been a pleasure to work with.

I would also like to thank my examiners Ash Asudeh and Gisbert Fanselow for their thorough and constructive feedback. The following people have also made valuable comments on this thesis (or parts thereof): John Coleman, John Hainsworth, William Snyder, and two Language and Speech reviewers. I have received valuable feedback from the participants of my SPLaT talk in February 2013, from the participants of the workshop "Understanding Acceptability Judgments" at the University of Potsdam, and from many participants of the CLS 51 talk that Jana Häussler and I gave.

The support of the FES and the AHRC is gratefully acknowledged. The responsibility for any remaining errors is solely my own.

Declaration of Authorship

The present thesis was written by myself and the work contained therein is my own work, unless explicitly stated otherwise in the text.

Chapter 1 ("Lab or Armchair? The Benefits of Formal Acceptability Judgements") is based on collaborative work with Dr Jana Häussler of the University of Potsdam, but it is entirely written by me. Dr Häussler and I intend to independently publish our joint work in the near future.

Chapter 4 ("The TOST as a Method of Similarity Testing in Linguistics") is based on collaborative work with Dr Johannes Kizach of Aarhus University, but it is entirely written by me. Dr Kizach and I intend to independently publish our joint work in the near future.

.....

(Tom S Juzek)

Contents

1 Lab or Armchair? 1

1.1 Introduction 1

1.2 Further Background 3

1.2.1 Defining Introspection 4

1.2.2 Issues with Introspection 9

1.2.3 Sources of Error: Severity and Solutions 16

1.2.4 Comparing Informal and Formal Results 22

1.2.5 Sprouse et al. (2013) 23

1.3 Our LI Corpus 37

1.4 The Experiment 45

1.4.1 Experimental Design 46

1.4.2 Analyses 57

1.4.3 Ratings 62

1.4.4 Results 63

1.5 Discussion 65

1.6 Conclusion 71

1.7 Chapter 1: Appendices 73

1.7.1 Background of the Authors 73

1.7.2 Experimental Stimuli 90

2 Textual and Auditory Stimuli in Acceptability Judgement Tasks 101

2.1 Introduction 101

2.1.1 Investigating Modality 103

2.2 Three Experiments 109

2.2.1 Experiment 1: The Original Experiment 110

2.2.2 Experiment 2: Timed versus Untimed Questionnaire 120

2.2.3 Experiment 3: Informal versus Formal 123

2.3 Statistical Analysis: Linear Mixed Effect Model 136

2.4 Discussion 140

2.5 Conclusion 145

2.6 Chapter 2: Appendices 147

2.6.1 Experimental Stimuli 147

2.6.2 R Output 149

3 Comparing Data Transformations for Syntactic Judgement Data 155

3.1 Introduction 155

3.2 Further Background 157

3.2.1 Making Methodological Choices for Acceptability Judgement Tasks 157

3.2.2 The Use of Z-Scores in Linguistics 159

3.2.3 Measurement Methods 161

3.2.4 Common Data Transformations 166

3.3 Practical Considerations: A Study 172

3.3.1 Experimental Design 173

3.3.2 Statistical Analysis 184

3.3.3 Ratings 187

3.3.4 The Results for the Resampled Data 191

3.4 Discussion 196

3.5 Conclusion 197

3.6 Chapter 3: Appendices 199

3.6.1 Experimental Stimuli 199

3.6.2 R Output 201

4 The TOST as a Method of Similarity Testing in Linguistics 205

4.1 Introduction 205

4.2 How The TOST Works 207

4.3 Using Data Simulations To Observe δ . . . 213

4.4 Calibration Phase: From Observing to Predicting δ . . . 225

4.5 Validation Phase . . . 229

4.6 TOSTs vs t-Tests . . . 231

4.7 Conclusion . . . 233

5 Concluding Remarks 235

6 Appendix: Detecting Non-Cooperative Participants in Acceptability Judgement Tasks 239

6.1 Introduction . . . 239

6.2 Further Background . . . 240

6.2.1 Core Concepts . . . 240

6.2.2 Prevalence, Demographics, and Impact . . . 241

6.2.3 Types of Non-Cooperative Behaviour . . . 245

6.3 Prevention, Discouragement, and Detection . . . 247

6.4 Concluding Remarks . . . 253

References 255

Chapter 1

Lab or Armchair?

The Benefits of Formal Acceptability Judgements

1.1 Introduction[1]

For decades, informal methods of obtaining acceptability judgements dominated syntactic theory, with researcher introspection being the most common informal method (in researcher introspection, the investigating linguist is his/her own informant). Particularly in the past 20 years, the dominance of informal methods has caused an ongoing debate about the empirical foundations of syntactic theory (among others, cf. Bard et al. 1996, Schütze, 1996, Edelman and Christiansen, 2003, den Dikken et al., 2007, Culicover and Jackendoff, 2010). Some researchers have defended researcher introspection on the grounds that the method has proven itself and that there are no reasons to assume that formal methods give better results (e.g. Phillips and Lasnik, 2003, Bornkessel-Schlesewsky and Schlesewsky, 2007,

[1]This chapter is based on joint work with Jana Häussler.

Grewendorf, 2007, Phillips, 2010, Sprouse and Almeida, 2012, Sprouse and Almeida, 2013, Sprouse et al., 2013), while others have voiced their concerns about the reliability of informal results (e.g. Wasow and Arnold, 2005, Featherston, 2007, Gibson and Fedorenko, 2010, Gibson and Fedorenko, 2013, Gibson et al., 2013).

Informal methods typically involve only a few participants (in the extreme case only one, as with researcher introspection) and often do not adhere to experimental standards, such as several lexicalisations per construction, randomisation, and concealing the study's purpose from the informants. The lack of such standards amplifies common sources of error: scale biases (due to the lack of a common scale), judgement errors, quantisation errors, and purpose biases (in case of researcher introspection, this might even become a "conflict of interest", as the investigating researcher may be attached to a certain theory). As we argue below, these sources of error cause the results from researcher introspection to be less reliable than results obtained by formal methods (primarily acceptability judgement tasks that adhere to experimental standards). Schütze (1996) provided strong theoretical arguments why this line of argumentation applied to linguistic data, as well. Although the field witnessed an increase in the use of formal methods since Schütze (1996) and Cowart (1997), data collection remains predominantly informal.

A possible reason for this over-reliance on informal methods might have been that a quantitative comparison of informal and formal acceptability judgements was long missing. Sprouse et al. (2013) filled this gap and presented quantitative results that suggest that informal and formal judgements agree to a large extent. However, Sprouse et al. focused on pairwise comparisons, i.e. they compared marked constructions to their unmarked counterparts (which are typically minimal pairs) and checked *for each pair* whether informal and formal results agreed in direction. Below, we argue that such an analysis is ineffective in detecting the potentially damaging effects of the aforementioned sources of error. Further, Sprouse et al.'s analysis

relies on the assumption that syntactic research is restricted to the comparison of sentence pairs. Although pairs play an important role, pairwise comparisons do not reflect best how syntactic research is conducted, nor how it should be conducted (cf. Section 1.2.5). A comparison of informal and formal methods should take this fact into account and should be able to detect the effects of the mentioned sources of error. Our aim is to provide exactly such a comparison. Consequently, we designed an experiment in which we randomly sampled constructions from the literature and then compared them at large (i.e. beyond pairs).

OVERVIEW Section 1.2 provides further background. We look in-depth at the key concept of introspection (1.2.1), at the sources of error affecting introspection, the severity of those factors and possible solutions to them (1.2.3). In Section 1.2.4, we ask how informal and formal results should compare based on theoretical arguments. Sprouse et al. (2013) provide a quantitative comparison and we discuss why their methodology is insufficient to detect the damaging effects of those sources of error (1.2.5). Our overall aim is to design an experiment that compares researcher introspection to experimental results, based on randomly sampled items. To do so, we need a corpus to sample from. This corpus is presented in Section 1.3. The actual experiment can be found in Section 1.4. In Section 1.4.1, we discuss the experimental design, Section 1.4.2 outlines our analyses, concrete hypotheses, and criteria to reject the null hypothesis. The experimental ratings are presented in Section 1.4.3, followed by the results of the analyses (1.4.4). Section 1.5 provides a detailed discussion of the results; this discussion is concluded in Section 1.6.

1.2 Further Background

In this section, we provide some further background on the core concepts of this chapter. In Section 1.2.1, we define introspection, explore the object of introspection,

ask whether some observers are better at it than others, and discuss the difference between acceptability and grammaticality. In Section 1.2.2, we discuss issues with introspection. Section 1.2.2 begins with an analogy between making acceptability judgements and nerf gun shooting, which we use throughout Section 1.2. Then, we examine the main sources of error that can occur when making acceptability judgements: judgement errors, quantisation errors, purpose biases, and scale biases. The severity of and possible solutions to those sources of error are discussed in Section 1.2.3. In Section 1.2.4, we look at how the use of informal vs formal methods might lead to differences in results. This is needed for the discussion of Sprouse et al. (2013) in Section 1.2.5, where we look into their methodology. We ask whether or not syntactic enquiry is restricted to the examination of sentence pairs. This is important in understanding the limitations of Sprouse et al.'s analysis. Section 1.2.5 concludes by outlining further issues with Sprouse et al. (2013).

1.2.1 Defining Introspection

The linguistic use of "introspection" subordinates to the wider use of the word in psychology. From James (1890; through Boring, 1953):

> The word introspection needs hardly to be defined – it means, of course, looking into our own minds and reporting what we there discover.

As such, the nature of introspection and its benefits and limitations are also subject to debate in psychology (among others, c.f. James, 1890, Dodge, 1912, Titchener, 1912, Boring, 1953, Lyons, 1988), philosophy (e.g. Nagel, 1974, Churchland, 1985, Jackson, 1986), and other fields. In the context of grammar theory, "introspection" typically refers to the situation when the investigating linguist is his/her own informant, i.e. he/she judges the acceptability of a syntactic construction him/herself.

However, to call only this practice "introspection" can be confusing, because participants of a formal acceptability judgement task rely on their introspection, as well: Each participant looks into his/her mind and observes his/her reaction when facing a certain sentence. So, to avoid confusion, we refer to the practice of a linguist being his/her own informant as "researcher introspection" (following the usage in consumer research, cf. e.g. Wallendorf and Brucks, 1993).

Below, we consider researcher introspection an "informal method" and we refer to its results as "informal results". It is informal, because it does not adhere to common experimental standards: The number of participants is very low, there is no variation in lexicalisations per construction, there is no randomisation, conditions are badly distributed such that participants might see an item in multiple conditions, and the purpose of the study is not concealed (cf. e.g. Cowart, 1997, or Gibson and Fedorenko, 2013, for a discussion of these standards). There are other informal practices apart from researcher introspection, e.g. when the linguist asks his/her colleagues for their intuitions on a certain phenomenon. This is different to researcher introspection, since the investigating linguist relies on other people's (introspective) judgements. In a way, this is an informal survey or experiment.

The situation where a researcher collects judgements from a pool of participants and adheres to experimental standards is what we refer to as a "formal acceptability judgement task". Of course, each participant in such an experiment makes his/her judgements introspectively. This is why when we speak of just "introspection", we strictly refer to both, the introspective process of a linguist and the introspective process of participants in an experimental setting.

There are two important questions about the nature of syntactic introspection: 1) What does the introspector observe? And: 2) Do linguists and naive speakers observe the same thing and if so, do they do so the same way?

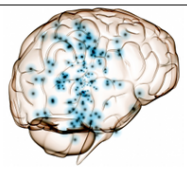
Option	a)	b)	c)
Input	*What did who buy?*	*What did who buy?*	*What did who buy?*
Object of Study	 *Intuitive Reaction*	 *Mental Processes*	 *Grammatical Model*
Output	*Observation* of the Intuitive Judgement	*Explanation* of the Intuitive Judgement	*Prediction* of the Intuitive Judgement

Figure 1.1: The different understandings of what the object of introspection could be, as specified above. (Using images by Jayel Aheram (the sad face), the Massachusetts General Hospital and Draper Labs (the brain), and Daniel Low (the book), all distributed under a CC-BY license.)

THE OBJECT OF INTROSPECTION Facing a certain syntactic stimulus, there are three broad possibilities as to what one might understand as the object of introspection[2]: a) One observes one's reaction to facing a certain syntactic stimulus. Conceptually, the reaction is completed before observation; this is almost an observation at the neuromuscular level. b) One observes one's mental processes when facing a certain syntactic stimulus. This is an observation of earlier stages in the reaction and such observation could help explain how one's reaction to the stimulus came about. c) One simulates the processes of a hypothesised grammatical formalism, theory, etc., and its outcome. Conceptually, c) is more abstract; it should be output-equivalent to a) and b), but the processes in c) can be detached from the subject's thought processes. Figure 1.1 illustrates those three possibilities.

Options b) and c) should not be viewed as the object of introspection. The third option, c), is not an introspective process at all. Assume that we maintain a certain theoretical model of how syntactic knowledge is represented in the mind and call this model e.g. the "V6IA2 Model". Any observation of the workings of the V6IA2

[2]We present this grouping in order to clarify our use of "acceptability" and "grammaticality".

Model when processing a certain stimulus (e.g. a sentence that violates a weak-island constraint) is not an act of introspection, nor is the observation of the outcome of such processing. We might even use a machine to perform this task. Of course, if we knew that the V6IA2 Model was an accurate representation of the syntactic processes in the mind, then observing how the V6IA2 Model processes inputs would be close to observing one's mental processes. However, it would still not be an observation of the mind itself.

As to b): There's no reason to believe that we can observe our mental processes accurately (and even if we could observe them, we would be observing them with other mental processes that might well affect the first set of processes). On seeing a weak-island violation, one does not observe the mental operations that are at work in case of such a violation. If we were able to observe such operations, the field would be much closer to a theory that explains how grammatical knowledge is represented in the mind. This is not the case; grammatical knowledge is *tacit* (cf. Chomsky, 1965:19) and the mental processes involved still elude syntacticians. This is why, when it comes to syntactic processes, the mind is best compared to a black box.[3]

This leaves option a): One observes one's own reaction to facing a certain syntactic stimulus. This reaction can also be called a "grammatical intuition" or, as it is commonly referred to, an "acceptability judgement"; and it has merely an observational character (i.e. it does not explain nor predict grammatical intuitions), it is merely a first reaction to a stimulus.

ARE THERE EXPERT OBSERVERS? This concerns the second question above (do linguists and naive participants observe the same thing?). Culbertson and Gross (2009) attempt to answer this question: Culbertson and Gross did not

[3]This is not to say that mental processes cannot be observed at all. Cognitive neuroscientists are slowly opening up the black box, using cutting edge technology. Our point is that this cannot be done introspectively.

find any differences in ratings between linguists who were familiar with experimental tasks and syntactically naive participants who were equally familiar with the task. This is a consequence of the above: If we think of the mind as a black box, then there is no reason why syntacticians should somehow have privileged access to the workings of the mind. To the best of our knowledge, there is no research that shows that the mind is less of a black box for syntacticians than it is for naive speakers.

One might object: In contrast to a naive speaker, syntacticians have access to explicit grammatical knowledge. For instance, linguists know that it is somewhat acceptable to extract a WH-argument out of a weak-island (e.g. "Who do you wonder whether we will meet at the opera?", from Haumann, 1997:177), but not of a WH-adjunct (e.g. "Why do you wonder whether Bill will quit his job", with extraction from the subordinate clause, ibid.), as Ross (1967) pointed out.

However, such explicit knowledge should not be confounded with the intuitive reaction one gets upon facing a certain syntactic stimulus. For instance, when faced with a violation of a strong-island constraint, explicit knowledge is not required for the processing of the construction and for having the intuition of "badness" of such a construction. Moreover, for a naive speaker, there is no easy way to explicate the tacit syntactic knowledge that is part of the mental processes that are involved for certain constructions.

ACCEPTABILITY VS GRAMMATICALITY Keeping these thoughts about introspection in mind, it is useful to distinguish (grammatical/syntactic) acceptability from grammaticality. In the following, when we use "acceptability", we refer to the intuitive response to a stimulus, as part of introspective considerations. We treat "grammaticality", on the other hand, as a technical notion, which we use to refer to the output of a syntactic theory or model. For instance, if an evaluated sentence is "good" or "well-formed" according to HPSG (Pollard and Sag, 1994),

then that sentence is "grammatical"; if it is "bad" or "ill-formed", then we refer to it as "ungrammatical". So, in this sense, checking for the grammaticality of a sentence equates to option c) in Figure 1.1.[4]

Acceptability and grammaticality are strongly connected, because grammars are constructed to reflect the human language behaviour of a speaker community. There are differences between the two, though, as exemplified by (1) (from Bever, 1970:316). Syntacticians would hold that (1) is grammatical. However, due to processing difficulties, (1) might be seen as unacceptable by some native speakers.[5]

(1) The horse raced past the barn fell.

That is, if one accepts the Chomskian distinction between competence and performance, then grammaticality can be conceived as being free of any performance-"noise" (memory limitations, priming effects, etc.), while acceptability is subject to such noise. (Among others, Fanselow and Frisch (2006) look into various processing factors that can cause such mismatches as in (1).)

1.2.2 Issues with Introspection

There are a variety of issues that come with introspective judgements. These include: judgement errors, quantisation errors, purpose biases, scale biases, and small differences in grammars. Apart from the differences in grammars, these issues are part of classical test theory and have been discussed in the literature as such (e.g. Winer et al., 1971, or Myers, 1972; both from psychology). Further, these issues,

[4]Mental realists would claim that checking the grammaticality of a sentence is best reflected by option b); however, for technological reasons, there is little hope of complete or accurate observation.

[5]In order to get to the grammaticality of (1), a syntactician would have to look at various, similar sentences, such as "the horse fell", "the horse that was raced past the barn fell", "the tractor left behind the barn disappeared", "the tractor that was left behind the barn disappeared", etc.

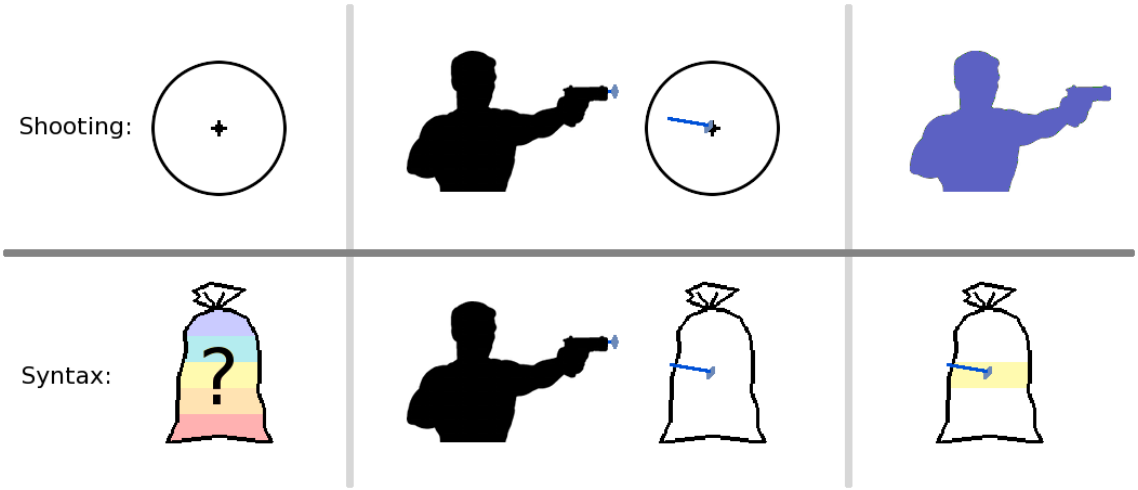


Figure 1.2: Illustrating the analogy between shooting (top row) and researcher introspection (bottom row). Top left: In shooting, we know the target (black cross). Top middle: The shooter takes a shot (blue nerf dart) at the target. Top right: From the outcome, we can draw conclusions about the shooter. In this example, we can conclude that the shooter has a very good aim. Bottom left: In "introspective shooting", different areas of the target represent different degrees of "goodness". In our examples, we mainly use a 5-point scale, where "blue" means "good/natural/acceptable", "red" means "bad/unnatural/unacceptable", and the other colours represent values in between. Bottom middle: We ask "introspective shooters" to take shots for a certain grammatical construction. Bottom right: From the outcome, we learn where the shooter thinks the targets for different stimuli are.

bar quantisation errors, are also part of the linguistic discourse (for excellent introductions, see Schütze, 1996, and Cowart, 1997). As those sources of error are important for our purposes, we discuss them here in greater detail.

ANALOGY: INTROSPECTION AND SHOOTING In order to illustrate the nature of the different issues, we employ an analogy in which we compare the process of introspection to nerf gun shooting at sandbags. Assume that judging syntactic constructions introspectively is like taking shots at a sandbag with a nerf gun. Different areas of the sandbag signify different degrees of "goodness" and different constructions will have their target points in different places of the sandbag (see Figure 1.2; this analogy is similar to the archery analogy in Gonick and Smith, 1993).

Figure 1.3: Illustrating judgement errors, using the shooting analogy. In an ideal situation, one shot would be sufficient (left; single shot in red). However, a subject's aim is not 100% steady, due to judgement errors (and the other factors discussed below, which we are dismissing for the moment). If a single subject took numerous shots at the same target, then the hits will most likely show some spread (right; other shots in blue). In this example, the original (red) shot is probably a little bit of an outlier.

JUDGEMENT ERRORS Whenever a speaker, whether trained or naive, makes a judgement, this judgement is subject to noise (again, in Chomskian terms this could be thought of as performance-noise). This noise consists of, but is not limited to, memory limitations, possibly time pressure, attention span issues, distractors, priming effects, etc. As a consequence, if the same stimulus is given to the same subject several times, then the ratings will most likely come out different. Using the image of shooting again, Figure 1.3 illustrates this.

QUANTISATION ERRORS Participants are also subject to quantisation errors. These can be thought of as "rescaling" errors: A subject is given a scale that has a different number of degrees than his/her "inherent" scale, which forces the subject to choose the degree that is an approximation to what he/she feels best corresponds to his/her "inherent" scale. The rescaled rating would then not completely reflect the mental reality of the subject. For instance, assume that a subject faces a certain stimulus and that he/she would have rated that stimulus with a "6" on the 7-point scale that he/she prefers. However, the researcher forced the subject to use a 5-point scale. On this 5-point scale, the subject might feel that there is no category that represents his/her rating well, but he/she is forced to give a rating anyway. The subject might give e.g. a "4" or "5", but neither rating would

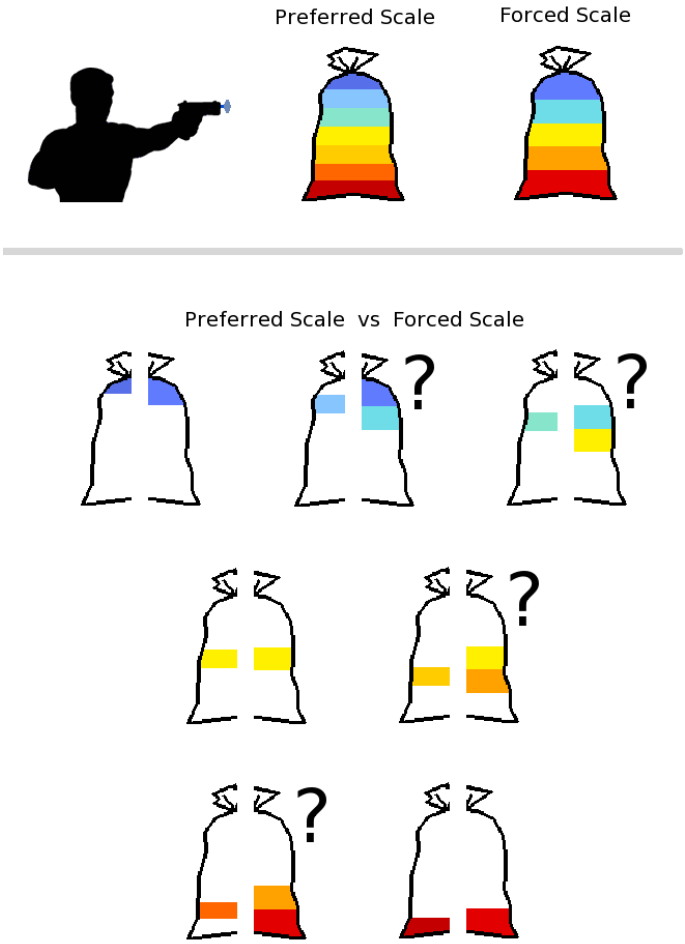


Figure 1.4: Illustrating quantisation errors, using the shooting analogy from above. Assume that a subject prefers a 7-point scale, but the researcher forces the subject to use a 5-point scale. As shown on the bottom part of the figure, for many degrees of his/her "inherent" scale (left half of a sandbag), it is not clear how they correspond to the degrees of the forced scale (right half of a sandbag). For instance, if the subject would have given a rating of "6" (out of 7) on his/her preferred scale, then it is not clear how such a rating is best expressed on the forced scale.

Figure 1.5: Illustrating purpose biases caused by knowledge of the purpose of a study. Using the shooting analogy from above, the target on the left shows where a hypothetical subject would have aimed at, if he/she had been oblivious about the purpose of the study ("Unbiased"). Before shooting at the target on the right, the subject has learnt about the purpose of the study and he/she changed his/her aim slightly, possibly to accommodate the researcher ("Biased").

be a precise representation of what the subject "had in mind" (cf. Figure 1.4).

PURPOSE BIASES Ideally, subjects are oblivious to the aim of the study. If a subject knows about the purpose of a study, there is a real danger that he/she may alter his/her ratings in order to accommodate the researcher. Such accommodation effects concern most psychological experiments (cf. e.g. Kelman, 1967, or Korn, 1997), including acceptability judgement tasks (cf. e.g. Cowart, 1997:87, Gibson and Fedorenko, 2013:89). Figure 1.5 illustrates accommodation effects. In the case of informal judgement tasks, the investigating linguist knows about the purpose of the study by definition: He/She has designed that very study. This can lead to similar biases.

SCALE BIASES Even if two subjects use the same scale (e.g. a 5-point scale), the meaning of any given category ("1", "2", "3", "4", 5" or "*", "?*", "??", "?", OK/unmarked) can vary from subject to subject. For instance, for one subject, the end-points of the scale might be the dominant categories and the middle value is rarely used; while for another subject, it is the very opposite. And a third

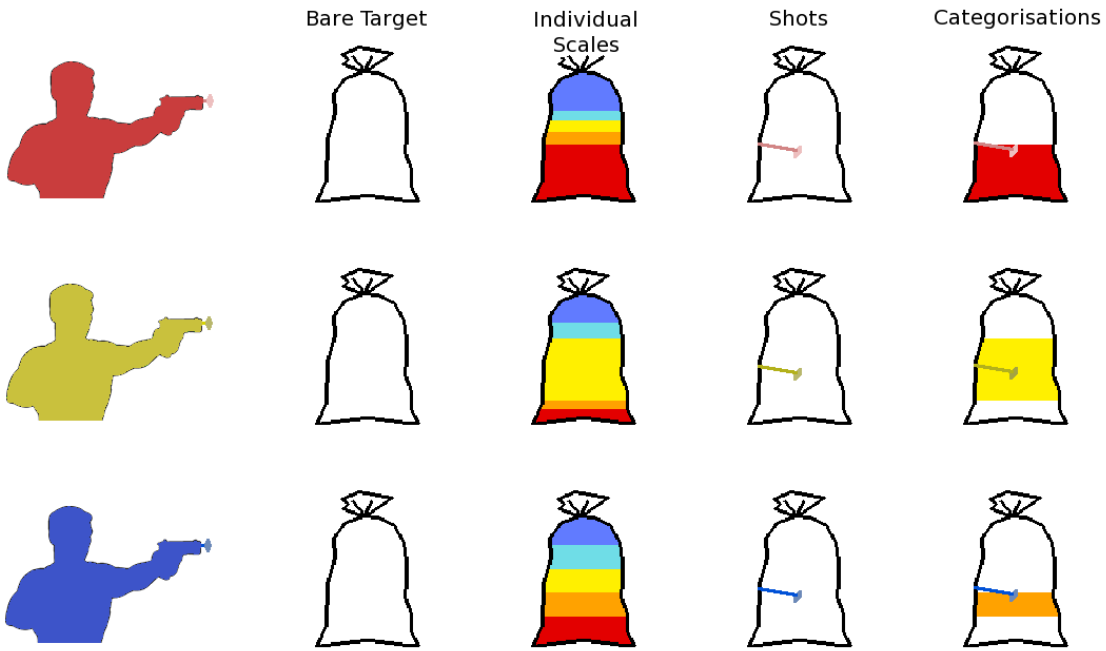


Figure 1.6: Illustrating scale biases, using the shooting image from above. Scale biases occur when the categories of the scale mean different things to different subjects. Rows: Different subjects, marked by different colours. The bare target looks the same for all three subjects ("Bare Target"). However, the subjects have a different understanding of the scale ("Individual Scales"). Although the shots by our three subjects look the same at first glance, i.e. the dart hit somewhat below the centre, the ratings mean something different to each subject (illustrated by the categorisations on the very right).

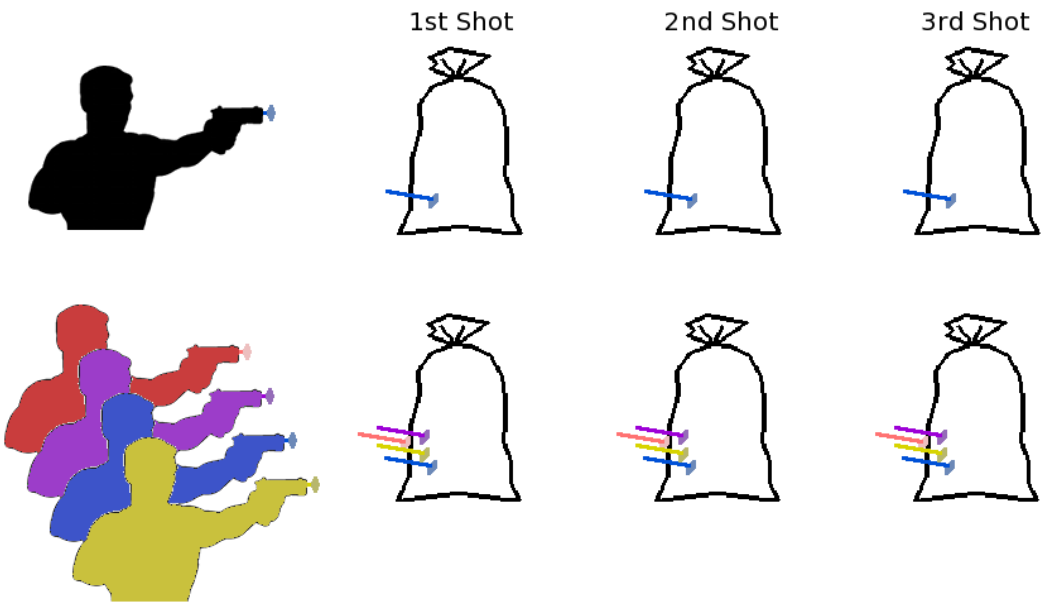


Figure 1.7: Illustrating differences in grammars, using the shooting analogy from above. Upper row: A single subject. Lower Row: Four different subjects. Columns: Number of trials. If there was no "noise" due to judgement errors, etc., then a subject could take multiple "shots" and the results would always be the same (upper row). However, even in such an ideal scenario, we would most likely observe differences between subjects (different subjects are marked by different colours), due to small differences in their grammars (lower row).

subject might have a "neutral" scale: All categories are equally spread out. This is illustrated in Figure 1.6. Also, categories can be "shifted", i.e. the subjective middle value and the centre (i.e. the objective middle point of the sandbags) may not match. This, too, is illustrated in Figure 1.6: Subject 1 (upper row) has a small shift "upwards" and Subject 2 (middle row) a small shift "downwards". Subject 3 (bottom row), on the other hand, has a perfectly neutral scale. A strong scale bias can also be the reason why a subject leaves parts of the scale unused. Scale biases can hamper comparisons across subjects, particularly when only a few ratings are available (because one needs a few measurements common to all subjects in order to estimate scale biases).

DIFFERENCES IN GRAMMARS Even if all subjects in a formal ac-

ceptability judgement task were completely immune to performance-noise (memory limitations, etc.) and other sources of error, one might still observe differences in ratings. Such differences would reflect the fact that speakers of the same language can still have slightly different grammars, due to dialectal differences or due to diachronic language change across generations (see Wardhaugh and Fuller, 1986, or Meyerhoff, 2006, for an introduction to sociolinguistics, including dialectal and regional differences).

The "needs" construction in (2) exemplifies dialectal differences: While for most speakers of Standard American English (2) is marked, speakers from the northern midlands region, notably western Pennsylvania, find it acceptable (cf. Murray, Frazier, and Simon, 1996). Figure 1.7 illustrates the effect of differences in grammars.

(2) The laundry needs washed.

Of course, dialectal differences and differences due to language change are different from the sources of error mentioned above: Linguists do not wish to eradicate such differences, but they wish to explore them and to account for them.

However, for our purposes, dialectal differences and differences due to language change do not play a central role. We wish to compare informal and formal methods and when doing so, we only consider constructions that belong to Standard American English. This way, we were able to sample our subjects from all dialectal regions in the US. (Nonetheless, we indirectly check for dialectal differences in Section 1.4.4.)

1.2.3 Sources of Error: Severity and Solutions

With respect to the issues presented above, two questions arise: 1) How severe are those issues? 2) How can their effects be reduced?

SEVERITY OF THE ISSUES How strong is the impact of scale biases, judgement errors, quantisation errors, inter-speaker grammar differences, and purpose biases on the results? To the best of our knowledge, the impact of these factors has not yet been quantified for linguistic data. However, based on data from a previous experiment (N=66, 2376 data points), we can make the following observation: On a 7-point scale, participants diverged on average by 0.64 points (10.7%) from the mean when rating a construction multiple times (in our case four times). The number, though, is drawn from only nine different constructions; so the 0.64 point divergence is just a rough estimate.

REDUCING THE IMPACT The impact of the different factors (scale biases, judgement errors, quantisation errors, and purpose biases) can be reduced. For formal acceptability judgement tasks, there are two strategies: several specific solutions (separate "fixes" for each confounding factor) and a general solution (one fix for all sources of error). Our focus is on the general solution, because discussing the different specific solutions would be too much of a sidetrack and we can only briefly hint at some of them.

Specific strategies to mitigate the effects of the above mentioned factors on linguistic data have been discussed in the literature and Cowart (1997) is an excellent starting point for linguistic questions. The impact of most factors contributing to judgement errors can be reduced by careful experimental design (e.g. by controlling the test environment; eliminating priming effects by fully randomising the stimuli, etc.). Quantisation errors can be reduced by letting a subject choose his/her preferred scale.[6] Scale biases can be reduced by a well-balanced set of experimental items and by normalising the ratings. Accommodation effects can be reduced by making the purpose of the study unobvious.

[6]N.B.: If a scale is forced on a subject, then errors due to quantisation will be indistinguishable from small differences in grammars across subjects.

For our purposes, the general solution is more important. The impact of scale biases and judgement errors can be mitigated by one relatively simple measure: Increasing the number of subjects. By having a larger sample, the noise created by those two factors is usually offset.[7] Figures 1.8 to 1.10 illustrate this and Equation 1.1 expresses this (where ϵ stands for any kind of error with zero mean and finite variance and σ is the sample's standard deviation).

$$\epsilon = \frac{\sigma}{\sqrt{N}} \tag{1.1}$$

Increasing the number of subjects is not only a strategy to mitigate the above mentioned sources of error, it also enables the researcher to employ common statistical tests and to test for significance, as e.g. Wasow and Arnold (2005) and Gibson and Fedorenko (2013) point out. Being able to do so is a requirement in other scientific fields and there is no reason why linguistics should be an exception.

When it comes to mitigating the effects of the discussed sources of error the situation is more difficult for the informal methods. We have to distinguish between researcher introspection and other informal methods, e.g. an informal judgement task.

For researcher introspection, the specific strategies hardly apply. At best, the researcher could add some "balancing items" and then randomise the stimuli. This might have a bit of a calibration effect and help mitigate a scale bias to some extent. Practically, however, it is not possible to apply data normalisations, because authors typically provide too few judgments and the data sets are not comparable. And with respect to purpose biases, these cannot be offset at all: If the designer of a study participates in it, then his/her knowledge of the purpose of the study cannot be undone. Further, the general strategy of increasing the number of subjects is

[7] Unfortunately, quantisation errors are not reduced by this measure: Whether an experiment has one or thirty subjects, quantisation errors will occur to the same extent.

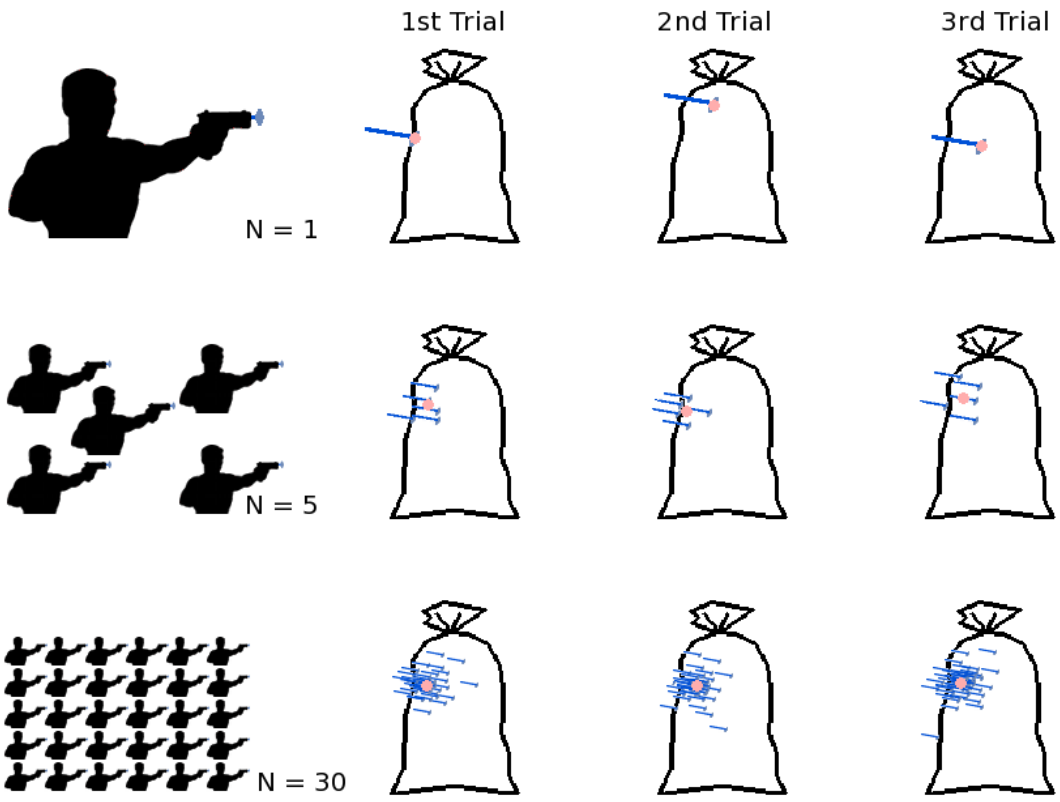


Figure 1.8: Illustrating how increasing the number of subjects in a formal acceptability judgement task reduces the effect of judgement errors, using the shooting analogy from above. Rows: Number of subjects. Columns: Number of trial. Each small dart (in blue) represents a single judgement by the different subjects. The averaged rating is in bright red. With an increasing number of subjects, the averaged ratings become more and more reliable throughout the trials.

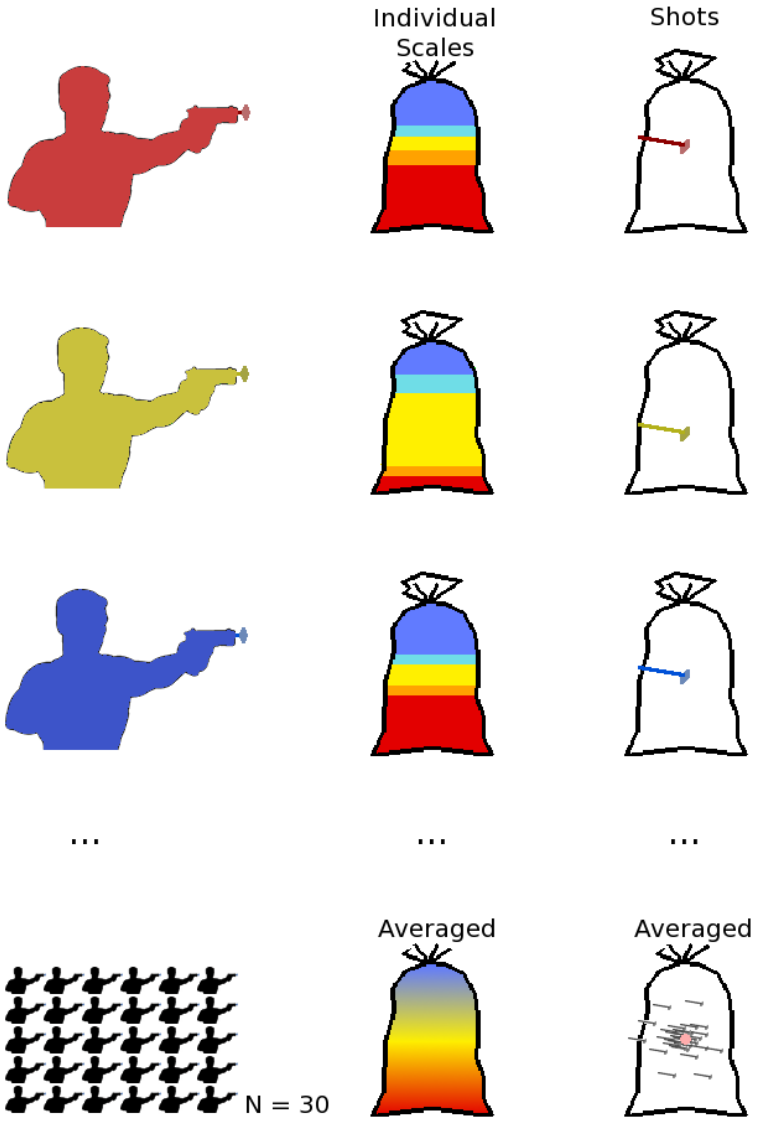


Figure 1.9: Illustrating how increasing the number of subjects in a formal acceptability judgement task reduces the effect of scale biases, even though we only have a single judgement per subject. This is achieved by getting the "averaged scale" for the sampled subjects. Row 1 to 3: Individual subjects, marked by different colours. Columns for row 1 to 3: Individual scales and shots. The different subjects have different understandings of the scale. Each subject takes a single shot (i.e. judgement) at their personal middle value (in yellow). Due to scale biases, the precise spot where the subjects' darts hit look slightly different to an outside observer (second column). However, if this is repeated for a pool of subjects (row 4), then the averaged rating (in red) should offset some of the effects of scale biases (columns for row 4), despite the fact that we did not normalise per subject. This is because we get an "averaged scale" for the sampled subjects. If we were to then compare averaged ratings from two experiments, they would probably come out as similar, as both ratings lie on similar averaged scales. It would be better, though, to collect several ratings per subject and to then normalise per subject (cf. Figure 1.10).

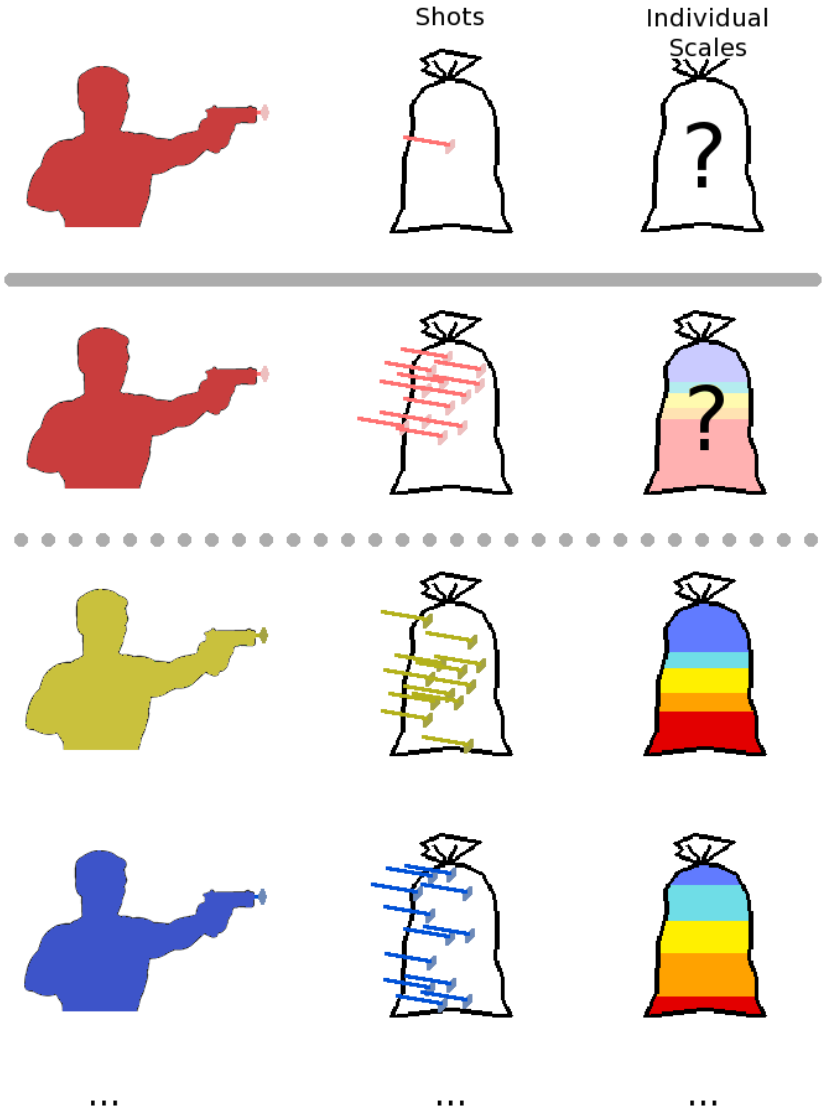


Figure 1.10: Another strategy to mitigate the effects of scale biases is to normalise the ratings for each subject. However, this requires several subjects and several judgements by each subject. To illustrate why, we use the shooting analogy from above. Rows: Subjects. Columns: Shots taken (left) and inferred scale (right). In the top row, the red subject only takes a single shot. From this, one cannot make any conclusive inferences about the subject's understanding about the scale. Even if we collect several judgements on various sentences by (only) that subject, we cannot be sure that the inferred scale is correct (right column, second row from the top). In this example, it could be the case that the red subject tends to not use the very bottom of the target. However, it could also be the case that the study was missing truly bad constructions and that there was no reason to use the very bottom of the target and scale. This is why we need further subjects (in yellow and blue). We observe their behaviour and find that they do use the entire scale. Thus, there are a good reasons to believe that red's scale is biased. (N.B.: This example assumes that the participants have similar grammar.)

not available for researcher introspection as per our definition, which states that in researcher introspection, the linguist is his/her own informant.

To other informal methods, e.g. an informal judgement task, these strategies could be applied. There is a catch, though: By applying these strategies, the informal turns formal. For instance, asking ten of one's students or colleagues would be an informal judgement task. Adding randomisation of stimuli, inclusion of fillers, a consistent test environment, etc., would turn this into a formal experiment.[8] This is fine, but by doing so, one would also negate one of the main reasons to use informal methods in the first place: ease of use.

This also indicates that the distinction between informal and formal methods is a vague concept that allows for cases in-between. And even for researcher introspection, there are interesting borderline cases: For instance, are co-authored papers (with two or more authors/informants) different from single-author papers? We wish focus on the extreme case in which the linguist "designed" his/her study and is the only subject to take it. However, for practical reasons, we also consider sentences from co-authored papers in the study in Section 1.4. We also consider items that were judged by several linguists over the years; for details, see Section 1.3.

1.2.4 Comparing Informal and Formal Results

The effects of scale biases, judgement errors, biases due to knowing a study's purpose, and the low number of subjects in researcher introspection have been discussed and criticised in the linguistic literature; particularly, Schütze (1996) did this in great detail. From a theoretical point of view, results from informal methods will be less reliable than results from formal methods. The claim about reliability is a consequence of the effects illustrated in Figures 1.8 to 1.10. The point can be made

[8] However, such an experiment would still make use of "convenience sampling", as the researcher has asked subjects that were "conveniently" available to take part in his/her study.

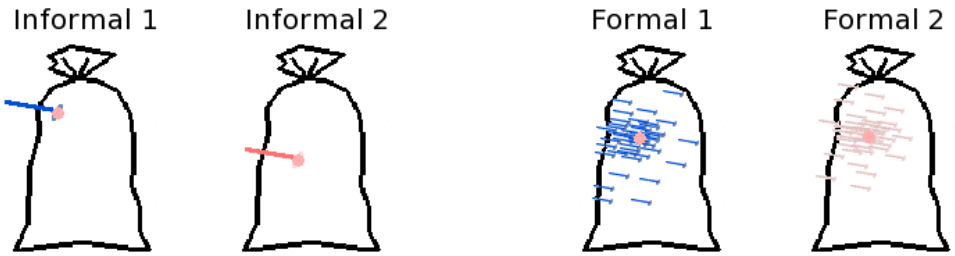


Figure 1.11: Using the shooting analogy from above, this figure illustrates how the results of two test-retest informal studies (left) are probably less reliable than of two test-retest formal studies (right). Ratings by individual subjects are the nerf darts in blue and red; the bright red dot indicates the averaged ratings.

even more sharply: Assume that two linguists were judging a certain construction introspectively ("Informal 1" and "Informal 2" in Figure 1.11). It is not too unlikely that the results will come out as somewhat different. However, if each of the two linguists conducted an experiment, asking e.g. 30 subjects to rate the same construction, then it is extremely likely that the results will be very similar (if the studies were methodologically sound; "Formal 1" and "Formal 2" in Figure 1.11).

From a practical point of view, there were several studies which compared the results from informal judgement tasks and from formal methods (e.g. Wasow and Arnold, 2005, and Gibson and Fedorenko, 2013). Those studies pointed out issues with results from informal methods and offered the same solution: Linguist should use formal methods. At the same time, many of those studies have been criticised for "merely" being case studies (Sprouse and Almeida, 2012, Sprouse and Almeida, 2013, Sprouse et al., 2013). A quantitative comparison of informal and formal acceptability judgements was long missing. Sprouse et al. (2013) filled this gap.

1.2.5 Sprouse et al. (2013)

Sprouse et al. (2013; henceforth *SSA*) quantitatively compared informal and formal judgements and their results suggest that the two judgement types concur to a

large extent. However, Sprouse et al. focused on pairwise comparisons, i.e. they compared marked constructions to their unmarked counterparts and checked for each pair whether informal and formal results agreed.

In the following, we briefly outline SSA's methodology and analysis and then, we will argue for two things: 1) Though pairs play an important role in syntactic research, pairwise comparisons do not reflect best how syntactic research is conducted. 2) Even if the results from informal methods are less reliable than results from formal methods, SSA's analysis would not have been able to detect this.

SSA'S METHODOLOGY SSA randomly selected 150 items that were introspectively marked (in most cases by a "*", but in a very few cases by "??", "?*", and "??(*)") from a pool of standard acceptability judgements, which SSA extracted from Linguistic Inquiry (henceforth *LI*) 2001 to 2010. For each introspectively marked item, they either extracted its unmarked counterpart (in a few cases, these included items marked with a "?") from LI or they created a suitable one themselves. For each pair, they created another seven similar pairs and they then tested those items in an online acceptability judgement task, using standard experimental methodology with respect to presentation of items, measurement methods used, etc.

SSA's main analysis was a test of directionality. Each informal pair from LI implied a certain direction of judgement. The marked item (henceforth **-item*) is worse than the unmarked item (henceforth *OK-item*) with respect to its acceptability. For the same pairs, SSA checked which direction the experimental ratings took. Would the introspectively marked item receive a lower or higher experimental rating than the unmarked item? For any given pair, if the direction for the experimental ratings was the same as for the ratings from researcher introspection, then there was a match in direction; otherwise, there was a mismatch. Applying this analysis to all pairs, SSA

found a match rate of about 95%. They concluded that syntactic theory based on informal acceptability judgements is empirically well-founded.

SENTENCE PAIRS However, SSA's analysis heavily relies on the assumption that syntactic research is restricted to the analysis of sentence pairs. It is true that the analysis of pairs plays an important role in syntactic research. However, we do not think that syntactic enquiry is *restricted* to it (and even if research in the field was restricted to only analysing sentence pairs, then it would be better to lift such a restriction). In the following, we present two arguments against SSA's assumption: First, restricting oneself to only pairs is cumbersome and leads to a severe logical conundrum and second, standard syntactic enquiry does already consider more than just pairs.

First, even in the discussion of constructions for which pairs could be sufficient, authors often consider further items, to put things into perspective. We think this is common practice. Consider the discussion of subject vs object control in (3) to (9). Syntacticians are interested in how examples of subject control, e.g. (3), compare to their baseline sentences, here (4), and to other variants like (5) and (6). But they are also interested in how sentences like (3) compare to other instances of subject control, e.g. (7). Further, it would be interesting to see how such examples of subject control compare to examples of object control, e.g. (8) and a possible baseline (9) (for more on subject and object control, cf. for instance Rosenbaum, 1967, Bowers, 1973, or Comrie, 1984).[9]

(6) should be rejected by most speakers. Under the intended subject control reading, (3) and (5) are accepted by many speakers; however, some speakers reject them (Comrie, 1985:55). (4), (7), (8), and (9) should be accepted.

[9]There are no diacritics in the following examples. The point we make is a methodological one and the grammatical status of (3) to (9) is secondary.

(3) Pete promised Sarah to get the mail.

(4) Pete promised Sarah that he will get the mail.

(5) Pete promised Sarah to shave himself.

(6) Pete promised Sarah to shave herself.

(7) Pete tried to get the mail.

(8) Pete asked Sarah to get the mail.

(9) Pete asked Sarah whether she could get the mail.

Of course, all these comparisons *could* be done in pairs (i.e. comparing (3) vs (4), (3) vs (5), (3) vs (6), etc.); but even for simple examples like this one, only doing pairwise comparisons is already quite cumbersome (as there are 21 unique pairs).

Further, even if we decide to stick to pairwise comparisons, then a real conundrum arises. Say we rated (4) > (6), (4) > (5), and (5) > (6). We should be able to form an ordered chain from these three comparisons: (4) > (5) > (6). Such a chain, however, is at odds with a strict notion of doing only pairwise comparisons. There are only two ways forward from here: 1) One strips "<" and ">" of their transitivity. This would be an absurd move. 2) Or one admits that there is no point in restricting oneself to only pairwise comparisons, as they can always be transformed into chains anyway.[10]

Second, often pairwise comparisons are not sufficient at all: In any discussion of weak/strong islands, weak/strong crossover effects, etc., linguists discuss at least three levels per definition. Consider the following quintuple in (10) to (14), which

[10]Many thanks to Greg Kochanski for pointing this argument out to us.

comes from Kluender (1992), cited in Hofmeister and Sag (2010). Kluender establishes the following ordering: (10) $\geq$ (11) $\geq$ (12) $\geq$ (13) $\geq$ (14), where "$\geq$" signifies "equally or more acceptable than".

(10) This is the paper that we really need to find someone who understands.

(11) This is the paper that we really need to find a linguist who understands.

(12) This is the paper that we really need to find the linguist who understands.

(13) This is the paper that we really need to find his advisor, who understands.

(14) This is the paper that we really need to find John, who understands.

(10) to (14) show gradience within a construction. A similar point can be made for weak island violations, e.g. (16), vs strong island violations, e.g. (17), and their baseline sentence (15) (from Szabolcsi, 2006; her diacritics).

(15) Which topic do you think that I talked about?

(16) ?*Which topic did John ask who was talking about?

(17) *Which topic did you leave because Mary talked about?

Such a use of gradience is an indicator that the field does not restrict itself to only pairwise comparisons. And the use of gradience is widespread in the field: The vast majority of papers in our corpus (see Section 1.3) involve more than two levels of acceptability.

Further, the fact that about 30% of the marked items in our LI corpus do not have

a good counterpart at all (i.e. they are unpaired) can be seen as evidence against the assumption that syntactic research is restricted to the analysis of pairs. One might object that such items have an implicit good counterpart. However, 30% of the unmarked items do not have a marked counterpart, either. It would be highly counterintuitive to assume *implicit* bad counterparts (as there is typically no single bad counterpart to an acceptable sentence).

We think that this is sufficient evidence to make our point that syntactic research is not restricted to the analysis of pairs. However, we can go one step further: Even if all syntactic research *was* restricted to the analysis of sentence pairs, then we would not want this to be the case, because the field as a whole would suffer. Judgments would be limited to the pairs in question and one could not easily compare several judgements even by the same author, let alone judgements across authors. If this was the case, any syntactic theory based on methods that only include pairwise comparisons would be extremely limited and most likely not adequate.

LIMITATIONS OF SSA'S ANALYSIS Although the reasons why SSA chose the analysis they did might be questionable, the analysis itself could still be adequate, viz. if it was able to detect most if not all relevant differences between the results of informal and formal methods. However, we do not think this is the case: A test of directionality, certainly how SSA implemented it, can only detect the most blatant differences.

The reason for this is that SSA used pairs that consisted of extrema (a *-item vs its OK-counterpart) and that the results from the formal methods were extremely reliable. The formal results were so reliable because for each construction, SSA surveyed 312 participants. Assuming that linguists and participants are equally prone to sources of error (judgement errors, etc.), the average rating from the participants

becomes more than 17 times more reliable than a judgement by a single person (i.e. by the linguist).[11] Thus, whenever the results of the informal and formal method differ in direction, this was most likely caused by a bad judgement by the linguist.[12]

However, since SSA were looking at extreme pairs (*-item vs OK-item), even the judgements by the linguists would have had to be really bad to cause a mismatch between informal and formal results (or the linguist must have been really unlucky to suffer an extreme blow from the above mentioned sources of error like judgement errors, etc.).

The following example makes this clearer and is a *proof-of-concept*-example. In the example, we first survey (fictional) ratings from formal methods, followed by informal ratings. This is done for illustrative reasons, but the methodology is analogous to SSA's methodology. We use the shooting analogy from Section 1.2.2. (In this analogy, we, the observers, do not know where the target lies. We let our subjects take shots to then use the positions of the nerf darts to derive the target. Taking a shot is equivalent to letting subjects give ratings and the position of a nerf dart is analogous to an actual rating.)

Assume we are interested in the target area for an apparently bad sentence and its good counterpart. We then ask 312 participants to take shots and we average their results. The red dot in Figure 1.12 is the average for the bad sentence and the blue dot for the good counterpart. Since we are dealing with a lot of subjects,

[11] 17 times, as the rate of judgement errors for the informal ratings is $\frac{1}{\sqrt{1}}$ and the rate for formal ratings is $\frac{1}{\sqrt{312}}$, i.e. 1 vs 0.057. See the error magnitude formulas on page 33 for further details.

[12] One might argue that it could be the case that the online participants are much more error-prone than the linguists. However, the online participants would have to be completely clueless to have a joint error rate that is similar to that of a single linguist.
But they are not clueless: SSA's participants probably had a high degree of task familiarity, which makes their judgements as reliable as judgements by linguists (also cf. "Are there expert observers?" in Section 1.2.1). We say "probably", because this depends on SSA's recruitment criteria on Amazon Mechanical Turk (which they do not specify). It is common practice, though, to only allow experienced participants to take part. Also cf. Section 1.4.1.

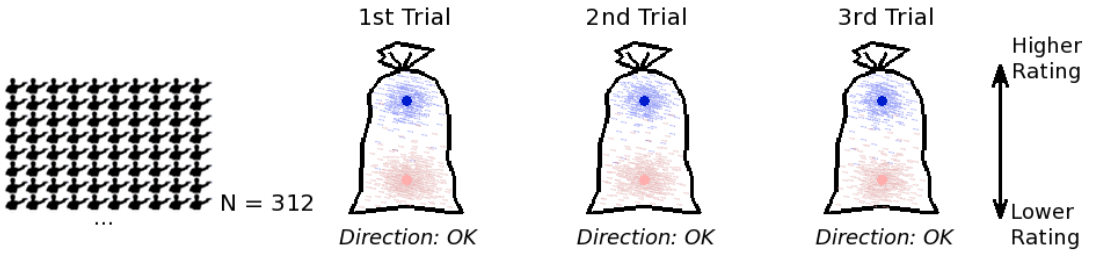


Figure 1.12: Illustrating why it is extremely unlikely for an extreme pair (*-item, in red, vs OK-item, in blue) to get the directionality wrong. In the shooting analogy from Section 1.2, we ask 312 participants to rate a certain sentence pair. Individual shots are marked by small darts, the averaged ratings by the big dots. We ask participants to take part in three trials. Due to the high number of participants, it is extremely likely that the results will look highly similar. Also, as we are considering extreme pairs (*-item vs OK-item), it is likely that a test of directionality will come out as positive (i.e. *-item $<$ OK-item).

the results would look very similar if we were to run the experiment a second and third time. And we can also test for directionality: If a shot is above another, then it received a higher rating. In all three trials in Figure 1.12, the averaged results for the bad sentence are worse than for the good sentence. This is no surprise, because we are looking at extreme pairs (i.e. bad item vs good counterpart) and the results are extremely reliable due to the high number of participants. In such a setting, to get a false positive (i.e. to get the direction wrong) is *extremely unlikely* (and a mismatch rate of 5% points at a considerable disagreement between linguists and the participants of a formal study).

We now want to compare the formal results to informal results. Assume we ask an ineffectual linguist to take a shot at both the bad and at the good sentence.[13] The shots are far away from where the target actually is, but they are still correct in directionality. This is because the extreme pairs leave quite some leeway. We then ask the linguist to try again. The results come out quite different, but they are still correct in directionality, as illustrated in Figure 1.13.

[13]We do not wish to say that the linguists featuring in SSA were incompetent. We only make this assumption to prove our point about how much leeway the linguists had.

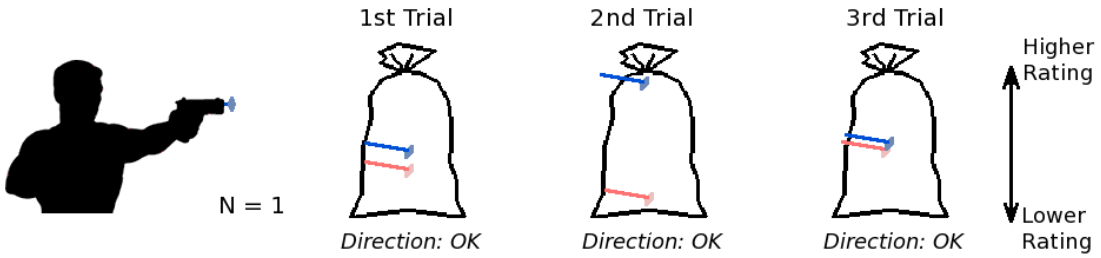


Figure 1.13: Illustrating why it is unlikely that even an incapable linguist gets the directionality of an extreme pair wrong, using the shooting analogy from Section 1.2. Assume that we ask a single linguist to rate a sentence pair (*-item, red dart, vs OK-item, blue dart) three times ("Trials"). Even if the ratings come out as quite different, they can still be OK in directionality, which illustrates that the linguist has a lot of leeway in his/her judgements.

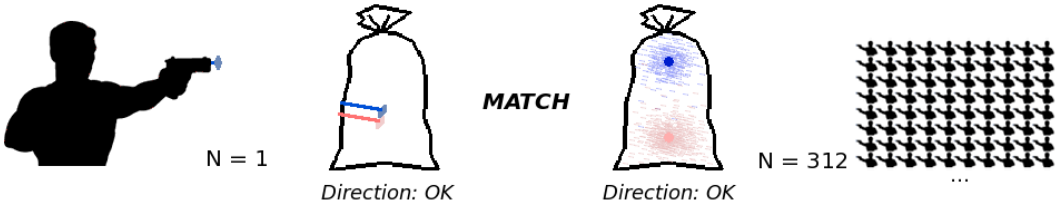


Figure 1.14: Illustrating why a comparison of directionality will most likely come out as correct, using the shooting analogy from Section 1.2. Assume that we compare the informal results from Figure 1.13 and formal results from Figure 1.12. With such a high number of participants (312) for the formal method, the formal results will be extremely reliable. Hence, any difference in directionality will most likely stem from a deviant judgement by the linguist (i.e. the informal result). Again, the linguist has a lot of leeway in his/her judgement, before he/she reaches the point at which the directionality of the averaged formal results do not match the directionality of the informal result.

If we then compare the informal and formal results under these assumptions, then it is very likely that we find that they match. Such a comparison is illustrated in Figure 1.14. If we repeated this for numerous pairs, most of those comparisons would probably come out as matches, see Figure 1.15.

In Figure 1.15, the overall match rate for the comparison of informal and formal results is rather high (9 out of 10). However, at the same time, the informal results are extremely unreliable, because for our example, we made sure that we only asked linguists who were ineffectual. We did this for two reasons: 1) To illustrate that a test

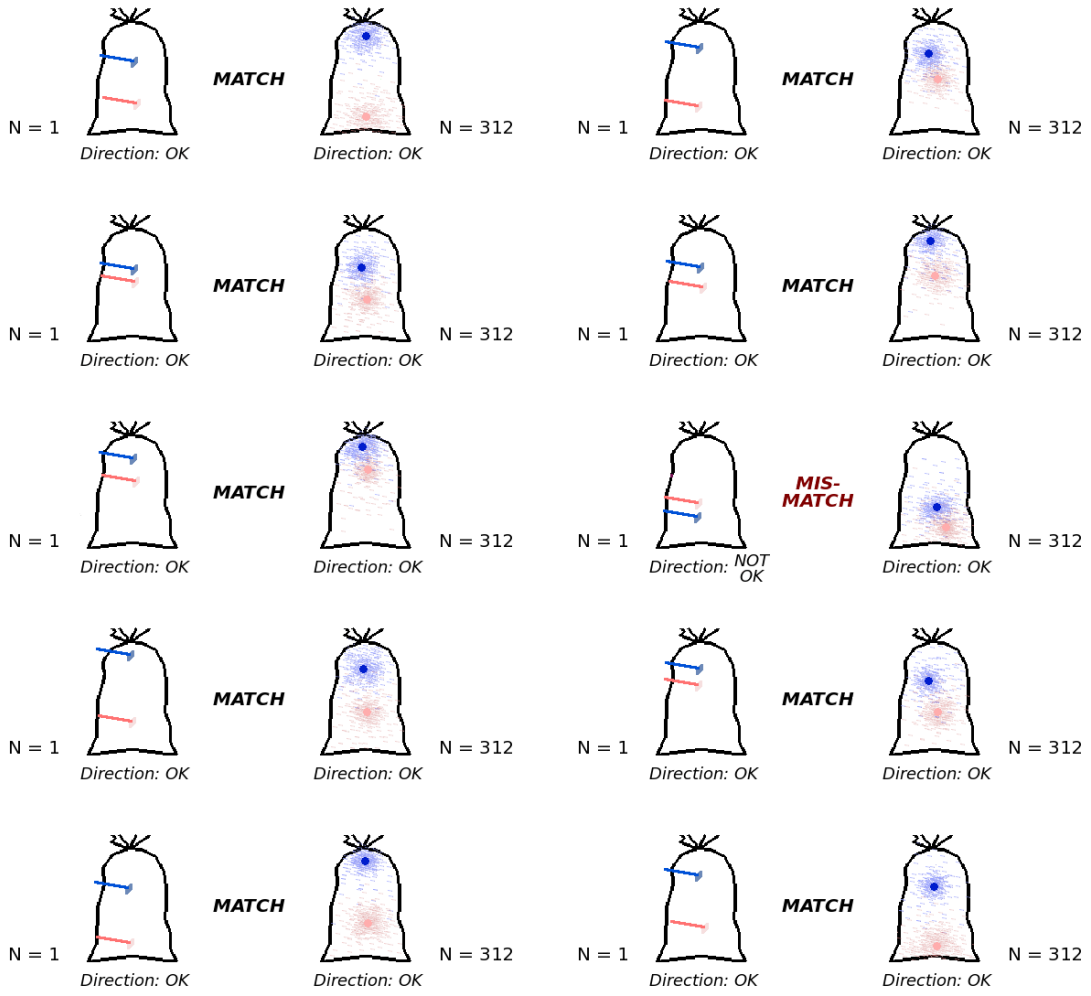


Figure 1.15: As Figure 1.14, but we perform this comparison of directionality for numerous constructions. For the reasons discussed above, we expect that the vast majority of these comparisons will come out as matches.

of directionality is unable to detect clear differences in reliability between informal and formal results. 2) To put the 95% match rate into context: the number appears more impressive than it really is.[14]

Further, even with a competent linguist, chances are that any mismatch is caused by a bad judgment on the linguist's side. The two formulas below express this. Equation 1.2 is the expected error magnitude for informal judgements (for a sentence, construction, etc.), Equation 1.3 is the expected error magnitude for formal judgements (and should be understood as an averaged rating, here averaged over the 312 participants in SSA's experimental sessions). The linguist's judgement is vulnerable to judgement error (the first term in the formulas), purpose bias (ϵ_p), scale bias (ϵ_s), and quantisation errors (ϵ_q).[15] The formal results ratings are far more reliable when it comes to judgement errors. Also, if the experiment is conducted properly, the ratings will not be affected by purpose biases. Further, scale biases will be greatly mitigated (expressed by the coefficient a, where $0 \leq a \leq 1$). Quantisation errors, though, will affect the formal results just as much.[16]

$$\epsilon^2_{informal} = \frac{1}{1} + \epsilon^2_p + \epsilon^2_s + \epsilon^2_q \tag{1.2}$$

[14] And it could or even should have been even higher.

[15] For co-authored papers, the denominator will be higher than 1. Also, some syntactic phenomena have been around for quite some time and have been judged by several linguists over the years; here, too, the denominator will be higher than 1.

[16] In some cases, a mismatch might be caused by the fact that the linguist considers more than just acceptability. Reconsider (1): "The horse raced past the barn fell". If we were to ask a syntactician for his/her judgement, he/she might give a high rating, but say something along the lines of "I know that it's not acceptable - but it certainly is grammatical". The linguist would hint at the distinction between acceptability and grammaticality as discussed in Section 1.2.1. If we were to give (1) to naive speakers and ask them to rate it, the averaged rating would probably be quite low. That is, we would get a mismatch between informal and formal ratings. There are two points we would like to make about this kind of mismatch. First, they are of interest to syntacticians, as they tell us something interesting about syntactic processing. But they are best detected through both careful thought by the researcher and empirical work. Second, mismatches surrounding constructions like (1) are the exception and in most cases, linguists and naive speakers give standard acceptability judgements.

$$\epsilon^2_{formal} = \frac{1}{312} + a^2 \cdot \epsilon^2_s + \epsilon^2_q \tag{1.3}$$

FURTHER ISSUES WITH SSA There are other potential issues with SSA's methodology: 1) SSA included items that come from papers authored by non-native speakers, i.e. non-native speakers might have judged these items. 2) They excluded papers that were not "predominantly" syntactic. 3) By focusing on marked items, SSA disregarded unmarked items without a counterpart. 4) For marked items without a counterpart, SSA made one up themselves, which leaves room for potential biases. 5) For each sentence pair, SSA created seven variants of the pair. This, too, leaves room for potential biases.

First, SSA included items from authors who are non-native speakers of American English. We think it would have been better to only include native speakers, because informal ratings from non-native speakers and native speakers are qualitatively different: Non-native speakers have to rely on other native speakers for their judgements (certainly for the more intricate examples). By doing so, they do not rely on researcher introspection but on something closer to an informal judgement task (which should improve the reliability of the results, as there is no purpose bias; also, a non-native researcher might ask several native speakers). On the other hand, researchers who are native speakers of the investigated language have a choice: They can rely on researcher introspection or they can ask other native speakers (e.g. in a classroom setting). However, authors who did ask others then typically include a note that this was the case.

It is likely that SSA are aware of this amalgamation of native and non-native speakers; since they do not claim that they have compared researcher introspection to formal methods. Instead, they say that they have compared *informal methods* to

formal methods.

Second, SSA only included items from papers that were "predominantly" syntactic, where "predominantly" in their definition means that 80% of the data points had to be syntactic judgements. However, we cannot think of any reason to exclude perfectly fine items, just because they come from a paper in which e.g. only 60% of all items are syntactic judgements. Reducing extraction effort cannot have been the reason for this choice: To determine whether a paper qualifies for the 80% threshold, one has to categorise a paper's items in the first place.

Third, in their random sampling procedure, SSA selected marked items (mainly *-items) and then found their good counterparts. However, this way, a large number of good items without a bad counterpart got disregarded, as they were categorically excluded from the sampling procedure. It is unclear how this will affect the results. (In their defence, SSA would point out that they think that syntactic research is mainly concerned with the analysis of pairs.)

Fourth, only 108 of the 150 marked items that SSA sampled had a good counterpart; 42 items (or 28%) did not. For these 42 items without a counterpart, SSA constructed one, saying that these had *implicit* counterparts. However, this leaves a backdoor for a potential bias: If SSA expected the results from informal and formal methods not to match, then they might have unconsciously created weaker counterparts, which would have reduced the contrast between pairs. If they expected concurrence, they might have unconsciously created stronger counterparts, which would have increased the contrast between pairs. We are not saying that this is what has actually happened; we are saying that this is a potential danger. However, it does not bode well that e.g. for SSA's Likert-Scale results, all five violations in

directionality came from sentence pairs for which the good counterpart was provided by the LI authors. That is, not a single violation occurred for those pairs for which the good counterpart was implicit (i.e. where SSA created the good counterpart themselves). (Assuming that the five violations are events that are independent from each other, then the probability for this to happen by chance is 17.6%.)

Fifth, for each item, SSA created seven variants in order to have more experimental items. For instance, (18) is the original item (from Martin, 2001) and (19) is the variant constructed by SSA.

(18) He seems to that Kim solved the problem.

(19) They appear to that Chris is the right person for the job.

This is an extreme example (the other variants are more similar). However, the same point from above applies here, too: If SSA expected a difference in informal and formal results, then they might have unconsciously created variants with lower contrast within a pair. If they expected concurrence, they might have unconsciously created variants that increase the contrast. (It would have been better if SSA had asked a third party to provide the variants.)

Because of SSA's limited analysis and the issues presented above, the question of how results from informal and formal judgement tasks compare is not settled. Consequently, we have designed a new study that includes a more adequate analysis and that avoids the issues presented above.

1.3 Our LI Corpus

Our goal is to provide a quantitative comparison of informal and formal judgement methods used in syntactic theory. For our materials, we wished to randomly sample items from the literature. Ideally, we would have used SSA's LI corpus for this. However, SSA did not publish their corpus, so we had to create our own.

THE CREATION OF OUR LI CORPUS To ensure a random sampling procedure, we created a corpus of introspective acceptability judgements (i.e. researcher introspection), based on all papers in Linguistic Inquiry (LI) from 2001 to 2010. This approach is similar to Sprouse et al. (2013), but differs in some important details.[17]

First, we only include authors who were native speakers of American English (for the reasons discussed in Section 1.2.5). We base the decision whether or not an author counts as a native speaker of American English on his/her biographic information (extracted from the authors' CVs, websites, and publications). We use the first information available and apply the following decision hierarchy.

Birthplace > Country of High School > Country of Undergraduate Studies > Country of Postgraduate Studies > Country of PhD > Country of First Position

The further left in this hierarchy a piece of information lies, the higher it ranks. High ranks take precedence in the evaluation of an author's background. For instance, if an author did his/her undergraduate studies in the US, but went to high school in

[17]N.B.:One of the reasons why we chose to collect items from LI is because we wanted to be able to compare our results to SSA's results. If we had chosen e.g. *Language* (which might actually have been a better choice), then any potential mismatch between SSA's results and ours could have been due to different corpora (but we would not have been able to tell).

Bermuda, then we label him/her as a non-native speaker of American English. In case of co-authorship, we include a paper if one of the authors turned out to be a native speaker of American English. A full list of all authors of LI papers between 2001 and 2010 can be found in Appendix 1.7.1.[18]

Further, we include all papers that come from native speakers (i.e. we had no restriction similar to SSA, who only included papers that were "predominantly" syntactic). This leaves us with 160 papers, from which we extracted all sentences that included some kind of judgement (whether of syntactic, semantic, or other nature; however, we do not consider dialectal items[19]). (Not all of those 160 papers feature in our LI corpus, because some of them do not include any judgements at all.)

CATEGORIES USED IN OUR CORPUS In a final step, we then categorised all extracted items. For our categorisation process, we used the categories from SSA. We have standard acceptability judgments, coreference judgments, interpretation judgments, judgments involving relatively few lexical items, and judgments involving prosodic manipulations. Below are the definitions by Sprouse et al. (2013:221).

> Standard acceptability judgments: These require only that the participant be presented with a sentence and asked to judge its acceptability on an arbitrary scale or in reference to another sentence.
>
> Coreference judgments: These are primarily used to probe binding relationships. Participants must be presented with a sentence that includes two or more noun phrases that are identified in some way. They are then asked to indicate whether the two noun phrases can or must refer to the

[18]There will be false positives in this list (and false negatives that did not make the list). It was most important to us to have transparent and objective criteria. If we had determined the background of an author instinctively, we might have had a better F1 score, but this would have been less scientific.

[19]We do not consider dialectal items, because this would force us to control for the dialectal background of our online participants, which would blow up the number of required participants considerably.

same entity.

Interpretation judgments: These are judgments based on the meaning of sentences, such as whether a sentence is ambiguous or unambiguous, or whether one quantifier has scope over another. These may require explicit training of participants to identify multiple potential meanings, and/or explicitly constructed contexts to elicit one or more potential meanings.

Judgments involving relatively few lexical items: These are acceptability judgments about phenomena that occur with relatively few lexical items, such that the construction of 8 substantially distinct tokens, as was done for the phenomena tested in this study, would likely be impossible. This is not to say that these phenomena cannot be tested in formal experiments, but participants in such experiments may require special instruction to guard against potential repetition confounds.

Judgments involving prosodic manipulations: These are acceptability judgments that are based on specific prosodic properties of the sentence. They require either the presentation of auditory materials or the use of some notational conventions for conveying the critical prosodic properties in writing (e.g., the use of capital letters to indicate emphasis).

Some items in the standard acceptability judgments category, although easy to classify, would be hard to test. For instance (20), from Adger and Ramchand (2003), is clearly a standard acceptability judgement, but it would be hard to test, because its deictic nature might throw participants off and influence the ratings.

(20) I consider the best pictures of Mary to be these.

This is why we added another category, viz. *non-testable standard acceptability judgments*. This category includes items that were categorised as *standard acceptability judgments* as defined above with the addition that they would be hard to test for the following reasons: They have deictic references, they include strong language, there are unintended alternative readings available (particularly repair readings), or they include colloquial language that might be stigmatised.

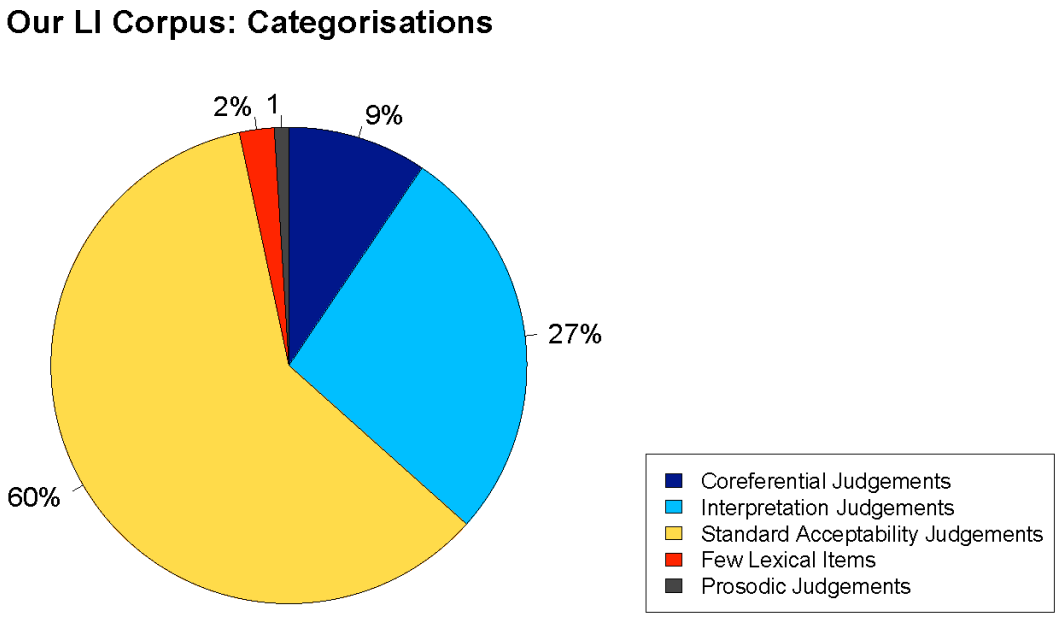


Figure 1.16: The categories that we use in our LI corpus and their share of our corpus. Here, each category includes the non-testable items. Starting with *Coreferential Judgements*, the chart gives the categories in clockwise direction.

THE STRUCTURE OF OUR LI CORPUS In total, we extracted 4334 items that include some kind of judgement. 2619 of these are standard acceptability judgements, of which we deem 2539 testable in an experiment (97%). See Figure 1.16 for details.

Of the 2619 standard acceptability judgments, 486 items come from authors whose introspective judgements were binary (i.e. items are either marked with a "*" or not marked by any diacritic, to which we refer to as *OK*). 2133 items come from authors whose judgements were gradient and include some form of "?", e.g. "?*", "??", etc. See Figure 1.17 for details. As to the 486 binarily judged items, they come from 37 authors and the top three authors are: Bowers (135 items), Landau (59 items), Stroik (36 items). As to the 2133 gradiently judged items, they come from 52 authors and the top three authors are: Culicover and Jackendoff (349 items), Martin (154 items), Landau (107 items).[20]

[20]In principle, any "binary paper" could be an underlying "gradient paper" in which no in-

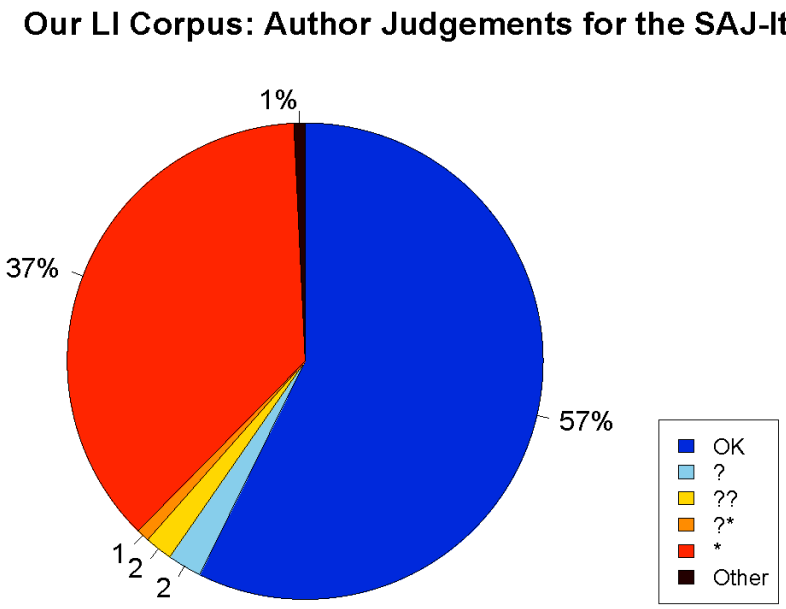


Figure 1.17: The ratings from researcher introspection for the 2619 standard acceptability judgments in our LI corpus. "Other" includes "#", "#?", "OK/*", "%", "*?", "(?)", "(??)", "??/*", "???/*". Starting with *OK*, the chart gives the categories in clockwise direction.

Also, as argued above, the fact that the vast majority of items come from authors who provide gradient judgements reinforces the idea that syntactic research does not restrict its methods to a strict analysis of pairs. If a strict comparison of pairs only was all that syntacticians did, then why would a syntactician use three or more categories (by including "?", etc.)? Also, the distinction between "binary" and "gradient" authors has some impact on the experimental design, as they require slightly different measurement techniques (see Section 1.4 for details).

The vast majority of the 2619 standard acceptability judgments are OK/unmarked, viz. 1501 items. 965 items are marked by a "*" and 153 items received a category

between category was used. In such a case, we would not be able to tell the difference between a true binary paper and a pseudo binary paper. It is likely that this thinking applies to some of the binary papers, because the binary papers are on average considerably shorter than gradient papers (12 vs 41 items, respectively).
Further, authors can appear in both categories, as with e.g. Landau. We decided to accept this, as we did not want to decide whether an author is "underlyingly" completely gradient or whether the author might have changed his/her view over the years.

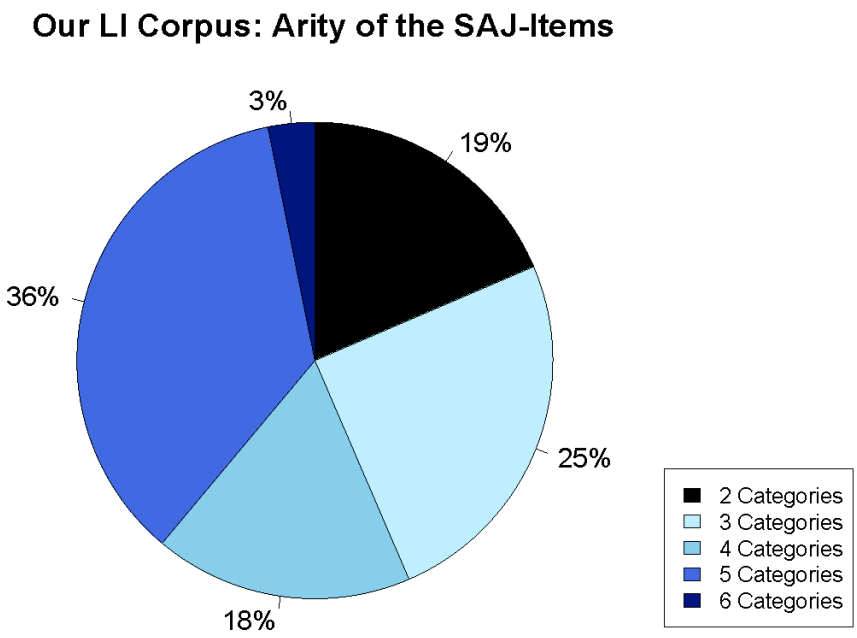


Figure 1.18: The arity for the 2619 standard acceptability judgments in our LI corpus (i.e. for any given item, this shows how many judgement categories were used in the LI article from which the item was extracted). Starting with *2 Categories*, the chart gives the categories in clockwise direction.

in between, indicated by "?*", "??", etc. See Figure 1.18 for details.

These numbers diverge from SSA's LI corpus, which should not come as a surprise: We exclude some authors that were included by SSA (i.e. non-native speakers of American English) and we include some authors that were excluded by SSA (i.e. authors of papers that are not predominantly syntactic).

FURTHER THOUGHTS ON INTROSPECTION There is the question whether the ratings for the 2619 standard acceptability judgements were given introspectively as defined above. There are three issues in particular: First, some items come from co-authored papers. Second, critics might argue that even for truly *new* items (i.e. the sentence/phenomenon was judged for the first time), researcher introspection was only a first step of a longer process. Third, critics might argue that the contribution of researcher introspection is even lower, as most items

are "tradiert" (i.e. their status was reliably established over decades) and for such items, there is no need for further judgement (introspective or otherwise).[21]

First, a good deal of items come from co-authored papers (this applies to 48.4% of all items). The problem with such items is that they were rated by several linguists. However, we include these items, because this is still very close to the original idea of *researcher* introspection, including all the problems that come with it.[22]

Another concern is that even for new items, researcher introspection is only a first step of a longer process. Most authors will have discussed their work with colleagues and their students, which will have affected their judgement in some cases. They will have also received feedback by reviewers.

The problem is that one simply cannot tell what the "longer process" looked like and that this process might have differed considerably from item to item. This hints at an important issue with this methodology: It is really opaque. We do not know how exactly the ratings came about.

So, what we effectively do is to subsume potential input by colleagues and students under *environmental influences*, because it is the author in question who made the final judgement; which is why it seems adequate to call the entire process "researcher introspection".

Linked to this is the third point: One might argue that the role that researcher introspection plays is even smaller, because most items that are "tradiert", i.e.

[21] Many thanks to Ash Asudeh and Gisbert Fanselow for pushing these points.

[22] Further, if anything, including co-authored papers will mitigate a potential mismatch between informal and formal results. So, we can still make inferences about "true" researcher introspection: If we find a mismatch below, it would be at least as big for "true" researcher introspection. It might be worth rerunning the experiment below, though, once with items from single-author papers and once for items from co-authored papers.

linguists established their status reliably over decades. Such items are not really being judged when they are discussed in recent papers; instead, researchers draw on a community consensus. This is why it is wrong to call the methodology "researcher introspection" and why we are also wrong to exclude non-native authors of American English.

We tried to quantify this: For 15.3% of all standard acceptability judgements, the authors cite someone else's work; a good deal of these items are replicas of existing items or very similar to them. However, it is likely that authors still "sanction" such "make alikes", i.e. the author introspectively confirms the judgement he/she finds in the literature. This sanctioning is based on an introspective judgement.

Many items in the remaining 84.7% will be original instantiations of syntactic phenomena that were discussed previously (to whatever degree); a few of them might even be *new*. For most new items, researcher introspection is used (which non-native authors cannot do) and the points from Section 1.2.2 apply.

As to (presumably original) instantiations of already discussed phenomena[23]: They need to be "sanctioned", too (which non-native authors cannot do, either), as there could be confounding factors, etc. The authors are likely to have previous papers in mind. Assuming that the previous papers were not the first to discuss the phenomenon in question, the authors of the previous paper also "sanctioned" the judgements they used. This reasoning continues until one gets to the original paper(s), in which the phenomenon was first discussed. In the original paper(s), the status of the phenomenon was determined introspectively. Effectively, we are looking at a *chain of researcher introspection*. The problem with this is that all the issues with "pure" researcher introspection apply to this chain of researcher introspection, too.

This is why we decided to not include papers by non-native authors; and why we decided to call all of these methodological variants "researcher introspection",

[23] *Presumably* original, because it is really hard to verify the actual status of so many items.

because in our view, this is the underlying process.[24]

1.4 The Experiment

We wish to provide a quantitative comparison of informal and formal judgement methods used in syntactic theory. Researcher introspection (i.e. the linguist is his/her own informant) is the main informal judgement method and features in our experiment as the source of informal results. We used an online acceptability judgement task to get formal results. For our materials, we randomly sampled 100 items from our LI corpus. 50 of these are *-items, 50 are OK-items (these are not paired). We did this twice, once for "binary" items (these come from authors whose introspective judgements were binary; the items in black in Figure 1.18) and once for "gradient" items (these come from authors whose introspective judgements were gradient; items in blue shades in Figure 1.18). We then check to which degree informal and formal results agree. We compare items at large, i.e. we do not restrict ourselves to the analysis of pairs. To do so, we use a point-biserial correlation measure (a correlation measure used when one variable consists of binomial, the other of interval data; see below for details) and a threshold test (similar to one of the tests used in Sprouse et al., 2013). Our informal null hypothesis is as follows.

(H_0) Informal and formal methods concur strongly; i.e. there is no substantial difference between informal judgements (that were done introspectively by LI authors) and formal judgements (collected in an online acceptability judgement task).

[24]The reader might disagree with our reasoning. If so, it is important to note, though, that this is just a label; a label for the predominant method of syntactic enquiry. We might call it *the process of community agreement* (which consists of researcher introspection + optional colleague introspection + optional student introspection + optional reviewer introspection). *The process of community agreement* still suffers from the issues discussed in Section 1.2.2. And the results in the next section apply to it, too, whatever name one chooses for it.

Why do we focus on end-points (*-items vs OK-items), while ignoring in-between categories (e.g. "??", "?", etc.)? The reason for this is that this way, any positive result (i.e. rejecting the null hypothesis) will make the strongest possible point about differences between informal and formal results.

1.4.1 Experimental Design

MATERIALS From the 2539 testable data points, we randomly sampled two sets of 100 sentences from our LI corpus, without replacement, using R (R Core Team, 2015). One set consists of items from papers by authors whose judgements are binary and the other came from papers in which judgements are gradient. In both sets, 50 items are *-sentences and 50 are OK-sentences. For each batch of 50 items, we randomly sampled another five items as "wildcards". We included the wildcard items to address the possibility that some of the 50 items were not standard acceptability judgements after all (and thus miscategorised with respect to the categories mentioned in the subsection "Categories Used In Our Corpus" above). As we decided to check the sampled items again, we thought that we might have to drop some of the sampled items. If this did happen, then the wildcard would "move up" to the batch of the selected 50 items, so that we would not have to redo the entire sampling procedure. The batches are as follows:

Set 1:
50 *-items from "binary" authors: 2 wildcards used.
50 OK-items from "binary" authors: 0 wildcards used.

Set 2:
50 *-items from "gradient" authors: 1 wildcard used.
50 OK-items from "gradient" authors: 0 wildcards used.

In all three cases in which we had to use a wildcard, this was done because the

testability of the item was in question (and thus, we re-categorised them as non-testable standard acceptability judgements). Further, we chose a 50-50 split for two reasons: First, to keep the proportions between the "binary" and the "gradient" identical, and second, to not cause any imbalances that we then would have to offset by additional fillers (including too many bad or too many good items can have an impact on the ratings, cf. Cowart, 1997). A full list of selected items can be found in Appendix 1.7.2.

FILLERS We chose to not include fillers. Fillers are typically included for two main reasons: 1) They should distract from the purpose of the study and make its structure less obvious. 2) They should offset possible imbalances within the set of the critical items (e.g. by reminding participants what the endpoints of the scale look like). As to the first point, we would not have distracted from the purpose of our study by including fillers, because we are interested in a meta-issue and not in concrete constructions. As to the second point, our set of critical items is well-balanced already. Thus, we saw no need for fillers.

PROCEDURE We wish to compare informal and formal judgements. Retrieving informal judgements was part of the random sampling procedure described above. To get the formal judgements, we tested the extracted items in an online acceptability judgement task. For the formal judgements, we decided to mirror the binary vs gradient split of the informal judgements (i.e. 100 of our items come from authors whose judgements were binary, the other 100 from authors whose judgements were gradient). Consequently, we ran the entire experiment twice: In the first run, participants used a gradient rating scale (in our case a 7-point Likert scale) and in the second, a binary scale (a 2-point Likert scale).

The main reason for matching conditions like this (binary author judgements & 2-

point scale in an online experiment and gradient author judgements & 7-point scale in an online experiment) is to ensure that authors and online participants rated the items under similar conditions. In psychology, differences between results from a 2-point scale and results from a gradient scale are well established: Ghiselli (1939) found that a 2-point scale makes the results more favourable (i.e. participants tend to give higher ratings). In his review, Cox (1980) discussed potential information loss due to using a 2-point scale (vs gradient scales). And Weijters et al. (2010) found that (aggregated) data from a 2-point scale have a tendency towards end-points. It is not unreasonable to expect such differences for linguistic data, as well: The results might become more contrastive, as illustrated in Figure 1.19.[25]

A downside of this parallelism could be this: The authors using a binary scale were free to choose their own scale and they chose a binary scale. Choosing one's own scale reduces quantisation errors (cf. Section 1.2.2). The online participants, on the other side, are most likely forced to use a scale that has considerably fewer degrees than they would have preferred (typically, participants prefer five or more degrees, cf. Bard et al., 1996:45). This will cause an increase in quantisation errors, which again will cause their judgements to be biased.

Further, we split both sub-experiments into two sessions in order to avoid fatigue effects (i.e. participants get tired of the task and the quality of their ratings becomes lower; see below for details). To do this, we randomly selected 25 of the 50 *-items and 25 of the 50 OK-items for a first session; the remaining 50 items (25 *-items and 25 OK-items) were tested in a second session. This gives us four sessions:[26]

Session 1, "Binary Part 1": Testing the first 50 items (25 *-items and 25

[25]N.B.: The underlying data structure of Figure 1.19a and 1.19b could differ considerably. Consider any in-between item, e.g. Item 7. In case of a 2-point scale (1.19a), the online ratings would take the form of [1, 1, 0, 1, 1, 0, ...], averaging about 0.65. In case of a 7-point scale (1.19b), the online ratings could take two forms. One might observe a bimodal distribution of ratings in the form of [6, 7, 2, 6, 6, 1, ...] or one might observe a unimodal distribution with lots of middle values in the form of [4, 4, 5, 4, 5, 4, ...]. Both would come out at about 4.3.]

[26]N.B.: There were different participants in each session.

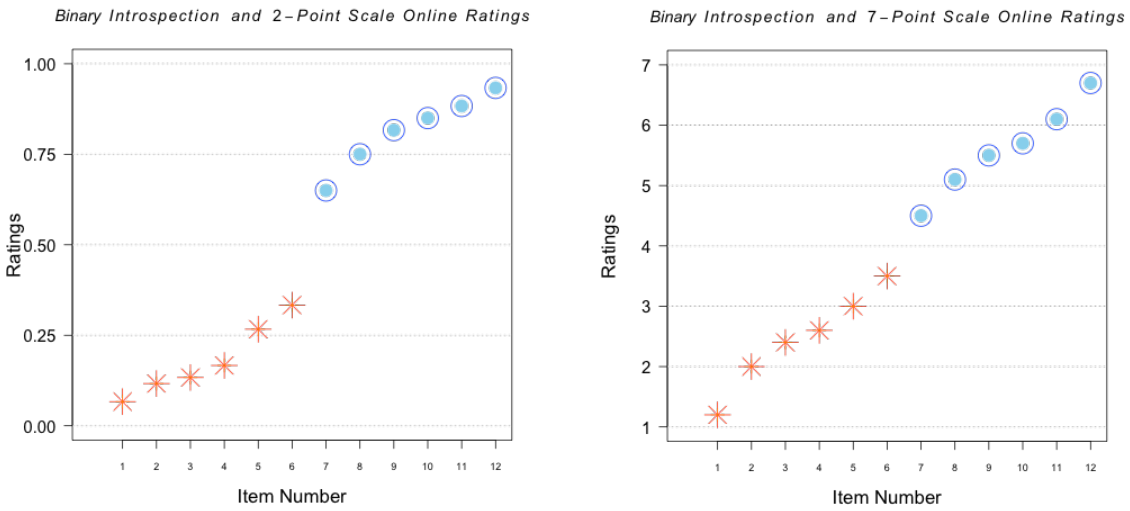


Figure 1.19: Using hypothetical data, an illustration of why we used a binary scale in the online experiment when testing "binary" items (i.e. items that came from authors whose introspective judgements were binary) and a gradient scale when testing "gradient" items (i.e. items that came from authors whose introspective judgements were gradient). Let us assume we are dealing with six rather marked sentences (red asterisks) and six somewhat unmarked sentences (blue circles). Our hypothetical authors have correctly judged the marked items as "*" and the unmarked items as OK. We now put the twelve items to test in an online experiment. Using a binary scale (left; 1.19a), our 30 participants are forced to choose one of two values, so that the data might become quite contrastive. We then repeat this procedure; however, this time we use a gradient scale in the online experiment (right; 1.19b). The 30 participants might give nuanced ratings and the results form somewhat of a continuum (averaged ratings are given on the x-axis).

OK-items) that come from "binary" authors.

Session 2, "Binary Part 2": Testing the remaining 50 items (25 *-items and 25 OK-items) that come from "binary" authors.

Session 3, "Gradient Part 1": Testing the first 50 items (25 *-items and 25 OK-items) that come from "gradient" authors.

Session 4, "Gradient Part 2": Testing the remaining 50 items (25 *-items and 25 OK-items) that come from "gradient" authors.

Our instructions were as follows. Participants were told that they will judge items with regard to how natural or unnatural they appear to them. We added to this that "natural/unnatural" should be understood with respect to the item's grammaticality and that participants "should not be bothered with meaning or punctuation". We further noted that we are interested in the participant's intuition and that "there is no right or wrong". We then introduced the rating scales; in both sub-experiments, participants made their ratings by pressing buttons. The binary rating scale is presented in Figure 1.20. We noted that the lower value corresponds to "fully unnatural/ungrammatical" (the red button) and the upper value to "fully natural/grammatical" (the green button). We included the smiley faces for the colourblind.[27]

The instructions for the gradient rating scale were very similar: The lower value corresponds to "fully unnatural/ungrammatical" (the red button) and the upper value to "fully natural/grammatical" (the green button). The remaining buttons denote values "in between". Figure 1.21 shows our gradient rating scale.

In both sub-experiments, the first four items were calibration items, which we included to give our participants an idea of the range of possible goodness and badness. The calibration items included two good items ((21) and (22); these are from USA

[27]To further assist the colourblind, it would have been preferable to use red and blue buttons.

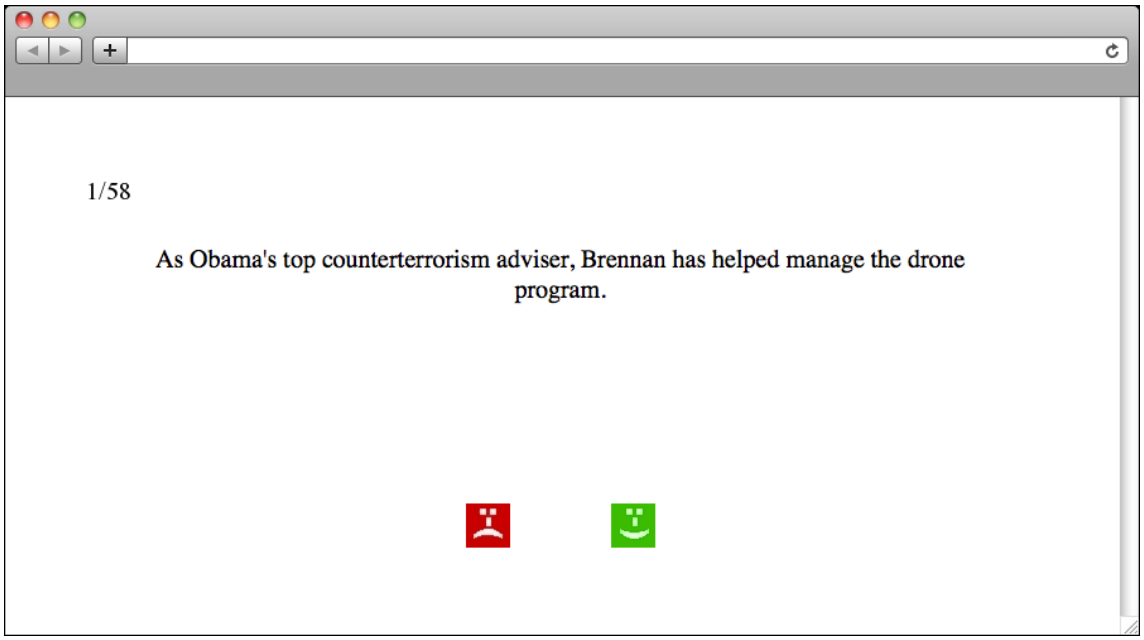


Figure 1.20: The experiment with a Binary Likert Scale.

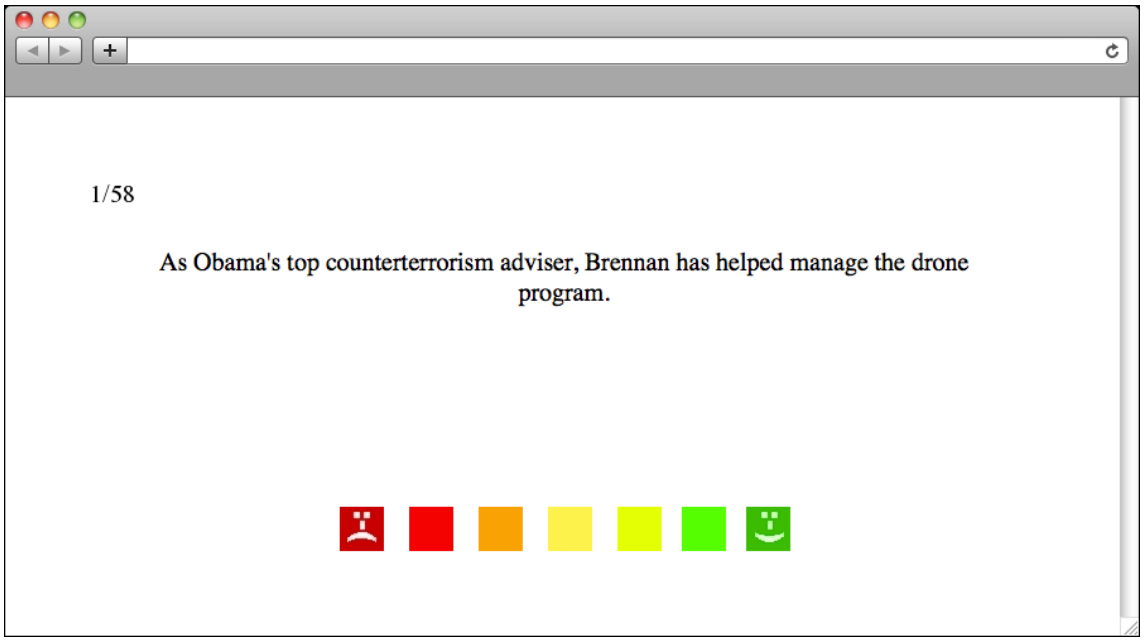


Figure 1.21: The experiment with a gradient Likert Scale.

Today and received high ratings in the study presented in Chapter 2) and two bad ones ((23) and (24); these are modelled after sentences in Ferreira and Swets, 2005, and received low ratings in the study presented in Chapter 2).

(21) As Obama's top counterterrorism adviser, Brennan has helped manage the drone program.

(22) Iran has proposed restarting talks as early as next month.

(23) This is a donkey that I don't know where it lives.

(24) This is the man that I don't know where he comes from.

Then, the critical items followed. Besides collecting ratings, we also collected reaction times, which we use to detect non-cooperative participants (the reaction time of an item is the time from loading the item until the time at which the rating is given). We further included an on-line warning mechanism that notified participants when their behaviour became uncooperative. This was determined through their reaction times. We deem a reaction time of under 400 ms as uncooperative. A first pop-up window, warning the participant, appeared if a participant fell below the 400-ms-threshold four times. A second pop-up window appeared after the twelfth violation. We chose 400 ms, because even for the shortest item of all sentences, 400 ms is less than half of its expected reading time (the expected reading times are calculated using a related formula from Bader and Häussler, 2010:289). We included this warning mechanism because in our pilot study, we had several non-cooperative participants.

PARTICIPANTS For each session (Binary Part 1, Binary Part 2, Gradient

Part 1, Gradient Part 2), we recruited 40 participants using Amazon Mechanical Turk (4×40 participants; 160 participants in total). To be able to take part, potential participants had to have an Amazon Mechanical Turk approval rate of at least 98% and to have finished at least 5000 approved tasks. We set these criteria to ensure that only reliable participants could take part.[28] We are interested in native speakers of American English; however, to avoid accommodation effects, we did not disclose this criterion to our participants. So, native speakers of any language could participate. And to increase the chances of native speakers of American English taking part in our study, recruiting took place between 21:00 and 02:00 GMT. Payment was such that it resulted in an hourly rate of about $10. Participants had to be non-linguists, which was mentioned in the introduction. We also mentioned that our study was approved by and followed the guidelines of the University of Oxford's Central University Research Ethics Committee.

After the experiment, we anonymously gathered the following information: a participant's age, gender, and home country; we also asked our participants where they predominantly lived the first ten years of their lives. If the home country and the place where a participant lived the first ten years of his/her life are the United States of America, then we consider that participant as a native speaker of American English (again, we did not ask for their first language directly, to avoid accommodation effects).

Our pools of participants have the following demographics (after exclusions):

> Session 1, "Binary Part 1": 32 participants included; mean age: 35.28 years (11.97); gender distribution: 20 females and 12 males.
>
> Session 2, "Binary Part 2": 36 participants included; mean age: 34.16

[28] N.B.: These criteria are more lenient than what is required to become an Amazon Mechanical Turk "Master Worker".

years (11.59); gender distribution: 9 females and 27 males.

Session 3, "Gradient Part 1": 30 participants included; mean age: 36.27 years (10.27); gender distribution: 13 female and 17 male participants.

Session 4, "Gradient Part 2": 35 participants included; mean age: 34.80 years (9.82); gender distribution: 18 female and 17 male participants.

EXCLUSION CRITERIA We have several exclusion criteria. These are: participating multiple times (0 exclusions), not being native speakers of American English (14 exclusions), returning incomplete results (3 exclusions), having extreme reaction times (5 exclusions), or failing on the booby trap items (5 exclusions). We applied these criteria in the listed order. In the following, we discuss our criteria in detail.

Participating multiple times: Participants could only take part once, to make sure that no single participant is overrepresented. Within a session, it was technically not possible for a participant to take part twice. However, it would have been possible for a participant to take part in multiple sessions. We explicitly asked participants not to do so and reminded them that we could easily detect violators through their unique Amazon Mechanical Turk ID number. Consequently, not a single participant attempted to take part in multiple sessions.

Not being a native speaker of American English: We did not state this explicitly to our participants, but we only consider results that come from participants who we believe to be native speakers of American English (see above for how we determine this).

Returning incomplete results: We disregard any results from a participant who did not return a complete set of ratings.

Having extreme reaction times: Participants that have either extremely low or extremely high reaction times are excluded. Extremely low reaction times are an indicator of being non-cooperative: These participants just "click their way through", submitting ratings of little value. Extremely high reaction times can be an indication of distractedness, possibly affecting the quality of responses as well. We use the following formulas to exclude participants who are either extremely slow or extremely fast. In our definition, "extremely low reaction times" are those that fall below the lower threshold (θ_{lower}) and "extremely high reaction times" are those that fall above the upper threshold (θ_{upper}). $\mu_{1/2}$ is a participant's median reaction time; $\overline{\mu_{1/2_{1-N}}}$ is the mean of all participants' median reaction times. σ is the standard deviation of all participants' median reaction times. The formulas are justified in Appendix 6.

$$\theta_{lower} = \overline{\mu_{1/2_{1-N}}} - 1.5\sigma \tag{1.4}$$

$$\theta_{upper} = \overline{\mu_{1/2_{1-N}}} + 4\sigma \tag{1.5}$$

$$\text{where } \overline{\mu_{1/2_{1-N}}} = \frac{\mu_{1/2_1} + \mu_{1/2_2} + \ldots + \mu_{1/2_N}}{N} \text{ and } \sigma = stdev(\mu_{1/2_{1-N}})$$

Failing on booby trap items: We include "booby trap" items to detect distracted participants, who otherwise would have normal reaction times. We include two booby trap items which should sound bad to speakers of American English ((25) and (26); modelled after Hansen et al., 1996) and two items which should sound

good to speakers of American English ((27) and (28); modelled after Brians, 2014, and Kövecses, 2000, respectively).

(25) Peter wanted that we should come early.

(26) My knowledges of chemistry are rather weak.

(27) My son's grades have gotten better since he moved out of the fraternity.

(28) The professor requested that Dillon submit his research paper before the end of the month.

These items were randomly interspersed in the final two thirds of the questionnaires. If the two bad booby trap items received a higher average rating than the two good ones, then we exclude that participant.

In total, we exclude 27 out of 160 participants. This exclusion rate (16.9%) is in line with the exclusion rate reported in Munro et al. (2010). After exclusions, we have a total of 6650 data points ((32 participants from Session 1 + 36 participants from Session 2 + 30 participants from Session 3 + 35 participants from Session 4) $\times$ 50 sentences).[29]

[29] Another way of booby trapping is to ask for a very specific response ("Please click the red button.", "Please click on the happy smiley face.", etc.). The advantage of this is that there is a clearly defined correct reply; while our booby trapping only provides an approximation (it leaves the possibility that someone truly thought that (25) and (26) sound worse than (27) and (28)). Many thanks to Ash Asudeh for pointing this out to us.

1.4.2 Analyses

In this section, we discuss the analyses we chose and the concrete hypotheses with respect to them. However, we begin this section by specifying our experimental conditions.

EXPERIMENTAL CONDITIONS We wish to compare ratings from researcher introspection to ratings from an online experiment. We do this for introspectively marked sentences (*-items) and unmarked sentences (OK-items) and we further distinguish between "binary" and "gradient" authors (see Section 1.4.1 for details). This gives us our eight experimental conditions. The first four conditions ($C_{intro*bin}$, $C_{intro\,OK\,bin}$, $C_{intro*gra}$, and $C_{intro\,OK\,gra}$) are the ratings from researcher introspection.

> ($C_{intro*bin}$): Ratings from researcher introspection for the *-sentences from "binary" authors.
>
> ($C_{intro\,OK\,bin}$): Ratings from researcher introspection for the OK-sentences from "binary" authors.
>
> ($C_{intro*gra}$): Ratings from researcher introspection for the *-sentences from "gradient" authors.
>
> ($C_{intro\,OK\,gra}$): Ratings from researcher introspection for the OK-sentences from "gradient" authors

Collecting the ratings for the items from researcher introspection was very straightforward: As we extracted items from LI, they already came with an author's introspective rating (e.g. as to $C_{intro*bin}$, the introspective rating for any marked sentence is, of course, a "*").

For the same items, we collected experimental ratings online. For instance, the items

in ($C_{intro * bin}$) and ($C_{online * bin}$) will be the same, so that we can compare these two conditions directly.

($C_{online * bin}$): Binary online ratings for the *-sentences from "binary" authors.

($C_{online\,OK\,bin}$): Binary online ratings for the OK-sentences from "binary" authors.

($C_{online * gra}$): Gradient online ratings for the *-sentences from "gradient" authors.

($C_{online\,OK\,gra}$): Gradient online ratings for the OK-sentences from "gradient" authors.

There are two crucial comparisons. First, we are particularly interested in a comparison of ($C_{intro * bin}$) and ($C_{intro\,OK\,bin}$) to ($C_{online * bin}$) and ($C_{online\,OK\,bin}$). This is what we refer to as the "binary comparison" or the "binary condition". Second, we are also interested in a comparison of ($C_{intro * gra}$) and ($C_{intro\,OK\,gra}$) to ($C_{online * gra}$) and ($C_{intro\,OK\,gra}$). This is the the "gradient condition". The latter comparison might even be more important, because the vast majority of items in LI come from authors who use a gradient scale (cf. Section 1.3).

ANALYSIS 1: POINT-BISERIAL CORRELATION MEASURE The introspective data are binomial data; the Z-scores of the online ratings are interval data. A point-biserial correlation measure would fit this data structure well (cf. Jackson, 2011). The measure is calculated as per Equation 1.6. In our case, μ_1 is the mean of the OK-items, μ_2 the mean of the *-items. σ_1 is the standard deviation of the OK-items, σ_2 the standard deviation of the *-items.

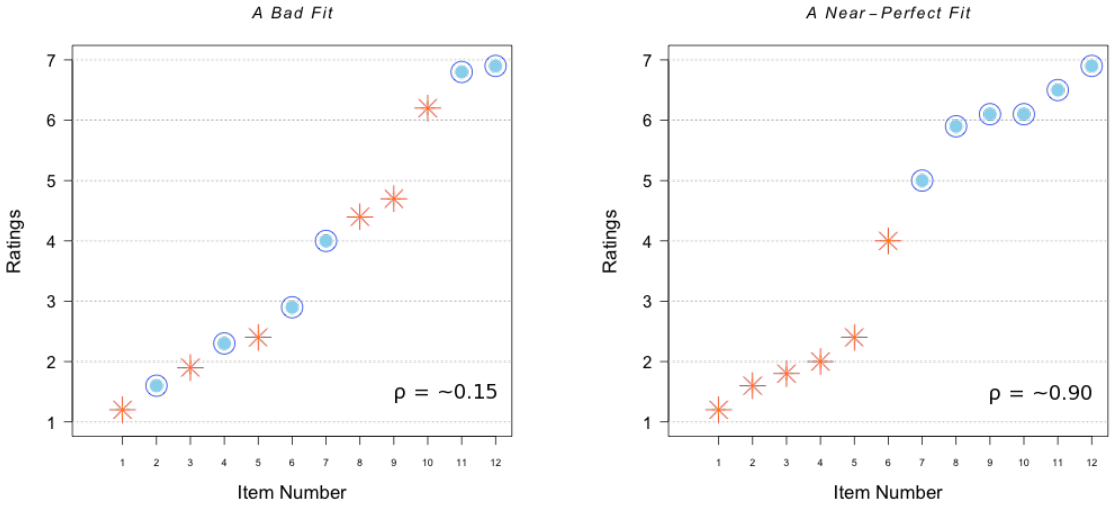


Figure 1.22: Based on hypothetical data, an illustration of a bad fit (left; 1.22a) and of a good fit (right; 1.22b) between informal and formal judgements, using the point-biserial correlation measure. Items are aligned on the x-axis, according to their online rating on a 7-point scale (y-axis). The author judgements are coded as follows: red asterisks = "*"; blue circles = OK. For the bad fit simulation, the correlation coefficient ρ is around 0.15. For the good fit simulation, ρ is around 0.90.

$$\rho = \frac{(\mu_1 - \mu_2) \cdot \sqrt{\frac{N_1 \cdot N_2}{(N_{1+2})^2}}}{\sigma_{N_{1+2}}} \tag{1.6}$$

$$\text{where } \sigma_{N_{1+2}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

The result is a correlation coefficient, ρ, which can take values between "−1" (perfect negative correlation) and "1" (perfect positive correlation), with "0" denoting "no correlation". With respect to our data, a near-ideal correlation between the judgements by the linguists and online judgements means that all the OK-sentences receive high ratings and all the *-sentences low ratings, as illustrated in Figure 1.22b. A bad correlation is illustrated in Figure 1.22a.

There is another advantage to a correlation measure like this: We can check for individual differences (i.e. to which degree individual online participants agree with

the LI authors). We return to this point later-on.

SSA reported a match rate of about 95%; so one would expect a *strong* or *very strong* correlation. However, it is a non-trivial task to define "strong" or "very strong". This is why the primary purpose of this correlation measure is to give us an idea of the effect size and of individual differences amongst participants. In terms of significance testing, the threshold test below is more important.[30]

ANALYSIS 2: THRESHOLD TEST Our second analysis is a threshold test, which also featured in SSA (p. 234). In this test, the items are lined up by their online ratings, similar to the hypothetical ratings in Figure 1.23a and 1.23b. Then, a threshold is defined; items below that threshold ought to be bad ("*") and items above it good (OK). Any *-item above the threshold counts as a violation, similarly, any OK-item below the threshold. We then find the optimal threshold; the optimal threshold takes a value at which the lowest number of violations occur (sometimes, there are several optimal values for the threshold[31]). In our example, Figure 1.23b shows the optimal threshold, i.e. the threshold of the least violations. Figure 1.23a shows a non-optimal threshold: It gives four violations (instead of two violations). The number of violations divided by the total number of items gives a violation percentage. The inverse of the violation percentage is the match rate. In the exam-

[30] We nonetheless define a criterion for the correlation measure (as not setting one might be interpreted as us dodging the question).

In the social sciences, even coefficients of about 0.5 can be interpreted as a strong correlation (among others, cf. Cohen, 1988, Meyer et al., 2001, Hemphill, 2003), whereas in the natural sciences a strong coefficient typically lies above 0.8 (cf. e.g. Walker and Almond, 2010:156), and very strong coefficient can be as high as 0.9 (cf. e.g. Riddiough and Thomas, 1998:128). Given that the social sciences consider questions like "how do post-high school grades relate to job performance" or "how do gender and weight interact" (from Meyer et al., 2001:131) and that our research question is a technical question ("Is Method A highly similar to Method B?"), we use the definitions for the natural sciences. Thus, we consider a correlation coefficient equal to or above 0.80 as evidence for (H_0) and against (H_1) and a correlation coefficient below 0.80 as evidence against (H_0) and as evidence for (H_1) (see Analysis 2 for a reminder of (H_0) and (H_1)).

[31] Further, if the optimal threshold varied from one experiment to another, then syntactic theories should ideally be able to explain why this threshold fluctuates.

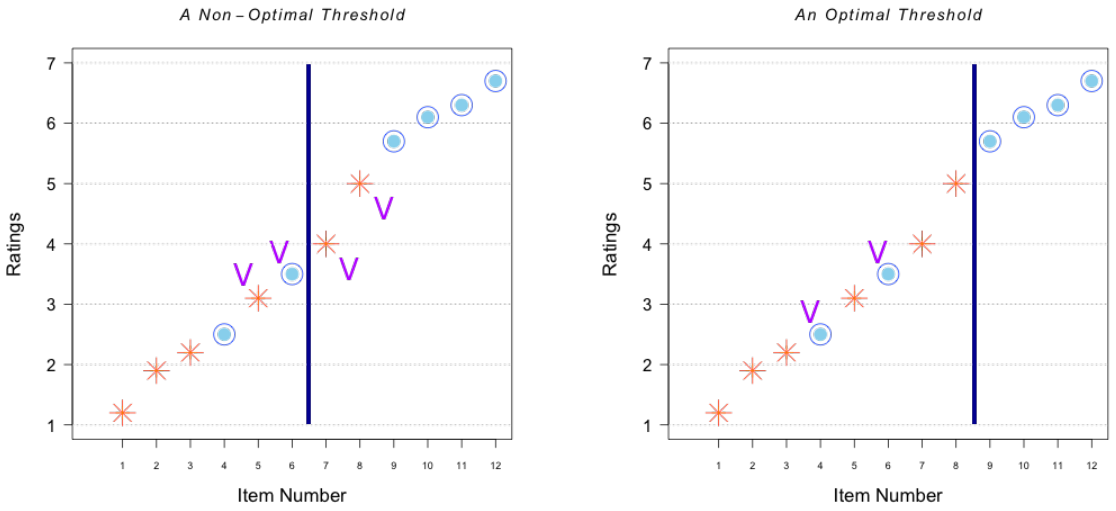


Figure 1.23: Illustrating the threshold analysis on hypothetical data. Items (x-axis) are in an ascending order by their corresponding online ratings (y-axis), including both *-items (red asterisks) and OK-items (blue circles). The (hypothetical) ratings are given on a 7-point scale and the threshold is marked in blue. Items below the threshold ought to be marked, items above it unmarked. Unmarked items below the threshold and marked items above it are violations. In 1.23b (right), there are two violations; this is the optimal threshold. On the left (1.23a), there are four violations; this is an example of an non-optimal threshold.

ple, the optimal case is a match rate of 83% (i.e. the percentage of violations is 17%).

At this point, we can explicate which outcomes will count as evidence for or against (H_0). For the reader's convenience, (H_0) and (H_1) are repeated here.

(H_0) Informal and formal methods concur strongly; i.e. there is no substantial difference between informal judgements (that were done introspectively by LI authors) and formal judgements (collected in an online acceptability judgement task).

(H_1) Results from informal judgements (that were done introspectively by LI authors) and formal judgements (collected in an online acceptability judgement task) do not concur strongly.

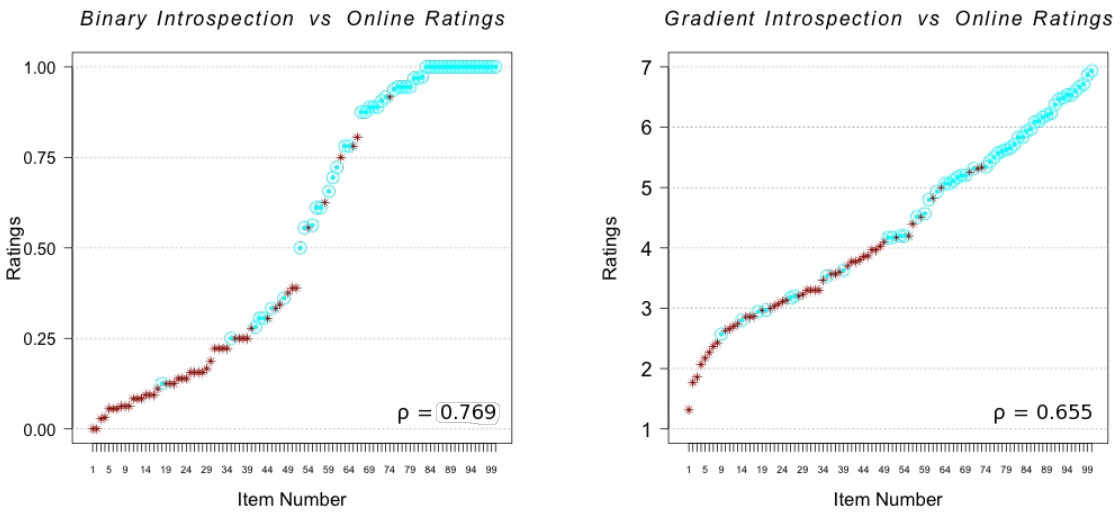


Figure 1.24: Figure 1.24a (left) shows the averaged 2-point ratings for items that come from LI articles with binary judgements. Figure 1.24b (right) shows the averaged 7-point ratings for items that come from articles with gradient judgements. Items (x-axis) are in an ascending order by their corresponding online ratings (y-axis), including both *-items (dark red asterisks) and OK-items (light blue circles).

If informal and formal judgements concur, one could easily expect a match rate of 90% or more, particularly for the gradient condition. This is what we consider as evidence for (H_0). Any value below that is considered as evidence against (H_0) and evidence for (H_1).

1.4.3 Ratings

Figure 1.24a shows the online ratings for the 50 *-items and 50 OK-items, taken from articles with binary author judgements; the online ratings were measured using a 2-point scale. Figure 1.24b shows the same for items from authors whose judgements were gradient; the online ratings were measured using a 7-point scale.

1.4.4 Results

POINT-BISERIAL CORRELATION MEASURE As one would expect, the *-items tend to get low experimental ratings and the OK-items tend to get high ratings. However, the experimental ratings for both the *- and the OK-items cover large parts of the scale. Moreover, for gradiently judged items, the ratings do not cluster at the two endpoints, contrary to what one might expect. In fact, 43 of the gradiently judged items receive a rating between "3" and "5", which makes them "in-between" items (for the binarily judged items, 14 are in-between; for more on this issue, see Question 3 in Section 1.5). As a consequence, the point-biserial correlations are far from being perfect: For items from authors with binary judgements, the coefficient is 0.769; for items from authors with gradient judgements, it is 0.655 (cf. Figure 1.24a and 1.24b, respectively).

William Snyder (personal communication) made the following important point: "(...) different U.S. English speakers can (...) have genuinely different judgements, as a result of regional and idiolectal differences in their grammars. Averaging across many subjects should reduce the effects of noise in individuals' judgements, but it could also have the effect of introducing some noise, as a result of mixing the judgements of systematically different grammars".

William Snyder suggested including individual correlations. Consequently, we took the ratings of each participant individually and then calculated the correlation coefficient. For the binarily judged items, the median correlation coefficient comes out at 0.594 (mean: 0.574, standard deviation: 0.139), for the gradiently judged items, the median correlation coefficient comes out at 0.469 (mean: 0.460, standard deviation: 0.156). So, if anything, taking the "grande" correlation coefficient introduced some noise that made the results look better than they might actually be.[32]

[32]N.B.: The decrease is not entirely unexpected: The correlation coefficient should decrease to

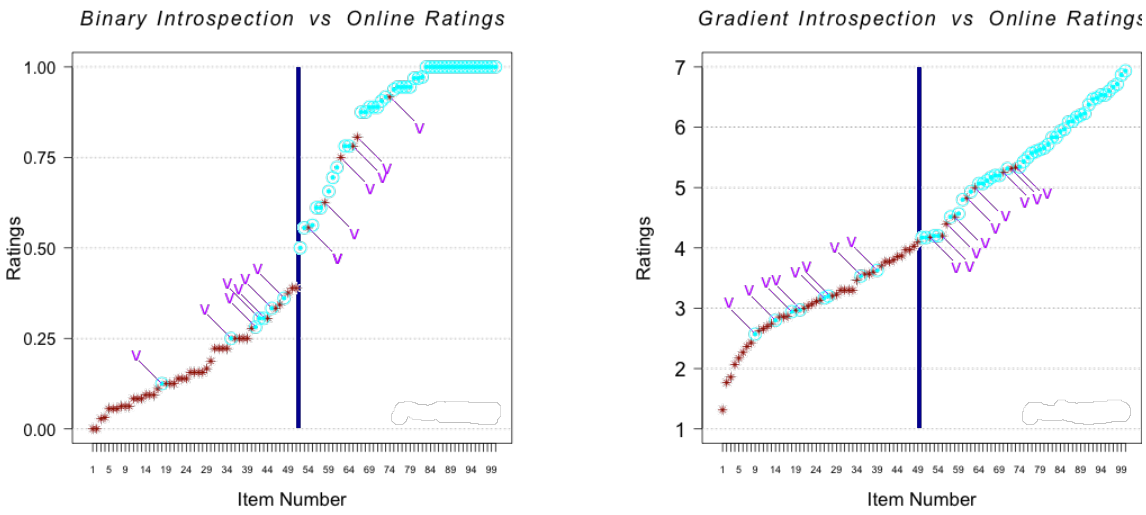


Figure 1.25: Figure 1.25a (left) shows the optimal threshold for items that come from LI articles with binary judgements. Figure 1.25b (right) shows the optimal threshold for items that come from articles with gradient judgements. Items (x-axis) are in an ascending order by their corresponding online ratings (y-axis), including both *-items (dark red asterisks) and OK-items (light blue circles).

THRESHOLD TEST The results for the threshold test point in the same direction. For items from authors with binary judgements, the match rate is 87%; for items from authors with gradient judgements, it is 83%. Figure 1.25a shows the optimal threshold for the binary condition, Figure 1.25b for the gradient condition. For both threshold analyses (binary and gradient), the results are below 90%, which is why we reject (H_0) and accept (H_1) in both cases.

In both analyses, the point-biserial correlation measure and the threshold test, we rejected (H_0) and accepted (H_1), for both the binary and the gradient condition. Thus, we conclude that the results from informal judgements (that were done introspectively by LI authors) and formal judgements (collected in an online acceptability judgement task) do not concur strongly.

some degree, as individual online participants are as susceptible to judgement errors as individual LI authors. Many thanks to Greg Kochanski for pointing this out to us.

1.5 Discussion

A few questions arise. 1) Is a correlation coefficient of 0.769 and 0.655 and a violation count of 13% and 17% in the threshold analysis that bad? 2) Why is the correlation coefficient lower for items that come from LI articles with gradient judgements than for items from articles with binary judgements? 3) What's the deal with all those "in-between" ratings? 4) Could the mismatch between informal and formal results have been caused by "bad" formal results? 5) Could the results be taken as evidence that syntactic research is conducted in sentence pairs after all? 6) Were Sprouse et al. (2013) wrong then? 7) Do these results apply to other subfields of linguistics (i.e. judgement tasks in semantics, phonetics, pragmatics) as well? 8) Does this mean that one should reject informal results in general?

1) *Is a correlation coefficient of 0.769 and 0.655 and a violation count of 13% and 17% in the threshold analysis that bad?* The results are not too bad. However, it does not justify the narrative that informal and formal results concur strongly. As Gibson et al. (2013) point out with respect to SSA's initial results (the 95% match rate): Even such a strong match (i.e. 95%) is not necessarily sufficient for reliable theory building. But our results suggest that the actual match between informal and formal data is considerably lower, which makes a strong case against informal results as the foundation for reliable theory building. This also leads to another question about the reliability of syntactic theories; we will come back to this in Question 3).

2) *Why is the correlation coefficient lower for items that come from LI articles with gradient judgements than for items from articles with binary judgements?* In fact, one might have expected the very opposite. Recall that we are looking at endpoints

(*-items and OK-items). For authors who rate on a binary scale, these two options take up the entire scale; any in-between item will take one of these values (i.e. "*" or OK). Figure 1.26a illustrates this (for hypothetical data). Thus, it is not unreasonable to expect that *-items and OK-items will "mingle" to some extent when it comes to their online ratings. The gradient results could have formed more of an S-shaped curve with a gap between the marked and unmarked items. Given this expectation, the gap would have been filled by somewhat marked/questionable items ("?*", "??", "?", etc.). This would have increased the coefficient in comparison to the binary results. Figure 1.26b illustrates this (also for hypothetical data).[33] We are not entirely certain why the results come out the way they do. This requires further research; but it also leads to the next question.

3) *What's the deal with all those "in-between" ratings?* As mentioned, a lot of the gradiently judged items received an "in-between" rating by the online participants. However, to look into this issue, we need to first answer what "in-between" means. As the gradient condition assumes at least three degrees, we can divide the scale into three parts: [1, 3[(corresponding to "*"), [3, 5] ("?"), and]5, 7] (OK). Assuming these intervals, 43 of our 100 gradiently judged items received a rating corresponding to "?" (i.e. they received online ratings between "3" and "5"). This is illustrated in Figure 1.27.

However, the LI authors judged those 43 "in-between" items as either "*" or as OK. Figure 1.27 shows how strong the mismatch between informal and formal data really is. As argued above, the formal results are extremely reliable. And as the individual correlation coefficients show, this was not caused by individual differences, either. So, how can a mismatch of such a degree be explained? As mentioned, judgement errors, purpose biases, etc. on the linguists' side will have played some role. However,

[33] Many thanks to Tom Wasow for pointing this out to us.

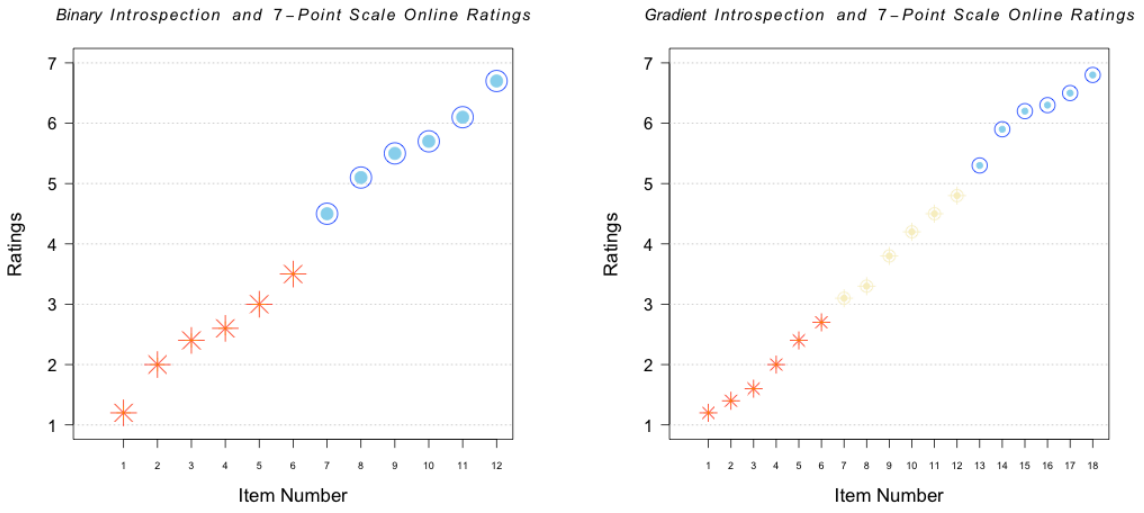


Figure 1.26: On the x-axis, hypothetical items are ordered by their formal ratings (y-axis), all given on a 7-point scale. The author judgements for the items are indicated by red asterisks (marked) and blue circles (unmarked). Left (1.26a): For items that come from LI articles with binary judgements, one would expect that the online ratings for marked and unmarked items form an almost seamless transition. This is because the two values, "*" and OK, take up the entire scale for the ratings given by researcher introspection. Right (1.26b): For the items that come from LI articles with gradient judgements (which are also hypothetical), one might have expected that they form more of an S-shaped curve, leaving a gap between the marked and unmarked items (to fully get this effect, imagine this figure without the orange items). This is because the two values, "*" and OK, take up only parts of the scale: Other values are "in-between" judgements ("?*", "??", "?", etc.), inserted in pale orange. One might expect that these fill a space between introspectively marked and unmarked items.

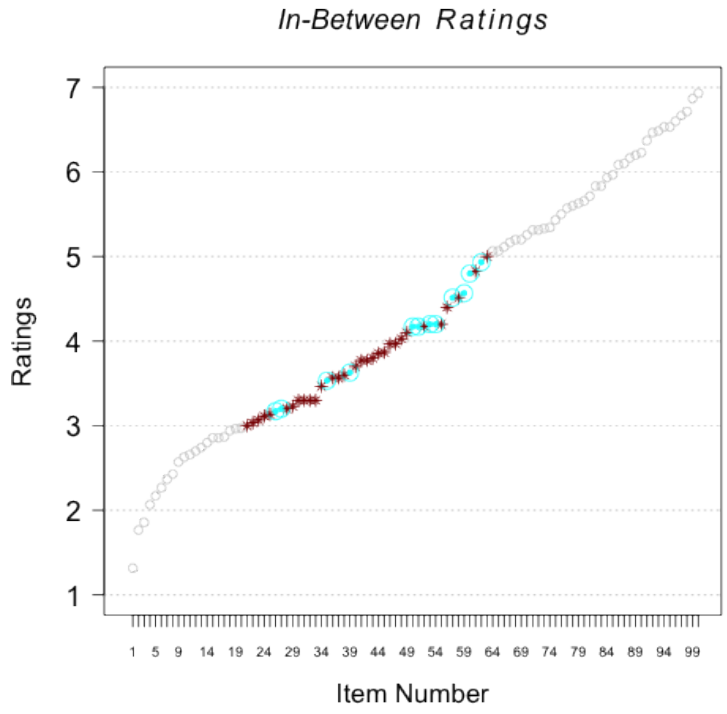


Figure 1.27: Showing the ratings for "in-between" items that come from articles with gradient judgements. Items (x-axis) are in an ascending order by their corresponding online ratings (y-axis), including both *-items (dark red asterisks) and OK-items (light blue circles). This graph highlights the items that received averaged online ratings between "3" and "5". The remaining items are in grey.

it seems unlikely that those sources of error can explain the entire mismatch. In our view, there is another factor contributing to the mismatch: The LI authors struggled to embrace gradience to its full extent (despite the fact that they used various gradient scales).

This also implies that informal data are inadequate as they fail to fully reflect linguistic reality. However, if syntactic theories heavily rely on informal data and these data fail to fully reflect linguistic reality, then how can those theories correctly model this linguistic reality? They can't. Any theory based on such inadequate, idealised data will most likely be empirically inadequate itself.[34]

4) *Could the mismatch between informal and formal results have been caused by "bad" formal results?* We do not think that this is the case: The formal results are extremely reliable, simply because we adhered to experimental standards. If we were to re-run the entire experiments, it is very likely that the results would come out highly similar.

5) *Could the results be taken as evidence that syntactic research is conducted in sentence pairs after all?* The argument would be that the results are underwhelming, because the ratings from researcher introspection were never meant to be a "global" comparison in the first place. Above, however, we argued against the assumption that syntactic research is conducted by only comparing sentence pairs. Further, as mentioned before, even if syntactic research was conducted in such a way, then this is holding the field back. Theory building should consider various syntactic phenomena in a cross comparison and thus, empirical enquiry should happen at large.

[34]Many thanks to Tom Wasow and Greg Kochanski for raising Question 3 and for insisting on its relevance.

6) *Were Sprouse et al. (2013) wrong then?* This question misses the point; no, they are not wrong. However, as argued above, their analysis is, to quite some degree, "blind" towards the effects of scale biases, judgement errors, quantisation errors, and purpose biases. This is because they focus on directionality and on sentence pairs of *-items vs OK-items. Consequently, they only detected the most blatant differences, but did not pick up on the effect size of the actual mismatch between informal and formal results.

7) *Do these results apply to other subfields of linguistics (i.e. judgement tasks in semantics, phonetics, pragmatics) as well?* It would not be unreasonable to assume that this is the case, but further research would have to look into this.

8) *Does this mean that one should reject informal results in general?* No, this would be an over-interpretation of our results and we wish to make two qualifications.

First, we think that Pullum's point for corpus linguistics (Pullum, 2003) also holds for acceptability judgement tasks: If one is concerned with the broad degree of acceptability (i.e. "unacceptable" vs "acceptable") for "clear-cut" cases (e.g. "Pete went home"), one does most likely not need an acceptability judgement task. And for "clearer" cases, an acceptability judgement task with a small number of participants is sufficient (cf. Mahowald et al., submitted).[35] The problem is that many constructions discussed in the literature are concerned with subtle differences in acceptability (e.g. weak/strong island effects), for which formal results are beneficial.[36]

Second, the mismatch between informal and formal methods is based on results from researcher introspection and from experimental results. From this, we cannot make strong conclusions about other informal methods, e.g. informal surveys. We expect

[35] Of course, "clear-cut" and "clearer" are vague concepts and it is for the researcher to determine whether or not these apply.

[36] This point might hold to an even greater extent for semantic acceptability judgements.

that there would be a smaller mismatch between results from informal surveys and results from formal experiments. The reason for this is that in contrast to researcher introspection, informal surveys typically rely on several participants, all of whom are typically oblivious to the purpose of the study.

Hence, our main concern is for those syntacticians who rely in their theory building on subtle differences based on their own judgements. We are less concerned with researchers who rely on informants to investigate the general syntax of e.g. Hiaki.[37]

1.6 Conclusion

The rather moderate correlation coefficients and the weaker than expected threshold analyses, even for the gradient set, suggest that there is a non-trivial mismatch between author judgements and experimental results.

In principle, the LI authors or the participants could be "wrong" in their judgements. However, the results from the online experiment are extremely reliable, due to the high number of participants. By increasing the number of participants to standard levels (e.g. N = 30) and averaging the results, the impact of scale biases (i.e. inter-speaker differences in the application of the scale) and judgement errors will be reduced to a great extent. Scale biases can be further mitigated by then normalising the individual ratings (i.e. using Z-scores).

As for the linguists, while they will have pondered on their judgements with great care, judgement errors and scale biases will be more pronounced, simply because their results are non-aggregated. Further, although giving judgements on a gradient scale, it seems that the linguists struggle to capture the effects of gradience to their full extent. As argued, it is the low number of data points and the failure to capture

[37] Thanks to Heidi Harley for pushing this point.

gradience that causes the mismatch between informal and formal results. This can have potentially damaging consequences, specifically for syntactic theory building, and raises the question: If a theory is built on questionable data, how solid can such a theory be? Increasing the number of participants, i.e. resorting to experimental methods, is an effective way to reduce the effects of the mentioned sources of error. On a theoretical level, this line of argumentation has already been laid out by Schütze (1996). The present study provides quantitative support for Schütze's arguments and shows that formal methods, and formal acceptability judgement tasks in particular, should have a significant place in syntactic theory.

1.7 Chapter 1: Appendices

1.7.1 Background of the Authors

Year	Issue	Article No.	Author(s)	Individual Verdict	Overall Verdict
2001	1	1	M. Babyonyshev J. Ganger D. Pesetsky K. Wexler	US US US US	INCL: US
2001	1	2	P. Boersma B. Hayes	NL US	INCL: US partly
2001	1	3	G. Grewendorf	GER	EXCL: NON-US
2001	1	4	J. Lidz	US	INCL: US
2001	1	5	R. Martin	US	INCL: US
2001	1	6	M. Aronoff S. Cho	US US	INCL: US
2001	1	7	S. Franks Ž. Bošković	US BOS/HER	INCL: US partly
2001	1	8	N. Richards	US	INCL: US
2001	2	1	A. Alexiadou E. Anagnostopoulou	GR GR.	EXCL: NON-US
2001	2	2	B. Bruening	US	INCL: US
2001	2	3	G. Longobardi	IT	EXCL: NON-US
2001	2	4	J. Nunes	US	INCL: US
2001	2	5	C. Boeckx S. Stjepanović	US US	INCL: US
2001	2	6	H. Lasnik	US	INCL: US

2001	2	7	T. Stroik	US	INCL: US
2001	3	1	J. Aoun L. Choueiri N. Hornstein	LEB US/FR CA	INCL: US partly
2001	3	2	G. Fanselow	GER	EXCL: NON-US
2001	3	3	E. Reuland	NL	EXCL: NON-US
2001	3	4	P. W. Culicover R. Jackendoff	US US	INCL: US
2001	3	5	S. Davis Bu. A. Zawaydeh	US US	INCL: US
2001	3	6	L. López	US	INCL: US
2001	3	7	J. Sabel J. Wolfgang	GER GER	EXCL: NON-US
2001	3	8	C. J.-W. Zwart	NL	EXCL: NON-US
2001	4	1	D. Embick R. Noyer	US US	INCL: US
2001	4	2	H. Kishimoto	JAP	EXCL: NON-US
2001	4	3	J. F. Bailyn	US	INCL: US
2001	4	4	L. Burzio	IT	EXCL: NON-US
2001	4	5	K. Hale	US	INCL: US
2001	4	6	L. López	US	INCL: US
2001	4	7	P. Ackema	UK	EXCL: NON-US
2001	4	8	J. C. Johnston I. Park	AUS AUS	EXCL: NON-US
2001	4	9	A. Meinunger	GER	EXCL: NON-US
2002	1	1	C. Collins	US	INCL: US
2002	1	2	M. den Dikken	NL	EXCL: NON-US

			A. Giannakidou	GR	
2002	1	3	D. Fox	US	INCL: US
2002	1	4	K. Johnson	US	INCL: US
2002	1	5	J. Camacho	US	INCL: US
2002	1	6	V. Hill	RO	EXCL: NON-US
2002	1	7	J. T. Runner	US	INCL: US
2002	2	1	J. Bowers	US	INCL: US
2002	2	2	F. Keller A. Asudeh	GER CA	EXCL: NON-US (CA)
2002	2	3	M. Polinsky E. Potsdam	US US	INCL: US
2002	2	4	U. Sauerland P. Elbourne	US UK	INCL: US partly
2002	2	5	M. C. Baker	US	INCL: US
2002	2	6	Ž. Bošković	BOS/HER	EXCL: NON-US
2002	2	7	K. Namai	JAP	EXCL: NON-US
2002	3	1	Ž. Bošković	BOS/HER	EXCL: NON-US
2002	3	2	P. Branigan M. MacKenzie	CA CA	EXCL: NON-US (CA)
2002	3	3	R.-M. Déchaine M. Wiltschko	CA A	EXCL: NON-US (CA)
2002	3	4	J. M. Fitzpatrick	US	INCL: US
2002	3	5	I. Landau	US	INCL: US
2002	3	6	Y. Winter	IS/NL	EXCL: NON-US
2002	3	7	B. Citko	US	INCL: US
2002	3	8	V. Dayal	IND	EXCL: NON-US
2002	3	9	C. Zoll	US	INCL: US

2002	4	1	A. Neeleman H. van de Koot	NL NL	EXCL: NON-US
2002	4	2	D. Takahashi	JAP	EXCL: NON-US
2002	4	3	G. Cinque	IT	EXCL: NON-US
2002	4	4	H. Tanaka	JAP	EXCL: NON-US
2002	4	5	M. Baltin	US	INCL: US
2002	4	6	H. Harley	CA	EXCL: NON-US (CA)
2002	4	7	M. McGinnis	CA	EXCL: NON-US (CA)
2002	4	8	J. Rubach	PL	EXCL: NON-US
2003	1	1	D. Basilico	US	INCL: US
2003	1	2	C. Phillips	UK	EXCL: NON-US
2003	1	3	B. Vaux	US	INCL: US
2003	1	4	A. Simpson T. Bhattacharya	UK IND	EXCL: NON-US
2003	1	5	D. Fox H. Lasnik	US US	INCL: US
2003	1	6	J. A. Legate	US	INCL: US
2003	1	7	X. Liejiong	CN	EXCL: NON-US
2003	2	1	K. von Fintel S. Iatridou	GER US	INCL: US partly
2003	2	2	C. Reiss	US	INCL: US
2003	2	3	C. Zoll	US	INCL: US
2003	2	4	C. Boeckx N. Hornstein	US CA	INCL: US partly
2003	2	5	P. Panagiotidis	GR	EXCL: NON-US
2003	2	6	I. Caponigro C. T. Schütze	IT CA	EXCL: NON-US (CA)

2003	2	7	U. Sauerland	US	INCL: US
2003	2	8	H. Tanaka	JAP	EXCL: NON-US
2003	3	1	D. Adger G. Ramchand	UK US	INCL: US partly
2003	3	2	T. Beasley K. Crosswhite	UK US	INCL: US partly
2003	3	3	V. Carstens	US	INCL: US
2003	3	4	I. De Crousaz U. Shlonsky	CH IS	EXCL: NON-US
2003	3	5	C. Heycock R. Zamparelli	UK IT	EXCL: NON-US
2003	3	6	I. Landau	US	INCL: US
2003	3	7	T. Hirose	JAP	EXCL: NON-US
2003	3	8	J. A. Legate	US	INCL: US
2003	3	9	Y. Takano	JAP	EXCL: NON-US
2003	4	1	Ž. Bošković H. Lasnik	BOS/HER US	INCL: US partly
2003	4	2	S. Chung	US	INCL: US
2003	4	3	J. Rubach	PL	EXCL: NON-US
2003	4	4	E. Aldridge	US	INCL: US
2003	4	5	L. Haegeman	BE	EXCL: NON-US
2003	4	6	H. Lasnik M.-K. Park	US US	INCL: US
2003	4	7	E. J. Rubin	US	INCL: US
2003	4	8	Y. Sharvit	US	INCL: US
2004	1	1	R. Bhatt R. Pancheva	IND US	INCL: US partly

2004	1	2	M. McGinnis	CA	EXCL: NON-US (CA)
2004	1	3	S. Beck K. Johnson	GER US	INCL: US partly
2004	1	4	E. G. Ruys	NL	EXCL: NON-US
2004	1	5	C. Cecchetto R. Oniga	IT IT	EXCL: NON-US
2004	1	6	A. Neeleman K. Szendröi	NL UK	EXCL: NON-US
2004	1	7	H. Rullman	NL	EXCL: NON-US
2004	1	8	Y. Takano	JAP	EXCL: NON-US
2004	2	1	C.-h. Han M. Romero	KOR ES	EXCL: NON-US
2004	2	2	B. Tesar	US	INCL: US
2004	2	3	H. Harley	CA	EXCL: NON-US (CA)
2004	2	4	R. K. Larson F. Marušič	US SLO	INCL: US partly
2004	2	5	J. Ouhalla	UK	EXCL: NON-US
2004	2	6	J. Roodenburg	NL	EXCL: NON-US
2004	2	7	C.-h. Han J.-B. Kim	KOR KOR	EXCL: NON-US
2004	2	8	I. Hazout	IS	EXCL: NON-US
2004	2	9	B. Schwarz	US	INCL: US
2004	3	1	D. Embick	US	INCL: US
2004	3	2	I. Hazout	IS	EXCL: NON-US
2004	3	3	C. Boeckx N. Hornstein	US CA	INCL: US partly
2004	3	4	N. Richards	US	INCL: US

2004	3	5	Y. Aydemir	TUR	EXCL: NON-US
2004	3	6	D. Fox J. Nissenbaum	US US	INCL: US
2004	3	7	S. A. Minkoff	US	INCL: US
2004	3	8	M. L. Rivero	US	INCL: US
2004	3	9	N. Sobin	US	INCL: US
2004	3	10	D. Watson E. Gibson	US US	INCL: US
2004	4	1	A. Cardinaletti U. Shlonsky	IT IS	EXCL: NON-US
2004	4	2	A. Watanabe	JAP	EXCL: NON-US
2004	4	3	Ž. Bošković	BOS/HER	EXCL: NON-US
2004	4	4	M. R. Marlo N. J. Pharris	US US	INCL: US
2004	4	5	J. Rubach	PL	EXCL: NON-US
2004	4	6	M. Akiyama	JAP	EXCL: NON-US
2004	4	7	R. Bhatt A. Joshi	IND IND	EXCL: NON-US
2004	4	8	M. Gordon	US	INCL: US
2004	4	9	L. Haegeman	BE	EXCL: NON-US
2005	1	1	N. Chomsky	US	INCL: US
2005	1	2	M. Halle	LAT	EXCL: NON-US
2005	1	3	I. Oltra-Massuet K. Arregi	ES ES	EXCL: NON-US
2005	1	4	J. Hankamer L. Mikkelsen	US DEN	INCL: US partly
2005	1	5	P. Lasersohn	US	INCL: US

2005	1	6	K. Nishiyama	JAP	EXCL: NON-US
2005	1	7	H. L. Soh	CA	EXCL: NON-US (CA)
2005	1	8	Y. Ueno	JAP	EXCL: NON-US
2005	2	1	D. Adger G. Ramchand	UK US	INCL: US partly
2005	2	2	J. Harris M. Halle	US LAT	INCL: US partly
2005	2	3	D. LaCharité C. Paradis	CA CA	EXCL: NON-US (CA)
2005	2	4	D. Büring	GER	EXCL: NON-US
2005	2	5	A. Ira Nevins	UK	EXCL: NON-US
2005	2	6	C. Collins	US	INCL: US
2005	2	7	J. Lin	TW	EXCL: NON-US
2005	2	8	U. Sauerland	US	INCL: US
2005	3	1	G. Cinque	IT	EXCL: NON-US
2005	3	2	P. Elbourne	UK	EXCL: NON-US
2005	3	3	M. Kural	TUR	EXCL: NON-US
2005	3	4	T. Reinhart T. Siloni	IS IS	EXCL: NON-US
2005	3	5	C. Boeckx N. Hornstein	US CA	INCL: US partly
2005	3	6	E. Cowper	FR	EXCL: NON-US
2005	3	7	P. C. Gordon R. Hendrick	US US	INCL: US
2005	3	8	A. Gualmini S. Crain	IT US	INCL: US partly
2005	4	1	B. Citko	US	INCL: US

2005	4	2	M. den Dikken	NL	EXCL: NON-US
2005	4	3	A. Holmberg	UK	EXCL: NON-US
2005	4	4	A. Rackowski N. Richards	US US	INCL: US
2005	4	5	H. Koopman	NL	EXCL: NON-US
2005	4	6	J.-M. Authier L. Reed	US US	INCL: US
2005	4	7	E. Ritter S. Thomas Rosen	CA US	INCL: US partly
2006	1	1	E. Benmamoun H. Lorimor	MOR US	INCL: US partly
2006	1	2	B. Bruening	US	INCL: US
2006	1	3	C. Clifton, Jr. G. Fanselow L. Frazier	US GER US	INCL: US partly
2006	1	4	O. Matushansky	FR	EXCL: NON-US
2006	1	5	E. Woolford	US	INCL: US
2006	1	6	M. Barrie	CA	EXCL: NON-US (CA)
2006	1	7	E. Benmamoun	MOR	EXCL: NON-US
2006	1	8	C. Boeckx H. Lasnik	US US	INCL: US
2006	1	9	T. Ishii	JAP	EXCL: NON-US
2006	1	10	P. Panagiotidis S. Tsiplakou	GR CYP	EXCL: NON-US
2006	2	1	O. Bat-El	IS	EXCL: NON-US
2006	2	2	E. Thompson	US	INCL: US
2006	2	3	M. de Vries	NL	EXCL: NON-US

2006	2	4	W. J. Idsardi	CA	EXCL: NON-US (CA)
2006	2	5	D. Isac	RO	EXCL: NON-US
2006	2	6	H. A. Sigurdsson	ICE	EXCL: NON-US
2006	2	7	O. Adesola	US	INCL: US
2006	2	8	I. Hirata	JAP	EXCL: NON-US
2006	2	9	M. Kawai	US	INCL: US
2006	2	10	H. Kishimoto	JAP	EXCL: NON-US
2006	2	11	W. P. Lawrence	JAP	EXCL: NON-US
2006	3	1	M. Halle O. Matushansky	LAT FR	EXCL: NON-US
2006	3	2	S. Rosenthall	CA	EXCL: NON-US (CA)
2006	3	3	M. Becker	US	INCL: US
2006	3	4	A. Grosu J. Horvath	US US	INCL: US
2006	3	5	L. Haegeman	BE	EXCL: NON-US
2006	3	6	J. Beavers A. Koontz-Garboden	US US	INCL: US
2006	3	7	M. L. Borroff	US	INCL: US
2006	3	8	Ž. Bošković	BOS/HER	EXCL: NON-US
2006	4	1	G. Chierchia	IT	EXCL: NON-US
2006	4	2	C. Boeckx N. Hornstein	US CA	INCL: US partly
2006	4	3	S. Miyagawa	US	INCL: US
2006	4	4	K. Nakatani	JAP	EXCL: NON-US
2006	4	5	E. Williams	US	INCL: US
2006	4	6	J. van Craenenbroeck M. den Dikken	BE NL	EXCL: NON-US

2006	4	7	C. Dye	US	INCL: US
2006	4	8	H. Nakajima	JAP	EXCL: NON-US
2006	4	9	M. Rezac	FR/ES	EXCL: NON-US
2007	1	1	C.-h. Han J. Lidz J. Musolino	KOR US FR	INCL: US partly
2007	1	2	H. Ko	KOR	EXCL: NON-US
2007	1	3	J. Rubach	PL	EXCL: NON-US
2007	1	4	B. Bruening	US	INCL: US
2007	1	5	J. van Craenenbroeck L. Haegeman	BE BE	EXCL: NON-US
2007	1	6	H. L. Soh	CA	EXCL: NON-US
2007	1	7	T. Toda	JAP	EXCL: NON-US
2007	2	1	R. Folli H. Harley	IT CA	EXCL: NON-US
2007	2	2	B. Hyde	US	INCL: US
2007	2	3	R. Bhatt V. Dayal	IND IND	EXCL: NON-US
2007	2	4	M. den Dikken	NL	EXCL: NON-US
2007	2	5	A. Giorgi	IT	EXCL: NON-US
2007	2	6	D. Adger	UK	EXCL: NON-US
2007	2	7	D. A. de A. Almeida M. Yoshida	BRA JAP	EXCL: NON-US
2007	2	8	Y. A. Haddad	LEB	EXCL: NON-US
2007	2	9	C.-h. Han C. Lee	KOR KOR	EXCL: NON-US
2007	2	10	G. Ó. Hansson	ICE	EXCL: NON-US

2007	2	11	N. Hornstein	CA	EXCL: NON-US (CA)
2007	3	1	C. Agüero-Bautista	CA	EXCL: NON-US (CA)
2007	3	2	K. von Fintel S. Iatridou	GER US	INCL: US partly
2007	3	3	I. Landau	US	INCL: US
2007	3	4	J. Aoun J. Nunes	LEB US	INCL: US
2007	3	5	B. Haddican	US	INCL: US
2007	3	6	T. Hirose	CA	EXCL: NON-US (CA)
2007	3	7	A. Meinunger	GER	EXCL: NON-US
2007	3	8	M. D. Richards	UK	EXCL: NON-US
2007	3	9	J. Sprouse	US	INCL: US
2007	3	10	R. Sybesma	NL	EXCL: NON-US
2007	4	1	Ž. Bošković	BOS/HER	EXCL: NON-US
2007	4	2	J. Mascaró	ES	EXCL: NON-US
2007	4	3	S. Miyagawa K. Arikawa	US US	INCL: US
2007	4	4	A. Neeleman K. Szendröi	NL UK	EXCL: NON-US
2007	4	5	W. Zonneveld	NL	EXCL: NON-US
2007	4	6	K. Flack	US	INCL: US
2007	4	7	M. Gouskova	US	INCL: US
2007	4	8	D. Kallulli	ALB	EXCL: NON-US
2008	1	1	D. Embick A. Marantz	US US	INCL: US
2008	1	2	J. A. Legate	US	INCL: US
2008	1	3	A. Darzi	IRAN	EXCL: NON-US

2008	1	4	I. Hazout	IS	EXCL: NON-US
2008	1	5	J. E. MacDonald	US	INCL: US
2008	1	6	N. Sobin	US	INCL: US
2008	1	7	L. McNally	US	INCL: US
2008	1	8	J. Merchant	US	INCL: US
2008	1	9	S. Stjepanović	US	INCL: US
2008	2	1	P. Elbourne	UK	EXCL: NON-US
2008	2	2	A. Idrissi J.-F. Prunet R. Béland	UAE CA CA	EXCL: NON-US (CA)
2008	2	3	B. Copley	US	INCL: US
2008	2	4	M. Gračanin-Yuksek	CRO	EXCL: NON-US
2008	2	5	K.-s. Kim	KOR	EXCL: NON-US
2008	2	6	D. Takahashi	JAP	EXCL: NON-US
2008	2	7	J. Gajewski	US	INCL: US
2008	2	8	J. Pater	CA	EXCL: NON-US (CA)
2008	2	9	R. Vermeulen	UK	EXCL: NON-US
2008	3	1	N. Friedmann G. Taranto L. P. Shapiro D. Swinney	IS US US US	INCL: US partly
2008	3	2	B. Hayes C. Wilson	US US	INCL: US
2008	3	3	K. É. Kiss	HUN	EXCL: NON-US
2008	3	4	R. D'Alessandro I. Roberts	IT UK	EXCL: NON-US
2008	3	5	S. Madigan	US	INCL: US

2008	3	6	J. Sabbagh	US	INCL: US
2008	3	7	S. Wurmbrand	GER	EXCL: NON-US
2008	4	1	A. Cardinaletti L. Repetti	IT US	INCL: US partly
2008	4	2	H. de Hoop A. L. Malchukov	NL RU	EXCL: NON-US
2008	4	3	R. Kennedy	CA	EXCL: NON-US (CA)
2008	4	4	M. C. Baker	US	INCL: US
2008	4	5	B. Citko	US	INCL: US
2008	4	6	M. Romero	ES	EXCL: NON-US
2008	4	7	G. Bouma H. de Hoop	SWE NL	EXCL: NON-US
2008	4	8	B. Spector	FR	EXCL: NON-US
2008	4	9	J. Sprouse	US	INCL: US
2009	1	1	E. O. Aboh	NL	EXCL: NON-US
2009	1	2	S. Béjar M. Rezac	CA FR	EXCL: NON-US (CA)
2009	1	3	F. Heck	GER	EXCL: NON-US
2009	1	4	J. D. Bobaljik I. Landau	CA US	INCL: US partly
2009	1	5	V. Gribanova	US	INCL: US
2009	1	6	I. Caponigro L. Pearl	IT US	INCL: US partly
2009	1	7	J. Coon	US	INCL: US
2009	1	8	A. Stepanov P. Stateva	SLO SLO	EXCL: NON-US
2009	2	1	A. Kratzer	GER	EXCL: NON-US

2009	2	2	L. López	US	INCL: US
2009	2	3	J. Heinz	US	INCL: US
			G. M. Kobele	US	
			J. Riggle	US	
2009	2	4	K. Johnson	US	INCL: US
2009	2	5	J. Sprouse	US	INCL: US
2009	2	6	I. Landau	US	INCL: US
2009	2	7	W. Lechner	A	EXCL: NON-US
2009	2	8	C. Potts	US	INCL: US (partly)
			A. Asudeh	...	
			S, Cable, Y. Hara	...	
			E. McCready	...	
			L. Alonso-Ovalle	...	
			R. Bhatt, C. Davis	...	
			A. Kratzer, T. Roeper	...	
			M. Walkow	...	
2009	3	1	S. A. Kripke	US	INCL: US
2009	3	2	S. Takahashi	JAP	INCL: US partly
			S. Hulsey	US	
2009	3	3	B. Bruening	US	INCL: US
2009	3	4	A. Conroy	US	INCL: US
			E. Takahashi	US	
			J. Lidz,	US	
			C. Phillips	UK	
2009	3	5	Y. Wu	UK	EXCL: NON-US
			A. Bodomo	CN	
2009	3	6	J. D. Haugen	US	INCL: US

2009	3	7	R. Truswell E. Titov	UK UK	EXCL: NON-US
2009	3	8	R. Truswell	UK	EXCL: NON-US
2009	4	1	G. Hicks	UK	EXCL: NON-US
2009	4	2	E.-S. Kim D. Pulleyblank	? NIG	EXCL: NON-US
2009	4	3	O. Preminger	IS	EXCL: NON-US
2009	4	4	P. Boersma	NL	EXCL: NON-US
2009	4	5	A. Williams	US	INCL: US
2009	4	6	A. Kornai	HU	EXCL: NON-US
2009	4	7	I. Sichel	IS	EXCL: NON-US
2010	1	1	E. Manetta	US	INCL: US
2010	1	2	G. Müller	GER	EXCL: NON-US
2010	1	3	Y. Takano	JAP	EXCL: NON-US
2010	1	4	C. Boeckx N. Hornstein J. Nunes	US CA US	INCL: US partly
2010	1	5	N. Sobin	US	INCL: US
2010	1	6	P. Jurgec	CA	EXCL: NON-US
2010	1	7	F. Niinuma	US	INCL: US
2010	1	8	R. Walker	CA	EXCL: NON-US (CA)
2010	2	1	L. Haegeman T. Lohndal	BE NOR	EXCL: NON-US
2010	2	2	H. Ko T. Ionin K. Wexler	KOR US US	INCL: US partly
2010	2	3	M. Lahrouchi	MOR	EXCL: NON-US

2010	2	4	B, Bruening	US	INCL: US
2010	2	5	J.-. Lin	TW	EXCL: NON-US
2010	2	6	M. Baltin	US	INCL: US
2010	2	7	B. Wiland	PL	EXCL: NON-US
2010	2	8	M. Yoshida	JAP	EXCL: NON-US
2010	3	1	I. Landau	US	INCL: US
2010	3	2	T. McFadden A. Alexiadou	US GR	INCL: US partly
2010	3	3	D. Pescarini	IT	EXCL: NON-US
2010	3	4	E. Torrego	ES	EXCL: NON-US
2010	3	5	I. Hazout	IS	EXCL: NON-US
2010	3	6	F. Costantini	IT	EXCL: NON-US
2010	3	7	M. Rezac	FR	EXCL: NON-US
2010	3	8	L. Vicente	ES	EXCL: NON-US
2010	4	1	B. Bruening	US	INCL: US
2010	4	2	S. Cable	US	INCL: US
2010	4	3	L. Haegeman	BE	EXCL: NON-US
2010	4	4	Jeffrey Heinz	US	INCL: US
2010	4	5	M. Koizumi K. Tamaoka	JAP JAP	EXCL: NON-US
2010	4	6	C. Clifton, Jr. L. Frazier	US US	INCL: US
2010	4	7	A. Drummond N. Hornstein H. Lasnik	UK CA US	INCL: US partly
2010	4	8	E. Keshet	US	INCL: US
2010	4	9	R. K. Larson	US	INCL: US

2010	4	10	M. L. Rivero A. Arregui E. Frackowiak	US BRA CA	INCL: US partly

1.7.2 Experimental Stimuli

		Binary Marked Items
LI Issue	No. in Paper	Sentence
32.2.4	48b	Mary drove Rio and John flew to Sao Paulo.
32.2.4	FN35ib	Which books about herself did John file before Mary read?
32.2.7	4b	Max may have been studying, but Mo may have done so too.
32.2.7	Fn_9_i_x	Mary likes Sam, and Chris does so too.
33.2.1	20b	How could there possibly such a serious misunderstanding arise?
33.2.1	49c	Arrested by the police are believed there to have been several linguists.
33.2.1	7e_1	It torrentially rained.
33.2.1	7a_4	John rolled perfectly the ball.
33.2.1	Fn10_4	It was to leave that John persuaded Mary.
33.2.1	Fn17	What seems is that John is sick.
33.3.4	23a	Where do you wonder John saw what?
34.2.4	13b	It was attempted to conceal oneself.
35.1.3	12b	Who did you believe a friend of satisfied?
36.3.8	10b	The boy who did not major in linguistics learned any Romance language.
37.3.6	6d_1	The doctor saw himself from Houston.

39.1.8	11_2	Many of them have turned in their take-home already, but they haven't yet all their paper.
39.1.8	4a	Some brought roses, and lilies were by others.
39.1.8	4d	Beth's mother invited more people to her wedding than were by Beth herself!
40.2.6	6a_1	We thought about John that something terrible had happened.
41.3.1	11a2	Rice was bought for John at the same time.
41.3.1	32b	Mary hated it when we said to Jim to behave herself.
41.3.1	38c_1	Bill found it embarrassing to him to discuss sex.
41.3.1	FN15ii1_1	Mary convinced of her innocence.
41.4.2	36b	I wonder pictures of whom John bought?
41.4.2	50c	A how big party will you throw?
32.2.4	70b	Who filed which report without reading?
32.2.4	FN6ib	There seem cats to be in the garden.
32.2.7	5	Chris has left already, and Pat has done too.
33.1.4	32	What has Betsy purchased and Sally will talk about?
33.2.1	31b_1	There might seem mice to be in the cupboard.
33.2.1	68b	This bureaucrat bribes easily to avoid the draft.
33.2.1	7a_3	John rolled perfectly the ball down the hill.
33.2.1	7d_1	Mary raucously laughed.
33.2.1	Fn14_4	There hit the wall a car.
33.3.4	22b	What did you say that who bought?
33.3.4	23b	What do you wonder where John saw?
34.2.4	14c_1	John said that Bill attempted to sneak each other into the party.
36.3.5	FN2ii	I assure you John to be the best.
36.3.8	4	Is the man who beating the donkey is mean?

37.3.6	6d_2	The doctor saw himself who stopped by last week.
39.1.8	3a	Roses were brought by some, and others did lilies.
39.1.8	4c	Laypeople respect Hundertwasser's work more than his ideas are by architects.
39.3.1	16	The good little monkey clapped silly.
41.3.1	10b	The game was played shoeless.
41.3.1	11b2	John and Mary heard that rice was bought at the same time.
41.3.1	38a_1	John finds it amusing to him to watch the prisoners suffer.
41.3.1	6b_1	John ate raw.
41.4.2	34a	A fish that is how big do you want?
41.4.2	3b2	Whose sisters is coming to your party?
32.2.4	9	Was kissed.

		Binary Unmarked Items
LI Issue	No. in Paper	Sentence
32.1.1	15a	There is a boy in the house.
32.1.8	8a	Susan gave a goldfish to Mary.
32.2.6	4b	You will believe Bob.
32.2.7	3d	Chris should leave soon, and Sam should do so too.
32.2.7	5	Chris has left already, and Pat has done so too.
32.4.6	4b	We spoke to someone.
33.2.1	18f	There were several books on the table.
33.2.1	21c_1	It seems that John is sick.
33.2.1	44a	A crow was sitting on the fence.
33.2.1	49c	Arrested by the police are believed to have been several linguists.

33.2.1	56a	Mary watered the tulips flat.
33.2.1	68a	This bureaucrat was bribed to avoid the draft.
33.2.1	75_10	I know a bureaucrat bribed to avoid the draft.
33.2.1	Fn10_1	It is to drive racing cars that John prefers.
33.3.4	18b	Who would John think that Bill could see?
35.3.9	1b_1	There are frogs in the pond.
36.3.5	3a	I sent the package to John.
37.3.6	6a	I borrowed your red jacket from Macy's that Sandy bought for you.
39.2.1	78	Neither of them can, but Bob wants to sail round the world and Alice wants to climb Kilimanjaro.
39.3.1	12	Simon rolled the ball.
41.3.1	3b1	It is impossible for me to be visited.
41.3.1	46b1	This proposal deserves reconsideration for resubmission.
41.3.1	FN15ii2_1	Mary convinced Bill of her innocence.
41.4.2	FN14b	I've met the man whose book you read.
37.2.7	9a	Who bought what?
32.1.8	12	Mary got a goldfish from Susan.
32.2.4	67a	This is the book which I was given by Ted after reading.
32.2.6	8	Mary said she can't swim, even though she really can swim.
32.2.7	4a	Max may have been studying, but Mo may have been doing so too.
32.2.7	Fn_9_ii_b	To like your parents is often difficult, but to do so as a teenager can be especially difficult.
33.2.1	18b	There appeared a ghastly face at the window.
33.2.1	19a	There will be someone in the garden.
33.2.1	42	Down the hill will roll the cart.

33.2.1	49a	Sitting on the fence seems to be a crow.
33.2.1	49d	I expect on the table to be a pile of books.
33.2.1	59a	The ball was thrown perfectly.
33.2.1	75_1	I saw a bureaucrat bribed to avoid the draft.
33.2.1	7d_2	Mary laughed raucously.
33.2.1	Fn36_2	This bureaucrat got himself bribed deliberately.
34.2.4	14a_1	John approved Bill's initial attempts to sneak each other into the party.
35.3.9	3a_1	There is a frog and some fish in the pond.
37.1.5	39	The key unlocked the door.
37.3.6	7a	Mary had her office painted, and Jane had hers re-modeled.
39.2.1	96	John read every book that Bill did.
39.3.1	13	The ball was rolled by Simon.
41.3.1	40b2	Bill found out it was embarrassing to him to discuss sex.
41.3.1	9a	It was decided to leave.
41.4.2	49	In what sense was he a doctor?
33.2.1	21b_1	It rained.
33.2.1	18g	There are many people skating on the lake this winter.

		Gradient Marked Items
LI Issue	No. in Paper	Sentence
32.3.4	34e_1	John offered Susan to leave.
32.1.5	91b_1	John appears to hit Bill right now.
32.3.4	22a_2	John told Sue to wash oneself.
32.3.4	22a_4	John told Sue to wash themselves.
32.3.4	22b_3	John told Sue when to wash himself.

32.3.4	34e_3	John pledged Susan to leave.
32.3.4	40c_6	John gave Susan some kind of order to take care of himself.
32.3.4	46c_3	John pleaded to take care of oneself.
32.3.4	49a_ii_2	John pleaded with Sally to be allowed to take care of herself.
32.4.6	16b	Sue estimated Bill.
33.2.4	25b	There are a cat and a dog in the yard.
34.1.5	26	She said that a biography of one of the Marx brothers is going to be published this year, but I don't remember which she did.
34.3.6	24c_2	John's friends think it is illegal to feed himself.
34.4.6	1a_2	Which Marx brother did she say a biography of will appear this year?
35.2.4	45b_1	Rose saw every taller man than my father.
35.3.6	14a_2	I asked John how many ideas about himself Mary is likely to have.
35.3.7	15b	Joshua wants itself to destroy the machine.
36.4.1	34a	I wonder who took what from Mary and gave a book to Jeremy.
37.3.4	27a_8	John is at most as much as Bill is tall.
37.3.7	17b	John is not very good the student.
37.3.7	5a_2	Mugsy Boags was very tall a basketball player.
38.3.2	20_2	If you want good cheese, you can only go to the North End.
38.3.3	33b_1	October 1st, he came back.
40.3.2	46b	Mary discovered the book about himself yesterday that Bob wrote.
40.3.2	FN18iiib	I saw the yesterday picture of himself that John liked.
32.1.5	16b	It seems to Naomi to have solved the problem.

32.1.5	FN42iib	John believed Mary to hit Bill.
32.3.4	22a_3	John told Sue to wash himself.
32.3.4	22a_8	John told Sue for Harry to wash himself.
32.3.4	29a_15	Helen pleased Bernie after compromising oneself.
32.3.4	35c_2	John is obligated to Susan to take care of herself.
32.3.4	41c_6	John said to take care of himself.
32.3.4	48a_2	John beseeched for Harriet to leave.
32.4.6	10c	We proved to the authorities Smith to be the thief.
33.1.3	60a_2	I read something yesterday John had recommended.
33.2.4	26d	There were a committee holding a meeting in here.
34.3.6	14b	We urged John's friends to talk about himself.
34.3.6 & 35.3.3	5a & 6a	John was hoped to win the game.
34.4.6	1b_1	Which Marx brother did she say that a biographer of interviewed her?
35.3.1	62b	John pounded the yam yesterday to a very fine and juicy pulp.
35.3.7	13a_1	Those stories about Sarai caused herself to become notorious.
35.3.7	FN3ib	The woman who took a picture of itself hit the Hope diamond.
37.3.4	24a_1	John is less than Bill is fit tall.
37.3.4	33a_4	John has at most as many as Bill has houses.
37.3.7	17c	John and Mary are not very good some students.
38.3.2	17_1	If you want good cheese, you only ought go to the North End.
38.3.3	31b_1	An hour, they slept, and then went to work.
38.3.3	5b	John lives a town in Canada.
40.3.2	55b	I will buy whichever books tomorrow that John likes.
32.1.5	FN42iia	John believes Mary to hit Bill.

		Gradient Unmarked Items
LI Issue	No. in Paper	Sentence
34.3.1	35b	The teacher is Jenny.
32.1.5	42b	John proved that Mary is sick.
32.1.5	86	There will arrive a man tomorrow.
32.3.4	12f_1	Vera left the party in order for Fred not to get embarrassed.
32.3.4	19d	Sally tried to seduce Stuart, and Liz did the same thing with Dan.
32.3.4	49a_ii_1	John pleaded with Sally to be allowed to take care of himself.
32.3.4	4c_1	Tabs tended to be kept on Bill.
33.1.7	4g	John took every picture that Bill did.
33.2.4	26a	A committee was holding a meeting in here.
33.2.4	Fn 4_i_a	Everyone isn't here yet.
34.1.1	96b	The children are almost all sleeping.
34.2.1	29	Who thinks that Susan talked with who?
34.4.1	7c	At that time, what did they believe that Peter fixed?
34.4.6	7b	Every investigator of one of these languages seems to his supervisor to be brilliant, but I won't tell you which of the languages.
35.2.4	45a2	Alice met no man taller than my father.
35.2.4	5a1	The visible stars include Capella, Betelgeuse, and Sirius.
35.3.6	20b1	I asked John how many books about himself Mary thinks are in the library.
37.3.3	15b3	John turns out to be winning.
37.3.3	4a2	I prefer for Sam to do the dishes.
37.3.7	13b	That's not that big of a deal.

38.3.2	15c	The skies need only to darken a little bit and my dog runs under the table.
39.1.6	5	Mary is eating snails, but Bill could never do so.
40.1.6	19b3	How did Lily sleep?
40.1.6	25a1	Lily won't marry who the king chooses.
41.1.7	13	Which paper did you file without reading?
40.1.5	47	Which man did you persuade to read which book?
32.1.5	48a	Lasnik's class was canceled but DP Saito's will be offered.
32.1.5	FN19ib	Jenny remembered bringing the wine.
32.3.4	19c_1	John tried to win and Bill did likewise.
32.3.4	41a_2	John shouted for Harriet to leave.
32.3.4	49a_iii_1	John prayed to Athena to be allowed to take care of himself.
33.1.7	10a	The picture of himself in Newsweek dominated John's thoughts.
33.2.4	11a	One translator each is likely to be assigned to the athletes.
33.2.4	27a	Was there a team drinking each other under the table?
34.1.1	34a	Right now, I know how to solve these kinds of problems but in a few days I won't.
34.2.1	25	Who must Bill have said that Susan married?
34.4.1	6b	How did they believe, and Mary claim, that Peter had murdered John?
34.4.1	FN6iiic	John hasn't, but Bill may be, questioning our motives.
35.1.1	FN6iia2	John is much taller than Mary than Bill is.
35.2.4	45a3	Alice met each man taller than my father.
35.3.1	8	This door was built closed.
36.4.6	1a2	The careful track that she's keeping of her expenses pleases me.

37.3.3	17b1	Max continued to write a paper.
37.3.4	16a	John is taller than Bill is.
37.3.7	5a1	Mugsy Boags wasn't very tall a basketball player.
38.3.3	FN10ii	Where she found this wreck of a car is unclear.
39.4.5	2a	John, my best friend, is here.
40.1.6	23acf9	Jack came after the person he is in love with.
40.2.1	16b	I let myself grow quiet.
32.1.5	89a	Kim seems to be intelligent.

Chapter 2

Textual and Auditory Stimuli in Acceptability Judgement Tasks

2.1 Introduction[1]

Syntacticians are increasingly embracing the distinction between spoken and written language (cf. e.g. O'Donnell, 1974, Tannen, 1982, Chafe, 1992, Miller and Weinert, 1998, Leech, 2000), based on the observation that certain syntactic constructions occur primarily in spoken language ($S - W$ in Figure 2.1) and others primarily in written language ($W - S$) (however, the majority of constructions occur in either modality, $S \cap W$).[2]

Differences and similarities between $S - W$ and $W - S$ matter to various linguistic subfields. Wasow et al. (to appear) is an example from theoretical syntax: By com-

[1]This chapter is based on a previous paper (Juzek, unpublished manuscript). The main difference to the previous paper is that instead of using separate t-tests and TOSTs for the three experiments, we now analyse all experiments with one linear mixed effect model.

[2]Our working definition of a construction is as follows: A construction is "[a]ny pattern, at whatever level of generality, in which units are connected in syntax" (from Matthews, 2007:75). Under this definition, most sentences comprise several constructions. For instance, in "Pete looked for his friend", the basic units "his" and "friend" are part of an NP-construction, and the NP is part of the PP-construction "for + NP", etc.

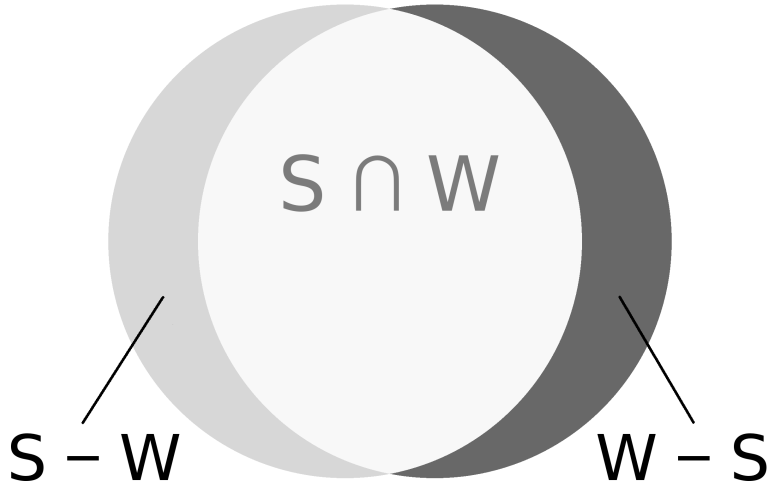


Figure 2.1: An illustration of the types of modality a construction can be part of. If a construction is a member of $S - W$, then it primarily occurs in spoken language. If it is a member of $W - S$, then it occurs primarily in written language. Constructions that are member of the intersection occur in both spoken and written language to a similar degree. This distinction should be understood in a probabilistic sense: Constructions in $S - W$ can be used in written language and constructions in $W - S$ can be used in spoken language, but people rarely do so.

paring modalities, they found that a speaker's choice of whether or not to add the optional "to" in the Do-Be construction (e.g. "All he wanted to do is (to) enjoy an ice cold Fanta") also depends on prosodic factors.[3] Other examples - to mention just two - include research from sociolinguistics, e.g. Biber and Gray (2011), and applied linguistics, e.g. Vandergriff (2005). Biber and Gray (2011) compared the grammatical complexity of conversation and academic writing and found that "conversation is structurally complex and elaborated, to an even greater extent than academic writing for some grammatical features" (ibid.:59). The results of Biber and Gray's cross-modality comparison challenge the stereotypical view that academic writing is more complex than conversation. Vandergriff (2005) discusses how educational policy for teaching German should deal with subordinate clauses introduced by the subconjuction "Weil". The use of "Weil"-subordinate clauses seems to be changing

[3]Interestingly, Wasow et al. also observed that to some degree, prosodic factors even affected the results in written language. They interpreted this as support for Fodor's Implicit Prosody Hypothesis, i.e. the claim that speakers employ prosody even when reading silently (cf. Fodor, 1998).

in spoken but not in written language, viz. from verb final to verb second.[4]

2.1.1 Investigating Modality

Three types of questions surround the issue of modality: observational issues (for any given syntactic construction C: Is it member of $S - W$, $W - S$, or $S \cap W$?), theoretical issues (where do the differences in modality, i.e. $S \Delta W$, stem from?), and methodological questions (which methods can be used to determine to which set a given construction C belongs?). Our investigation focuses on methodological questions; but we will touch on observational and theoretical issues *en passe*.

Arguably, the most common method used to look into differences in modality is a corpus analysis, as corpora typically declare their source. There are corpora of spoken language (e.g. the Fisher Corpus, cf. Cieri et al., 2004), solely written ones (e.g. the New York Times Archive[5]), and hybrid corpora, which include both spoken and written items and specify an item's modality accordingly (e.g. COCA, cf. Davies, 2008-present).

Another prevalent method in experimental syntax is acceptability judgement tasks. In an acceptability judgement task (AJT), the researcher asks a group of native speakers how acceptable or natural a certain construction is (cf. Schütze, 1996, and Cowart, 1997). The native speakers express their judgements through ratings on various scales (cf. e.g. Bard et al., 1996, or Weskott and Fanselow, 2011, for details on and a comparison of various scales). AJTs were introduced to counter the practice of researcher introspection (in researcher introspection, the linguist judges the construction in question him-/herself).[6] Further, with the emergence

[4]Verb final: "Ich mag Fanta, weil sie so süßschmeckt" ("I like Fanta, because it so sweet tastes"); Verb second: "Ich mag Fanta, weil sie schmeckt so süß" ("I like Fanta, because it tastes so sweet").

[5]http://www.nytimes.com/content/help/search/archives/archives.html

[6]Among others, Wasow and Arnold (2005) made a strong case against introspection as the primary method of syntactic enquiry. However, some researchers argue against the need for AJTs, e.g. Sprouse et al. (2013). For further details and a defence of AJTs, also see Chapter 1.

of the internet, it became increasingly easy to conduct AJTs (see Gibson et al., 2011, for an introduction to online AJTs). Arguably, because of their advantages over researcher introspection and because of the possibility of running them online, AJTs are more popular than ever before.

Interestingly, acceptability judgement tasks are rarely used to investigate questions of modality (notable exceptions are e.g. Ferreira and Swets, 2005, on resumptive pronouns, Bader, unpublished manuscript, on detecting case errors and related effects of garden path sentences, and Frazier, 2012, working on a theory that describes human interpretation of natural language). It was this absence of auditory AJTs that led us to a more fundamental question:

(RQ) Can acceptability judgement tasks be used to determine a construction's modality?

(RQ) motivates this research, and to fully appreciate its relevance, it is useful to recall why syntacticians use AJTs in the first place. Syntacticians are interested in the grammaticality of different syntactic constructions. There are several indicators of grammaticality and most syntacticians would consider relative frequency[7] as one of them (as measured through a corpus analysis).[8] Acceptability (measured through AJTs) is commonly considered as another indicator of grammaticality. Frequency and acceptability often coincide (i.e. speakers tend to like what they say and tend not to say what they do not like, cf. Sprouse, 2013), but this is not always the case: Sometimes speakers say things they do not like (as in the case of resumptive pronouns, cf. Ferreira and Swets, 2005) or they like things they do not say. Those cases of divergence are of increased interest to syntacticians (for instance, Asudeh,

[7]Such frequency is relative to the frequencies of other corresponding syntactic phenomena.

[8]As Ash Asudeh points out, one should distinguish between *low* and *high* frequency. High frequency can be seen as an indication of grammaticality; low frequency, however, is not necessarily an indication of ungrammaticality.

2011, argues that the reason why resumptive pronouns are being produced but not accepted is that language production proceeds incrementally and locally, but language perception is more global). But to reveal cases of divergence, both a corpus analysis and an acceptability judgement experiment are required.

Related to the last case (speakers like things they do not say) is another advantage of AJTs: the problem of negative evidence. Certain constructions are of strong interest to syntacticians, but their grammaticality can be hard to determine, as they rarely occur in prevalent corpora, for instance long distance dependency constructions with non-bridging verbs ("What did she say that Fred had done?" vs "What did she simper that Fred had done?"; cf. Erteschik-Shir, 1973; these examples are from Erteschik-Shir, 2005:1). This is where AJTs come in handy: Syntacticians can still survey the acceptability of different rare constructions.

Cases where frequency and acceptability diverge and cases where there is a need to investigate rare constructions might be relevant in the context of modality, as well. For instance, it is conceivable that certain constructions only diverge in frequency and acceptability in spoken language, but not in written language; or that there are constructions that occur only in written language, but we still want to learn more about their acceptability in spoken language.

To investigate the matter further, there are two factors that we need to control for: a) the type of construction and b) the mode of presentation. a) requires that we choose constructions that differ in modality. In terms of Figure 2.1, we should choose constructions that are members either of $S - W$ or $W - S$, but not of the intersection of S and W. b) requires that we compare different modes of presentation in the actual judgement task (i.e. we need to distribute both an auditory and a textual questionnaire).

A construction's modality and the mode of presentation in a questionnaire can interact in several ways, two of which are of particular interest. One possibility is that spoken constructions might receive substantially higher acceptability ratings in auditory judgement tasks than they do in textual tasks; and similarly, written constructions might receive substantially higher ratings in textual tasks than they do in auditory ones. Another possibility is that the mode of presentation has no substantial effect on the ratings of either type of construction.

If modality and mode affect acceptability, then this would have three major consequences. First, if grammatical acceptability judgements are susceptible to issues of modality, it would raise the question whether grammaticality also depends on modality and whether syntactic theories need to model this.

Second, when looking into spoken or written constructions, syntacticians would have to pay special attention when choosing the right type of questionnaire (i.e. whether the linguist will use an auditory or textual questionnaire). This is because the "wrong" questionnaire might not be able to detect true differences.

Thirdly, AJTs could be used to determine the modal preference of a given construction C, i.e. to determine whether it belongs to either $S - W$, $W - S$, or $S \cap W$ in terms of Figure 2.1. If C receives significantly higher ratings in an auditory questionnaire than in a textual questionnaire, then it would be part of $S - W$. If the reverse is the case, then it would be part of $W - S$. If no substantial difference was observed, then it would be part of $S \cap W$. For constructions that actually occur often enough to be captured in a corpus, a corpus analysis might be an easier way to classify C, but for rare constructions, AJTs have the edge.

On the other hand, if 2) was the case, then syntacticians would not have to pay attention to the mode of presentation when running an AJT; but they could also not use such a task to classify constructions in the manner described above.

Either outcome would be of interest and to the best of our knowledge, such an investigation has not yet been done systematically.[9] In our attempt to answer our research question (can acceptability judgement tasks be used to investigate differences in modality?), our default hypothesis is (H_0) (below), which posits equality. Our alternative hypothesis (H_1) posits differences (however, (H_0) and (H_1) do not exhaust the spare of possibilities; see Footnote 10 for details).

(H_0) Irrespective of the mode of presentation of a questionnaire, the ratings for any given construction, no matter whether it occurs mainly in spoken or mainly in written language, come out the same.

(H_1) "Spoken constructions" (i.e. syntactic constructions that mainly occur in spoken language) receive higher acceptability ratings when using an auditory questionnaire instead of a textual questionnaire, and "written constructions" (i.e. constructions that mainly occur in written language) receive higher acceptability ratings when using a textual questionnaire rather than an auditory questionnaire.

Below, we express the hypotheses in a formal way, marked with an "f"-subscript. C_s are constructions that mainly occur in spoken language, C_w mainly occur in written language. Q_a represents auditory questionnaires, Q_t textual questionnaires. ϵ represents uncertainty and experimental error. In (H_{0f}), C_s in Q_a and C_s in Q_t are regarded as similar, if their absolute difference is smaller than ϵ; similarly for C_w in Q_a and C_w in Q_t. In (H_{1f}), C_s in Q_a and C_s in Q_t are regarded as different, if their difference equals or is bigger than ϵ; similarly for C_w in Q_a and C_w in Q_t.

[9] N.B.: Frazier looked into a similar question as a sideline of her (2012) project.

(H_{0f}) $\quad |C_s \text{ in } Q_a - C_s \text{ in } Q_t| < \epsilon \quad \wedge \quad |C_w \text{ in } Q_a - C_w \text{ in } Q_t| < \epsilon$

(H_{1f}) $\quad (C_s \text{ in } Q_a - C_s \text{ in } Q_t) \geq \epsilon \quad \wedge \quad (C_w \text{ in } Q_a - C_w \text{ in } Q_t) \geq \epsilon$

Even with ϵ, (H_{0f}) and (H_{1f}) are hypothetical in the sense that they assume idealised conditions. Ideally, the mode of presentation would be the only parameter that differs in an experimental setting, *ceteris paribus* (especially with respect to the investigated items, the sample, the experimental conditions, etc.).[10]

Interestingly, (H_1) is the default assumption in the literature and it is also our starting intuition in which we follow Ferreira and Swets (2005:275): "But perhaps sentences like (4a)[11] are from a spoken register; that is, they are not sentences that tend to occur in written English, so an auditory judgment task might yield more meaningful data."

However, to the best of our knowledge, there is no hard evidence for (H_1) in the literature and if we could provide such evidence, then this would be a worthwhile result. Support for (H_0) and against (H_1) would be an interesting result, as well. If (H_0) is true, then acceptability judgement tasks could not be used to determine the modality of a construction (i.e. which part of Figure 2.1 the construction belongs to). However, such a result would indicate that acceptability judgement tasks give results that are fairly robust, certainly against changes in modes of presentation.

[10] (H_0) and (H_1) are not exhaustive and further alternative hypotheses are conceivable. For instance, consider $(H_1\prime)$ below, which is the reverse of (H_1). The formal representation of $(H_1\prime)$ is $(H_{1f}\prime)$. $(H_1\prime)$ and other combinations are possible, but somewhat obscure.

$(H_1\prime)$ Spoken constructions receive lower ratings in auditory questionnaires than in textual questionnaires and written constructions receive higher ratings in auditory questionnaires than in textual questionnaires.

$(H_{1f}\prime)$ $\quad (C_s \text{ in } Q_a - C_s \text{ in } Q_t) \geq -\epsilon \quad \wedge \quad (C_w \text{ in } Q_a - C_w \text{ in } Q_t) \geq -\epsilon$

[11] "[This is a] [donkey] [that] [I don't know] [where it lives]". This is one of Ferreira and Swets' examples of a sentence with a resumptive pronoun.

OVERVIEW To put these hypotheses to a test, we have designed an experiment comprising two online questionnaires, one presented auditorily, the other textually. Both questionnaires included two constructions that mainly occur in spoken language and two constructions that mainly occur in written language. This is the original experiment, which is presented in Section 2.2.1. In the second experiment in Section 2.2.2, we ask whether it makes a difference if the textual questionnaire is "timed" or not. In the third experiment in Section 2.2.3, we introduce another factor: formality. In Section 2.3, we present a linear mixed effect model that we use for all three experiments. The results are discussed in Section 2.4. Section 2.5 concludes this chapter.

2.2 Three Experiments

To put (H_0) and (H_1) to a test, we have designed three experiments. The original experiment in Section 2.2.1 comprises two online questionnaires, one with an auditory mode of presentation, the other with a textual mode of presentation. Both questionnaires include two constructions that mainly occur in spoken language and two constructions that mainly occur in written language.

To mirror the auditory questionnaire, the items in the textual questionnaire were "timed", i.e. they disappeared after a certain while. However, Robin Melnick and Tom Wasow pointed out to us that this diverges from a canonical textual questionnaire. So, in a follow-up, we added a canonical textual questionnaire (i.e. "untimed"). This is the second experiment in Section 2.2.2.

We did not find any obvious differences between "untimed" and "timed" questionnaires. However, it could be the case that we did not observe any differences because we did not control for a possible confounding factor: formality. Formality and mode of presentation could interact and not controlling for formality might conceal dif-

ferences across modes of presentation. This is why we added the third experiment, which does control for formality. The third experiment is presented in Section 2.2.3.

In a previous version of this chapter, we analyse the three experiments separately, using t-tests and a common similarity test, the TOST (cf. Chapter 4). Arguably, the updated analysis simplifies the earlier analysis into a single statistical test and uses the data more effectively: We apply a linear mixed effect model to the data of all three experiments. So, while the ratings for each experiment are presented in the respective sub-sections (Sections 2.2.1, 2.2.2, and 2.2.3), the analysis of all three experiments is presented in Section 2.3.

2.2.1 Experiment 1: The Original Experiment

The first experiment comprises two online questionnaires, one with an auditory mode of presentation, the other with a textual mode of presentation. Below, we outline the experimental design in detail and present the ratings. In the general experimental design of all three experiments, we follow recommendations by Schütze (1996) and Cowart (1997) as closely as possible.

MATERIALS The sentences with resumptive pronouns are modelled after Ferreira and Swets (2005) and we chose them, because Prince (1990:482) observes that, although considered ungrammatical, (2): resumptive pronouns do in fact occur in spoken language. The other critical constructions come from Miller and Weinert (1998). Based on the literature and on the analysis of various corpora, they observe that (1): alternative if-clauses mainly occur in spoken language and that (3): sentence-initial gerunds and (4): wh-infinitives mainly occur in written language.

(1) If she would come to see things for herself, she would change her mind im-

mediately.

(2) We are afraid of things that we don't know what they are.

(3) Their being unaware of the situation really annoyed Rob.

(4) We found a splendid house in which to spend our holiday.

These four constructions are our critical constructions. For each construction, we have four instances, all of which are similar in their syntactic structure. (5) and (6) are an example of structural similarity.

(5) We are afraid of things that we don't know what they are.

(6) You fear things that you don't understand what they are.

We further included grammatically good and bad sentences as reference points (these are our reference constructions). Both are taken from news sources, but while "good" sentences are mostly unaltered (we sometimes shortened them), "bad" sentences are modified, particularly in their word order, in agreement, and by dropping function words. For both types, we took examples from a spoken source (NPR Radio) and from a written source (mainly from USA Today). (7) and (8) are examples of "good" sentences from a spoken and from a written source, respectively. Similarly, (9) (spoken source) and (10) (written source) are examples of "bad" sentences.

(7) You can also make the argument that they didn't approve anything.

(8) The project would more than double the population of Benewah County, home to 9,000 people.

(9) If was key word "hope" in 2009, perhaps "change" key word 2013.

(10) Is William travel on behalf of Queen and is on his second official trip to country?

Altogether, we have eight constructions: Alternative if-clauses, resumptive pronouns, sentence-initial gerunds, wh-infinitives, good sentences from a spoken source, good sentences from a written source, bad sentences from a spoken source, and bad sentences from a written source. A full list of all sentences can be found in Appendix 2.6.1.[12]

We further included fillers, which made up about 70% of a questionnaire. About 2/3 of the fillers were good and the remaining portion was equally split between bad sentences (which are similar to the bad sentences described above) and mediocre sentences (which were only slightly modified, in a manner similar to the bad sentences).

We made audio recordings of all items, including the fillers. Two native speakers of American English, one male and the other female, did the recordings for us. Both were PhD students in linguistics at Stanford University at the time. They were instructed to make each item sound as natural as possible and they were given indefinite time to familiarise themselves with the items. They were also allowed to re-record any item as often as they required. We cut out initial and final silences in the resulting sound files.

[12] At this point, it would be nice to compare the four critical constructions with respect to their relative frequencies in natural language, determined through a corpus analysis. In our view, this could be done for alternative if-clauses and sentence-initial gerunds. However, it would be harder to provide frequencies for wh-infinitives and fairly hard for resumptive pronouns. Resumptive pronouns rarely occur, so that they would require an analysis of an exceptionally large corpus. At the same time their syntactic structure is fairly complex and, to the best of our knowledge, no large corpus provides adequate search tools to allow resumptive pronouns to be detected easily.

PROCEDURE Each participant was asked to do both the auditory and the textual questionnaire, leaving at least 24 hours between the two. The two questionnaires were similar, but had no overlap in items, as we randomly put two of the four items of each construction into the auditory questionnaire and the other two into the textual questionnaire. Fillers were similarly randomised. Each participant had an individually randomised set of questionnaires, but other than this random sampling of sentences, the questionnaires of all participants were similar. The order of the questionnaires (auditory first vs textual first) was random, as well. Items were rated on a 5-point scale, by pressing buttons (cf. Weskott and Fanselow, 2011, for a comparison of various measurement methods). There were five possible answers, each represented by a button: "fully unacceptable", "rather unacceptable", "in between", "rather acceptable", and "fully acceptable" (Figure 2.3 shows the design of the rating website[13]).

Besides collecting ratings, we also collected reaction times (the reaction time of an item is the time from loading the item until the time at which the rating is given). Our instructions asked participants "to evaluate grammatical acceptability of certain sentences", followed by the question "how natural do they sound to you with respect to their grammaticality?", but participants were also advised that they should "not be bothered with meaning or punctuation". They were further told that "this is about your intuition and there is no right or wrong". With these instructions we wish to strike a balance between not confusing our participants by scientific jargon and sufficiently introducing them to the task. The introduction included a test-run in which participants rated a good, a mediocre, and a bad sentence (as determined through a pilot). The rating website was designed to be neutral in appearance. Figure 2.2 shows the very first screen of the website.

[13]The design does not provide an adequate rating tool for the colourblind. In the experiment in Section 2.2.3, we fix this by including smiley faces (see Figure 2.8 and Figure 2.9 below). In hindsight, the best solution would have been to use red and blue buttons.

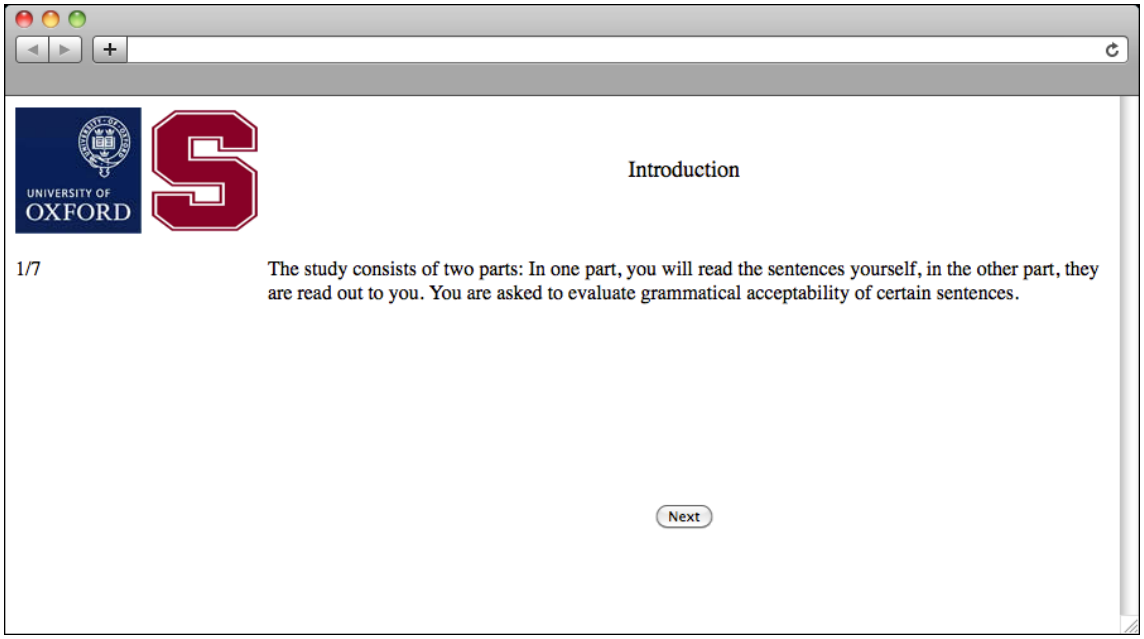


Figure 2.2: The start screen of our experiment.

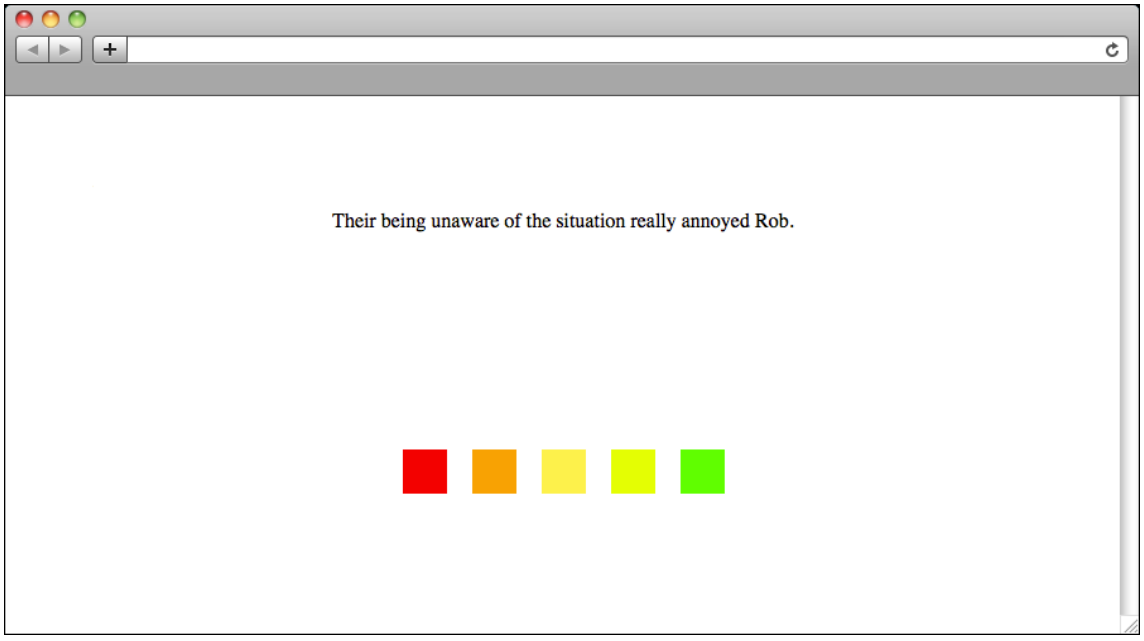


Figure 2.3: Illustrating the design of the website that we used to run the experiments (here for the textual questionnaire).

It could be the case that any differences in ratings between the two modes are simply due to the fact that audio recordings "pass in time" and that textual items typically remained on the screen until a rating was given. If memory limitations are a major factor, then they would affect the auditory questionnaire, but not the textual questionnaire (for a review of the effects of memory decay and interference effects, cf. e.g. Berman, Jonides, and Lewis, 2009). This is why we made textual items "pass in time", as well. Each textual item I disappeared after a time I_t, where I_c is the character count of an item and AR_1 and AR_1 are the length of our two audio recordings (see Equation 2.1).

$$I_t = \frac{225 + I_c \cdot 25}{2} + \frac{AR_1 + AR_2}{2} \quad ms \tag{2.1}$$

We used a formula that averaged the expected reading time (the first term; related to a formula from Bader and Häussler, 2010:289) and the actual reading times from our audio recordings (the second term). We included the second term because the original formula (the first term only) was not designed to reflect the reading time of "scrambled" sentences (e.g. sentence (9) above). Participants might miss a part of an item at times; so, they were allowed to replay an item once (this was the case for both auditory and textual items) and if they still weren't sure of an item's rating, they could skip it altogether (however, this occurred in less than 2% of all cases).

PARTICIPANTS We asked 40 participants to take part in our study, all of whom were non-linguistics students at Stanford University at the time. 31 of them (mean age: 24.52 years (3.55); gender distribution: 25 females and 6 males) finished both parts. Their results are reported below. Payment was $10. For any given participant we calculate a participant's median reaction time ($\mu_{1/2}$). We then calculate the mean and the standard deviation of all the medians ($\overline{\mu_{1/2_{1-N}}}$). In the next step, we determine a lower and an upper threshold (equations 2.2 and 2.3

below). The lower threshold is the average median minus one and a half standard deviations of the medians. The upper threshold is the average median plus four standard deviations of the medians. If a participant's median reaction time falls below the lower threshold or goes above the upper threshold, then that participant will be excluded.[14]

$$\theta_{lower} = \overline{\mu_{1/2_{1-N}}} - 1.5\sigma \tag{2.2}$$

$$\theta_{upper} = \overline{\mu_{1/2_{1-N}}} + 4\sigma \tag{2.3}$$

$$\text{where } \overline{\mu_{1/2_{1-N}}} = \frac{\mu_{1/2_1}+\mu_{1/2_2}+\ldots+\mu_{1/2_N}}{N} \text{ and } \sigma = stdev(\mu_{1/2_{1-N}})$$

This approach is designed to detect different kinds of non-cooperative behaviour and is justified in Appendix 6. However, we did not have to exclude any of our participants. Further, there was no need to add a criterion that excludes non-native speakers of American English: Since we recruited them in person, we were able to ask potential participants for their mother tongue during the recruitment process.

In total, we have 1010 data points (2 questionnaires $\times$ (4 primary constructions + 4 reference constructions) $\times$ 2 tokens $\times$ 32 participants $-$ 14 skips), excluding fillers.[15] The analysis is presented in Section 2.3.

RATINGS The raw ratings for the first experiment are given in Table 2.1 and Table 2.2. The Z-scores are given in Table 2.3 and Table 2.4 (the Z-scores are

[14] This approach is similar to the exclusion criteria in Chapter 1 and Chapter 3.

[15] "Token" refers to an instance of the different constructions; i.e. for each construction, we have four instances (cf. Appendix 2.6.1), two of which were in each questionnaire.

calculated by participant and only using the ratings for the critical items[16]) and are illustrated in Figure 2.4. A further note on Figure 2.4: A lighter shade means that the ratings originated from the auditory questionnaire, a darker shade means that the ratings originated from the textual questionnaire. For instance, the first bar from the left represents Good Constructions from a written source and the ratings were collected with the auditory questionnaire. The second bar represents Good Constructions from a written source collected with the textual questionnaire.

Raw Ratings	Wh-Inf. Written Source	Altern.-Ifs Spoken Source	Init.-Gerunds Written Source	Resumptive-PN Spoken Source
Auditory Questionnaire	4.08 (1.00)	3.78 (1.28)	3.60 (1.22)	2.05 (1.00)
Textual Questionnaire	4.23 (1.01)	3.82 (1.23)	3.50 (1.37)	1.83 (0.77)

Table 2.1: The mean ratings for the four critical constructions for both questionnaires (including the population standard deviations in parentheses).

Raw Ratings	Good Sent. Written Source	Good Sent. Spoken Source	Bad Set. Written Source	Bad Sent. Spoken Source
Auditory Questionnaire	4.53 (0.78)	4.24 (1.05)	1.73 (1.03)	1.77 (1.12)
Textual Questionnaire	4.53 (0.89)	4.55 (0.75)	1.77 (1.17)	1.58 (1.02)

Table 2.2: The mean ratings for the four reference constructions for both questionnaires (including the population standard deviations in parentheses).

Z-Scores	Wh-Inf. Written Source	Altern.-Ifs Spoken Source	Init.-Gerunds Written Source	Resumptive-PN Spoken Source
Auditory Questionnaire	0.58 (0.54)	0.39 (0.75)	0.27 (0.71)	−0.77 (0.58)
Textual Questionnaire	0.67 (0.62)	0.44 (0.70)	0.21 (0.71)	−0.83 (0.42)

Table 2.3: The Z-scores for the four critical constructions for both questionnaires (including the population standard deviations in parentheses).

[16]See Section 2.3, "The Model", for the reason why we only considered the critical items for the Z-scores.

Z-Scores	Good Sent. Written Source	Good Sent. Spoken Source	Bad Set. Written Source	Bad Sent. Spoken Source
Auditory Questionnaire	0.91 (0.49)	0.68 (0.67)	−1.03 (0.49)	−1.06 (0.55)
Textual Questionnaire	0.85 (0.55)	0.86 (0.49)	−0.98 (0.58)	−1.07 (0.38)

Table 2.4: The Z-scores for the four reference constructions for both questionnaires (including the population standard deviations in parentheses).

Wh-Infinitives Written Source	WHI WS	Good Sentences Written Source	GD WS
Alternative-If-Clauses Spoken Source	IF SS	Good Sentences Spoken Source	GD SS
Sentence Initial Gerunds Written Source	GER WS	Bad Sentences Written Source	BAD WS
Resumptive Pronouns Spoken Source	RES SS	Bad Sentences Spoken Source	GD SS

Table 2.5: A key to the abbreviations used in the following figures.

At first glance, the ratings look quite similar across modality. This does not bode well for (H_1). Further, if (H_1) was true, then some of the observed differences take the wrong direction. For instance, the alternative if-clauses (IF SS) should have received higher ratings in the auditory questionnaire for (H_1) to be true; but the opposite is the case (similarly, GD WS, GD SS, and GER WS). This does not bode well for (H_1), but, of course, a statistical test is needed to confirm this.

ISSUES The following was pointed out to us by Robin Melnick and Tom Wasow: While tests on the data of Experiment 1 might allow conclusions about time-constrained textual questionnaires, the results may not apply to canonical written questionnaires (i.e. a questionnaire in which items stay on the screen until the participant rates the item). That is, it is possible that the results of a canonical written questionnaire differ to those of the auditory questionnaire and to those of a time-constrained written questionnaire. Why? Because in a time-constrained questionnaire, participants have to rely on their memory if they want to ponder on an

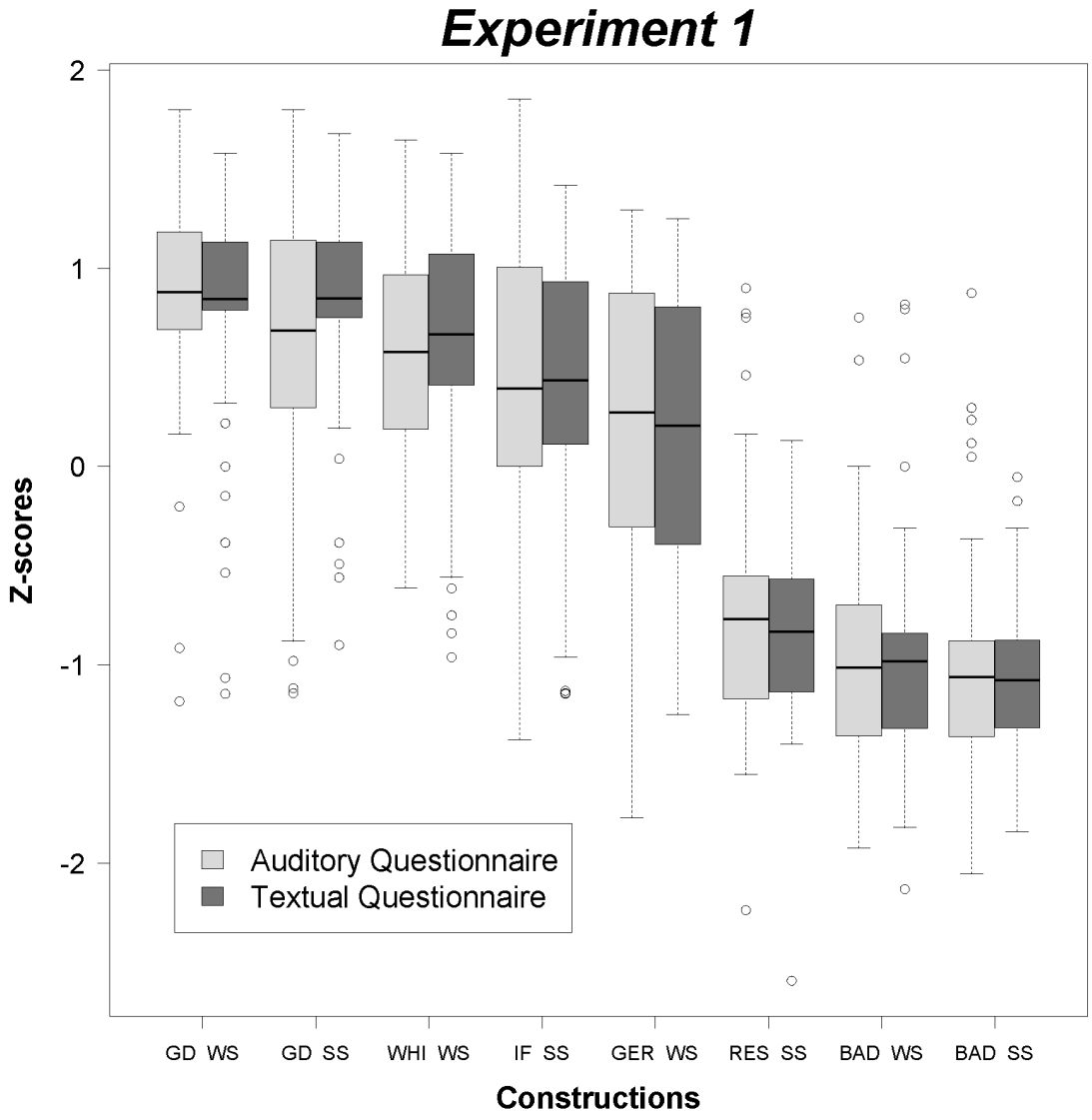


Figure 2.4: The Z-scores (y-axis) for the eight constructions (x-axis) for both questionnaires (a higher score denotes higher acceptability). The shading indicates the mode of presentation. See Table 2.5 on page 118 for the key to the abbreviations used in this figure (as a reminder: "WS" means that the construction comes from a written source and "SS" that it comes from a spoken source).

item, as items disappear after a short while. In a canonical written questionnaire, participants do not have to rely on their memory if they wish to ponder on an item, as items stay on the screen until a rating is given.

Thus, to investigate the matter further, we did a second experiment, which we ran about one month after the original experiment.

2.2.2 Experiment 2: Timed versus Untimed Questionnaire

In the second experiment, we added a non-time-constrained textual questionnaire and gave it to the participants of the first study. One could think of this as adding a third questionnaire to the original study *post hoc*. We then compared the results for this "canonical" written questionnaire to the results of the textual questionnaire of our original study.

MATERIALS, PROCEDURE, AND PARTICIPANTS By and large, everything is identical to the textual questionnaire from our original study, except that this time, there are no time-constraints. The materials are the same as in the original study (see Section 2.4) and we applied the same randomisation procedure as described above (from each construction, two randomly chosen items were put in a questionnaire; this was done for each participant), so that the third questionnaire was unique in item arrangement, but underlyingly identical to the other two questionnaires.[17] Other procedural parameters are also identical (the use of a 5-point scale, the fillers and the filler rate, the collection of both ratings and reaction times, the instructions given, the design of the website, the exclusion criterion for participants; for details, please see Section 2.4).

[17] However, this also means that participants saw some of the items for a second time. We deemed this harmless, because there was more than a month's time between the original experiment and this follow-up. The alternative would have been to conceive more items.

In our original study, the same participants filled in the two questionnaires. Consequently, we asked the same 31 participants who completed the first experiment to take part in our follow-up. 18 of them decided to take part (mean age: 24.17 years (3.88); gender distribution: 13 females and 5 males). Payment was $7, which is a slightly higher hourly rate than before (this was done to encourage participation).

In total, we have 287 additional data points (1 questionnaire × (4 primary constructions + 4 reference constructions) × 2 tokens × 18 participants − 1 skip), excluding fillers. An analysis of all experiments is presented in Section 2.3.

RATINGS The raw ratings are given in Table 2.6 and Table 2.7. The Z-scores are given in Table 2.8 and Table 2.9 and are illustrated in Figure 2.5.

Raw Ratings	Wh-Inf. Written Source	Altern.-Ifs Spoken Source	Init.-Gerunds Written Source	Resumptive-PN Spoken Source
Untimed Textual Questionnaire	3.92 (1.23)	3.67 (1.24)	3.47 (1.08)	2.00 (0.79)
Timed Textual Questionnaire	4.23 (1.01)	3.82 (1.23)	3.50 (1.37)	1.83 (0.77)

Table 2.6: The mean ratings for the primary constructions for the time-constrained ("timed") and the non-constrained questionnaire ("untimed"), including the population standard deviations (in parentheses). The ratings for the timed questionnaire are duplicated from Table 2.1 in Section 2.4, for ease of comparison.

Raw Ratings	Good Sent. Written Source	Good Sent. Spoken Source	Bad Set. Written Source	Bad Sent. Spoken Source
Untimed Textual Questionnaire	4.22 (0.92)	4.29 (0.83)	1.53 (0.91)	1.40 (0.88)
Timed Textual Questionnaire	4.53 (0.89)	4.55 (0.75)	1.77 (1.17)	1.58 (1.02)

Table 2.7: The mean ratings for the reference constructions for the time-constrained ("timed") and the non-constrained questionnaire ("untimed"), including the population standard deviations in parentheses. The ratings for the timed questionnaire are duplicated from Table 2.2 in Section 2.4, for ease of comparison.

Z-Scores	Wh-Inf. Written Source	Altern.-Ifs Spoken Source	Init.-Gerunds Written Source	Resumptive-PN Spoken Source
Untimed Textual Questionnaire	0.56 (0.76)	0.37 (0.73)	0.26 (0.79)	−0.69 (0.37)
Timed Textual Questionnaire	0.67 (0.62)	0.44 (0.70)	0.21 (0.71)	−0.83 (0.42)

Table 2.8: The Z-scores for the primary constructions for the time-constrained ("timed") and the non-constrained questionnaire ("untimed"), including the population standard deviations in parentheses. The ratings for the timed questionnaire are duplicated from Table 2.3 in Section 2.4, for ease of comparison.

Z-Scores	Good Sent. Written Source	Good Sent. Spoken Source	Bad Set. Written Source	Bad Sent. Spoken Source
Untimed Textual Questionnaire	0.77 (0.56)	0.83 (0.48)	−1.02 (0.58)	−1.10 (0.55)
Timed Textual Questionnaire	0.85 (0.55)	0.86 (0.49)	−0.98 (0.58)	−1.07 (0.38)

Table 2.9: The Z-scores for the reference constructions for the time-constrained ("timed") and the non-constrained questionnaire ("untimed"), including the population standard deviations in parentheses. The ratings for the timed questionnaire are duplicated from Table 2.4 in Section 2.4, for ease of comparison.

FURTHER ISSUES Again, at first glance, the ratings look very similar and it looks as if the mode of presentation does not make much of a difference (we analyse the ratings properly in Section 2.3). However, one might argue as follows: It could be the case that the use of alternative if-clauses, resumptive pronouns, wh-infinitives, sentence-initial gerunds, etc. is tied to the *formality* of the communicative situation (and not the modality of the communication, i.e. spoken vs written; many thanks to Gisbert Fanselow and Lyn Frazier for pointing this issue out to us). Since spoken communication tends to happen in informal settings and written communication in formal settings, these conditions often coincide. This would also explain why we did not observe substantial differences above: All the questionnaires are neutral in formality. And maybe we would have observed differences, if we had included formality as a factor in the experiment. This is why we added another follow-up (which we ran about a year after the original experiment).

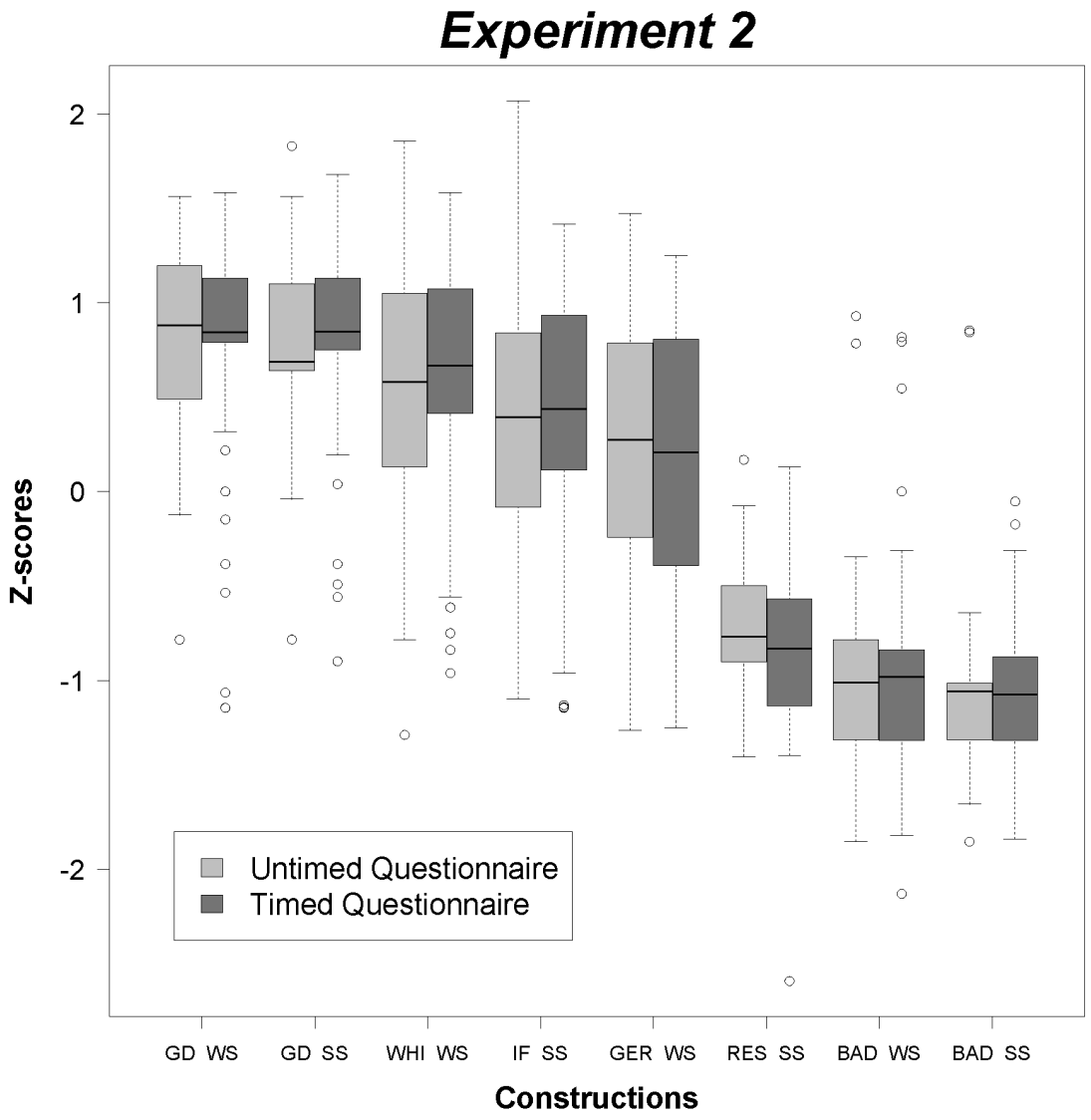


Figure 2.5: The Z-scores (y-axis) for the eight constructions (x-axis) for both questionnaires of the follow-up. A higher score denotes higher acceptability. The shading indicates the type of questionnaire (light grey: items remained on the screen until a rating was given; dark grey: items disappeared after a short time). See Table 2.5 on page 118 for the key to the abbreviations used in this figure.

2.2.3 Experiment 3: Informal versus Formal

The third experiment is by and large a repeat of the original experiment in Section 2.4, with another dimension added: formality. Defining "formality" is a non-trivial task and our notion of formality is based on a multi-dimensional continuum by Koch and Oesterreicher (1985). Two defining endpoints in this continuum are *Sprache der*

Nähe ("language of closeness") and *Sprache der Distanz* ("language of distance"). They list the following characteristics of the *Sprache der Nähe* and *Sprache der Distanz* (from Koch and Oesterreicher, 1985:21; emphasis by Koch and Oesterreicher).

> Die Kombination 'Dialog', 'freier Sprecherwechsel', 'Vertrautheit der Partner', 'face-to-face- Interaktion', 'freie Themenentwicklung', 'keine Öffentlichkeit', 'Spontaneität', 'starkes Beteiligtsein', 'Situationsverschränkung', etc. charakterisiert den Pol 'gesprochen'. Die ihm entsprechende Kommunikationsform lät sich am besten auf den Begriff **Sprache der Nähe** bringen. Analog charakterisiert die Kombination von 'Monolog', 'kein Sprecherwechsel', 'Fremdheit der Partner', 'räumliche und zeitliche Trennung', 'festes Thema', 'völlige Öffentlichkeit', 'Reflektiertheit', 'geringes Beteiligtsein', 'Situationsentbindung', etc. den Pol 'geschrieben'. Die ihm entsprechende Kommunikationsform definieren wir als **Sprache der Distanz**![18]

Although spoken language often concurs with *closeness* and written language with *distance*, they are not identical. Spoken communication can be both close (e.g. chatting away with a friend) and distant (e.g. talking to the King of the Netherlands at an official occasion) and the same is true for written communication (e.g. chatting with a friend via an instant messenger vs having an exchange of letters with the tax department).

For our purposes, most of the factors mentioned in the quote are fixed. However, there is some leeway when it comes to the relationship between the researcher and the participants and to the way any communication is delivered (viz. the design of our experimental website). The relationship between the researcher and the participants can be influenced by the researcher's social status. Social status is determined through (perceived) income, education, class/lifestyle, and prestige and

[18] "The combination of 'dialogue', 'turn taking', 'familiarity between speakers', 'face-to-face interaction', 'free development of topics', 'no publicness', 'spontaneity', 'high degree of involvement', 'entanglement of situations', etc. characterises the pole 'spoken' [language]. The type of communication appropriate for this pole is best described by the term **language of closeness**. Analogously, the combination of 'monologue', 'a lack of turn taking', 'non-familiarity of speakers', 'spatial and temporal disconnect', 'fixed topic', 'complete publicness', 'reflection', 'low degree of involvement', 'disentanglement of situations', etc. characterise the pole 'written' [language]. We define the appropriate type of communication for this pole as **language of distance**." Our translation.

can affect interpersonal behaviour, including language use (Wolf, 1985, and Guy, 1985). We could simulate a closer vs more distant relationship by presenting the researcher differently to the participants. The other way is to alter the design of the website: An informal design can signify closeness and a formal design distance.

Accordingly, we add two more conditions to the original experiment. The *informal condition* tries to establish an informal relationship through presenting the researcher as having a lower social status and using an informal design for the website. The *formal condition* tries to establish a formal relationship through presenting the researcher of higher social status and use a formal design for the website. This, then, results in four questionnaires: 1) an informal auditory questionnaire, 2) a formal auditory questionnaire, 3) an informal textual questionnaire, and 4) a formal textual questionnaire. With this in mind, we can adjust (H_0) and (H_1).

(H_0) Irrespective of the mode of presentation of a questionnaire and irrespective of a questionnaire's formality, the ratings for any given construction, no matter whether it occurs mainly in spoken or mainly in written language, come out the same.

(H_1) "Spoken constructions" (i.e. syntactic constructions that mainly occur in spoken language) receive higher acceptability ratings when using an informal auditory questionnaire instead of a formal textual questionnaire and "written constructions" (i.e. constructions that mainly occur in written language) receive higher acceptability ratings when using a formal textual questionnaire rather than an informal auditory questionnaire.

Below, we express the hypotheses in a formal way. C_s are constructions that mainly occur in spoken language, C_w mainly occur in written language. Q_{ai} represents informal auditory questionnaires, Q_{tf} formal textual questionnaires. ϵ represents

uncertainty and experimental error.

$$(\mathrm{H}_{0f}) \qquad |\mathrm{C}_s \text{ in } \mathrm{Q}_{ai} - \mathrm{C}_s \text{ in } \mathrm{Q}_{tf}| < \epsilon \quad \wedge \quad |\mathrm{C}_w \text{ in } \mathrm{Q}_{ai} - \mathrm{C}_w \text{ in } \mathrm{Q}_{tf}| < \epsilon$$

$$(\mathrm{H}_{1f}) \qquad (\mathrm{C}_s \text{ in } \mathrm{Q}_{ai} - \mathrm{C}_s \text{ in } \mathrm{Q}_{tf}) \geq \epsilon \quad \wedge \quad (\mathrm{C}_w \text{ in } \mathrm{Q}_{ai} - \mathrm{C}_w \text{ in } \mathrm{Q}_{tf}) \geq \epsilon$$

MATERIALS AND PROCEDURE The materials were the same as in the original study (see Section 2.4) and we applied the same randomisation procedure as described above. Thus, the four questionnaires were unique in item arrangement, but underlyingly identical to each other and to the questionnaires of the original experiment in Section 2.2. Many other procedural parameters were also identical (i.e. the use of a 5-point scale, the collection of both ratings and reaction times, etc.; for details, please see Section 2.4). We also kept the audio recordings. (Perhaps we could have included conditions like "sounding informal" vs "sounding formal", but we are not sure we could have made these categories meaningful. So, as before, the audio recordings are meant to sound as natural as possible.)

There were a few minor differences: We increased the number of critical items per questionnaire. Previously, the critical items were split between questionnaires, because the same participant took multiple questionnaires. But now, any given participant filled in only one questionnaire, so that we could give all critical items to each participant. At the same time, we took the filler rate down to 50%, which is relevant below. Further, we had to adjust the exclusion criteria for the participants, as the method of recruiting participants had changed (see below for details).

To establish lower vs higher social status, we added an additional introduction page to the website. In the informal condition, we began with the salutation "Hi" and presented the researcher simply as a student ("Tom") who was doing some research for his unspecified degree at an unspecified university. The website still showed

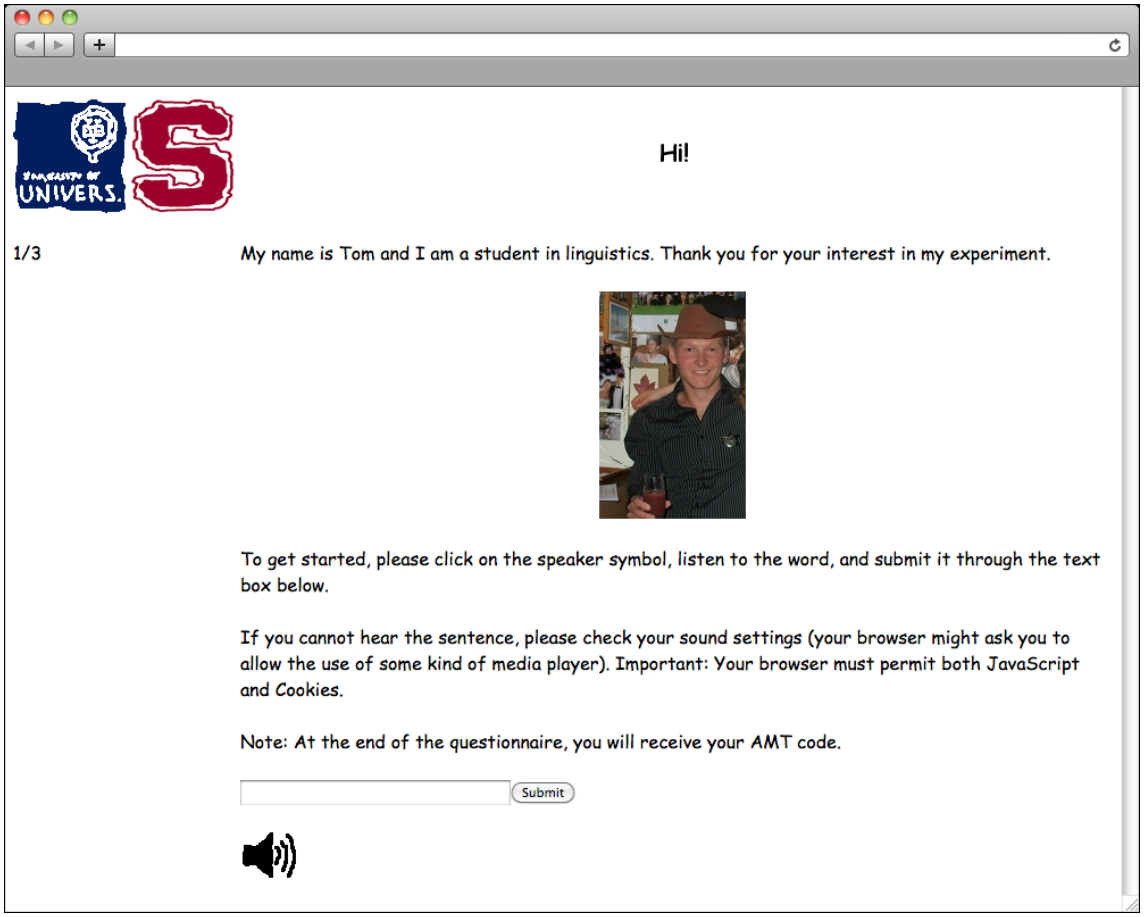


Figure 2.6: The introduction to the experiment for the informal condition.

the university crests, which were barely legible, though. Of course, the small print on the consent form provided all the required information, including the fact that our study was approved by and followed the guidelines of the University of Oxford's Central University Research Ethics Committee. We added a photo from an informal occasion (see Figure 2.6). In the formal situation, we presented the researcher as a scientist who was working on an important experiment for his PhD at the University of Oxford. We added a photo from a formal occasion (see Figure 2.7).

Further, the designs of the informal and formal websites differed. In the informal condition, everything was "doodled" and the text font used was Comic Sans. In the formal condition, the website was clean and precise and Times New Roman was used. In this, we followed Hong, Li, Lin et al. (2001) for designing informal and formal websites. Structurally, however, the informal and formal websites were

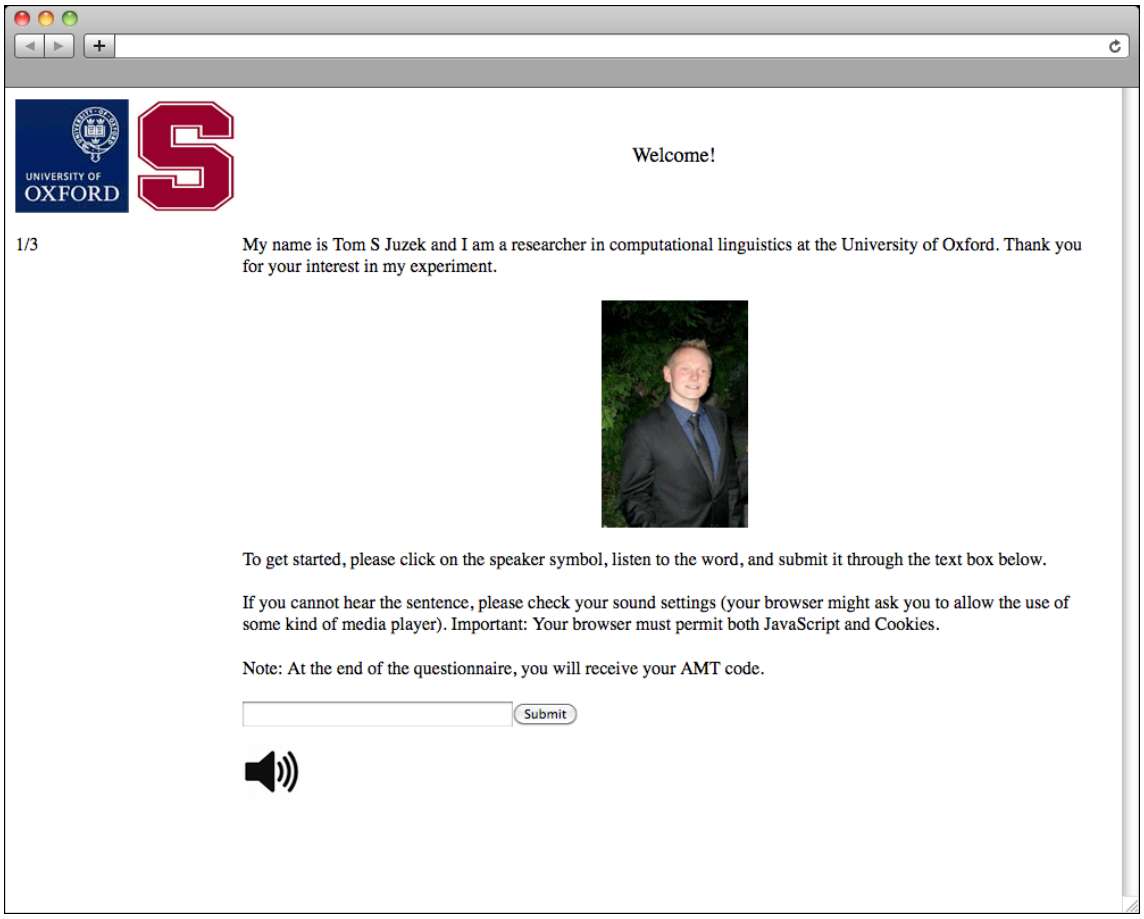


Figure 2.7: The introduction to the experiment for the formal condition.

Figure 2.8: Illustrating the informal design of the websites that we used to run the experiments (here for the textual questionnaire).

identical in order to prevent the possibility that other sources of error were created along the way. Figures 2.8 and 2.9 show the design of the informal website and the formal website, respectively. We included the smiley faces for the colourblind.[19]

PARTICIPANTS Since the required number of participants has doubled in comparison to the original experiment in Section 2.2, we had to sample our participants anew. For each of the four questionnaires (informal auditory questionnaire, formal auditory questionnaire, informal textual questionnaire, formal textual questionnaire), we recruited 40 participants with Amazon Mechanical Turk (4×40 participants; 160 participants in total). To be able to take part, potential participants had to have an Amazon Mechanical Turk approval rate of at least 98% and to have finished at least 5000 approved tasks. We set these criteria to ensure that only reliable participants could take part.[20] We were only interested in native speakers

[19] To further assist the colourblind, it would have been preferable to use red and blue buttons.

[20] N.B.: These criteria are more lenient than what it requires to become an Amazon Mechanical Turk "Master Worker". Further, these criteria are similar to the criteria in Chapter 1 and Chapter 3.

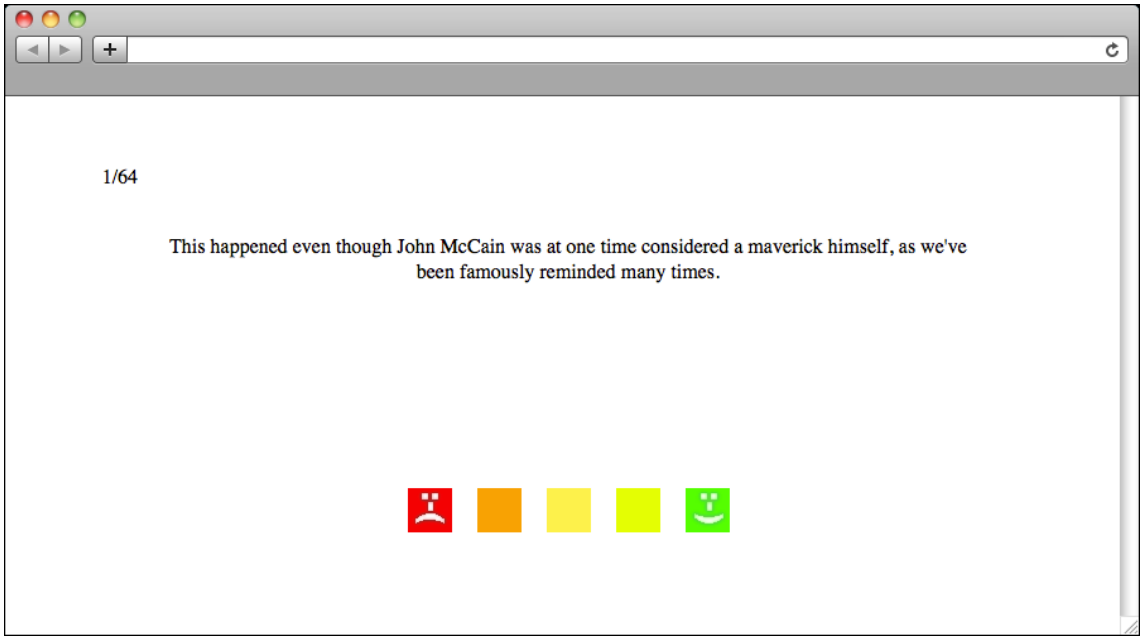


Figure 2.9: Illustrating the formal design of the websites that we used to run the experiments (here for the textual questionnaire).

of American English. To avoid accommodation effects, though, we did not disclose this requirement to our participants. So, native speakers of any language could take part. However, to increase the chances of native speakers of American English participating in our study, recruiting took place between 21:00 and 02:00 GMT. Payment was such that it came out at an hourly rate of about $10. Participants had to be non-linguists, which was mentioned in the introduction. After the experiment, we anonymously gathered the following information: a participant's age, gender, and home country; we also asked our participants where they predominantly lived the first ten years of their lives. Our pools of participants have the following demographics (after exclusions):

Session 1, "Informal Auditory Questionnaire": 32 participants included; mean age: 32.50 years (9.71); gender distribution: 12 females and 20 males.

Session 2, "Informal Textual Questionnaire": 33 participants included;

mean age: 36.06 years (10.02); gender distribution: 17 females and 16 males.

Session 3, "Formal Auditory Questionnaire": 30 participants included; mean age: 35.83 years (12.07); gender distribution: 14 females and 16 males.

Session 4, "Formal Textual Questionnaire": 32 participants included; mean age: 33.13 years (8.74); gender distribution: 13 females and 19 males.

EXCLUSION CRITERIA For this experiment, we have four exclusion criteria. These are: participating multiple times (0 exclusions), not being native speakers of American English (13 exclusions), returning incomplete results (5 exclusions), and having extreme reaction times (15 exclusions). We apply these criteria in the listed order. Details are discussed in the following.

Participating multiple times: We asked participants to take part only once, to make sure that no single participant is overrepresented. We reminded them that we could easily detect violators through their unique Amazon Mechanical Turk ID number. Consequently, not a single participant attempted to take part in multiple sessions.

Not being a native speaker of American English: We did not state this explicitly to our participants, but we only consider results that come from participants who we believe to be native speakers of American English. If the home country and the place where a participant lived the first ten years of his/her life are the United States of America, then we consider that participant as a native speaker of American English.

Returning incomplete results: We disregard any results from a participant who did

not return a complete set of ratings.

Having extreme reaction times: This is identical to the criterion in the previous experiments above.

In total, we have 4049 additional data points ((32 participants in the informal auditory questionnaire + 33 participants in the informal textual questionnaire + 30 participants in the formal auditory questionnaire + 32 participants in the formal textual questionnaire) × (4 primary constructions + 4 reference constructions) × 4 tokens − 15 skips), excluding fillers.

RATINGS The raw ratings are given in Table 2.10 and Table 2.11. The Z-scores are given in Table 2.12 and 2.13 and are illustrated in Figure 2.10. In Figure 2.10, lighter shades mark the informal condition, darker shades the formal condition. The mode of presentation is encoded similarly: The auditory questionnaire is marked by a lighter shade, the textual questionnaire by a darker shade. For instance, the very first bar on the left represents the ratings of GD WS for the informal auditory questionnaire. The second bar represents the GD WS ratings for the informal textual questionnaire.

The ratings look quite similar across modality and degree of formality. Also, the ratings come out very similar to the ratings in the first and second experiment. However, there is one exception: the resumptive pronouns. The ratings in the third experiment are considerably higher (around −0.50) than the ratings in the first or second experiment (around −0.75). For a discussion of possible reasons, please see Section 2.4. In the next section, we analyse the data of all three experiments with one model.

Raw Ratings	Wh-Inf. Written Source	Altern.-Ifs Spoken Source	Init.-Gerunds Written Source	Resumptive-PN Spoken Source
Informal Auditory Questionnaire	4.54 (0.83)	4.07 (1.08)	3.77 (1.25)	2.92 (1.31)
Informal Textual Questionnaire	4.33 (0.95)	4.33 (0.91)	3.75 (1.26)	2.65 (1.11)
Formal Auditory Questionnaire	4.48 (0.89)	4.13 (1.08)	3.93 (1.11)	2.90 (1.27)
Formal Textual Questionnaire	4.56 (0.81)	4.08 (1.08)	3.80 (1.25)	2.95 (1.33)

Table 2.10: The mean ratings for the four primary constructions for both questionnaires (including the population standard deviations in parentheses).

Raw Ratings	Good Sent. Written Source	Good Sent. Spoken Source	Bad Set. Written Source	Bad Sent. Spoken Source
Informal Auditory Questionnaire	4.63 (0.78)	4.71 (0.75)	2.24 (1.13)	2.02 (1.21)
Informal Textual Questionnaire	4.52 (0.80)	4.60 (0.70)	2.30 (1.33)	1.81 (1.18)
Formal Auditory Questionnaire	4.66 (0.71)	4.69 (0.62)	2.14 (1.24)	2.09 (1.35)
Formal Textual Questionnaire	4.64 (0.77)	4.73 (0.75)	2.27 (1.16)	2.13 (1.30)

Table 2.11: The mean ratings for the four secondary constructions for both questionnaires (including the population standard deviations in parentheses).

Z-Scores	Wh-Inf. Written Source	Altern.-Ifs Spoken Source	Init.-Gerunds Written Source	Resumptive-PN Spoken Source
Informal Auditory Questionnaire	0.65 (0.51)	0.32 (0.68)	0.12 (0.77)	−0.46 (0.77)
Informal Textual Questionnaire	0.54 (0.68)	0.54 (0.54)	0.15 (0.83)	−0.62 (0.67)
Formal Auditory Questionnaire	0.59 (0.62)	0.35 (0.71)	0.20 (0.74)	−0.49 (0.80)
Formal Textual Questionnaire	0.64 (0.53)	0.29 (0.71)	0.12 (0.76)	−0.46 (0.78)

Table 2.12: The Z-scores for the four primary constructions for both questionnaires (including the population standard deviations in parentheses).

Z-Scores	Good Sent. Written Source	Good Sent. Spoken Source	Bad Set. Written Source	Bad Sent. Spoken Source
Informal Auditory Questionnaire	0.72 (0.50)	0.78 (0.49)	−0.97 (0.77)	−1.14 (0.83)
Informal Textual Questionnaire	0.67 (0.50)	0.73 (0.51)	−0.84 (0.87)	−1.17 (0.77)
Formal Auditory Questionnaire	0.70 (0.52)	0.71 (0.52)	−1.03 (0.76)	−1.05 (0.84)
Formal Textual Questionnaire	0.70 (0.52)	0.78 (0.48)	−0.98 (0.78)	−1.08 (0.87)

Table 2.13: The Z-scores for the four secondary constructions for both questionnaires (including the population standard deviations in parentheses).

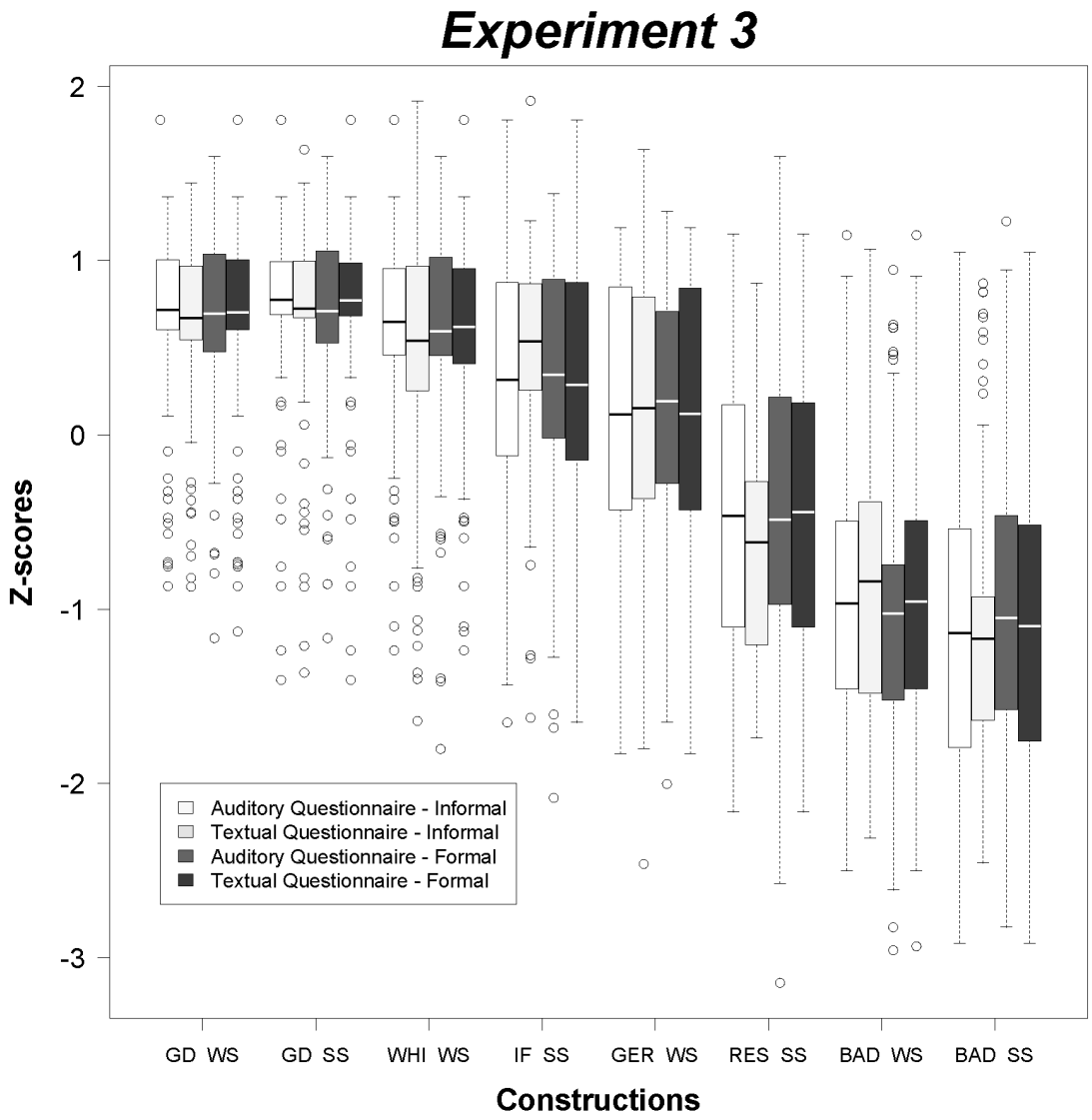


Figure 2.10: The Z-scores (y-axis) for the eight constructions (x-axis) for both questionnaires. A higher score denotes higher acceptability. The shading indicates the mode of presentation and degree of formality. See Table 2.5 on page 118 for the key to the abbreviations used in this figure.

2.3 Statistical Analysis: Linear Mixed Effect Model

In a previous version of this chapter, we analyse all three experiments separately, using t-tests and a common similarity test, the TOST (cf. Chapter 4). This way, however, the results are limited to each experiment and we had to run numerous tests on various sub-hypotheses, which might have confused some readers. Further, the criteria that defined support for (H_1) are opaque. And we could not check for any interactions between various factors, because the previous analyses only look at the means of the different constructions. This is why we decided to re-analyse the results, so that there is one analysis for the data of all three experiments. We decided to use a linear mixed effect model for this.

THE MODEL For the model, we used the lme4 package (Bates, Maechler, and Bolker, 2014) for R (R Core Team, 2015). We use a linear mixed effect model with random factors (*lmer*) to determine how an item's rating depends on the mode of a questionnaire (i.e. auditory vs timed textual vs canonical textual; we cross this with a construction's source) and the formality of a questionnaire (formal vs informal). These are our fixed effects. We have the following random effects: Differences across participants (i.e. how much variance can be explained by the fact that participants evaluated our constructions differently) and differences within constructions (this is a measure of how consistent the construction categories were, i.e. how consistent the four instances of each construction were).

For the data from the first and second experiment, we also include: Position within questionnaire (i.e. does it make a difference whether or a not an item appeared at the beginning vs at the end of a questionnaire; this is to measure fatigue), order of questionnaires (i.e. are items rated differently in the first vs the second questionnaire; this is to measure possible repetition effects), and previous sentence (i.e. does the previous sentence influence the rating of the current item; this measures possible

order effects).[21]

We use Z-scores for the model. Typically, Z-scores are calculated by taking all ratings into account, including the fillers. Unfortunately, we used slightly different fillers for the first and second experiment (Section 2.2.1 and 2.2.2) and the third experiment (Section 2.2.3).[22] As a consequence, we calculated the Z-scores based on the critical items only (since the critical items were the same). (Also see Section 3.2.4 in Chapter 3 for a detailed discussion of how the choice of experimental items can affect the Z-scores.)

To check whether the Z-scores based on just the critical items are valid, we compared the REML criterion of the Z-score model to the REML criterion of the model based on raw ratings (i.e. the original 5-point scale ratings, ranging from "1" to "5").[23] For the Z-score model, the REML criterion was 10528.7, for the model based on raw ratings, the REML criterion was 14936.3. This indicates that the Z-score model is in fact considerably stronger than the model based on raw ratings, which is a sign that the Z-scores are valid, indeed. Further, we checked for normality and homoscedasticity by visually inspecting the residual plots.

RESULTS A summary of our results can be found in Table 2.14 for the random effects and Table 2.15 for the fixed effects.

As to the random factors, Differences Across Participants shows some variance (0.027), but the effect size is rather small. The same applies to Differences Within Constructions. As mentioned in Section 2.2.1, the concrete sentences for the vari-

[21] We did not include these factors for the third experiment, because we decided to analyse them after having run the experiments. While extracting the additional data was quite doable for the first and second experiment, it was much harder for the third experiment. This has to do with how the outputs were coded by the rating website.

[22] We did this because we wished to have a slightly cleaner set of fillers in the third experiment. This would not have done any harm with respect to our original analyses based on t-tests and TOSTs.

[23] REML (restricted maximum likelihood) is the likelihood for a good fit of a model. A model with a lower REML criterion provides a better fit (cf. Ané, 2010).

Parameter	Variance	Standard Deviation
Differences Across Participants	0.027	0.165
Differences Within Constructions	0.042	0.205
Position Within Questionnaires	0.003	0.058
Order of Questionnaires	0.000	0.000
Previous Sentence	0.001	0.027
Residual	0.383	0.619

Table 2.14: The random effects of our model, their variance, and their standard deviation.

Parameter	Estimate	Standard Error	t-Value
Intercept GD Construction Questionnaire Type: Auditory Formality: Neutral	0.771	0.102	7.537
Questionnaire Type: Timed Textual	0.007	0.027	0.263
Questionnaire Type: Canonical Textual	0.015	0.061	0.240
Formality: Informal	−0.079	0.091	−0.862
Formality: Formal	−0.080	0.092	−0.869

Table 2.15: The fixed effects of our model, their estimates, standard errors, and t-values. (The full output for our R model can be found in the Appendix 2.6.2.)

ous constructions differ slightly; cf. e.g. (5) and (6). Thus, the "Differences Within Constructions"-condition is an important measure of how well chosen our categories are and it is good that the variance is quite low (0.042).

Effects by an item's position within a questionnaire, the order of the questionnaires, and the effect of the previous sentence are negligible factors (with their variances below 0.005). Thus, factors like fatigue effects ("Position in Questionnaires"), repetition effects ("Order of Questionnaires"), and order effects ("Previous Sentence") have very little impact on our results.

The residual variance is considerably higher than the variance of the specified factors and it accounts for 83.99% of the variance of the random factors. However, a residual variance of 0.383 also implies that the model explains a good deal of the data.

As to the fixed effects, the rating of an item hardly changes across modes of presentation (the estimate for auditory questionnaires is 0.771, for timed textual questionnaires 0.778, and canonical textual questionnaires 0.786). The factor of formality has virtually no impact (the estimate for the formal condition is 0.691 and the informal condition 0.692).[24]

To determine whether the mode of presentation and the formality of a questionnaire have a *significant* impact on the ratings, we compare our model to the respective null models. For example: The null model used to test for the significance of a questionnaire's formality equals the critical model excluding the factor formality. We then compare the null model formality to the critical model, using an ANOVA (cf. Winter, 2014:13, for details on this approach).

[24]The reader might wonder why the estimate of the formal and the informal condition comes out at about 0.69, while the estimate for the neutral condition comes out at 0.771. The reason for this is that the data for the formal and the informal condition come from the third experiment, while the data for the neutral condition come from the first and second experiment. As mentioned in Section 2.2.3, the participants of the first and second experiment and of the third experiment were sampled from different populations. This, in our view, leads to somewhat different mean ratings for some of the constructions, e.g. for the resumptive pronouns.

Neither the mode of presentation ($\chi^2 = 0.282$, p = 0.991) nor the formality of a questionnaire ($\chi^2 = 4.588$, p = 0.600) have a significant effect.[25]

Due to the lack of significance, there are no reasons to reject (H_0) and to adopt (H_1).

2.4 Discussion

The results suggest that neither the mode of presentation nor a questionnaire's formality have a significant impact on an item's rating. This applies to both the canonical textual questionnaire and the timed textual questionnaire. These findings can be seen as a limitation: Syntacticians cannot expect to use acceptability judgement tasks to study negative evidence (i.e. looking into the acceptability of rare constructions); they have to use other methods to do so, e.g. elicitation tasks. Further, AJTs are not useful to classify constructions with respect to their modality (e.g. in the manner of Miller and Weinert, 1998). For this, syntacticians have to resort to other methods.

These results also show that acceptability judgement tasks provide a robust measurement of grammatical acceptability: One does not have to pay attention to the mode of presentation of a questionnaire, i.e. one can use a textual questionnaire for constructions from spoken language and an auditory questionnaire for constructions from written language.

The results also make a wider point: Grammatical acceptability seems to be a fairly robust concept. Not only is a participant's grammatical intuition hardly influenced by a change in modality or formality, other factors like fatigue effects, repetition effects, and order effects have little bearing on the results, either.

[25] Here, formality of a questionnaire only includes the conditions "formal" and "informal". If we include the condition "neutral", then $\chi^2 = 19.194$, p = 0.084.

This is linked to another point: Despite the fact that the task description was somewhat vague (we asked participants "to evaluate grammatical acceptability of certain sentences" and "how natural do they sound to you with respect to their grammaticality?"), participants seem to have a good grasp both of the task and of the concept of grammatical acceptability.

Our research question can also be seen as an indirect comparison of corpus analyses and acceptability judgement tasks. Corpus analyses can be used to investigate differences in modality. We find that acceptability judgement tasks are not useful for this purpose. How can the differences between corpus analyses and acceptability judgement tasks be explained? And related to this: Do our findings have implications for Figure 2.1 (i.e. the distinction between constructions that primarily occur in spoken language, constructions that primarily occur in written language, and constructions that occur in both modalities)?

As to the first question: It is clear that while e.g. sentence initial gerunds are used in written language, they are rarely used in spoken language. We think the reason why this difference does not show up in acceptability judgement tasks hints at an interesting difference between corpus analyses and acceptability judgement tasks. Corpus analyses are clearly an observation of language production, measuring frequency; acceptability judgement tasks, on the other hand, are rather a product of language perception, measuring grammatical acceptability. This might explain why frequency and acceptability do not always agree; and it might also explain the observed difference between corpus analyses and acceptability judgement tasks with respect to modality. However, these thoughts are somewhat speculative and require further research.

With this in mind, we need to qualify Figure 2.1. The differences between $S - W$,

$W - S$, and $S \cap W$ seem to apply to language production (and are reflected by frequencies), but not to the perception of grammatical acceptability.

This, however, does not mean that people are insensitive to differences in modality when perceiving language in an acceptability judgement task. In our task, we asked our participants to evaluate exactly that: *grammatical acceptability*. If we asked our participants to evaluate something like "social acceptability" or "perceived poshness" and presented the textual questionnaire as a document of written language and the auditory questionnaire as a document of spoken language, then there is a good chance that we would have found differences between modalities. This is not an outrageous claim: People quickly detect if someone talks like a course book.

There are various concerns. First, there is a noticeable difference in ratings between the first and second experiment and the third experiment: The resumptive pronouns received considerably higher ratings in the third experiment. We think this difference is caused by sampling from different populations. For the first and second experiment, we sampled students who studied at Stanford University at the time. For the third experiment, we sampled Amazon Mechanical Turk users.

There are various differences between the two populations. Stanford students are young and highly educated (on average, they are younger than our Amazon Mechanical Turk participants, cf. the "Participants" subsections for the three experiments). But there is another difference: Our Amazon Mechanical Turk users were required to have previous task approvals and a high approval rate, i.e. they have a good deal of experimental experience. So, while our Amazon Mechanical Turk users are highly familiar with judgement tasks, this might or might not be the case for the Stanford students.

Culbertson and Gross (2009) found that task familiarity is an important factor, out-

weighing other factors like linguistic experience. It is not unreasonable to believe that the differences in ratings are also due to differences in task familiarity.

Second, the reader might worry about the fact that we add the linear model *post hoc*. The reason why we add a new analysis is because the initial analysis (using t-tests and TOSTs) was cumbersome (as it involves various sub-hypotheses and a lot of tests) and opaque (the old criteria that define support for (H_1) are somewhat arbitrary). Further, as we were only looking at means of constructions, we could not check for any interactions between various factors. The linear mixed effect model presented here is a simpler and more transparent analysis.

This is why we firmly hold the view that adding the model to further understand the ratings did not do any harm. The following would have been true bad practice: Unsatisfied with our findings, we would have poked around for the "right" analysis. Once found, we would then have only reported those results (cf. Simmons et al., 2011, for a discussion of bad practices that increase false-positive research). Clearly, we did not do this.

Third, one might question how "informal" our informal condition in Section 2.2.3 is. A truly informal study could look a lot more informal than the experiment in Section 2.2.3. One might, for example, invite the participants to one's home and administer a pen and paper questionnaire in a relaxed atmosphere; all this, of course, in a controlled and scientific way.

That is true; but we wish to add to our defence: We intended to hand out an informal version of a typical textual questionnaire that could be taken online. Against the backdrop of these requirements, we think that there was little leeway and we made good use of it.

Fourth, some readers might worry about the number of constructions we are investigating. Are four constructions, two from each modality, sufficient? Why didn't we randomly sample constructions from the literature, as done in Sprouse et al. (2013) or in Chapter 3?

The problem with such an approach is that there is no readily available pool from which to sample possible constructions. For many constructions discussed in the literature, it is not clear whether they belong to $S - W$, $W - S$, or $S \cap W$ (cf. Figure 2.1). It is true that if our results had shown the very opposite (i.e. assume that the mode of presentation had made a difference for the four constructions under investigation), then more research would have been needed to solidify such an outcome. However, with the results that we actually find, one could argue as follows: Those four constructions are "extreme" in a sense that they clearly belong to either spoken or written language. Since we have a negative result with these "extreme" constructions, it is unlikely that we would have had a positive result with less "extreme" constructions.

Related to this is the worry about our choice of constructions. The constructions were implicitly chosen such that they were non-dialectal and not socially stigmatised (explicitly, they were such that the literature attests differences in usage across modalities). That is, it could be the case that the criteria for choosing our constructions might have increased the probability of a negative result and that other constructions might have led to different results.

For instance, the "wanna" construction is of interest to syntactic theory with respect to the discussion about the existence of traces (e.g. "I wanna/want to play the bagpipe", cf. Postal and Pullum, 1978). However, because of stigmatisation effects,

it might be hard to get adequate acceptability ratings with textual questionnaires and an auditory questionnaire might lead to different ratings. We do not include such cases, as any observed effect would probably have been a point about colloquialisms, social stigmatisation, and modality (and not so much about syntax and modality). However, it might still be worth looking into this matter further.

Also, the constructions are not perfectly balanced with respect to their source. Of the two constructions from spoken sources, there is one construction that receives high ratings and one that receives low ratings (the alternative if-clauses and the resumptive pronouns, respectively). However, for the two constructions from written sources, both constructions receive high ratings (i.e. the wh-infinitives and the sentence-initial gerunds). Our analysis, though, should not have been affected by this asymmetry.

2.5 Conclusion

We have asked whether acceptability judgement tasks can be used to investigate differences in modality and designed a study to test this question. Our study consists of two questionnaires, one with an auditory mode of presentation, the other with a textual mode of presentation. In both questionnaires, we included constructions that either mainly occur in spoken language or mainly in written language. We compared those constructions across questionnaires and found that the mode of presentation has no significant effect on the ratings of the constructions under investigation.

Our textual questionnaire is unusual in the sense that items were time-constrained (i.e. they disappeared after a while) and the question arose whether our results hold for canonical textual questionnaires, as well. Consequently, we added a third questionnaire in a second experiment, which had no such time-constraints. We then compared the time-constrained textual questionnaire to the canonical one and

found no difference between ratings across questionnaires. From this we concluded that our results regarding the mode of presentation extends to canonical textual questionnaires, as well.

In a third experiment, we also controlled for formality, i.e. introducing informal and formal questionnaires. However, the factor formality does not affect the results, either.

We also found that other factors like fatigue effects, repetition effects, and order effects have little impact on the ratings. By and large, it seems, grammatical acceptability measured through acceptability judgement tasks is a fairly robust concept, certainly when it comes to changes in experimental design.

2.6 Chapter 2: Appendices

2.6.1 Experimental Stimuli

Alternative If-Clauses (Mainly Spoken Language)
If she would come to see things for herself, she would change her mind immediately.
If I would go to New York myself, I might get the chance to meet Lady Gaga.
If you would go on such a journey, you would understand what he is talking about.
If we would spend every day connecting with people, we could learn more about the world, too.
Sentences with Resumptive Pronouns (Mainly Spoken Language)
We are afraid of things that we don't know what they are.
You fear things that you don't understand what they are.
This is a donkey that I don't know where it lives.
This is the man that I don't know where he comes from.
Sentence Initial Gerunds (Mainly Written Language)
Their being unaware of the situation really annoyed Rob.
Her being so happy makes all of us happy, too.
Our being late will upset Sarah.
My being focused on the current project does impress our boss.
Wh-Infinitives (Mainly Written Language)
We found a splendid house in which to spend our holiday.
I finally found a store in which to buy new running shoes.

He discovered a nice hotel in which to stay the next two weeks.
They moved to another state in which to sue the company.
Good Sentences (Spoken Source)
You can also make the argument that they didn't approve anything.
And in his absence, the Senate Republicans pushed through a redistricting plan for the State Senate.
It is our job to figure out what happened and do everything we can to prevent it from happening again.
Stephen Colbert always threatens to run for office in South Carolina, but of course, always with a wink.
Good Sentences (Written Source)
The French embassy in the capital warned that the nuclear cloud could reach the city within hours.
As Obama's top counterterrorism adviser, Brennan has helped manage the drone program.
The project would more than double the population of Benewah County, home to 9,000 people.
Iran has proposed restarting talks as early as next month.
Bad Sentences (Spoken Source)
If was key word "hope" in 2009, perhaps "change" key word 2013.
I feel was really thrown to wind caution these times and really far more aggressive he were.
First term President Obama start out with numbers of envoy, I think.
1986 the last time was really it happened.

Bad Sentences (Written Source)
Is William travel on behalf of Queen and is on his second official trip to country?
Consumer and investors remains warying of budget battles ongoing in Washington.
Resident stock enough food also would be required and water to last a year.
Compound include houses, school, hotel and firearms factory and museum.

2.6.2 R Output

```
#Our model for the Z-scores.

Linear mixed model fit by REML ['lmerMod']
Formula: crit_z_score ~ source * mode + formality * construction + (1 |
    sentence) + (1 | position_q:construction) + (1 | prev_construction) +
    (1 | quest_no:construction) + (1 | construction:subject)
   Data: data_all_CRIT

REML criterion at convergence: 10528.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.0403 -0.5347  0.0939  0.5606  3.8798

Random effects:
```

```
 Groups                   Name        Variance  Std.Dev.
 construction:subject     (Intercept) 0.0271737 0.16484
 sentence                 (Intercept) 0.0421181 0.20523
 position_q:construction  (Intercept) 0.0033374 0.05777
 quest_no:construction    (Intercept) 0.0000000 0.00000
 prev_construction        (Intercept) 0.0007477 0.02734
 Residual                             0.3828357 0.61874
Number of obs: 5348, groups:
construction:subject, 954; sentence, 40; position_q:construction, 30;
quest_no:construction, 18; prev_construction, 12

Fixed effects:
                                  Estimate Std. Error t value
(Intercept)                       0.771488   0.102354   7.537
sourceWS                          0.059127   0.089302   0.662
modectx                           0.014584   0.060767   0.240
modetxt                           0.007119   0.027046   0.263
formalityformal                  -0.079732   0.091710  -0.869
formalityinformal                -0.078764   0.091421  -0.862
constructionBAD                  -1.845352   0.127773 -14.442
constructionGER                  -0.598153   0.157214  -3.805
constructionIF                   -0.364200   0.157427  -2.313
constructionRES                  -1.557339   0.157130  -9.911
constructionWHI                  -0.226414   0.157270  -1.440
sourceWS:modectx                 -0.039931   0.083722  -0.477
sourceWS:modetxt                 -0.013564   0.036599  -0.371
formalityformal:constructionBAD   0.090233   0.159459   0.566
formalityinformal:constructionBAD 0.096003   0.159108   0.603
```

```
formalityformal:constructionGER    0.007226   0.130200   0.056
formalityinformal:constructionGER -0.014712   0.129642  -0.113
formalityformal:constructionIF    -0.014649   0.130477  -0.112
formalityinformal:constructionIF   0.097298   0.129923   0.749
formalityformal:constructionRES    0.389889   0.130160   2.995
formalityinformal:constructionRES  0.321176   0.129578   2.479
formalityformal:constructionWHI    0.095951   0.130245   0.737
formalityinformal:constructionWHI  0.071582   0.129689   0.552

Correlation of fixed effects could have been required in summary()

Correlation of Fixed Effects:
            (Intr) sorcWS modctx modtxt frmltyf frmltyn cnsBAD
 cnsGER cnstIF cnsRES
sourceWS    -0.437
modectx     -0.112  0.090
modetxt     -0.133  0.139  0.221
formltyfrml -0.332  0.001  0.063 -0.002
frmltynfrml -0.333  0.001  0.063 -0.001  0.858
cnstrctnBAD -0.626  0.003 -0.002 -0.002  0.248   0.249
cnstrctnGER -0.390 -0.268  0.017  0.000  0.200   0.201   0.405
constrctnIF -0.623  0.267 -0.026  0.003  0.201   0.202   0.407
 0.252
cnstrctnRES -0.624  0.267 -0.024  0.002  0.201   0.202   0.408
 0.253  0.408
cnstrctnWHI -0.390 -0.268  0.016  0.000  0.200   0.201   0.405
 0.408  0.252  0.253
srcWS:mdctx  0.051 -0.128 -0.694 -0.147  0.005   0.005   0.004
```

```
 -0.028  0.035  0.034
srcWS:mdtxt  0.091 -0.207 -0.150 -0.675 -0.001  -0.002   0.001
 0.000 -0.003 -0.002
frmltyf:BAD  0.179 -0.002  0.002  0.001 -0.510  -0.428  -0.543
 -0.115 -0.116 -0.117
frmltyn:BAD  0.179 -0.002  0.002  0.001 -0.428  -0.509  -0.544
 -0.115 -0.116 -0.117
frmltyf:GER  0.219 -0.005 -0.040  0.002 -0.623  -0.523  -0.174
 -0.344 -0.139 -0.139
frmltyn:GER  0.220 -0.005 -0.040  0.001 -0.524  -0.624  -0.175
 -0.346 -0.140 -0.140
frmltyfr:IF  0.215  0.005  0.052 -0.005 -0.623  -0.523  -0.174
 -0.139 -0.347 -0.144
frmltynf:IF  0.216  0.005  0.052 -0.004 -0.523  -0.623  -0.175
 -0.140 -0.348 -0.144
frmltyf:RES  0.215  0.005  0.049 -0.003 -0.624  -0.524  -0.174
 -0.140 -0.144 -0.343
frmltyn:RES  0.216  0.006  0.050 -0.003 -0.525  -0.624  -0.175
 -0.140 -0.145 -0.345
frmltyf:WHI  0.219 -0.005 -0.040  0.001 -0.623  -0.523  -0.174
 -0.143 -0.139 -0.139
frmltyn:WHI  0.220 -0.005 -0.040  0.000 -0.523  -0.623  -0.175
 -0.144 -0.139 -0.140
            cnsWHI srcWS:mdc srcWS:mdt frmltyf:BAD frmltyn:BAD
 frmltyf:GER frmltyn:GER
sourceWS
modectx
modetxt
```

```
formltyfrml

frmltynfrml

cnstrctnBAD

cnstrctnGER

constrctnIF

cnstrctnRES

cnstrctnWHI

srcWS:mdctx -0.029

srcWS:mdtxt  0.002  0.218

frmltyf:BAD -0.115 -0.003    -0.001

frmltyn:BAD -0.115 -0.003    -0.001     0.907

frmltyf:GER -0.143  0.061    -0.002     0.359       0.301

frmltyn:GER -0.144  0.062    -0.001     0.301       0.358

       0.815

frmltyfr:IF -0.139 -0.070     0.005     0.358       0.300

       0.434       0.364

frmltynf:IF -0.140 -0.071     0.004     0.301       0.358

       0.364       0.434

frmltyf:RES -0.139 -0.069     0.003     0.359       0.301

       0.435       0.365

frmltyn:RES -0.140 -0.069     0.003     0.302       0.359

       0.365       0.436

frmltyf:WHI -0.345  0.063    -0.004     0.358       0.301

       0.443       0.373

frmltyn:WHI -0.346  0.063    -0.002     0.301       0.358

       0.373       0.444

            frmltyf:IF frmltyn:IF frmltyf:RES frmltyn:RES frmltyf:WHI

sourceWS
```

modectx					
modetxt					
formltyfrml					
frmltynfrml					
cnstrctnBAD					
cnstrctnGER					
constrctnIF					
cnstrctnRES					
cnstrctnWHI					
srcWS:mdctx					
srcWS:mdtxt					
frmltyf:BAD					
frmltyn:BAD					
frmltyf:GER					
frmltyn:GER					
frmltyfr:IF					
frmltynf:IF	0.816				
frmltyf:RES	0.444	0.374			
frmltyn:RES	0.374	0.444	0.814		
frmltyf:WHI	0.433	0.364	0.435	0.365	
frmltyn:WHI	0.364	0.434	0.365	0.435	0.815

Chapter 3

Comparing Data Transformations for Syntactic Judgement Data

3.1 Introduction

In the past 20 years, syntactic theory has seen an increase in the use of experimental methods. The main motivation for this change is the sentiment that the prevalent method of syntactic enquiry, researcher introspection, is inferior to other more formal methods (in researcher introspection, the investigating linguist is his/her own informant). Common formal methods include acceptability judgement tasks, elicitation tasks, eye tracking studies, etc. They are regarded as formal, because they adhere to certain scientific standards (for details, cf. the discussion throughout Chapter 1), which makes their results reproducible and more reliable. Two types of questions surround the use of formal methods: 1) *Why* should formal methods be used? 2) *How* are formal methods best used?

The debate around the first question centres around another, closely connected

question: Is researcher introspection an adequate method of syntactic enquiry? In Chapter 1, we argue that acceptability judgement tasks are indeed preferable to researcher introspection (based on a quantitative comparison of the two methods).

If one accepts the need for formal methods, then certain "how" questions arise. This chapter is concerned with such a "how" question and our focus is on acceptability judgement tasks (acceptability judgement tasks and corpus analyses are the most common formal methods of syntactic enquiry). In an acceptability judgement task, the researcher asks his/her participants to judge the acceptability of certain stimuli. The judgements are based on the participants' intuitions.[1] In this chapter, we ask how different data transformations (scaled ratings, ordinal data, and Z-scores) compare to each other in terms of their ability to detect true differences.

"How" questions are relevant, because making the right methodological choices can help prevent both false positives (i.e. detecting differences that do not exist) and false negatives (i.e. failing to detect real differences). Both false positives and false negatives will result in a degraded empirical basis, which will have a negative effect on syntactic theory building.

However, not all "how" choices are equally important. Some choices do affect the results significantly (e.g. Culbertson and Gross, 2009, report that a participant's task familiarity has a significant impact on the results), other factors do not (e.g. Weskott and Fanselow, 2011, argue that the choice of measurement method has little effect on the results). In a previous project (Juzek, unpublished manuscript), we showed that applying Z-scores to one's data is a methodological choice that does improve the results significantly. In the present chapter, we address the main weaknesses of

[1]Such intuitive responses contrasts with the notion of grammaticality, which we treat as a technical notion that refers to the output of a syntactic theory or model (also see Section 1.2.1 in Chapter 1).

the previous project.[2] We also include ordinal data.

OVERVIEW We provide further background in Section 3.2. In 3.2.1, we present the main methodological choices one faces when running an acceptability judgement task. Part of the relevance of this chapter comes from the fact that syntacticians make little use of Z-scores (or ordinal data), as shown in Section 3.2.2. This is a problem, because Z-scores can improve the quality of one's data by reducing the effects of scale biases. We wish to provide empirical evidence for this. However, before doing so, we first need to properly introduce the most common measurement methods (Section 3.2.3) and the most common data transformations (Section 3.2.4). In Section 3.3, we give empirical evidence of how the different data transformations compare. We randomly sampled 36 sentences from the literature and collected judgements on them in an acceptability judgement task, using common measurement methods. We then apply the different data transformations and test how well they perform in detecting the differences between the 36 sentences. Section 3.4 discusses the findings and the chapter concludes in Section 3.5.

3.2 Further Background

3.2.1 Making Methodological Choices for Acceptability Judgement Tasks

Syntacticians who wish to conduct an acceptability judgement task face various methodological choices. The impact of a broad variety of methodological choices has been the subject of debate in classical test theory and psychology in general (Box, Hunter, and Hunter, 1978, and Foster and Parker, 1995, are good starting points). In

[2]Our thanks go to two Language and Speech reviewers for their thorough feedback.

many instances, findings from test theory and psychology "trickle down" to linguistics and its subfields (e.g. based on findings in psychology, linguists acknowledge the importance of randomisation of item order, including fillers to conceal the study's purpose, etc.; cf. e.g. Cowart, 1997).

However, there is also a syntax-specific discourse. This concerns, for instance, potential differences between different measurement methods on syntactic judgements: E.g. Bard et al. (1996) argued in favour of Magnitude Estimation for syntactic purposes; Featherston (2008) and Featherston (2009) made a case for the "Thermometer Method"; Sprouse (2009) questioned how well the mechanism of Magnitude Estimation works for syntactic acceptability; and Weskott and Fanselow (2011) quantitatively showed that there are few differences between common measurement methods. Other syntax-specific discussions concern a participant's confidence in his/her judgement (cf. Fanselow, Häussler, and Weskott, 2013), the number of participants needed (Mahowald et al., submitted, showed that as few as seven participants can be sufficient for reliable judgement data), or potential differences between offline and online questionnaires (e.g. Munro et al., 2010). Cowart (1997) is an excellent starting point to better understand various "how" questions concerning syntactic acceptability judgement tasks.

In a previous project (Juzek, unpublished manuscript), we argued for using Z-scores on syntactic judgement data by quantitatively comparing non-normalised data to data normalised using Z-scores. The project faced several lines of criticism. First, it was argued that the benefits of Z-scores are well-established and that Z-scores are already in standard use. Second, our quantitative comparison of non-normalised and normalised data was based on only a few selected sentences. This, it was argued, limited our results. And third, we did not include other data transformations, in particular ordinal data.

We agree with the second and third point and below, we address these weaknesses (by randomly sampling sentences from the literature and by including ordinal data). However, as to the first point, this is a misguided argument. The argument consists of two claims: First, the benefits of Z-scores are well-established. And second, Z-scores are in standard use already. Both claims are true, however, only for psychology in general. From this, one cannot make claims about linguistics. While it is true that many findings from general psychology apply to syntactic enquiry, this is not a necessity, as e.g. the debate around the advantages of Magnitude Estimation for experimental syntax shows. Thus, with respect to the benefits of Z-scores, it could well be the case that using Z-scores has a considerable impact on the results from studies in the area of market research but a comparably small impact on the results from syntactic studies. The second claim (Z-scores are in standard use already) is based on the same unfortunate inference from general psychology to linguistics. To illustrate why, we looked at the talks given at two recent linguistics conferences: The LSA 2013 meeting and the Linguistic Evidence 2014 conference.

3.2.2 The Use of Z-Scores in Linguistics

At the LSA 2013 meeting, there were 46 talks with extended abstracts from all linguistic areas. Sixteen included experimental data ("experimental" should not be confused with "quantitative"; e.g. we did not count corpus analyses as experimental), three of which used judgement data (one syntax talk, one psycholinguistics talk, and one phonetics talk). None of those three papers used Z-scores. At the Linguistic Evidence 2014 conference, there were seventeen talks with extended abstracts, of which 14 included experimental data. Eight talks included judgement data (seven syntactic talks and one phonetics talk). Four of those eight used Z-scores (three syntax talks and one phonetics talk). Figure 3.1 illustrates this.

Of course, we cannot draw a definitive conclusion about the field from looking at

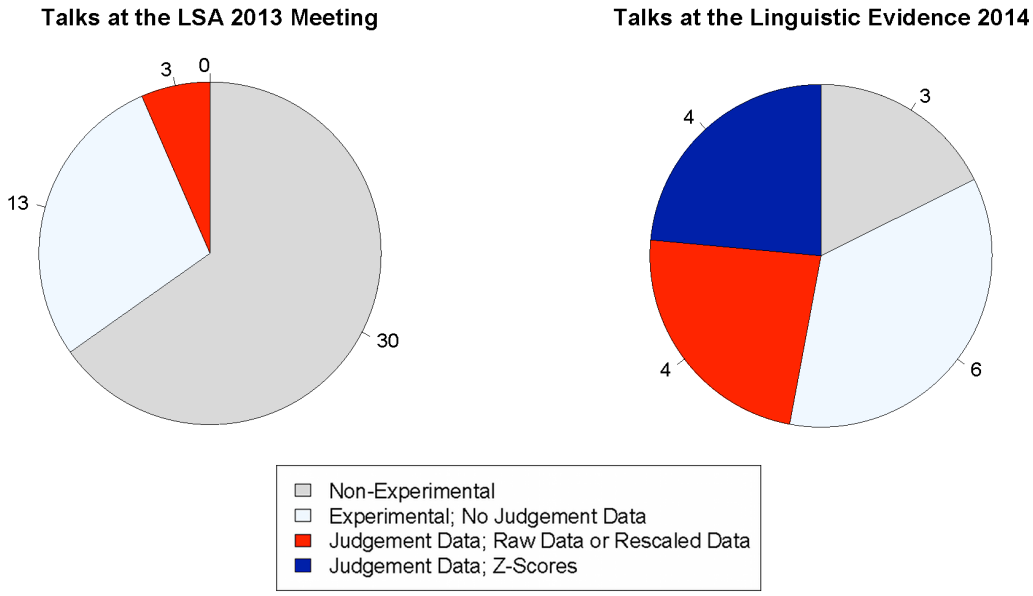


Figure 3.1: An illustration of the prevalence of Z-scores used on judgement data in linguistics. We looked at the talks given at the LSA 2013 meeting (left; 3.1a) and at the Linguistic Evidence 2014 conference (right; 3.1b) and counted how many of them were experimental (light grey, red, and blue) and how many of them used judgement data (red and blue). For the talks using judgement data, we checked whether they used Z-scores (blue) or not (red). Starting with *Non-Experimental*, the chart gives the categories in clockwise direction. (N.B.: Ordinal data was not used at all.)

talks from only two conferences; however, having looked at 63 talks in total should give us a good idea of the state of the field. If using Z-scores was truly standard practice in linguistics, then these numbers would look different: Almost all talks including judgement data would be using Z-scores. We will not speculate why this is not the case. Whatever the reason, a quantitative comparison of different data transformations (including Z-scores) is a worthwhile endeavour.

Our focus is on syntactic data, so we decided to collect judgement data for such a comparison. We also decided to collect our data using several measurement methods (Likert Scales, the Thermometer Method, and Magnitude Estimation), because it is possible that certain transformations affect the results more or less, depending on the measurement method that was used.

3.2.3 Measurement Methods

In this section we introduce four of the major measurement methods that can be used in acceptability judgement tasks: Likert Scales (binary and gradient), Forced Choice, the Thermometer Method, and Magnitude Estimation. Apart from Forced Choice, these methods feature in the experiment in Section 3.3. Our main concern is with their workings and the motivation why they were introduced to linguistics.

BINARY LIKERT SCALES When using a Likert Scale (named after Rensis Likert, cf. Likert, 1932), participants rate stimuli on a pre-set scale, chosen by the researcher. A binary Likert Scale is a special case of a Likert Scale in which only two degrees, e.g. “0” and “1” or “no” and “yes”, are used. Some syntacticians prefer binary scales for theoretical reasons: They view grammaticality as a binary concept (about 20% of the standard acceptability judgements in Linguistic Inquiry for the years 2001 to 2010 were judged using binary judgements; cf. Chapter 1). A researcher

who advocates a binary concept of grammaticality might use binary categories in researcher introspection (using "*" and *unmarked*). And, if that researcher was to conduct an experiment, he/she might make his/her participants use a binary Likert Scale, too, so that participants face a similar strong good-vs-bad-choice just as the linguist did. Figure 3.2a illustrates a binary Likert Scale (for a larger figure, see page 51).

GRADIENT LIKERT SCALES A gradient Likert Scale is a Likert Scale with more than two degrees. Common numbers of degrees range from three to ten, although there is no upper limit (according to Cox, 1980, and to Weijters et al., 2010, 7-point scales are the most common Likert Scales in market research; anecdotally, we think this also holds for linguistics). With respect to syntactic studies, one extreme of the scale represents "unnatural" or "unacceptable" and the other "natural" or "acceptable"; with the remaining values denoting various degrees in between (e.g. "rather unnatural" or "rather natural"). The motivation for using a gradient Likert Scale in contrast to a binary Likert Scale could be that the investigating linguist advocates gradience in both grammaticality and acceptability (for an in-depth discussion of gradience in grammar theory, see Keller, 2000, and Sorace and Keller, 2005). The motivation for linguists to use a Likert Scale in general (including binary Likert Scales) is that they are a reliable method (cf. Weskott and Fanselow, 2011) that is easy to use for both the researcher and the participants of a survey. Figure 3.2b illustrates a gradient Likert Scale (for a larger figure, see page 51).

FORCED CHOICE In this method, a number of stimuli are presented to the participants, who then have to make a choice for or against one of the items based on criteria set by the researcher. In syntactic experiments, typically two stimuli are presented and participants are asked to mark the item that is more

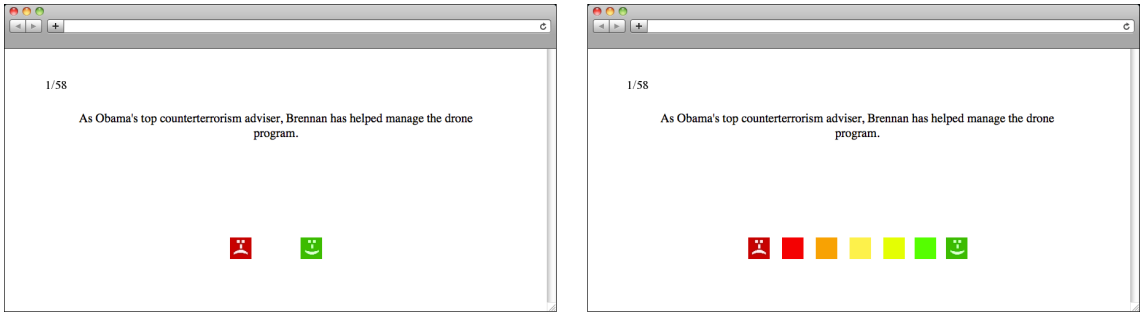


Figure 3.2: An acceptability judgement task using a 2-point scale (left; 3.2a) and a 7-point scale (right; 3.2b).

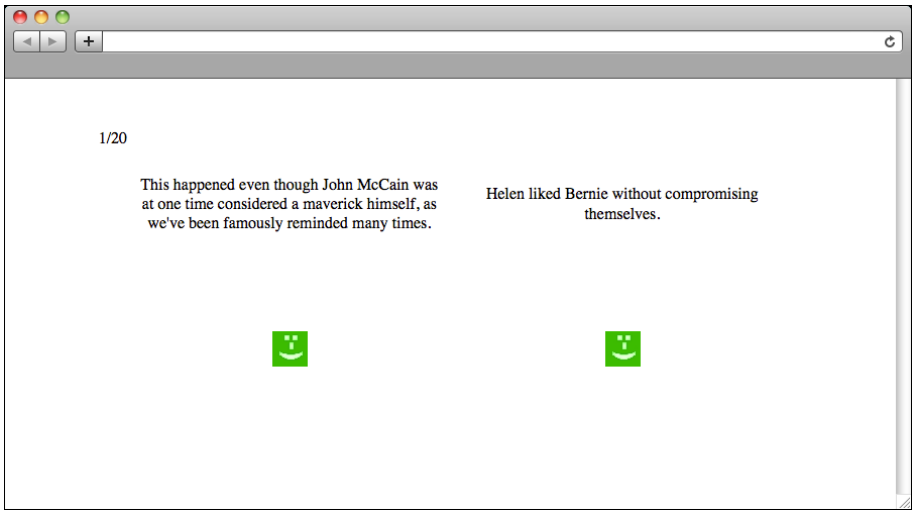


Figure 3.3: An acceptability judgement task using Forced Choice. In this example, the participant has to choose the sentence that sounds more acceptable to him/her.

acceptable. Forced Choice is typically preferred by syntacticians who think that syntactic research is best conducted by pairwise comparisons. We do not include Forced Choice in the comparison below. This is partly due to the fact that we do not think that syntactic theory should restrict itself to the analysis of sentence pairs (cf. the discussion in Chapter 1, Section 1.2.5) and partly because of another major disadvantage: The ratings of two Forced Choice experiments are hard to compare, unless the very same pairs are used in both experiments (which is unlikely across researchers). For instance, a mediocre sentence might be chosen 80% of the time if compared to a bad sentence, but only 20% of the time if compared to a good sentence. Figure 3.3 illustrates a Forced Choice study.

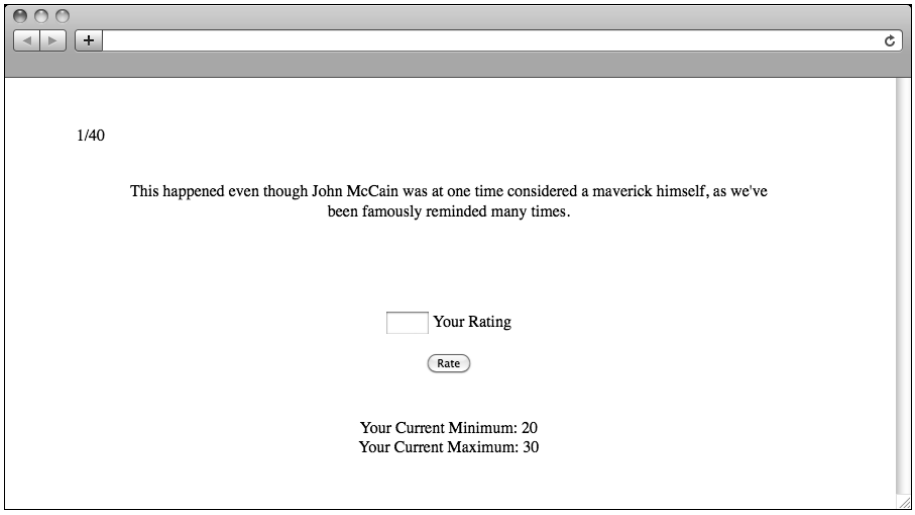


Figure 3.4: An acceptability judgement task using the Thermometer Method. The participant is asked to put his/her rating in relationship to an individual minimum and maximum (which are shown at the bottom of the website).

THERMOMETER METHOD The Thermometer Method is a self-anchoring scale, i.e. participants can choose their own scale (as described by Kilpatrick and Cantril, 1960). The analogy of a thermometer to introduce a self-anchoring scale was to our knowledge first used by Nugent (2004) in the area of social work practice. Featherston (2008) then introduced the method to linguistics. When using the Thermometer Method, participants choose a minimum and maximum value prior to starting the survey (e.g. "20" and "30", respectively), but they are allowed to exceed these pre-set extrema as they progress through the questionnaire (to give an example from a syntactic experiment: A truly bad sentence, worse than anything the participant imagined, might receive a "15", although the pre-set minimum was "20"). It might also happen that participants leave parts of the scale unused (e.g. with the pre-set range of "20" to "30", a certain participant might not rate anything worse than "23"). The possibility of participants choosing their own scales and number of degrees and the possibility of leaving parts of the scale unused is an interesting mixture of features of Magnitude Estimation and Likert Scales, respectively (in Magnitude Estimation, the former holds but not the latter, and vice versa for Likert Scales). The method is illustrated in Figure 3.4.

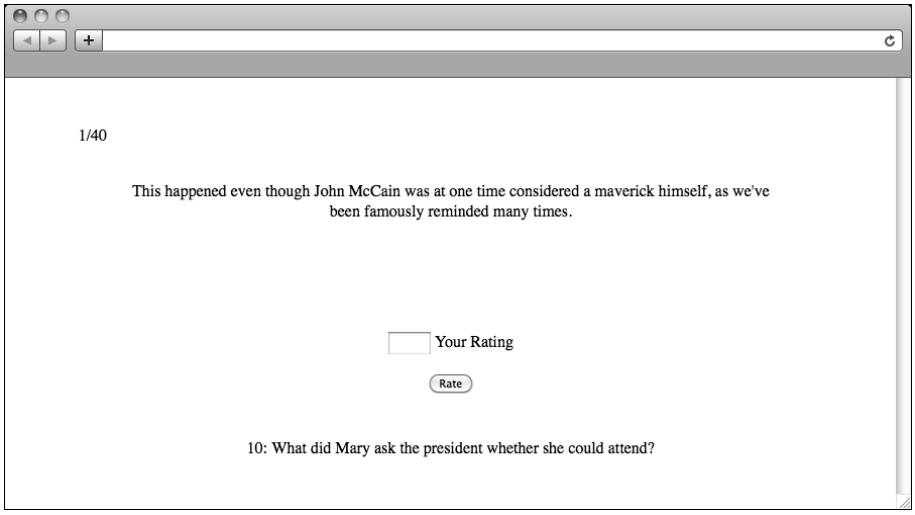


Figure 3.5: An acceptability judgement task using Magnitude Estimation. The participant is asked to put his/her rating in relationship to a reference sentence, the standard (which is shown at the bottom of the page; the "10" in front of the standard is the modulus, i.e. the concrete rating that the participant gave to the standard.

MAGNITUDE ESTIMATION In Magnitude Estimation, participants first give a rating to a reference sentence, the *standard*, and then put all other items in relation to the standard (the standard is the same sentence for all participants). This way, participants effectively choose their own scale. Magnitude Estimation was introduced to psychology by Stevens (1946) and to linguistics by Bard et al. (1996). Stevens introduced Magnitude Estimation as it reflects the nature of certain psychophysical concepts (e.g. lumen). These percepts have an absolute zero, no absolute maximum, and are perceived as gradient. There are issues, though: The standard has to be well chosen, as a low rating for the standard will result in higher ratings for the critical items and a standard that receives a high rating will push the ratings for the critical items down. This can make it hard to compare results across studies. (N.B.: The concrete rating that a participant gives to the standard is called the *modulus*.) Figure 3.5 illustrates a study using Magnitude Estimation.

There is a debate as to which measurement method works best for the concept of syntactic acceptability (e.g. Bard et al., 1996, Featherston, 2008, Featherston, 2009,

Sprouse, 2009, Weskott and Fanselow, 2008, Weskott and Fanselow, 2009). In our view, the debate reached a turning point with Weskott and Fanselow (2011), who conclude (p. 271):

> Given our results, we can outright reject this hypothesis [i.e. "that ME judgments as a measure of linguistic acceptability are more informative than binary or seven-point judgments"]. Not only did we not find any indication of an informativity difference in terms of the JUDGETYPE factor in the three experiments reported, but we also did not find an effect for which our measure of effect size, the eta-squared value, was substantially higher for the ME data than for the other two types of data [i.e. a binary Likert Scale and a gradient Likert Scale].

In our investigation of the different data transformations, we decided to include all major measurement methods (binary Likert Scale, gradient Likert Scale, the Thermometer Method, and Magnitude Estimation), because of the possibility that a certain measurement method is more or less affected by a certain data transformation (i.e. there might be interactions).

3.2.4 Common Data Transformations

In this section, we introduce common data transformations used on syntactic acceptability judgements. These include the different basic transformations for the measurement methods, Z-score normalisation, and transformation to ordinal data. All of these feature in Section 3.3.[3] However, the starting point for any data transformation is raw data, which is why we start by looking at the structure of each measurement method's raw data.

RAW DATA The structure of the raw data is quite distinct for each measurement method.

[3]This list is not exhaustive. For instance, we did not include Bayesian approaches, as they are less common in linguistic analysis.

For *binary Likert Scales*, concrete ratings will take one of two values, e.g. "0" and "1" or "no" and "yes". Theoretically, it is possible that a participant does not use both degrees of the scale, but this is unlikely. If this happens on a large scale (i.e. several participants only use one value), then this points to a severe flaw in the experimental design (e.g. missing extreme items, missing control items, etc.) or in the sampling procedure of the participants.

For *gradient Likert Scales*, concrete ratings will be within the range of the pre-set number of degrees. For instance, for a 7-point scale (our choice for a gradient Likert Scale in Section 3.3), ratings would range from "1" to "7". However, it is quite possible that not all degrees of the scale will be used. Sometimes, a certain value will not be used simply by chance (e.g. on a 100-point scale, it might very well happen that the value "42" does not get used at all). Sometimes, though, this behaviour is more systematic: A participant might not use any "in-between" values, because he/she rejects the idea of gradience. Or a participant does not use extreme values at all, because 1) he/she thinks that there were no extreme items in the questionnaire or 2) he/she has a scale bias (i.e. inter-speaker differences in the application of the scale).

In order to avoid 1), researchers need to make sure that they include extremely bad and extremely good items. For 2), see the paragraph on Z-scores below. Both binary Likert Scale data and gradient Likert Scale data can be statistically analysed in its raw form. Raw data for binary and gradient Likert Scales is statistically equivalent to their rescaled counterparts,[4] which is why we only included rescaled data (below "basic data") in the quantitative comparison in Section 3.3.

For the *Thermometer Method*, raw ratings can theoretically take any value (unless the researcher asks participants to not exceed certain boundaries), as participants can choose their own scales. This is why, similar to Magnitude Estimation, raw

[4]That is, the way they are commonly rescaled; see *basic transformations* below.

ratings from the Thermometer Method are harder to interpret and, with respect to statistical analyses, not very useful. Further, there might be unused parts of a Thermometer scale, similar to the raw ratings from Likert Scales. However, there is a crucial difference: Thermometer scales were chosen by the participants, Likert Scales were pre-set by the researcher.

For *Magnitude Estimation*, raw ratings can take any value above "0" (unless the researcher asks participants not to exceed an upper boundary), as participants can choose their own scales, too. The raw ratings can also vary greatly between participants, because participants might choose radically different moduli (e.g. one participant might have given the standard a rating of "10" and used ratings from "1" to "20", another participant, however, might have given a "500" to the standard and gave ratings ranging from "1" to "1000"). This is why, with respect to statistical analyses, raw ratings from Magnitude Estimation are of little value.

BASIC TRANSFORMATIONS For *binary Likert Scales* and *gradient Likert Scales*, the basic transformations are identical: Ratings are rescaled using the pre-set extrema, so that rescaled ratings range from "0" to "1". For binary Likert Scales, ratings only need this transformation if replies were non-numerical (e.g. "unnatural" and "natural"). The rescaled data is statistically equivalent to their raw counterpart (unlike raw data vs rescaled data from the *Thermometer Method* and *Magnitude Estimation*). The reason for this is that participants use their own scales. Thus, when rescaling a participant's Thermometer ratings, one needs to take into account that each participant has his/her individually set minimum and maximum.[5] The same is true for Magnitude Estimation: Each participant

[5]The Thermometer Method: The formula for any given "basic" rating r_s is as in Equation 3.1 below, where r_r is the raw rating, min_S is the pre-set or adjusted minimum (for the participant in question), and max_S is the pre-set or adjusted minimum (for the participant in question).

$$r_s = \frac{(r_r - min_S)}{(max_S - min_S)} \tag{3.1}$$

has an individual rating for the standard, so some form of data transformation is required. In the basic transformation for Magnitude Estimation one has to divide a participant's critical ratings by a participant's modulus.[6] Raw data that have been transformed by the mentioned basic transformations are what we call "basic data" in the following.

PROBLEMS WITH BASIC DATA There are two major issues with basic data: scale biases and varying intervals.

A *scale bias* is a subject's tendency towards certain parts of the scale, due to an individual interpretation of the values of the scale. Scale biases can make it hard to compare ratings across subjects and make the overall data less powerful in statistical analyses (Takahashi, 2009). For instance, assume that, on a 7-point scale, Subject A tends to give a lot of high ratings to the list of critical items, while Subject B tends towards lower ratings. Arguably, to Subject A a "7" means something different than to Subject B: For Subject A, a "7" might mean "very good", but for Subject B, it could be interpreted as "truly exceptional". A strong scale bias can also be the reason why a subject leaves parts of the scale unused. See Chapter 1, Section 1.2.2 for an in-depth discussion and illustration of scale biases. A calibration phase at the beginning of a questionnaire and well chosen, counterbalancing items should help reduce the effect of scale biases.

Another problem are *unequal intervals* between different parts of the scale. For instance, when using a 7-point scale, a subject might perceive the distance between "1" and "2" as different to the distance between "3" and "4". For some psychophysical concepts, including syntactic acceptability, it is not clear whether one can assume

[6]Magnitude Estimation: The formula for any given "basic" rating r_s is as in Equation 3.2, where r_r is the raw rating and mod_S is a participant's modulus.

$$r_s = \frac{r_r}{mod_S} \tag{3.2}$$

equidistant intervals (cf. e.g. Stevens, 1946, Poulton, 1989, Schütze, 1996, Bard et al., 1996).

These two issues were important reasons for why Stevens (1946) introduced Magnitude Estimation to psychology and why Bard et al. (1996) followed suit for linguistics. They are also the motivation for transforming judgement data to Z-scores or to ordinal data, as these two transformations mitigate the effects of scale biases and unequal intervals.

Z-SCORES Z-scores help mitigate scale biases and unequal intervals. To transform a rating to Z-scores (z in Equation 3.3 below), the following is done for each participant (the subscripted "S"): The mean of all ratings (μ) is subtracted from an item's rating (x) and then the difference is divided by the standard deviation of all ratings (σ) (from Takahashi, 2009:72).

$$z = \frac{x - \mu_S}{\sigma_S} \tag{3.3}$$

In a syntactic context, this normalisation is typically applied per participant (otherwise it would not mitigate the effects of scale biases and unequal intervals). For any given rating, the quotient will indicate how many standard deviations the rating lies away from the mean. At first, results can be harder to interpret, as normalised ratings can look cryptic (a Z-score near zero is close to the participant's average; a score > 1 indicates a high rating for that participant, a score < 1 indicates a low rating).

There are problems with Z-scores. First, they could destroy real tendencies: What if a participant had no scale bias, but these were his/her genuine ratings? Second, another problem is that ratings become more context sensitive: Including a large number of bad items (e.g. by including too many bad fillers) will drive the Z-scores

for all items up (including bad and mediocre items); vice-versa, if there is an excess of good items. This "context sensitivity" can make it hard to compare normalised results across studies that use different sets of items (it also requires that each participant gets a similar fraction of bad and good items). Thirdly, Z-scores fully undermine the mechanics of Magnitude Estimation (as the modulus is completely ignored). They also slightly subvert the Thermometer Method (if a participant did not use his/her entire scale, then Z-scores will ignore the lower and upper boundaries set by the participant) and Likert Scales (if a participant did not use the entire scale set by the researcher, Z-scores will ignore the lower and upper boundary).

ORDINAL DATA By ranking each participant's ratings, commonly in ascending order, raw data is transformed to ordinal data. Typically, this is done to solve the issue of unequal intervals. However, ordinal data are subject to a similar issue as Z-score data. First, they are context sensitive: Including a large number of bad items can push up the rank of a mediocre item, and vice versa by including too many good items. Second, ordinal data requires different statistical tests compared to basic data and Z-scores, because ordinal data are not normally distributed. And the method loses statistical power when there are many equal values in a slice of data, so it works better for results from measurement methods with numerous degrees (typically Magnitude Estimation and the Thermometer Method). If it is used with scales that have only a few degrees (e.g. a 2- or 3-point scale), it is best to use aggregated data (i.e. averaging multiple ratings on the same item or on very similar items, where the ratings were given by the same participant).

While it is generally assumed that using Z-scores or ordinal data on syntactic acceptability judgements has a positive effect on the quality of the data and the resulting statistical analyses, there is, to the best of our knowledge, no research that verifies

this quantitatively. In the next section, we attempt to do exactly that.

3.3 Practical Considerations: A Study

We designed a study that compares the outlined transformations (basic transformations, Z-scores, and ordinal data) across the most common measurement methods (binary Likert Scale, gradient Likert Scale, the Thermometer Method, and Magnitude Estimation). The study consists of an acceptability judgement task in which participants rated 36 randomly sampled sentences that differ in their syntactic structure, using the different measurement methods. We apply the different transformations to the ratings and check how "informative" the ratings are across the different measurement methods (for a definition of "informative", see below). The "grande" null hypothesis, (H_0), is that there are no differences between transformations across measurement methods. This implies the following two sub-hypotheses: ($H_{0\ transformations}$) and ($H_{0\ measurement\ methods}$).[7]

(H_0) There are no differences between transformations across measurement methods in terms of informativity.

($H_{0\ transformations}$) For any given measurement method, there are no differences in terms of informativity between the three data transformations that we consider (basic data, Z-scores, and ordinal data).

($H_{0\ measurement\ methods}$) For any given transformation, there are no differences in terms of informativity between data collected with a binary Likert Scale, a gradient Likert Scale, the Thermometer Method, and Magnitude Estimation.

[7]There are multiple (H_1)s for (H_0), ($H_{0\ transformations}$), and ($H_{0\ measurement\ methods}$). For space reasons, we do not list them here; but they are easily derived from the respective (H_0)s.

INFORMATIVITY For our purposes, "informativity" means that information is added by successfully distinguishing different sentences. In our study, we look at 36 sentences. These are structurally different from each other. Further, these items fall into three categories: Twelve sentences are marked, twelve are questionable, and twelve are unmarked (these are judgements given by the authors from whom we sampled the sentences; for details see Section 3.3.1). We assume that the acceptability of each sentence (e.g. the first marked sentence) is different from the 24 sentences in the other two categories (e.g. the twelve questionable sentences and the twelve unmarked sentences).

Below, we test for differences across sentences (but we only consider comparisons across categories). If a test between two sentences comes out positive (a difference is found), then we consider this to be a true positive and the test result is *informative* by our definition. If the test comes out negative (no difference is found), then this is a false negative and the result is *uninformative* by our definition. In this sense, "informativity" is close to the notion of statistical power.

3.3.1 Experimental Design

CONDITIONS Our aim is to compare data transformed by the outlined transformations (basic transformation, Z-scores, and ordinal data) across the four measurement methods (binary Likert Scale, gradient Likert Scale, the Thermometer Method, and Magnitude Estimation). For each measurement method, we prepared a questionnaire. This gives us the twelve combinations presented in Table 3.1 (ordered by measurement methods) and Table 3.2 (ordered by transformations).

MATERIALS We randomly sampled 36 sentences from the literature that differ in their syntactic structure (for a full list of experimental items, see Appendix

Measurement Method :: Data Transformation
Binary Likert Scale :: Basic Transformation
Binary Likert Scale :: Z-Scores
Binary Likert Scale :: Ordinal Data
Gradient Likert Scale :: Basic Transformation
Gradient Likert Scale :: Z-Scores
Gradient Likert Scale :: Ordinal Data
The Thermometer Method :: Basic Transformation
The Thermometer Method :: Z-Scores
The Thermometer Method :: Ordinal Data
Magnitude Estimation :: Basic Transformation
Magnitude Estimation :: Z-Scores
Magnitude Estimation :: Ordinal Data

Table 3.1: Our experimental conditions ordered by measurement methods.

Data Transformation :: Measurement Method
Basic Transformation :: Binary Likert Scale
Basic Transformation :: Gradient Likert Scale
Basic Transformation :: The Thermometer Method
Basic Transformation :: Magnitude Estimation
Z-Scores :: Binary Likert Scale
Z-Scores :: Gradient Likert Scale
Z-Scores :: The Thermometer Method
Z-Scores :: Magnitude Estimation
Ordinal Data :: Binary Likert Scale
Ordinal Data :: Gradient Likert Scale
Ordinal Data :: The Thermometer Method
Ordinal Data :: Magnitude Estimation

Table 3.2: Our experimental conditions ordered by data transformations.

3.6.1). The items are taken from the Linguistic Inquiry Corpus presented in Chapter 1. The corpus includes all standard acceptability judgements published in Linguistic Inquiry in the years 2001 to 2010; however, the corpus only includes items from authors whose first language is American English[8] (for details on the corpus, please see Chapter 1, Section 1.3). Twelve sentences are marked sentences (i.e. the authors introspectively judged those sentences as unacceptable; indicated by a "*"), twelve sentences are questionable (indicated by either "*?", "??", "?", and variants thereof), and twelve sentences are unmarked (no diacritic was used). However, the transformations across measurement methods are our critical conditions, so that the concrete items are secondary to this project. It is not relevant whether or not the critical items include e.g. a weak island violation or an agreement violation. The concrete ratings for a sentence are not relevant, either. For instance, whether the first questionable sentence of the draw ("?Montana was promised to be healthy by game time on Sunday", from Culicover and Jackendoff, 2001) receives an average rating of "3.14" (on a range from "1", "unnatural", to "7", "natural") or a "4.28" does not really matter for our purposes.

Further, we decided to not include fillers. Fillers mainly fulfil two functions: to distract from the study's purpose and to offset possible imbalances across critical items. As to the first point, fillers would not have helped to distract from the purpose of our study, as we are concerned with meta-issues (it is extremely unlikely that participants were able to guess the purpose of our study, as they were only allowed to take one questionnaire). Further, there is no need to counterbalance the critical items, as the set of items is well-balanced by design.

PARTICIPANTS There are four questionnaires, one for each measurement method. For each questionnaire, we recruited 40 participants through Amazon Me-

[8]Which suits our purposes well, because all of our online participants are native speakers of American English as well (see below).

chanical Turk (4×40 participants $= 160$ participants).[9] To be able to take part, potential participants had to have an Amazon Mechanical Turk approval rate of at least 98% and to have finished at least 5000 approved tasks. We set these criteria to ensure that only reliable participants could take part.[10] Although not explicitly stated in the introduction to the experiment, participants had to be native speakers of American English. We did not disclose this criterion to our participants in order to avoid accommodation effects. Recruiting took place between 21:00 and 02:00 GMT, to ensure that our participants were predominantly from North America. Payment was such that it best matched an hourly rate of $10. Further, participants had to be non-linguists. Prior to the experiment, participants were advised that our study was approved by and followed the guidelines of the University of Oxford's Central University Research Ethics Committee.

After the experiment, we anonymously collected the following personal data: a participant's age, gender, and home country; we also asked where the participants predominantly lived the first ten years of their lives. If participants listed the United States of America as their home country and as the place where they lived the first ten years of their lives, then we consider them as native speakers of American English.

Our pools of participants have the following demographics (after exclusions):

> Session 1, "Binary Likert Scale": 32 participants included; mean age: 35.81 years (10.60); gender distribution: 19 females and 13 males.
>
> Session 2, "Gradient Likert Scale": 30 participants included; mean age: 37.10 years (10.26); gender distribution: 15 females and 15 males.
>
> Session 3, "The Thermometer Method": 33 participants included; mean

[9] In this subsection and in the next subsection "Exclusion Criteria", the methodology used is nearly identical to the methodology of the experiments presented in the previous chapters. Thus, both subsections are a near repeat of the corresponding subsections in Chapter 1 and Chapter 2.

[10] N.B.: These criteria are more lenient than what it requires to become an Amazon Mechanical Turk "Master Worker".

age: 33.21 years (10.37); gender distribution: 15 female and 18 male participants.

Session 4, "Magnitude Estimation": 32 participants included; mean age: 34.19 years (8.23); gender distribution: 13 female and 19 male participants.

EXCLUSION CRITERIA As mentioned, participants are excluded for not being native speakers of American English (affecting 22 participants). A pilot study indicated that there could be a considerable number of non-cooperative participants, so we have additional exclusion criteria to ensure the quality of our results. These criteria are: participating multiple times (0 participants), returning incomplete results (1 participants), having extreme reaction times (6 participants), or failing on the calibration items (5 participants). We apply these criteria in the listed order. In the following, we discuss our criteria in detail.

Participating multiple times: As mentioned, we ran separate questionnaires for each measurement method (resulting in four questionnaires). Theoretically, a participant was able to take part in more than one questionnaire. This would have been suboptimal, for two reasons: First, such a participant would have been overrepresented. Second, the participant could have realised the purpose of our study. In our initial instructions on Amazon Mechanical Turk, we asked participants to not take part multiple times. As each participant has a unique Amazon Mechanical Turk ID, participants were well aware that we were able to easily check whether they took part multiple times. It is therefore not surprising that there was not a single participant who decided to ignore this request.

Not being a native speaker of American English: We did not state this explicitly to

our participants, but we only consider results that come from participants who we believe to be native speakers of American English (see above for how we determine this).

Returning incomplete results: We disregard any results from a participant who did not return a complete set of ratings.

Having extreme reaction times: We also collected reaction times (these are defined as the time from loading an item until the time at which the rating is given). Participants that have either extremely low or extremely high reaction times are excluded. Extremely low reaction times are an indicator of being non-cooperative: Such participants just "click their way through", submitting ratings of little value. Extremely high reaction times can be an indication of distractedness, possibly affecting the quality of responses as well. We use the following formulas to exclude participants who are either extremely slow or extremely fast (these formulas are identical to the formulas in Chapter 1 and Chapter 2). In our definition, "extremely low reaction times" are those that fall below the lower threshold (θ_{lower}) and "extremely high reaction times" are those that fall above the upper threshold (θ_{upper}). $\mu_{1/2}$ is a participant's median reaction time; $\overline{\mu_{1/2_{1-N}}}$ is the mean of all participants' median reaction times. σ is the standard deviation of all participants' median reaction times.

$$\theta_{lower} = \overline{\mu_{1/2_{1-N}}} - 1.5\sigma \tag{3.4}$$

$$\theta_{upper} = \overline{\mu_{1/2_{1-N}}} + 4\sigma \tag{3.5}$$

where $\overline{\mu_{1/2_{1-N}}} = \frac{\mu_{1/2_1} + \mu_{1/2_2} + \ldots + \mu_{1/2_N}}{N}$ and $\sigma = stdev(\mu_{1/2_{1-N}})$

The formula is justified in Appendix 6. This exclusion criterion affected 6 participants. However, this number could have been substantially higher if it had not been for the on-line warning mechanism described in the Procedure Section below.

Failing on calibration items: The first four items of each questionnaire were extremely bad and extremely good calibration items (taken from previous projects; see Appendix 3.6.1 for the concrete items). If a participant rated the two bad calibration items higher than the good items (i.e. on average), then we exclude that participant. In this sense, our calibration items fulfil two functions: Giving participants an expression of extreme items (calibration) and catching out non-cooperative participants (booby trapping).[11]

In total, we exclude 33 of our 160 participants (11 of whom were non-cooperative, while the others were non-NAE). We find that this exclusion rate (20.6%) is in line with the exclusion rate reported in Munro et al. (2010). For each questionnaire/measurement method, the following number of participants remained: binary Likert Scale (30), gradient Likert Scale (32), the Thermometer Method (33), Magnitude Estimation (32). This left us with 4572 ratings (36 sentences × (30 participants + 32 participants + 32 participants + 33 participants)).

PROCEDURE For all measurement methods, we first introduced the general task: Participants were advised that they will judge sentences by how "natural" or "unnatural" they seem (we added: with respect to their grammaticality). We also noted that participants should not be bothered with meaning or punctuation, that

[11]In hindsight, it would have been better to separate those two functions and to randomly intersperse the booby trap items throughout the questionnaire, similar to the study in Chapter 1. There might be participants who pay attention initially, but then "doze off" as they progress. Combined calibration and booby trap items at the beginning of the questionnaire will not detect such participants; however, separate booby trap items at a later stage of the questionnaire would probably have done so.

we are interested in their intuition, and that there is no right or wrong. Before the actual experiment, each measurement method was introduced through an example. In the following, we present the instructions for each measurement method. See Section 3.2.3 for illustrations of the different measurement methods.

Binary Likert Scale: You are asked to rate sentences as either "unnatural/ungrammatical" (red button with an unhappy smiley face) or as "natural/grammatical" (green button with an happy smiley face). (We included the smiley faces for the colourblind.)

Gradient Likert Scale: You will rate sentences on a "1" to "7" scale, "1" (red button with an unhappy smiley face) denoting "fully unnatural/ungrammatical" and "7" (green button with an happy smiley face) denoting "fully natural/grammatical", while "2" to "6" are ratings between those two extremes (dark orange to light green buttons).

The Thermometer Method: Our instructions for the Thermometer Method read as follows:

> You are asked to rate sentences on your own scale. Please define a minimum (denoting "unnatural/ungrammatical") and a maximum ("fully natural/grammatical") and then rate sentences with respect to these extrema. For example, you could set your minimum to "20" and your maximum to "30". A mediocre sentence would then get a rating of about "25".

Participants went through an example and were told that they could adjust their extrema "on the go". The instructions went on:

> If you encounter a sentence that is better than your maximum or worse

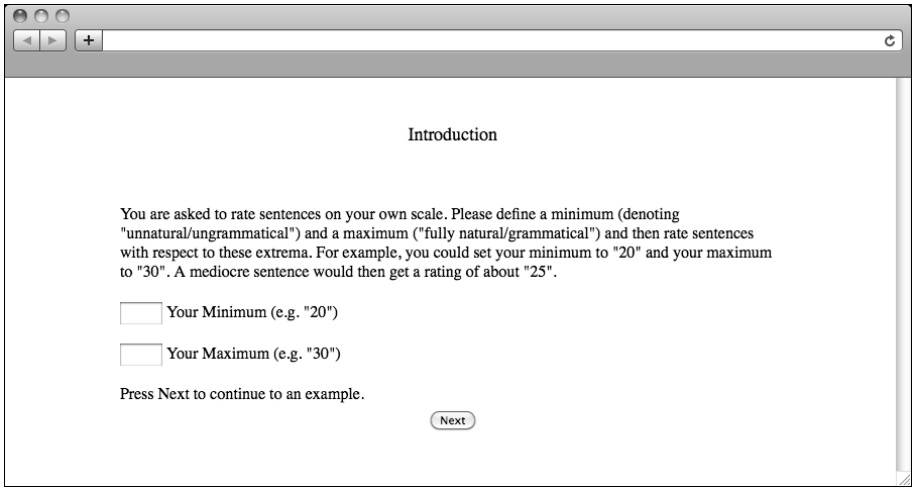


Figure 3.6: An illustration of how we introduced the Thermometer Method for rating the acceptability of sentences. Thereafter, participants went through an example, after which they could adjust their extrema if required. Then, they proceeded to the actual questionnaire.

> than your minimum, then you can change them. If it helps, you could imagine that the scale is like a *thermometer-type* instrument.

This was followed by a short example. Then, participants set their personal minimum and maximum and began the experiment.

Magnitude Estimation: We introduced Magnitude Estimation through an analogy, following Bard et al. (1996). Participants first rated the length of certain lines in comparison to a "reference line" and then this task was transferred to the acceptability of sentences. The instructions read (also see Figure 3.7 and Figure 3.8 for details):

> You are asked to rate sentences with respect to a "reference sentence". You assign a value of your choice (e.g. "10") to that reference sentence and then compare other sentences to it.
>
> Consider the following example, which illustrates the method by using lines. Please, assign a number to (RL) and put the following lines in relation to (RL). If you e.g. assigned "10" to (RL), then a line as long as (RL) should be assigned the value "10", too; a line twice as long as (RL)

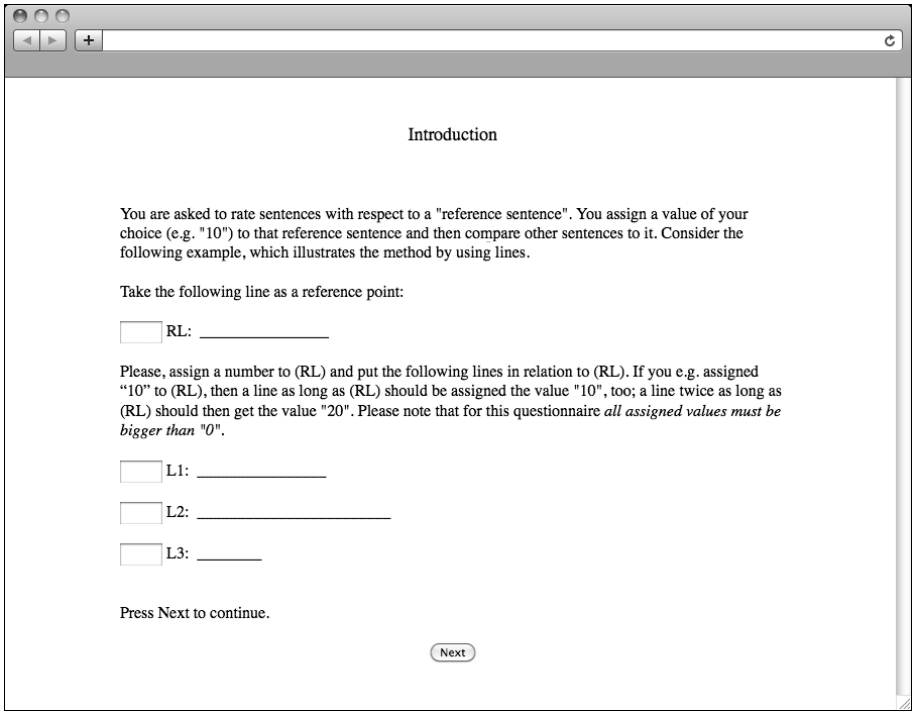


Figure 3.7: Introducing Magnitude Estimation, using the length of lines as an analogy.

> should then get the value "20". Please note that for this questionnaire all assigned values must be bigger than "0".

The line example was followed by an example for acceptability.

> Now, do the same with sentences. However, do not evaluate their length, but their grammatical acceptability. (RS) is your reference sentence.

In the study, we used the following sentence as our standard ("What did Mary ask the president whether she could attend?", similar to (13d) in Legate, 2010). In a previous experiment, this example of a weak island violation received an "in-between" rating ("2.91" of "5"), making it an ideal candidate for the standard (cf. Bard et al., 1996).

Regardless of the measurement method, the first four items of all questionnaires were calibration items. Then, the 36 critical items were presented in a random order (see

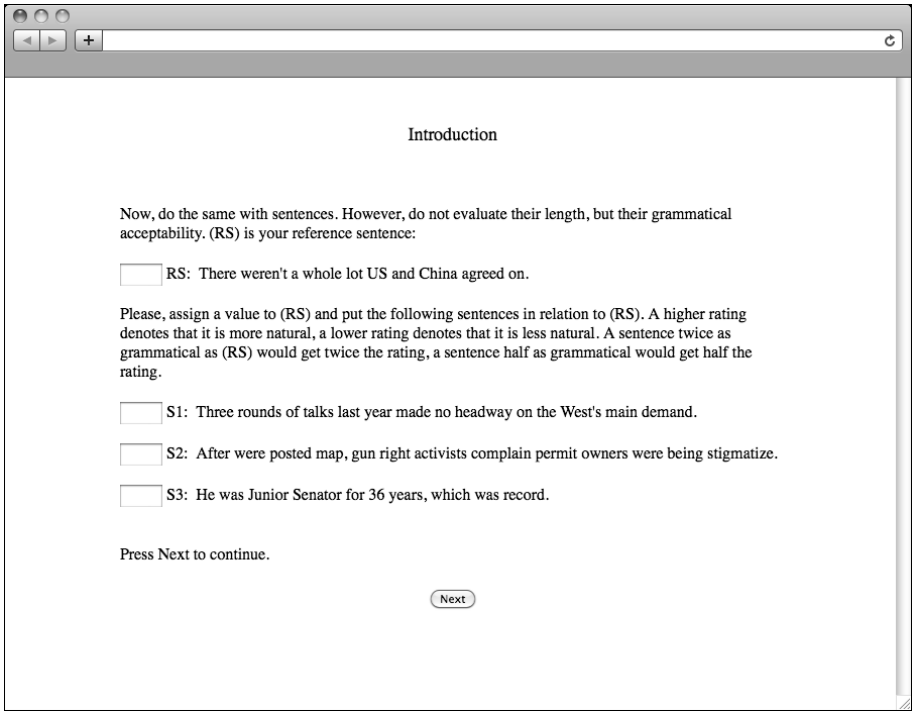


Figure 3.8: An illustration of how we introduced Magnitude Estimation for rating the acceptability of sentences.

Appendix 3.6.1 for the concrete items). Also, for all questionnaires, we included an on-line warning mechanism that produced a pop-up window if participants had "unrealistically fast" reaction times. This was done because in our pilot study, we were concerned that quite a few participants were non-cooperative. A first warning was produced if four ratings were submitted unrealistically fast, a second warning after the twelfth unrealistically fast rating. For both the binary and gradient Likert Scale, we consider a reaction time below 400 ms as unrealistically fast. This is because even for the shortest sentence in the questionnaire, 400 ms is less than half of its expected reading time (in the calculation of expected reading time, we use a formula related to Bader and Häussler, 2010:289). For both the Thermometer Method and Magnitude Estimation, we consider a reaction time of 1200 ms as unrealistically fast. This minimum value is higher than for Likert Scales, because using Magnitude Estimation and the Thermometer Method participants had to do the following for each item: Click on a text box, enter a numeric value, and then

submit it. (Using the Likert Scales, participants just had to click on a button.) In a pilot study, we determined that these additional actions take about 800 ms, which we add to the 400 ms for the minimum reading time (more on this warning mechanism in Appendix 6).

3.3.2 Statistical Analysis

By design, each of the 36 sentences used in the experiment should substantially differ in acceptability compared to the 24 sentences that come from a different judgement category (i.e. marked, questionable, and unmarked). Thus, the "best" combination of measurement method and data transformation would be the one that detects the most differences, using standard statistical tests (effectively, this is asking for the rate of true positives; this is $1 - \alpha$).

SERIES OF DIFFERENCE TESTS To determine the informativity of a combination, we use a series of difference tests. In this series, we test each sentence against the other 24 sentences that belong to the other judgement categories (e.g. all twelve marked sentences are tested against the twelve questionable and the twelve unmarked sentences). We use Wilcoxon Signed Rank Tests to test for differences, because only 30.8% of our data are normally distributed (determined by Shapiro-Wilks normality tests for the 36 constructions; we did this for each of our four sessions mentioned on page 176). Compared to our previous work, this number is quite low. In our view, there are two reasons for such a low figure. First, results from a binary Likert Scale are unlikely to come out as normally distributed. Second, we think that quite a few of the "bad" items "hit the floor" and quite a few of the "good" items "hit the ceiling"; cf. the Figures in Section 3.3.3. Naturally, floor and ceiling data are less likely to be normally distributed.

If a Wilcoxon Signed Rank Test comes out significant (showing a difference) with the standard significance level of $p < 0.05$, then this is a successful outcome (a true positive). Testing 36 sentences against 24 sentences each results in 432 unique tests. A test outcome can take one of two values: "0" or "1". It is these test outcomes (with the values "0" and "1") that are our data points.

We do this for each of the twelve experimental conditions (e.g. Basic Transformation :: Binary Likert Scale, Basic Transformation :: Gradient Likert Scale, etc.), which results in 5184 unique tests (resulting in 5184 data points).

In order to analyse the results, we wished to use a logistic regression model. Ideally, such a model would be based on more than 432 data points per condition (i.e. the 432 unique outcomes of Wilcoxon Signed Rank Tests), which is why we decided to include a random resampling procedure.

RANDOM RESAMPLING PROCEDURE From each of the four questionnaires, we created 30 new subsamples, consisting of 30 participants, randomly drawn with replacement (this is much like running the experiment 30 times for each measurement method, 120 times in total). The procedure is similar to drawing marbles with replacement from an urn. In this image, we have four urns (our four questionnaires, i.e. the four measurement methods), the four urns contain 30 to 33 marbles each (our participants), and each marble is made of 36 pieces (i.e. the 36 ratings each participant made). (One might have the concern that a pool of 30 to 33 participants could be too small for random resampling; however, as Kochanski, 2005, observes, a pool of 30 participants is just fine for such a procedure.) After resampling, we have 129600 "ratings" (4 measurement methods $\times$ 30 random subsamples $\times$ 30 participants $\times$ 36 sentences).

To each subsample, we reapply the series of difference tests. This way, we now have 12960 unique test outcomes per condition (e.g. Binary Likert Scale :: Basic Trans-

formation, Binary Likert Scale :: Z-scores, etc.), i.e. 155520 unique test outcomes in total. From this, we can also give a percentage of overall successful outcomes (i.e. a true positive rate) per experimental condition, per measurement method, and per data transformation. For instance, from this we can say that for the Basic Transformation :: Binary Likert Scale condition, there is a true positive rate of 73.94% (standard deviation: 2.02%).[12] For the percentages, see Table 3.4, Table 3.6, Figure 3.12, and Figure 3.13 in Section 3.3.4.

LOGISTIC REGRESSION Crucially, though, we can feed these data into a logistic regression model. For the model, we use the lme4 package (Bates, Maechler, and Bolker, 2014) for R (R Core Team, 2015). We use a general linear mixed effect model with random factors (*glmer*) to determine how the outcome of a comparison of sentences depends on the data type (i.e. which transformation was chosen) and on the measurement method (we cross these two effects). Since this is resampled data (and not fully independent), we treat the 30 trials of resampling as a random effect. The outcome of each Wilcoxon Signed Rank Test also depends on the concrete constructions that are being compared, so we include these as random effects, as well.

The model also tests for significance, which allows us to test our hypotheses. As a reminder: Our "grande" null hypothesis (H_0) posits that there is no difference between any transformations and measurement methods. From this, we can go one step further and derive concrete null hypotheses, specific to each of the 132 comparisons across the twelve critical conditions. For instance, when comparing Basic Transformation :: Binary Likert Scale vs Z-scores :: Binary Likert Scale, then the specific null hypothesis is that there is no difference in results from Basic data using a binary Likert Scale and results from Z-scores using a binary Likert Scale.

[12] The percentage describes how well the method could distinguish each of the 36 sentences from the 24 sentences that belong to the other two judgement categories and also takes the resampled data into account.

The specific (H_1) is that there *is* a difference between the two conditions. This way, we would get 132 derivatives from (H_0) and 132 corresponding alternative hypotheses (N.B.: there are 66 unique (H_0) derivatives and 66 unique (H_1) derivatives). We choose to not list all the specific sub-hypotheses for space reasons.

For practical reasons, we run the model twice, once with a focus on data types (as described above) and once for the twelve conditions. We do this so that the output becomes easier to interpret. In our view, these are two derivatives of the same model; but we still apply a Bonferroni correction in order to appease readers who think of these as two separate models.

3.3.3 Ratings

Figure 3.9 shows the Basic data for the 36 sentences for the different measurement methods, Figure 3.10 the Z-scores, and Figure 3.11 the ordinal data.

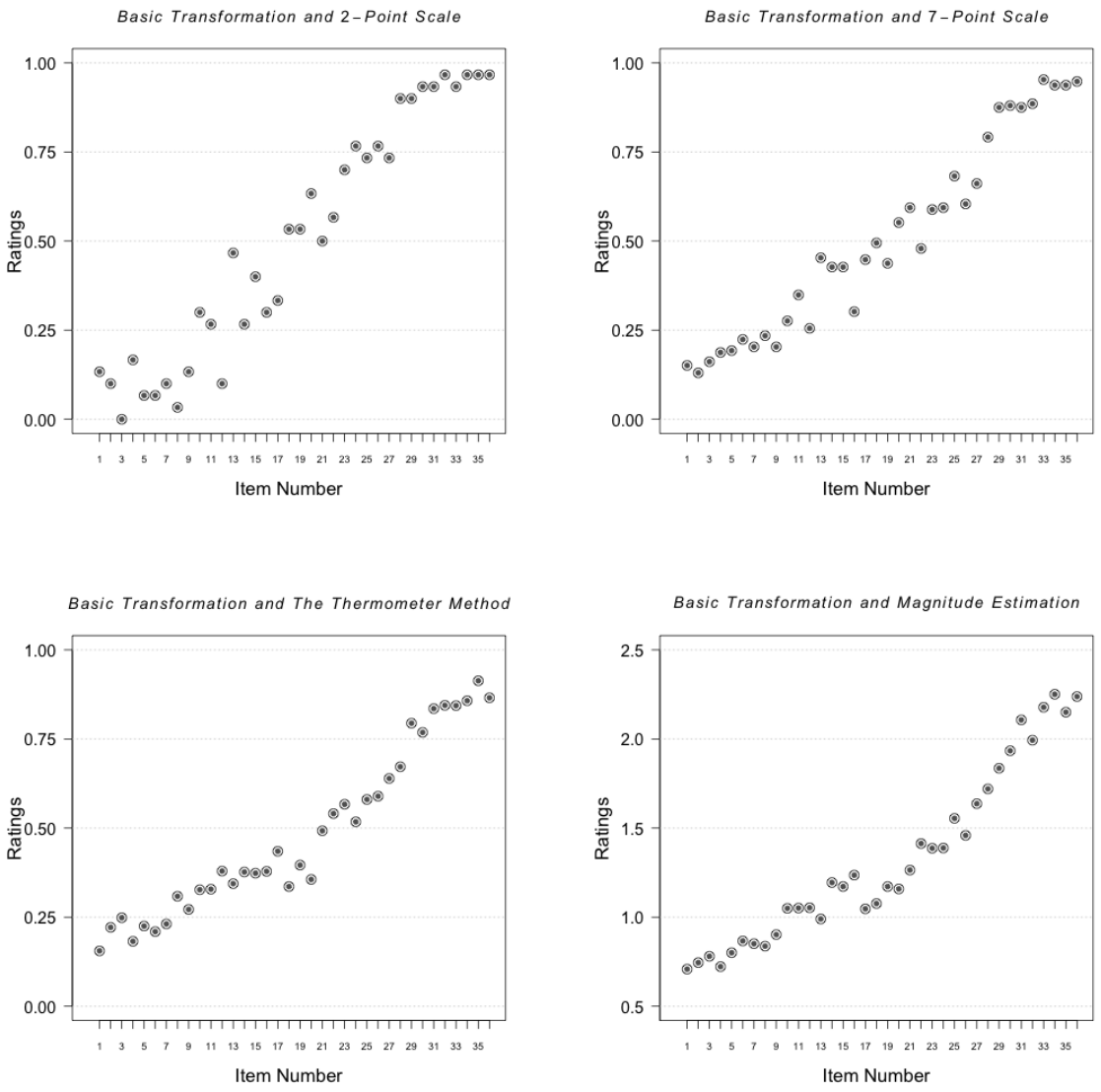


Figure 3.9: The "basic" ratings for the 36 sentences for the four measurement methods (ordered by average ranked ratings for the 36 sentences across all four measurement methods).

(N.B.: For the 2-point scale, the 7-point scale, and the Thermometer Method, we rescaled the ratings to a common scale ("0" to "1"). It was easy to rescale, because in these three measurement methods, the scale's endpoints (e.g. in case of the 7-point scale: "1" and "7") are used to calculate the ratings. This is not possible for Magnitude Estimation, as in Magnitude Estimation the ratings are calculated by dividing by the moduli. See Section 3.2.4 for details.)

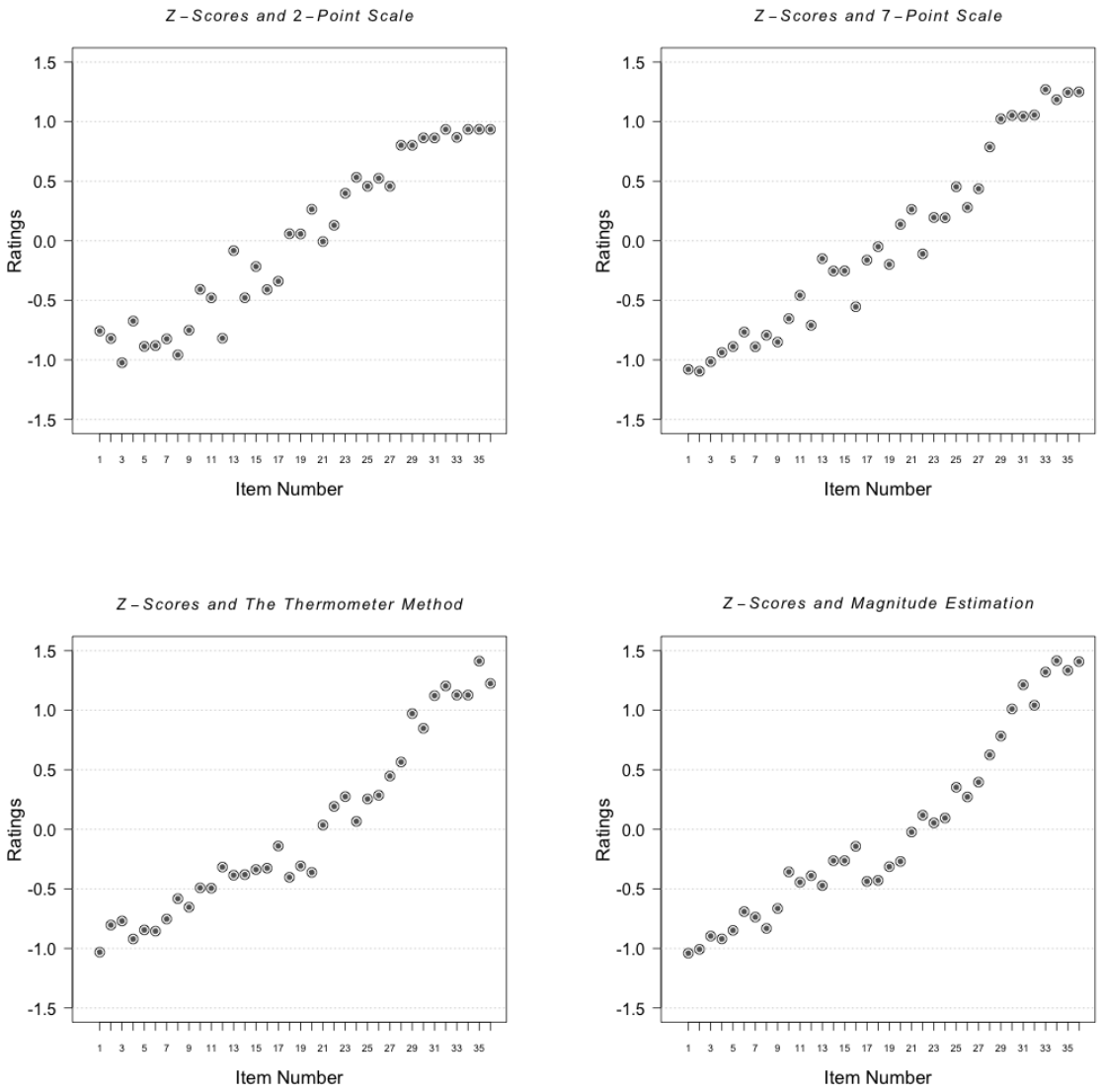


Figure 3.10: The Z-score ratings for the 36 sentences for the four measurement methods (ordered by average ranked ratings for the 36 sentences across all four measurement methods).

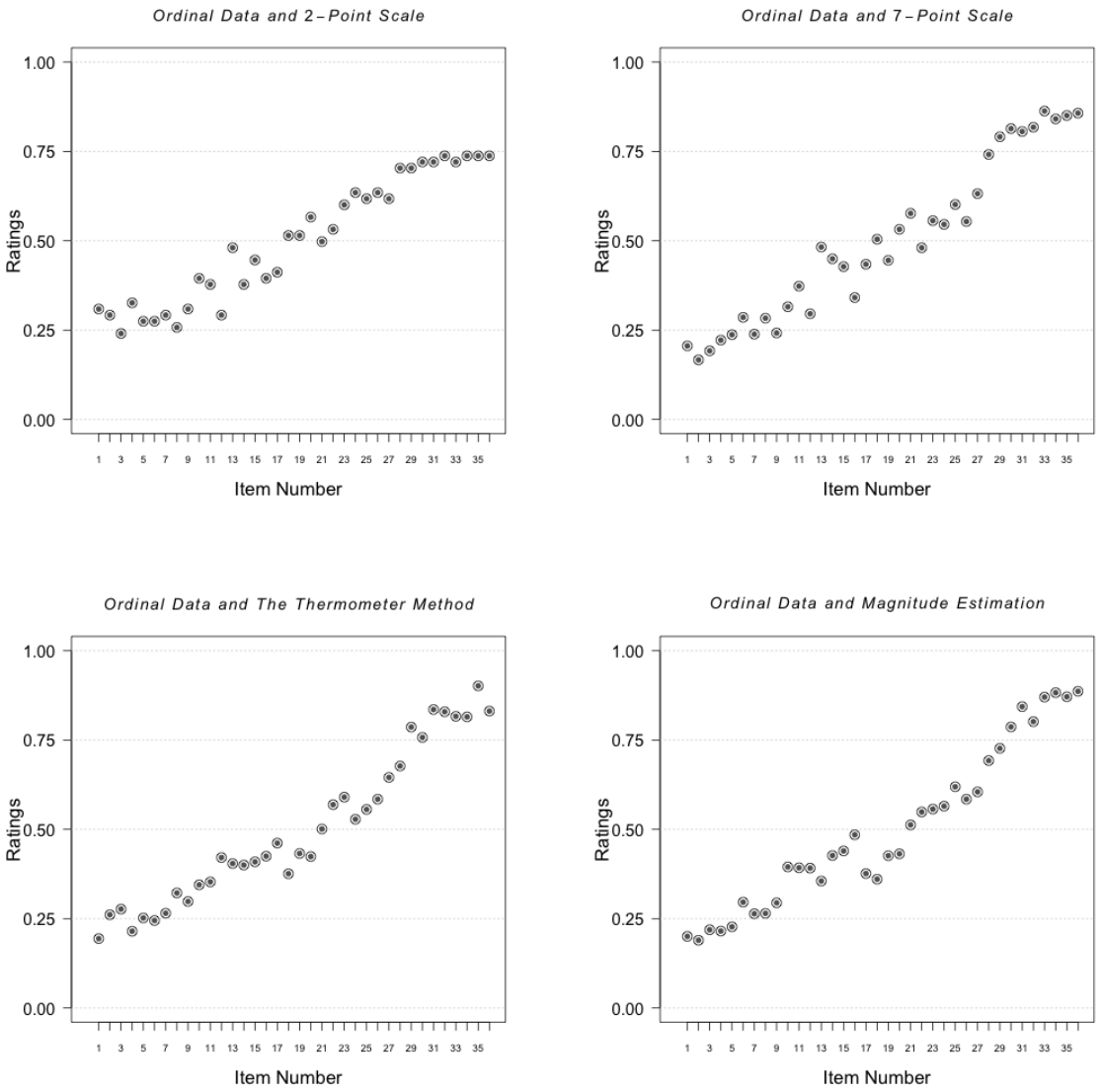


Figure 3.11: The ordinal ratings for the 36 sentences for the four measurement methods (ordered by average ranked ratings for the 36 sentences across all four measurement methods).

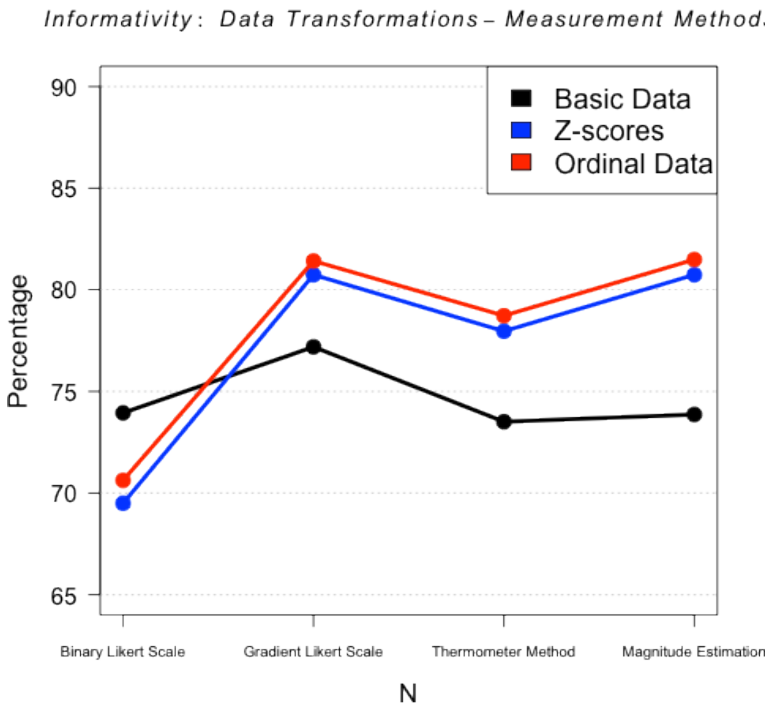


Figure 3.12: The percentages (y-axis) for the series of difference tests for the resampled data, ordered by measurement method by transformation (x-axis).

3.3.4 The Results for the Resampled Data

The rate of positive outcomes for the twelve conditions, as determined by the series of difference tests on the resampled data, are illustrated in Figure 3.12 and Figure 3.13. The results for our model focusing on an overall comparison are given in Table 3.3 (for the random effects) and Table 3.4 (for the fixed effects). The results for our model focusing on the twelve critical conditions are given in Table 3.5 (for the random effects) and Table 3.6 (for the fixed effects). Table 3.4 and Table 3.6 also give the rate of positive outcomes, as determined by the series of difference tests on the resampled data (this was not part of the model's output).

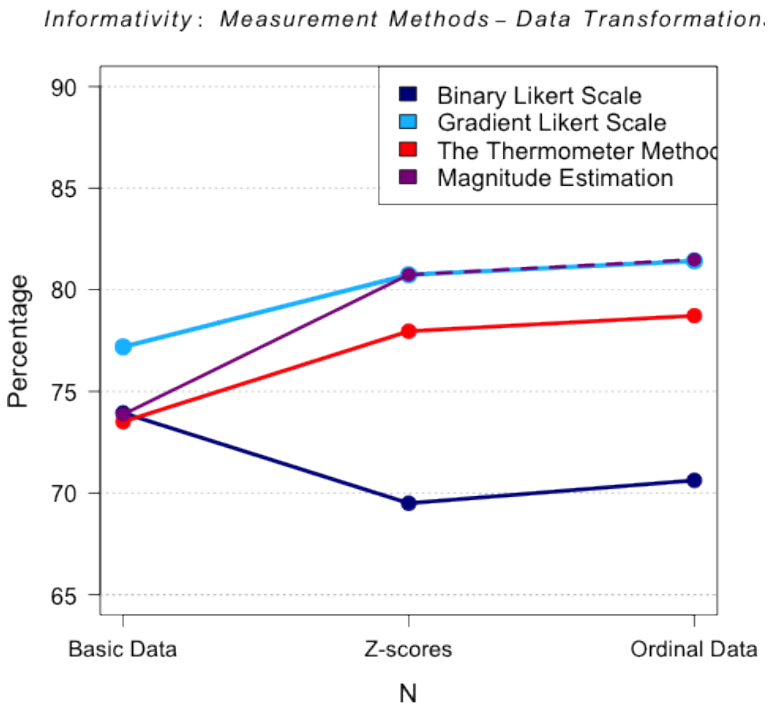


Figure 3.13: The percentages (y-axis) for the series of difference tests for the resampled data, ordered by transformation by measurement method (x-axis).

Parameter	Variance	Standard Deviation
Resampling Trial	0.001	0.023
First Construction of a Wilcoxon S. R. Test	0.482	0.694
Second Construction of a Wilcoxon S. R. Test	0.482	0.694

Table 3.3: The random effects, their variances, and standard deviations for our overall model.

As to the random factors, the trial number (Resampling Trial) has very little effect on the results. The First Construction of a Wilcoxon Signed Rank Test and the Second Construction of a Wilcoxon Signed Rank Test show some variance, but this is expected, because the constructions that we compare do matter (e.g. first marked construction vs first questionable construction; or first marked construction vs first unmarked construction). As to the fixed effects, the results show that Z-score data (estimate: 1.441) are significantly better than basic data (1.280) and so are ordinal data (1.495). This is not unexpected and in these respects, ($H_{0\ transformations}$) can be

Parameter	Estimate	Standard Error	z-Value	Significance	Overall Rate of True Positives
Basic Data	1.280	0.011	-10.234	vs Z-Scores: $p < 0.001$ vs Ordinal Data: $p < 0.001$	74.63% (1.14%)
Z-Scores	1.441	0.011	4.493	vs Basic Data: $p < 0.001$ vs Ordinal Data: $p < 0.001$	77.24% (1.06%)
Ordinal Data	1.495	0.161	9.292	vs Basic Data: $p < 0.001$ vs Z-Scores: $p < 0.001$	78.07% (1.03%)

Table 3.4: The fixed effects, their estimates, standard errors, z-values, and significance levels between conditions. This is the output for our overall model. (The full output for our R model can be found in Appendix 3.6.2.) The estimates are the natural log of the odds ratio of getting a successful outcome (i.e. a true positive in the Wilcoxon Signed Rank Test) and getting an unsuccessful outcome (i.e. a false negative). We also give the rate of positive outcomes in the series of difference tests (cf. Section 3.3.2) including the standard deviation across resampled trials (the latter was not part of the logistic regression model).

Parameter	Variance	Standard Deviation
Resampling Trial	0.001	0.023
First Construction of a Wilcoxon S. R. Test	0.491	0.701
Second Construction of a Wilcoxon S. R. Test	0.491	0.701

Table 3.5: The random effects, their variances, and standard deviations for our model that focuses on the twelve experimental conditions.

Parameter	Estimate	Standard Error	z-Value	Overall Rate of True Positives
Basic Data :: Binary Likert Scale	1.243	0.023	−11.16	73.94% (2.03%)
Basic Data :: Gradient Likert Scale	1.442	0.023	−2.10	77.19% (1.61%)
Basic Data :: The Thermometer Method	1.217	0.023	−12.33	73.51% (3.34%)
Basic Data :: Magnitude Estimation	1.238	0.023	−11.37	73.87% (3.65%)
Z-Scores :: Binary Likert Scale	0.991	0.022	−22.89	69.50% (2.20%)
Z-Scores :: Gradient Likert Scale	1.681	0.024	8.37	80.74% (1.65%)
Z-Scores :: The Thermometer Method	1.492	0.023	0.12	77.96% (3.44%)
Z-Scores :: Magnitude Estimation	1.681	0.024	8.37	80.74% (1.64%)
Ordinal Data :: Binary Likert Scale	1.050	0.022	−19.98	70.63% (1.87%)
Ordinal Data :: Gradient Likert Scale	1.730	0.024	10.45	81.42% (1.73%)
Ordinal Data :: The Thermometer Method	1.542	0.023	2.35	78.73% (2.60%)
Ordinal Data :: Magnitude Estimation	1.735	0.016	10.66	81.49% (2.04%)

Table 3.6: The fixed effects, their estimates, standard errors, and z-values. This is the output for the model that focuses on the twelve experimental conditions. (The full output for our R model can be found in Appendix 3.6.2.) The estimates are the natural log of the odds ratio of getting a successful outcome (i.e. a true positive in the Wilcoxon Signed Rank Test) and getting an unsuccessful outcome (i.e. a false negative). We also give the rate of positive outcomes in the series of difference tests (cf. Section 3.3.2) including the standard deviation across resampled trials (the latter was not part of the logistic regression model).

rejected. Additionally, ordinal data are slightly better than Z-scores. The difference is significant, though the effect size is rather moderate.

Comparing specific conditions (e.g. Z-scores :: binary Likert Scale vs Z-scores :: gradient Likert Scale; etc.) reveals interesting details. Many of the differences are significant[13], but it would be tedious to list all of the 66 unique comparisons. However, at this stage, we can certainly reject ($H_{0\ measurement\ methods}$), because we observe numerous significant differences between measurement methods. Having rejected both ($H_{0\ transformations}$) and ($H_{0\ measurement\ methods}$) implies that the "grande" null hypothesis, (H_0), can be rejected, too.

We discuss some outcomes that are of particular interest. Transforming data to Z-scores or to ordinal data has no positive effect on the results from a binary Likert Scale: Its basic data is not too bad (estimate: 1.243), but its Z-scores and ordinal data perform significantly worse than its basic data (0.991 and 1.053, respectively).[14] The gradient Likert Scale, on the other hand, has decent basic data (1.442) and the results significantly increase for data transformed to Z-scores (1.681) and to ordinal data (1.730), as one would expect. The results for the Thermometer Method show the same pattern (basic data: 1.217; Z-scores: 1.492; ordinal data: 1.542), but are somewhat worse in general. The results for Magnitude Estimation are split: Its basic data performs just fair (1.238), but the results for the Z-scores and the ordinal data surge to significantly higher levels (1.681 and 1.735, respectively).

[13] Any difference in estimates of 0.096 or larger is necessarily significant. This is because the highest standard error in Table 3.6 is 0.024. Applying the Bonferroni correction, this gives a definite significance threshold of ± 0.096.

[14] Ultimately, the results are based on a series of Wilcoxon Signed Rank Tests. It might be the case that this has "dampened" the results of the binary Likert Scale. Tied ratings reduce the likelihood of a positive test outcome and binary Likert Scales are particularly prone to ties in ratings. We are not sure whether or not the results would have come out better in a series of t-tests. Using t-tests is not advisable, however, because of the above mentioned issue with a lack of normal distributions. One way to fix this is to only look at normally distributed data sets. However, such data sets would not represent the full range of syntactic data.

The ordinal data for Magnitude Estimation come out top of all conditions in Table 3.6 and its Z-scores are third overall (although there is not a significant difference to their counterparts from the gradient Likert Scale).

3.4 Discussion

Our results confirm the assumption that Z-scores are more informative than basic data (i.e. with Z-scores, one is able to detect more true differences than with basic data). The same is true for ordinal data.

There are theoretical arguments for Z-scores (they mitigate a subject's scale bias) and ordinal data (they help overcome the assumption of unequal intervals between points of a scale). But there are also theoretical arguments against these transformations: Z-scores might destroy real tendencies (it could be the case that a subject does not have a bias and the ratings reflect what he/she really thinks). Further, both Z-score transformed data and ordinal data are "context sensitive" in the sense described in Section 3.2.4. Consequently, if Z-scores or ordinal data are used, it is even more important to carefully design one's experiment. The results are also harder to interpret than basic data; which is why one might wish to report both basic data and Z-scores or basic data and ordinal data.

Weskott and Fanselow (2011) found that ratings using Magnitude Estimation are not more informative than ratings using other measurement methods. Our results agree with theirs for basic data: Our results do not favour Magnitude Estimation, either. However, we find that Magnitude Estimation's Z-scores and its ordinal data are more informative than its counterparts from the Thermometer Method and the binary Likert Scale (but it is not significantly more informative than its counterpart from the gradient Likert Scale). The problem with combining Magnitude Estimation

with Z-scores or ordinal data is that the mechanisms of Magnitude Estimations are subverted: One then ignores the modulus, rendering the extra effort of using Magnitude Estimation unnecessary.

Further, Weijters et al. (2010) suggested that ratings using a binary Likert Scale are more contrastive compared to ratings from a gradient Likert Scale ("contrastive" in the sense of the discussion in Section 1.4.1 in Chapter 1). Looking at Figure 3.9, our results seem to agree with Weijters et al. (2010). Our results also suggest that ratings using a binary Likert Scale are less informative than ratings from a gradient Likert Scale. This agrees with Cox's (1980) literature review.

Lastly: We used a linguistic concept ("syntactic acceptability") for our experiment. It would be interesting to re-run our experiment with other psychological concepts (e.g. perceived musical harmony, perceived attractiveness, etc.).

3.5 Conclusion

We asked how different data transformations (basic data, Z-scores, and ordinal data) compare to each other in terms of informativity. The motivation for our inquiry is that linguists mainly use basic data in their research, although the benefits of Z-scores and ordinal data are well established in psychology. Z-scores and ordinal data can help mitigate scale biases and overcome the problem of unequal intervals that affects basic data.

Our results support findings from general psychology: Z-scores are, by and large, more informative (i.e. able to detect real differences between samples) than basic data. This holds for all measurement methods (the gradient Likert Scale, the Thermometer Method, and Magnitude Estimation), except for the binary Likert Scale. The picture for ordinal data is very similar. Looking at the concrete measurement

methods, gradient Likert Scales strike a good balance between simplicity of use and informativity.

However, Z-scores and ordinal data have potential downsides, too: They might make real tendencies look as if a bias was present, their results can be hard to interpret, and it might be hard to compare results across studies. To mitigate the first point (obscuring real tendencies), careful experimental design is needed. To mitigate the second point (difficulty in interpreting the results), the researcher might wish to report both basic data and either Z-scores or ordinal data. Adopting these two strategies should also mitigate the third point (difficulties in comparing results across studies).

Our bottom line is that linguists should at least seriously consider the use of Z-scores or ordinal data. If they choose to not use them, they should justify their choice. In general, we think the following is good practice: One should report both standard data (for interpretability) and the Z-scores or ordinal data; but one should use Z-scores or ordinal data for statistical tests.

3.6 Chapter 3: Appendices

3.6.1 Experimental Stimuli

		Marked Items (*)
LI Issue	No. in Paper	Sentence
35.1.1	48b	So many people ate faster yesterday that we were all done by 9 p.m. than we had expected.
32.3.4	20c_6	John tried to win and Bill did likewise to try to win.
32.3.4	29a_4	Helen liked Bernie without compromising themselves.
32.3.4	34e_5	John pledged to Susan to leave.
32.1.5	48b	A student may not select a major until the finishes all the general education courses.
32.1.5	91b	John appears to hit Bill right now.
39.1.5	FN14ii	The house was constructed for five months.
37.2.2	FN4ib_2	When did you read that book until?
32.3.4	22a_9	John told Sue for Harry to wash themselves.
32.3.4	29a_14	Helen pleased Bernie after compromising himself.
34.1.1	7b	The burglar saw the prisoner know French.
36.4.6	12a_2	No knight, who wore chain mail, left the castle.

		Questionable Items (?, ??, ?*)
LI Issue	No. in Paper	Sentence
32.3.4	39f_1	Montana was promised to be healthy by game time on Sunday.
37.1.8	3	Who did Sue convince Bill that she met the woman who kissed him?

39.1.6	27b_3	Mary solved the problem on Tuesday neatly, and Bill did so on Friday.
32.1.5	90a	John does not like math but Mary seems to.
37.2.2	FN20iib	I read the book for the same amount of time that you did.
38.3.2	114b	Everyone has not ever been to Paris.
34.3.8	7	Which story was the editor shown without anyone verifying?
41.4.7	4a	I saw yesterday Jim.
37.1.3	4b	Who believes that which man will win the prize?
39.1.6	26b_4	Mary solved the problem on Tuesday completely, and Bill did so partially.
37.2.10	2b_2	John shouted she was very hungry.
39.1.5	FN21i	John carried the bag toward the store and Frank did so toward the church.

		Unmarked Items
LI Issue	No. in Paper	Sentence
32.3.4	12a_3	John tries to give his kids a better life.
32.3.4	39g_3	Susan was promised to be enabled to take care of herself.
32.3.4	41a_10	John signaled to Sally for Harriet to leave.
32.3.4	49b_3	John yelled to Sally to be allowed to leave.
33.1.3	61c	I wanted to expect everyone you did to come to the party
33.2.1	18_8	There were many demonstrators arrested by the police.
40.2.1	50b	Nobody has lost their job yet.
40.2.6	6a_2	We thought about John that something terrible had happened to him.
37.3.7	6c_1	I doubt you can do anything about poverty.

35.3.3	19b	John saw Mary before entering the room.
32.1.5	FN24iia	I believe John to sing.
32.3.6	17a	Sue estimated Bill's weight to be 150 lbs.

3.6.2 R Output

```
#Our model for data transformations only.


Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: outcome ~ data_type + (1 | resampling_trial) +
(1 | constr1) + (1 | constr2)
   Data: data_axbin


      AIC       BIC    logLik  deviance  df.resid
 302216.2  302280.1 -151102.1  302204.2    311034


Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.8412  0.2103  0.3473  0.5935  1.4466


Random effects:
 Groups           Name        Variance  Std.Dev.
 constr2          (Intercept) 0.4820906 0.69433
 constr1          (Intercept) 0.4820774 0.69432
 resampling_trial (Intercept) 0.0005304 0.02303
```

```
Number of obs: 311040, groups:  constr2, 36; constr1, 36;
resampling_trial, 30

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.49459    0.16086   9.292  < 2e-16 ***
data_typeSTA -0.21451    0.01099 -19.526  < 2e-16 ***
data_typeZSC -0.05353    0.01116  -4.799  1.6e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Correlation of Fixed Effects:
            (Intr) dt_STA
data_typSTA -0.035
data_typZSC -0.034  0.514




#Our model for data transformations and measurement methods.

Generalized linear mixed model fit by maximum likelihood
(Laplace Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: outcome ~ mm_&_dt + (1 | resampling_trial) +
    (1 | constr1) + (1 | constr2)
   Data: data_axbin

      AIC       BIC    logLik  deviance  df.resid
```

```
 299439.9  299599.6 -149704.9  299409.9    311025

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.7175  0.1952  0.3551  0.5918  1.6900

Random effects:
 Groups           Name        Variance  Std.Dev.
 constr2          (Intercept) 0.4907330 0.70052
 constr1          (Intercept) 0.4907389 0.70053
 resampling_trial (Intercept) 0.0005486 0.02342
No. of obs: 311040, groups: cstr2, 36; cstr1, 36; res_tr, 30

Fixed effects:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.734939   0.162745   10.66   <2e-16 ***
mm_&_dtME_STA  -0.496700   0.022546  -22.03   <2e-16 ***
mm_&_dtME_ZSC  -0.054045   0.023565   -2.29   0.0218 *
mm_&_dt7PS_ORD -0.005077   0.023706   -0.21   0.8304
mm_&_dt7PS_STA -0.292944   0.022953  -12.76   <2e-16 ***
mm_&_dt7PS_ZSC -0.054045   0.023564   -2.29   0.0218 *
mm_&_dtTHM_ORD -0.192779   0.023191   -8.31   <2e-16 ***
mm_&_dtTHM_STA -0.517549   0.022515  -22.99   <2e-16 ***
mm_&_dtTHM_ZSC -0.243160   0.023067  -10.54   <2e-16 ***
mm_&_dt2PS_ORD -0.681855   0.022253  -30.64   <2e-16 ***
mm_&_dt2PS_STA -0.492145   0.022552  -21.82   <2e-16 ***
mm_&_dt2PS_ZSC -0.743793   0.022173  -33.55   <2e-16 ***
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Correlation of Fixed Effects:
        (Intr) m_ME_S m_ME_Z m_7PS_O m_7PS_S m_7PS_Z m_THM_O
m_ME_ST -0.074
m_ME_ZS -0.070  0.528
m_7PS_OR -0.070  0.525  0.502
m_7PS_ST -0.072  0.543  0.519  0.516
m_7PS_ZS -0.070  0.528  0.505  0.502  0.519
m_THM_OR -0.071  0.537  0.513  0.510  0.527  0.513
m_THM_ST -0.074  0.554  0.529  0.526  0.544  0.529  0.538
m_THM_ZS -0.071  0.540  0.516  0.513  0.530  0.516  0.525
m_2PS_OR -0.075  0.560  0.535  0.532  0.550  0.535  0.544
m_2PS_ST -0.074  0.552  0.528  0.525  0.542  0.528  0.537
m_2PS_ZS -0.075  0.562  0.537  0.534  0.552  0.537  0.546
            m__THEME_S m__THEME_Z m__TWOPS_O m__TWOPS_S
m__MAGES_ST
m__MAGES_ZS
m__SEVPS_OR
m__SEVPS_ST
m__SEVPS_ZS
m__THEME_OR
m__THEME_ST
m__THEME_ZS  0.541
m__TWOPS_OR  0.561      0.547
m__TWOPS_ST  0.553      0.540      0.560
m__TWOPS_ZS  0.563      0.549      0.570      0.562
```

Chapter 4

The TOST as a Method of Similarity Testing in Linguistics

4.1 Introduction[1]

Classical statistical tests are designed to test for differences and their null hypotheses (the hypotheses to be rejected) typically posit that the compared samples come from the same population. If classical tests come out negative, i.e. the null hypothesis cannot be rejected, then there is insufficient evidence to assume a difference between the tested samples. Insufficient evidence, however, is not sufficient evidence to assume equivalence (Altman and Bland, 1995) and, for that matter, it is not sufficient evidence to assume "similarity", either.

Linguistics relies to a great extent on classical tests (e.g. out of the 16 experimental talks at the LSA 2013 meeting, not a single one tested for similarity or equivalence). However, there are valuable linguistic questions where classical tests are insufficient.

[1]This chapter is based on joint work with Johannes Kizach.

Consider the following examples from phonetics, psycholinguistics/applied linguistics, and syntax:

(RQ_1) Can highly experienced L2 learners attain a native-like level of language production?

(RQ_2) At which age do teenagers typically reach adult-like reading times?

(RQ_3) Are resumptive pronouns perceived as equally bad across modalities?

For all three examples, a negative result in a classical test would only lead to the conclusion that there is no reason to assume a difference. With such results, research into RQ_{1-3} would probably go unreported, which disincentivises researchers to investigate such questions (Bakker, van Dijk, and Wikkerts, 2012, make this point for psychology, but it applies to linguistics just as much); and the field as a whole would miss out.

For RQ_{1-3}, a similarity or equivalence test would be more suitable (particularly, if one wishes to compare means and test for significance). The TOST (*t*wo *o*ne-*s*ided *t*-tests) is one of the most common similarity tests (cf. Richter and Richter, 2002) and in this chapter, we explore how the TOST can be best utilised by linguists (and possibly by researchers in other fields, as well). However, setting the parameter that controls the TOST, δ, can cause difficulties. The aim of this chapter is to provide guidelines for how to set δ to achieve the narrowest similarity range that meets common error rates of $1 - \beta = 80\%$ and $1 - \alpha = 95\%$ (where β is the rate of false negatives and α the rate of false positives).[2]

[2]N.B.: This chapter is limited to statistical similarity. Questions of functional similarity are not addressed.

OVERVIEW In Section 4.2, we look at the workings of the TOST. The TOST consists of two t-tests. Consequently, we begin Section 4.2 by examining how t-tests work. This is followed by an analysis of the mechanics of the TOST. We explore why setting the parameter δ is crucial for performing a TOST. However, there have not been any objective guidelines for how to set δ. Consequently, we set out to provide such guidelines. We wish to do so by observing how the error rate of the test depends on δ. Unfortunately, this requires more data than we could access, which is why we decided to simulate our data. Our approach to simulating data is outlined in Section 4.3. We discuss our random subsampling procedure in detail and explain why we had to apply a post-testing analysis of β. We had two stages for our simulations: an initial calibration phase and a validation phase. In the calibration phase (Section 4.4), we observe how δ comes out for our simulated data. Based on our observations, we derive a formula that can be used to predict δ. In Section 4.5, the validation phase, we compare how predicted δ compares to further observations of additional data. In Section 4.6, we quantitatively compare how outcomes of TOS-Tests compare to the outcomes of t-tests. The chapter concludes with Section 4.7.

4.2 How The TOST Works

The TOST is generally attributed to Westlake (1976) and Schuirmann (1981) and is one of the most common similarity tests (cf. Richter and Richter, 2002). The goal of a TOST is to determine whether two normally distributed samples come from populations that are sufficiently similar in means. As the name suggests, it is based on t-tests. Consequently, we begin this section by looking at t-tests in order to then outline the mechanics of a TOST.

T-TESTS A t-test's aim is to determine whether two normally distributed samples come from different populations, by testing whether the two samples' means are significantly different. It takes the magnitude of the difference between means, the size of the samples, and their standard deviations into account. The null hypothesis (H_0), the hypothesis to be rejected, is that there is no difference in means between the two samples and that they come from two populations that have the same mean ($\mu 1 = \mu 2$) and the alternative hypothesis (H_1) is that there is a difference in means and that the two samples come from different populations ($\mu 1 \neq \mu 2$).[3]

T-tests can be one-sided (where the direction of the effect is known, i.e. we test for either $\mu 1 < \mu 2$ or $\mu 1 > \mu 2$) or two-sided (which tests both directions an effect could take, i.e. we test for $\mu 1 < \mu 2 \wedge \mu 1 > \mu 2$); and they can be paired (the values that are compared come from the same group, e.g. pre-treatment vs post-treatment scores) or independent (the values that are compared come from different groups, e.g. two post-treatment scores; cf. e.g. Elliott and Woodward, 2006). The TOST uses one-sided t-tests and for now, we focus on independent data.[4] The formula for a one-sided t-test is given in the formula (f_{t-test}) below. (f_{t-test}) is used to calculate the t-statistics. Using a table of critical values, the t-statistics can then be used to determine the probability that two samples could be drawn from the same population. Typically, if that probability is below 5% (sometimes 1%), then the samples are thought to be significantly different. Figure 4.1 gives the p-distribution for the outcomes of an independent, one-sided t-test and (f_{t-test}) is the formula used to calculate the t-statistics.

[3]The samples could, of course, also come from the same population, which would lead to the two means being identical.

[4]Paired data can always be transformed into unpaired data by subtraction. Given a paired data set (x_i, y_i), one can transform it into ($x_i - y_i$, 0) and then compare the mean of the pairwise difference to zero. Many thanks to Greg Kochanski for pointing this out to us.

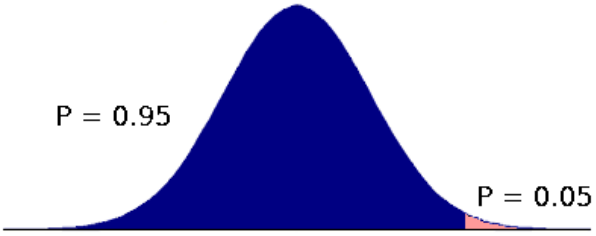


Figure 4.1: The probability-distribution for a t-test. The samples' difference in means, the samples' standard deviation, and the samples' sizes are used to calculate the t-statistics. Using a table of critical values, the probability p, i.e. the probability that the given distribution would generate a sample that has a t-score of t or beyond, can be calculated. If there is a significant difference (the probability for this is marked in red), then it is likely that the two samples come from two populations that differ in means.

$$(f_{t-test}): \quad t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{2}{n}}} \tag{4.1}$$

$$\text{where } s_p = \sqrt{\tfrac{1}{2}(s_{x1}^2 + s_{x2}^2)}$$

The test is one-sided, because we only test whether the second mean is bigger than the first mean. If the result of a t-test falls within the red area of the curve, the null hypothesis is rejected and we conclude that the samples are significantly different.

TOSTS In a TOST, the difference in means between two samples is tested against a lower and an upper boundary. δ has to be set by the researcher prior to testing. The lower boundary is $-\delta$ and the upper boundary is δ. Since the TOST uses two one-sided t-tests to test for this, the corresponding null hypotheses for the two t-tests are ($H_{0\,a}$): the difference in means is no different or even smaller than the lower boundary $-\delta$ and ($H_{0\,b}$): the difference in means is no different or even bigger than the upper boundary δ.

$(\mathrm{H}_{0\,a})$: $\mu_1 - \mu_2 \leq -\delta$ *or* $(\mathrm{H}_{0\,b})$: $\mu_1 - \mu_2 \geq \delta$

The alternative hypotheses for the two t-tests are $(\mathrm{H}_{1\,a})$ and $(\mathrm{H}_{1\,b})$. (H_1) is the overall alternative hypothesis of the TOST.

$(\mathrm{H}_{1\,a})$: $\mu_1 - \mu_2 > -\delta$ *and* $(\mathrm{H}_{1\,b})$: $\mu_1 - \mu_2 < \delta$

(H_1): $-\delta < \mu_1 - \mu_2 < \delta$

Rejecting both nulls is a positive test outcome, which indicates similarity within the range δ (i.e. one can reject the null hypothesis that the difference in the two means is bigger than δ; and one can be reasonably certain that the difference in means is smaller than δ). If δ has been set correctly, then one can conclude that it is likely that the two samples come from populations that are sufficiently similar in means. (f_{TOST}) is the formula to calculate a TOST, Figure 4.2 gives the p-distributions.

$$(f_{TOST}): \quad t = \frac{\bar{X}_1 - \bar{X}_2 \pm \delta}{s_p \sqrt{\frac{2}{n}}} \tag{4.2}$$

$$\text{where } s_p = \sqrt{\tfrac{1}{2}(s_{x1}^2 + s_{x2}^2)}$$

ISSUES An immediate question that comes to mind is: How to "best" set δ? We come to this question at the end of this subsection; however, we would like to first address four other issues.

First, the literature sometimes describes the TOST as an "equivalence" test. This is a misnomer: By definition the TOST does not test for equivalence, as it considers

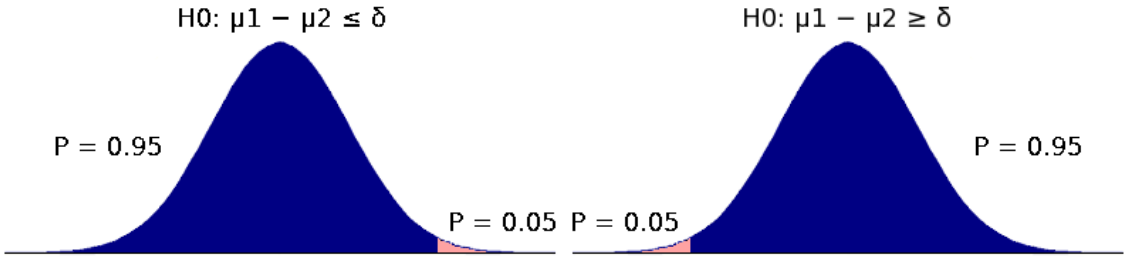


Figure 4.2: In a TOST, two one-sided t-tests are performed. In the first test (left), the samples are tested for differences against a lower threshold, in the second test (right), the samples are tested for differences against an upper threshold. If both results come out as significantly different, then one can assume similarity within the specified range that defines the thresholds.

a *range* (i.e. the difference in means of two samples $\pm\delta$). We think that calling the TOST a "similarity test" makes much more sense and in this chapter we adhere to this naming.[5]

Second, the assumptions of and requirements for a TOST are the same as for a t-test. Crucially, the data need to be normally distributed. This has been a requirement for the data sets we use below.

Third, the range that the TOST uses is sometimes expressed through absolute values (e.g. $\pm$0.5 points on a 5-point scale) and sometimes through relative values (e.g. $\pm$10%). The literature is not always clear about this. Our focus is on absolute differences, typically denoted by δ. In most cases, but not always, δ can be converted to relative differences (sometimes referred to as θ).

Fourth, to answer how to objectively set δ, it is important to know which factors influence the outcome of a TOST. Since TOSTs are based on t-tests, one can expect that all of the factors that influence the outcome of a t-test also influence the outcome of a TOST. Leaving δ aside, these factors include: The magnitude of the difference in means, the standard deviation of the samples, and the size of the samples.

[5]Many thanks to Greg Kochanski for making this point.

An increase in the observed difference in means will make it less likely that the two samples come from populations that are sufficiently similar in means. This will drive the p-values up (towards 1). A smaller standard deviation of the two samples will lead to the assumption that the underlying population has a smaller standard deviation, as well. Any differences between the two samples will then make it more likely that they come from different populations. This, too, will drive the p-values up. Similarly for smaller sample sizes: It will increase the denominator in (f_{t-test}) and (f_{TOST}), which will, by tendency, decrease the t-statistics. A lower t-statistics will lead to an increase in p-values.

This leaves δ. Unsurprisingly, choosing a smaller or a bigger range affects the results, as well. Decreasing δ will make it less likely that the two t-tests will come out as significantly different, making it less likely for the TOST to indicate similarity. Similarly, increasing δ will make it more likely that the two t-tests come out as significantly different, making it more likely for the TOST to indicate similarity.

This leads back to the question posed at the beginning of this subsection: How can δ be set objectively? To the best of our knowledge, there are no guidelines. Instead, δ is set by the researcher, typically based on her knowledge of previous research. However, as Clark (2009) observed, this practice leaves some room for subjectiveness, if not bias. Hence, our goal is to find an objective way to set δ.

Ideally, we would have access to a great number of pairs of samples of which we know that they come from the same population. Then we could use them to determine which values the "right" δ takes. The "right" δ value would be such that TOSTs come out positive (i.e. indicating similarity) at the common rates for statistical power ($1 - \beta = 80\%$ and $1 - \alpha = 95\%$). Figure 4.3 illustrates this approach.

The problem with this approach is that it would take at least a few hundred, if not

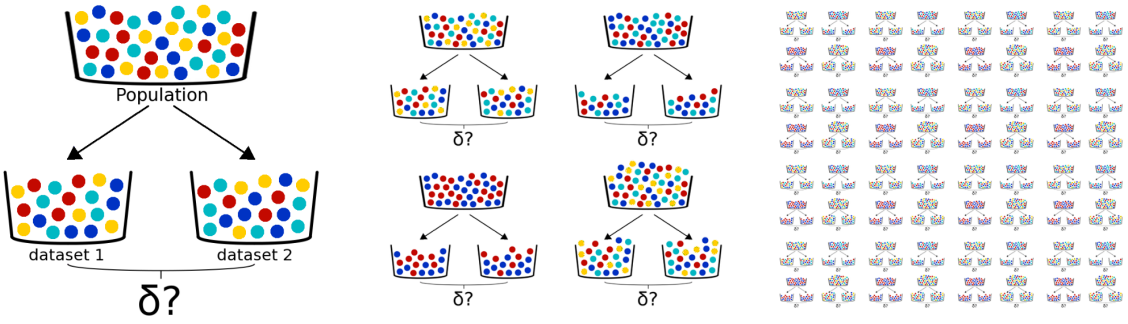


Figure 4.3: An illustration of how comparing pairs of real data sets could be used to outline a way to set δ. If we had two data sets of which we knew that they come from the same population, we could find the right δ, so that the TOST comes out positive (indicating similarity; left). If we did this for numerous data sets, then we might observe patterns across the different δ-values (middle and right).

a few thousand pairs of samples of which we know that they come from the same population in order to find a reliable answer. However, we do not have access to such a large body of similar pairs of samples. Hence, we decided to simulate these data.

4.3 Using Data Simulations To Observe δ

We simulate data that forms pairs of samples which we can then test in TOSTs in order to observe how δ comes out for our simulated data. Our simulations take various degrees of aggregation[6] (from 1 to 4) and various sample sizes (from 3 to 50000) into account.

[6]In our use of the term, an *aggregate* is the collection of multiple data points for a group of observations. In most cases the data points are simply averaged. Typically, data are aggregated in order to preserve the assumption of independence in statistical testing, whilst reducing the chance of extreme measurements for any given group of observations. Consider the following example from phonetics: Assume a researcher is interested in the voice onset time for the initial /d/ in "dude" in Standard American English. The researcher decides to ask five participants. However, to reduce the impact of extreme measurements within participants, the researcher records each participant multiple times. Feeding multiple data points that are on the same item and come from the same participant into his/her analysis can violate the assumption of independence, which is why the researcher decides to average the measurements for each participant first. The data are now *aggregated*. See Figure 4.6 on page 216 for an illustration of how we simulated aggregation.

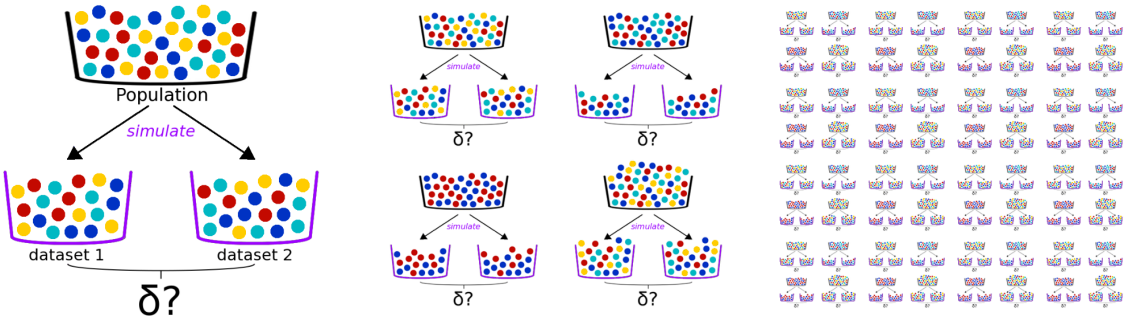


Figure 4.4: An illustration of our approach to the random subsampling procedure. We start with a given data set that functions as the population. We then randomly select two samples. This is our pair of samples of which we know that they come from the same population. We then run a TOST and set δ such that the outcome is positive. This procedure is repeated many times and for many different populations (middle and right).

RANDOM SUBSAMPLING The starting point for our simulations are real data sets, which serve as populations. From such a "population", we then randomly sample pairs of subsets. The sampling is done by randomly drawing data points with replacement[7] and these subsets represent the samples in this two-samples–one-population setting. For each population, we create numerous pairs of subsamples. We then run a TOST on each pair, adjusting δ in the process such that the outcome of the TOST becomes positive (indicating similarity) at the desired rates for statistical power. Figure 4.4 is a schematic illustration of this procedure.

We include various degrees of aggregation and various sample sizes. We systematically repeat the sampling procedure for the following aggregation values: 1 (non-aggregated), 2, 3, and 4 data points per group of observations[8]; and for the following sample sizes: 3, 5, 8, 12, 17, 23, 30, 38, 47, 57, 68, 80, 93, 107, 255, 530, 1037, 2558, 5052, 10155, 25202, and 50088.[9] Figure 4.5 illustrates the sampling procedure for those factors (mean of a sample, its standard deviation, its sample size, size of aggre-

[7] We do this so that we do not quickly run out of data points to draw from.

[8] For linguistic data, a participant is the typical group of observations. For other fields, groups of observations can vary widely. E.g. for weather data, a group of observations could be several temperature measurements for some brief period of time (e.g. measurements around 10pm).

[9] All the values are instances of the following sequence: $F_n = F_{n-1} + (n-1)$, where the seed is $F_1 = 2$. The use of this series is a relic from an earlier, alternative approach to setting δ.

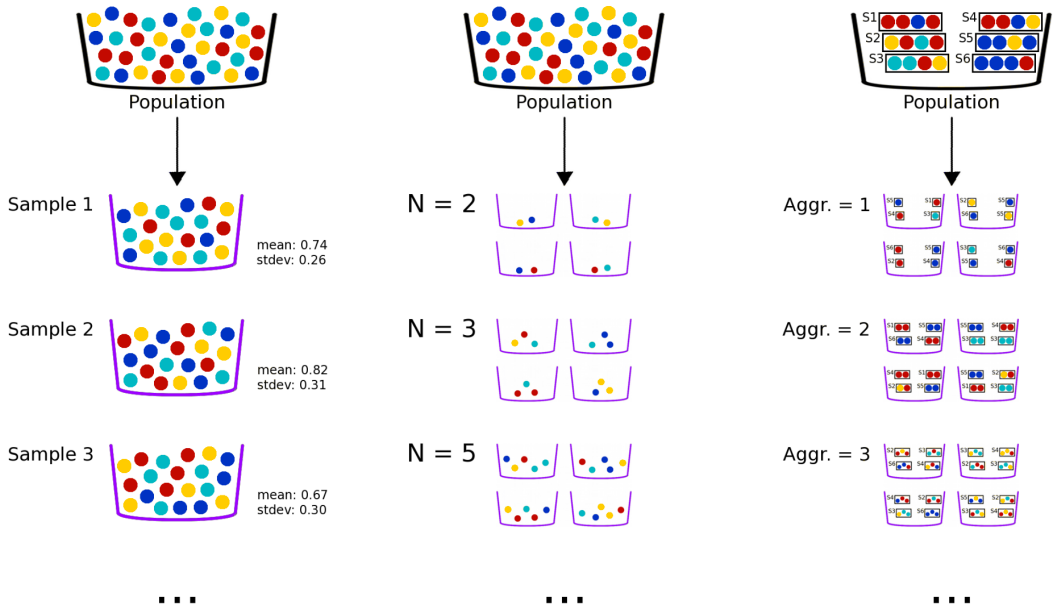


Figure 4.5: An illustration of how some factors vary by the process of random selection (e.g. a sample's standard deviation and mean; left). However, we can control for other factors, e.g. sample size (middle) and aggregation (right).

	Non-Aggregated	Aggregation: 2	Aggregation: 3	Aggregation: 4
N = 3	N = 3 & non-aggregated	N = 3 & aggregation = 2	N = 3 & aggregation = 3	N = 3 & aggregation = 4
N = 5	N = 5 & non-aggregated	N = 5 & aggregation = 2	N = 5 & aggregation = 3	N = 5 & aggregation = 4
...	...	...	...	...
N = 50088	N = 50088 & non-aggregated	N = 50088 & aggregation = 2	N = 50088 & aggregation = 3	N = 50088 & aggregation = 4

Table 4.1: An overview of the different parameter settings for the data simulation.

gates). Figure 4.6 illustrates how the random selection process works for aggregated data. Table 4.1 gives an overview of the different parameter settings for the data simulation.

ISSUES WITH THE β-RATE Initially, we intended to take the following approach: For each condition (represented by the cells in Table 4.1), we planned to sample 10^6 pairs of subsamples on which we would have run TOSTs. We would have observed how δ behaves over those different conditions and then tried to generalise. The standard α-rate ($1-\alpha = 95\%$) was set by adjusting the significance level of the

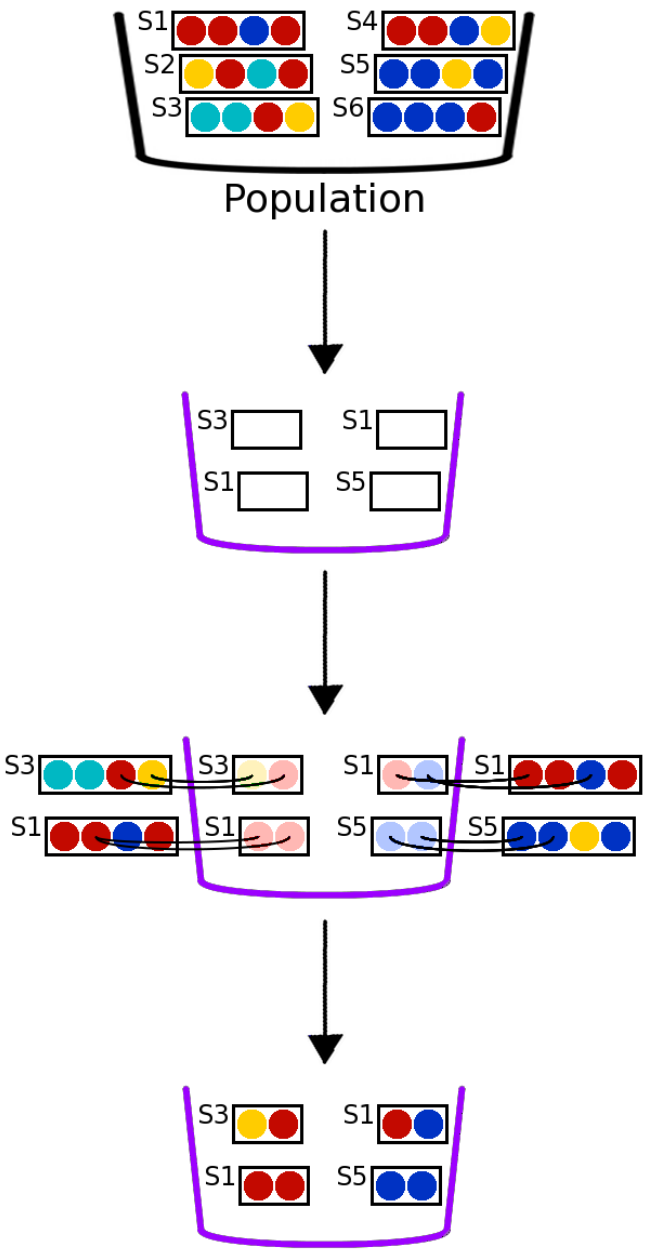


Figure 4.6: An illustration of how the process of random selection works for aggregated data. First, we randomly draw participants from the population. Then, from the corresponding participants, we randomly draw the data points.

t-tests accordingly. However, using the standard β-rate ($1-\beta = 80\%$), calculated by the formula below (modelled after Liu, 2014:32), only 30% to 60% of the performed TOSTs came out positive, depending on the data set.

$$Z_\beta = \frac{\sqrt{\frac{n}{2}} \cdot |\mu_1 - \mu_2|}{s_p} - Z_{\alpha/2} \tag{4.3}$$

$$\text{where } s_p = \sqrt{\tfrac{1}{2}(s_{x1}^2 + s_{x2}^2)}$$

The problem with a low β-rate is that it becomes less likely to find an existing effect. Thus, we take a different approach (below) to ensure a β-rate of 80%.

POST-TESTING β-ANALYSIS We set the β-rate through a post-testing analysis. The general approach is as follows[10]: We start at a certain δ. For each condition (e.g. N = 3 & non-aggregated; cf. Table 4.1), we run a certain number of tests, e.g. we would have a "batch" of 1000 pairs of subsamples. The size of the batches increases over successful outcomes. For each batch, we then run the TOSTs on the pairs of subsamples. From this, we calculate how many of those 1000 tests come out positive. If the outcome is "reasonably close" to 80%, we will have found the right δ-value. If it is considerably below or above 80%, then we adjust δ accordingly and resample another 1000 pairs of subsamples. The size of the batches increases incrementally and so does the criterion for being "reasonably close" to 80%. The general approach for setting β is illustrated in Figure 4.7.

For each condition (e.g. N = 3 & non-aggregated; cf. Table 4.1), the first δ-value of the very first batch (i.e. pairs of samples) is 10% of the range of the values observed in the population. For instance, if the data points of the population vary from -3

[10]For details, particularly on the batches and on a definition of "reasonably close", see page 219.

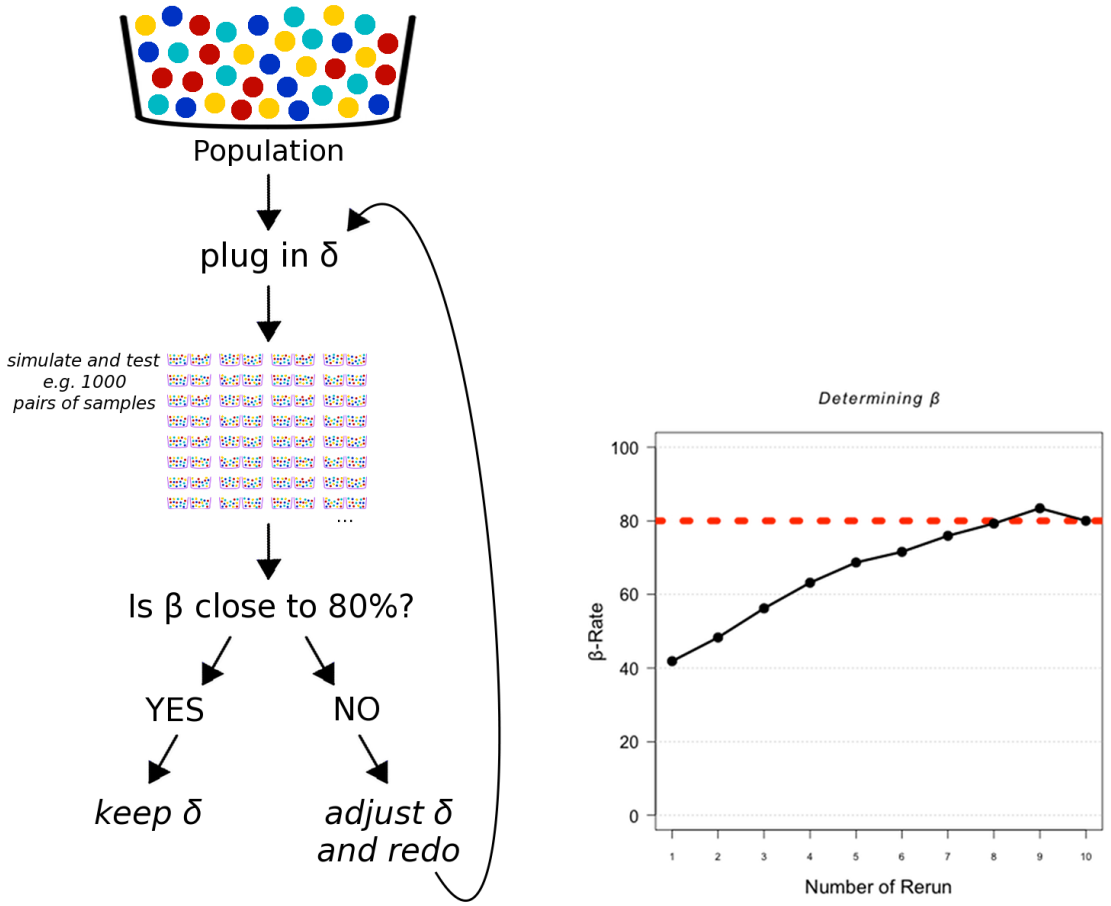


Figure 4.7: Left: An illustration of the procedure we use to set β. We simulate e.g. 1000 pairs of samples, which we then test. We plug in a δ-candidate and observe the outcomes. If about 80% of the tests are positive, then we keep that δ. If not, we adjust δ accordingly and redo this procedure. Right: An illustration of how adjusting δ over several re-runs brings us closer to the desired β-rate of 80%.

to 3, then the first δ-value is 0.6.

The first batch consists of ten pairs of subsamples. We run a TOST on each pair of subsamples of a batch. If the outcome is "reasonably close" to 80%, then the size of the batch is increased incrementally. If the outcome is considerably below or above 80%, then δ is adjusted accordingly, we randomly resample that batch, and we re-run the tests. Our batches are of the following sizes: 10, 10, 25, 50, 100, 250, 500, 1000, 2500, 5000, 10000, 25000, 50000, and 50000 again. For each batch, having several dozen re-runs (i.e. adjustments of δ) is not unusual. We stop at 50000, because we ran out of computational power; e.g. processing an entire run for the last batch takes about ten hours on a high end PC (as of mid-2014). Computational power is also the reason for the incremental approach in general: We could have started with a batch of 50000 and then applied the "adjust-and-re-run" procedure; but this would have been too time-consuming.

REASONABLY CLOSE The formula for a "reasonably close" outcome is given below and illustrated in Table 4.2. It was chosen because it provides a compromise between precision and saving computational resources. For the last increment (a batch of 50000 pairs of subsamples), the outcome is required to lie in a range from 39994 to 40006 positive outcomes; i.e. 80% of the outcomes $\pm 0.012\%$ have to be positive. ($\mathrm{f}_{thresholds}$) gives the concrete formulas for the lower and upper threshold. ($\mathrm{f}_{thresholds}$) is designed to be as precise as is practical.

$$(f_{thresholds}) : 0.8 \cdot batch\ size \pm ceil(\sqrt{0.99^{4\sqrt{batch\ size}}}) \tag{4.4}$$

Ideally, the outcomes would have been 80% exact (in order to mirror the common power rate of $1 - \beta = 80\%$), instead of allowing for a range. However, an exact

Batch Size	$0.8 \cdot$ Batch Size	Lower Boundary	Upper Boundary	Tolerance Percentage
10	8	7	9	$\pm 10\%$
25	20	19	21	$\pm 4\%$
50	40	39	41	$\pm 2\%$
100	80	79	81	$\pm 1\%$
250	200	198	202	$\pm 0.8\%$
500	400	396	404	$\pm 0.8\%$
1000	800	794	806	$\pm 0.6\%$
2500	2000	1990	2010	$\pm 0.4\%$
5000	4000	3987	4013	$\pm 0.26\%$
10000	8000	7986	8016	$\pm 0.14\%$
25000	20000	19989	20011	$\pm 0.044\%$
50000	40000	39994	40006	$\pm 0.012\%$

Table 4.2: An overview of the concrete lower and upper thresholds for the different batch sizes, including the tolerance as a percentage.

result (e.g. 40000 positive outcomes of 50000 tests) would have required a lot of computational power, as this would have led to a lot more re-runs. Second, because of the random subsampling procedure, there will always be some fluctuation of the outcomes. Hence, an outcome of "exactly" 80% would only be an approximation, too, as the 80% apply to only one randomly sampled batch, e.g. to 50000 pairs of subsamples from a certain population. For another 50000 pairs of subsamples from the same population, the outcome would have most likely been slightly different.

The rate at which δ is adjusted is also computed incrementally (cf. Equation 4.5). If the number of positive outcomes for a small batch size (e.g. 10) is not within the specified range (7 to 9), then the adjustments applied to δ are quite big. For a big batch size, the adjustments are smaller, because δ is already set fairly accurately and only needs slight adjustments. See Table 4.3 for the concrete rate of adjustment. This, too, is done to save computational power and increase precision.

Batch Size	Rate of δ-Adjustments
10	$\delta_{i+1} = \delta_i \cdot (0.99062 \; or \; 1.00938)$
25	$\delta_{i+1} = \delta_i \cdot (0.99096 \; or \; 1.00904)$
50	$\delta_{i+1} = \delta_i \cdot (0.99132 \; or \; 1.00868)$
100	$\delta_{i+1} = \delta_i \cdot (0.99182 \; or \; 1.00818)$
250	$\delta_{i+1} = \delta_i \cdot (0.99272 \; or \; 1.00728)$
500	$\delta_{i+1} = \delta_i \cdot (0.99362 \; or \; 1.00638)$
1000	$\delta_{i+1} = \delta_i \cdot (0.99470 \; or \; 1.00530)$
2500	$\delta_{i+1} = \delta_i \cdot (0.99634 \; or \; 1.00366)$
5000	$\delta_{i+1} = \delta_i \cdot (0.99759 \; or \; 1.00241)$
10000	$\delta_{i+1} = \delta_i \cdot (0.99866 \; or \; 1.00134)$
25000	$\delta_{i+1} = \delta_i \cdot (0.99958 \; or \; 1.00042)$
50000	$\delta_{i+1} = \delta_i \cdot (0.99989 \; or \; 1.00011)$

Table 4.3: An overview of the concrete adjustments of δ, depending on the batch size. If the current δ (δ_i) is too large, then it is multiplied by the first argument ($0.99\ldots$). If it is too small, then it is multiplied by the second argument ($1.00\ldots$).

$$(f_{\delta \; increments}) : \delta \cdot (1.0 + \epsilon \pm \frac{\sqrt{0.99^{4\sqrt{batch \; size}}}}{100}) \tag{4.5}$$

$$\text{where } \epsilon = randbetween(-0.0005, 0.0005)$$

For all conditions (e.g. N = 3 & non-aggregated; cf. Table 4.1), we record the δ-value of the very last batch (50000 pairs of subsamples). We apply this procedure to all our data sets (which served as populations) and for each data set, we get one δ-value per choice of N (which we later average; see below). In total, we simulated about 4.3×10^{12} data points from 48 different data sets. The code is written in C++.

TWO PHASES – OUR DATA SETS There are two stages: an initial *calibration phase* (Section 4.4) and a *validation phase* (Section 4.5). In the calibration phase, we use the observed δ-values ($\delta_{observed}$ or simply δ_o) to predict δ-values ($\delta_{predicted}$ or simply δ_f). In the validation phase, which uses different data sets than the calibration phase, we check how δ_f compares to δ_o. For each of the two stages,

Data Set	Source	Area	Data Type	Unit	N	Observ. per N
1	TSJ	Syntax	Judgement Data	7-point scale	66	4
2	TSJ	Syntax	Judgement Data	7-point scale (Z-scores)	66	4
3	TSJ	Syntax	Judgement Data	7-point scale	66	4
4	TSJ	Syntax	Judgement Data	7-point scale	132	2
5	TSJ	Syntax	Judgement Data	7-point scale (Z-scores)	66	4
6	TSJ	Phonetics	Frequencies	Hz	52	2
7	TSJ	Phonetics	Frequencies	Hz	52	2
8	TSJ	Syntax	Judgement Data	7-point scale (Z-scores)	66	4
9	TSJ	Syntax	Judgement Data	7-point scale	132	2
10	TSJ	Gen. Linguistics	Judgement Data	6-point scale	152	1
11	JK	Gen. Linguistics	Judgement Data	5-point scale	60	16
12	TSJ	Phonetics	Duration	ms	104	1

Table 4.4: The linguistic data sets used in the calibration phase.

we use 24 data sets, half of which are linguistic data sets and the other half non-linguistic data sets, taken from various scientific areas. Table 4.4 gives the linguistic data sets used in the calibration phase, Table 4.5 the non-linguistic data sets. Table 4.6 gives the linguistic data sets for the validation phase, Table 4.7 the non-linguistic ones.

Data Set	Source	Area	Data Type	Unit	N	Observ. per N
13	PLOS ONE	Medicine	Body Measurements	Scaled Unit	84	1
14	PLOS ONE	Medicine	Body Measurements	Scaled Unit	84	1
15	PLOS ONE	Psychology	Grades	10-point scale	73	1
16	PLOS ONE	Psychology	Grades	10-point scale	74	1
17	PLOS ONE	Engineering	Match Rates	Percentages	38	1
18	PLOS ONE	Engineering	Match Rates	Percentages	27	1
19	PLOS ONE	Engineering	Reaction Times	ms	38	1
20	PLOS ONE	Engineering	Reaction Times	ms	27	1
21	PLOS ONE	Engineering	Reaction Times	ms	42	1
22	PLOS ONE	Engineering	Reaction Times	ms	42	1
23	NASA	Earth Sciences	Temperature	C	87	1
24	NASA	Earth Sciences	Temperature	C	87	1

Table 4.5: The non-linguistic data sets used in the calibration phase.

Data Set	Source	Area	Data Type	Unit	N	Observ. per N
25	TSJ	Phonetics	Duration	ms	104	1
26	TSJ	Syntax	Judgement Data	7-point scale	132	2
27	TSJ	Syntax	Judgement Data	7-point scale	132	2
28	TSJ	Phonetics	Frequencies	Hz	52	2
29	TSJ	Phonetics	Frequencies	Hz	52	2
30	TSJ	Syntax	Judgement Data	7-point scale	66	4
31	TSJ	Syntax	Judgement Data	7-point scale	132	2
32	TSJ	Syntax	Judgement Data	7-point scale	66	4
33	TSJ	Syntax	Judgement Data	7-point scale (Z-scores)	66	4
34	TSJ	Phonetics	Judgement Data	6-point scale	127	1
35	TSJ	Phonetics	Judgement Data	6-point scale	136	1
36	TSJ	Phonetics	Duration	ms	104	1

Table 4.6: The linguistic data sets used in the validation phase.

Data Set	Source	Area	Data Type	Unit	N	Observ. per N
37	JK	Chemistry	Judgement Data	7-point scale	896	1
38	TSJ	Psychology	Reaction Times	ms	42	1
39	TSJ	Psychology	Reaction Times	ms	42	1
40	UK Government	Economics	Inflation	Percentages	50	1
41	UK Government	Environmental Studies	Rainfall	mm	166	1
42	UK Government	Environmental Studies	Temperature	C	166	1
43	US Government	Economics	Inflation	Percentages	30	12
44	UK Government	Sociology	Demographics	Percentages	333	1
45	German Government	Economics/ Sociology	Unemployment Rate	Percentages	34	1
46	US Government	Economics	Income	USD	67	1
47	US Government	Economics	Income	USD	67	1
48	JK	Chemistry	Judgement Data	7-point scale	896	1

Table 4.7: The non-linguistic data sets used in the validation phase.

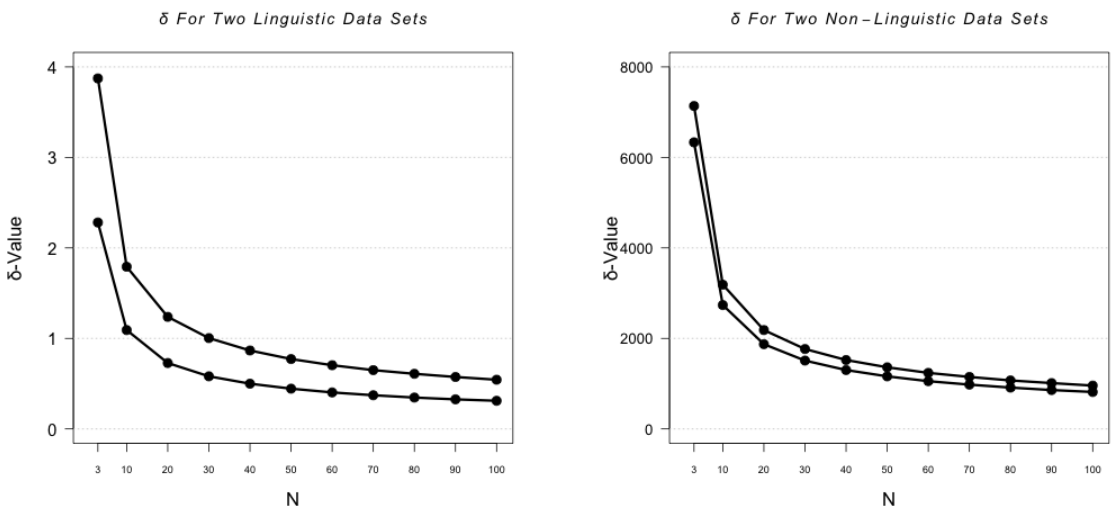


Figure 4.8: A comparison of observed δ_o for two linguistic data sets (left; Figure 4.8a) and for two non-linguistic data sets (right; Figure 4.8b). Both linguistic data sets consist of 7-point scale ratings. Nonetheless, δ_o differs considerably for the two data sets. The non-linguistic data sets consist of reaction times. They, too, differ from each other. Crucially, though, δ_o differs greatly across data sets with different units (e.g. the linguistic and the non-linguistic data sets).

4.4 Calibration Phase: From Observing to Predicting δ

Our overall goal is to use observed δ-values (δ_o) to predict δ (δ_f). However, the δ-values that we observe based on the 24 data sets used in this calibration phase can be hard to compare to each other. Figure 4.8 illustrates δ_o for two linguistic and two non-linguistic data sets.

We checked for any regularities between observed δ-values and other factors (mean, standard deviation, aggregation, sample size, etc.) and found a relationship between δ and the pooled standard deviation (s_p). As described above, the δ-value comes from observing entire batches of pairs of subsamples; but s_p is calculated for each pair of subsamples. We find that the quotient from the division of s_p by δ_o is near-constant for a given pooled subsample size N_p (also pooled from each pair of subsamples). We call the quotient of the division the *Tübingen Quotient* (τ in (f_1)

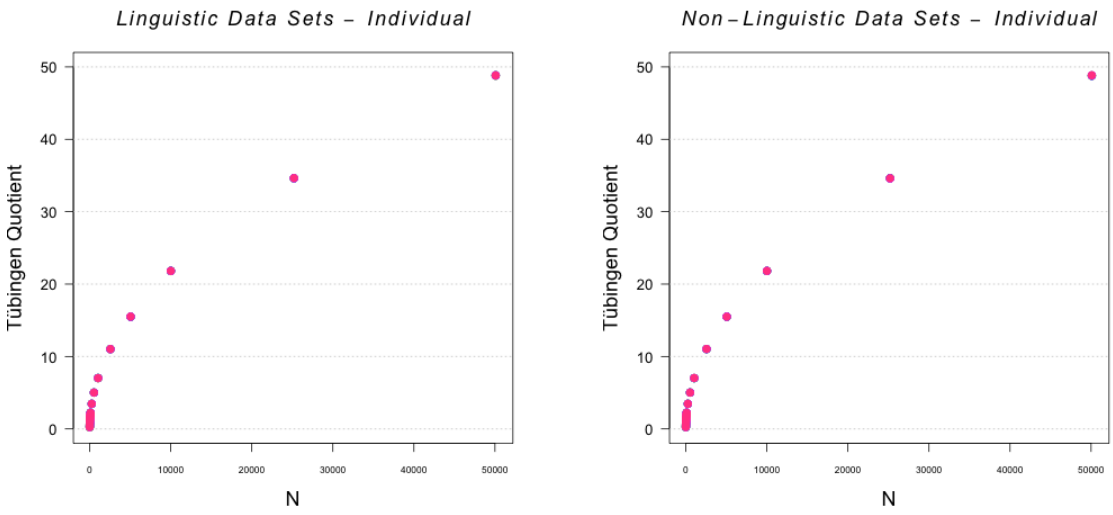


Figure 4.9: Illustrating how observed τ_o plots over increasing sample sizes N_p (x-axis). Left (4.9a): τ_o for the twelve linguistic data sets. Right (4.9b): τ_o for the twelve non-linguistic data sets. N.B.: The results for the individual data sets are so similar that it is hard to graphically show the individual differences.

below; τ is calculated from simulated data and thus labelled τ_o). For all linguistic data sets, we get a τ_o for any given N_p. τ_o comes out extremely similar (within a range of $\pm 0.1\%$); see Figure 4.9a. This is why we decided to average the results. The same is true for the non-linguistic data sets; see Figure 4.9b. The averaged results of the linguistic data sets are illustrated in Figure 4.10a. Figure 4.10b illustrates averaged τ_o for the non-linguistic data sets.

$$(f_1): \quad \tau = \frac{s_p}{\delta_o} \tag{4.6}$$

We tested for differences between the τ_o-values from the linguistic data sets and the τ_o-values from the non-linguistic data sets, using a two sided unpaired t-test. The test come out non-significant ($p > 0.05$). This is why we decided to treat the two as one, as illustrated in Figure 4.11.

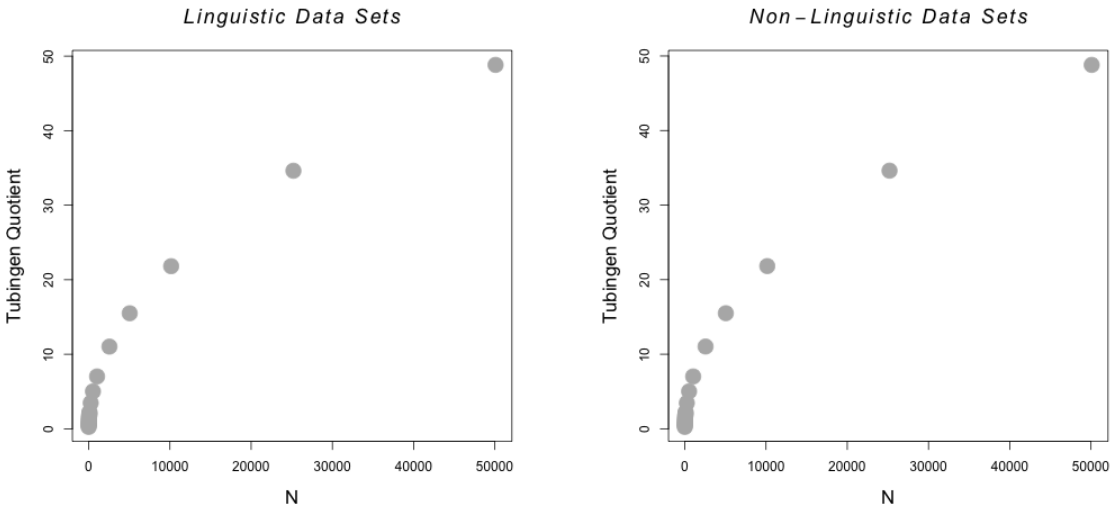


Figure 4.10: Illustrating how observed τ_o (averaged for the twelve linguistic data sets and for the twelve non-linguistic data sets; y-axis) plots over increasing sample sizes N_p (x-axis). Left (4.10a): τ_o for the linguistic data sets. Right (4.10b): τ_o for the non-linguistic data sets.

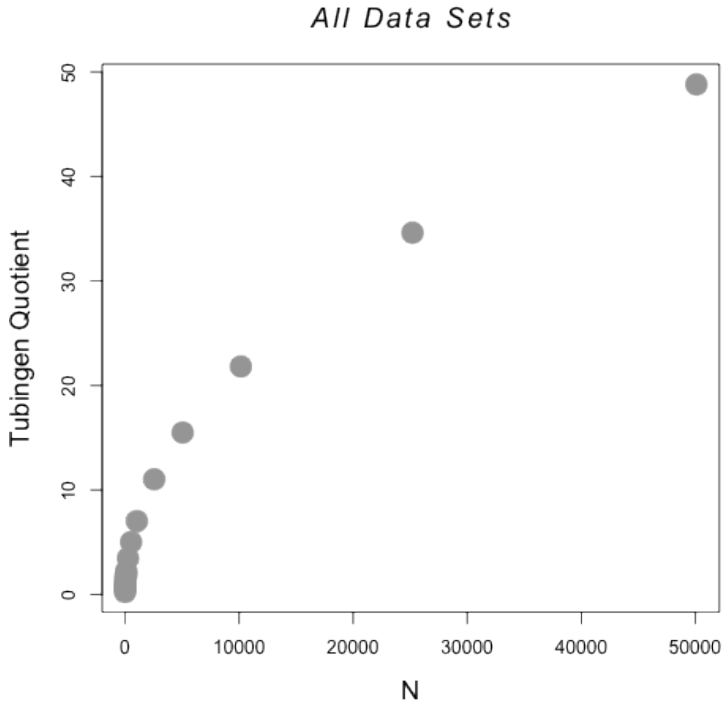


Figure 4.11: Illustrating how observed τ_o (y-axis) plots over increasing sample sizes N_p (x-axis) for all 24 data sets of the calibration phase.

We curve fitted the graph in Figure 4.11 (using Python) and arrived at (f_2), which *predicts* τ (τ_f = "predicted τ"; this contrasts to τ_o, as observed in our simulations).

$$(f_2): \quad \tau_f = \frac{\sqrt{N_p}}{4.58} \tag{4.7}$$

(f_2), and the 4.58 in particular, are our critical findings, because: *by reversing (f_1) to (f_3), (f_2) can be used to predict δ (δ_f).* (f_4) is (f_3) with τ_f resolved as per (f_2).

$$(f_3): \quad \delta_f = \frac{s_p}{\tau_f} \tag{4.8}$$

$$(f_4): \quad \delta_f = 4.58 \, \frac{s_p}{\sqrt{N_p}} \tag{4.9}$$

In a first comparison, τ_o and τ_f match within a range of $\pm 0.1\%$ for large parts of the range we tested. Figure 4.12 illustrates this. However, we observe slightly deviant early model behaviour ($N_p < 15$) and cannot exclude deviant late model behaviour ($N_p > 50000$).

δ can now be set objectively (given the assumption of $1-\alpha = 95\%$ and $1-\beta = 80\%$), which eliminates a potential source of subjectivity from the TOST. However, the problem with the comparison illustrated in Figure 4.12 is that we compare predicted τ_f to the data from which we extract τ_f, which is circular. This is why we add a validation phase, which includes another 24 data sets, half of which are linguistic data sets and the other half consists of non-linguistic data sets.

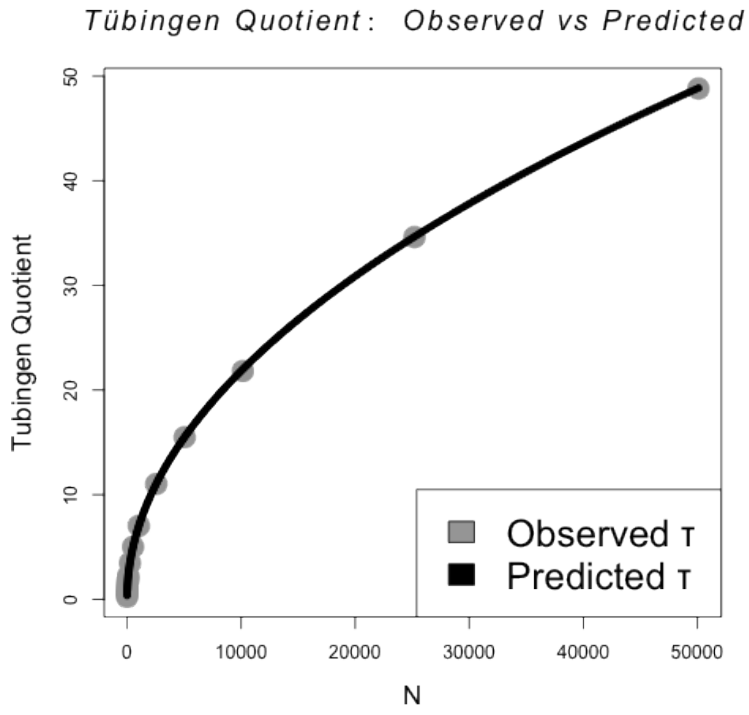


Figure 4.12: Illustrating how observed τ_o (grey; y-axis) compares to predicted τ_f (black; y-axis) over increasing sample sizes N_p (x-axis).

4.5 Validation Phase

In a next step, we compare predicted τ (τ_f; from (f_2)) to observed τ for further data sets ($\tau_{o\,2}$). $\tau_{o\,2}$ is determined through further simulations based on the data sets in Tables 4.6 and 4.7. The methodology for the simulations is the same as described in Section 4.4. The results of this comparison are illustrated in Figure 4.13.

On the whole, $\tau_{o\,2}$ and τ_f match within a range of $\pm 0.1\%$, as illustrated by Figure 4.14. Thus, the formulas (f_1) to (f_4) presented here hold in general. As before, we observe slightly deviant early model behaviour ($N_p < 15$) and cannot exclude deviant late model behaviour ($N_p > 50000$).

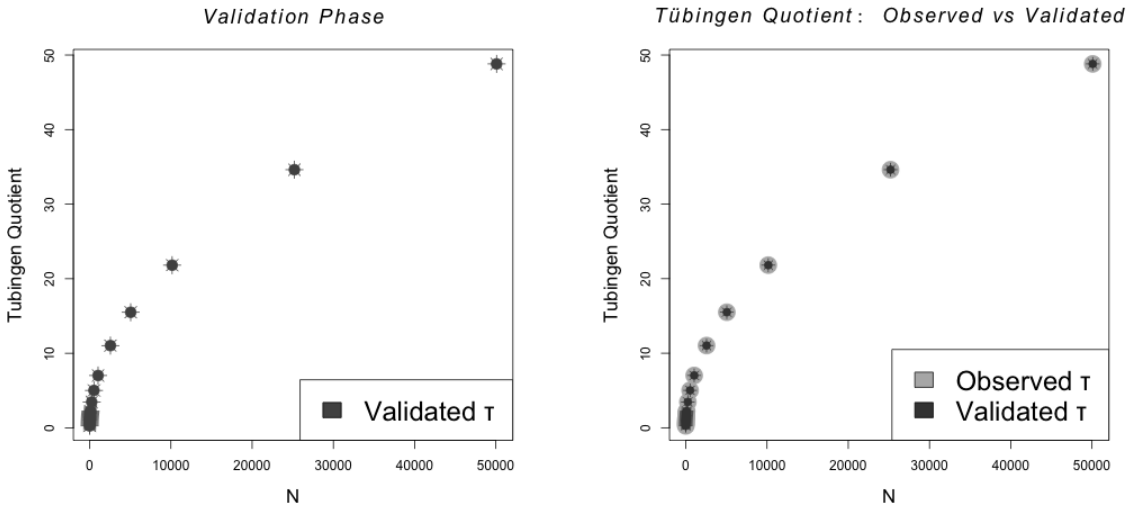


Figure 4.13: Left (4.13a): An illustration of newly observed $\tau_{o\,2}$ (dark grey asterisks; y-axis) over increasing sample sizes N_p (x-axis). $\tau_{o\,2}$ is an average taken from all 24 data sets of the validation phase. Right (4.13b): A comparison of $\tau_{o\,2}$ from the validation phase (dark grey asterisks) and τ_o from the calibration phase (light grey dots; from Figure 4.11).

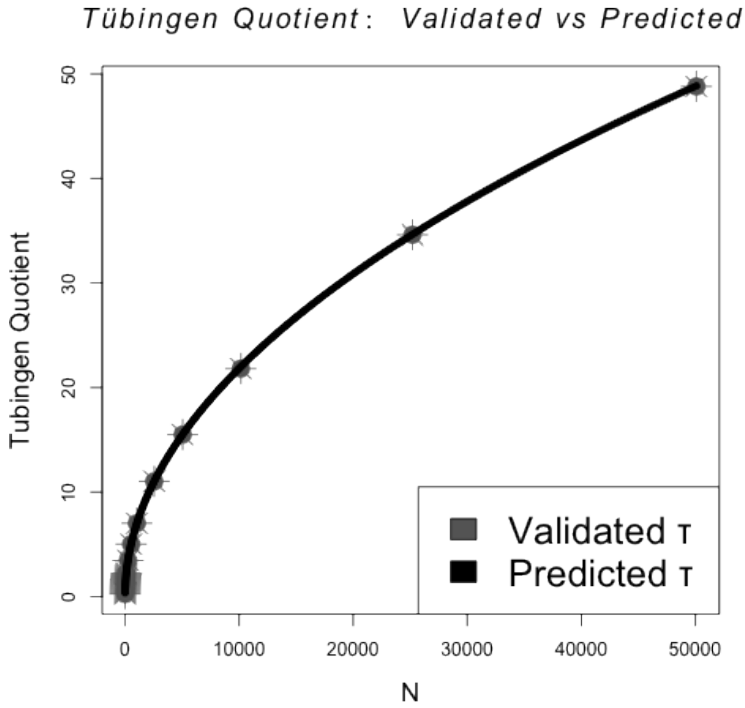


Figure 4.14: An illustration of how the newly observed $\tau_{o\,2}$ (dark grey asterisks; y-axis) compares to predicted τ_f (black; y-axis) over increasing sample sizes N_p (x-axis). ($\tau_{o\,2}$ is an average from the 24 data sets of the validation phase.)

4.6 TOSTs vs t-Tests

An accurate prediction of the Tübingen Quotient implies an accurate prediction of δ. However, the results might be hard to contextualise; this is why we also perform t-tests on our data.

For all of our 48 data sets (cf. Table 4.4, Table 4.5, Table 4.6, and Table 4.7), we randomly sample 10^6 pairs of subsamples following the methodology in Section 4.4. As for sample sizes, we consider the following Ns: N = 3, 4, 5, 10, 20, 30, 40, 50, 100, 250, 500, 1000, 2500, 5000, 10000. We only consider aggregated data with two observations. For each pair, we perform both a TOST and a t-test. For the TOST, we plug in δ-values according to (f_4). The δ-values are calculated for each pair of subsamples individually. Since we know that each pair of samples that we compare comes from the same population and since we know that the TOSTs are set to $1 - \beta = 80\%$ and $1 - \alpha = 95\%$, about 80% of all the TOSTs should return a positive result. We also perform t-tests (independent and one-tailed, as we check for directions prior to testing) on the very same pairs of subsamples. For the t-test, the α-rate stands at $\alpha = 5\%$. So, no more than 5% of the t-tests should come out positive (since we simulate *similar* data).

For each pair that we test, four outcomes exist: 1) a negative TOST result and a negative t-test result (*inconclusive outcome*); 2) a positive TOST result and a negative t-test result (*conclusive similarity*); 3) a negative TOST result and a positive t-test result (*conclusive dissimilarity*); 4) a positive TOST result and a positive t-test result (*contradictory outcome*). We report the averaged results (Table 4.8) and the individual results for one data set, viz. data set 8 (Table 4.9). Data set 8 is representative of the other data sets, which come out very similar.

The overall results come out as expected. About 80% of the TOSTs come out positive, hardly any of the t-tests come out positive, and - importantly - there are

N	Negative TOST Negative t-Test	Positive TOST Negative t-Test	Negative TOST Positive t-Test	Positive TOST Positive t-Test
10	298851	701070	79	0
20	244222	755769	9	0
30	228870	771127	3	0
40	221763	778235	2	0
50	217340	782658	2	0
100	208873	791127	0	0
500	205811	794189	0	0
1000	201440	798560	0	0
5000	201279	798721	0	0
10000	201212	798788	0	0

Table 4.8: The TOST and t-test results for 10^6 pairs of subsamples, which come from the same population. The table gives the average of all 48 data sets. Thus, each row gives the product of 48×10^6 pairs of subsamples.

N	Negative TOST Negative t-Test	Positive TOST Negative t-Test	Negative TOST Positive t-Test	Positive TOST Positive t-Test
10	263949	735987	64	0
20	228501	771491	8	0
30	219913	780085	2	0
40	215661	784338	1	0
50	213227	786772	1	0
100	206035	793965	0	0
500	205183	794817	0	0
1000	200975	799025	0	0
5000	201400	798600	0	0
10000	200569	799431	0	0

Table 4.9: The TOST and t-test results for 10^6 pairs of subsamples, which come from the same population. The table gives the results for data set 8 (cf. Table 4.4).

no clashes between TOSTs and t-tests (i.e. there are no contradictory outcomes). Further, the results indicate why data sets with small sample sizes are problematic. The TOST fails to fully capture the similarity of data sets with a low N. On the other side, the TOST does not fall for the "big-N" fallacy: Even over increasing sample sizes, the results remain very close to 80%. Altogether, these results illustrate nicely how potent a well-conducted TOST can be.

It would be interesting to reverse the setting, i.e. to test pairs of subsamples of which we know that they are dissimilar (by sampling pairs of subsamples from clearly different populations). In such a scenario, one would expect that 80% of the t-tests come out positive and less than 5% of the TOSTs. As this is a non-trivial task in various aspects, we have to leave this for further research.

4.7 Conclusion

In our view, the TOST is a useful tool in a linguist's repertoire, allowing the investigation of research questions that ask about similarity. So far, the lack of a clear method to objectively set δ, the controlling parameter of the TOST, might have been a barrier to the use of this test. The present work outlines such guidelines and we hope that they will help boost similarity testing in linguistics. However, there is an important limitation: Our results work best for samples with a size of $N \geq 15$.

Chapter 5

Concluding Remarks

The present work asked *why* quantitative methods, and acceptability judgement tasks in particular, are beneficial to syntactic enquiry (Chapter 1) and looked into aspects of *how* to best conduct syntactic judgement tasks (Chapter 2 and Chapter 3). In Chapter 4, we validated a common similarity test, the TOST. We included this, because we started out with judgement data (but then moved on to any kind of data). The relevance of all chapters comes from their goal to improve data reliability (which might help to improve syntactic modelling).

To look into the question of *why* formal methods are needed, we quantitatively compared informal methods (researcher introspection) and formal methods (acceptability judgement tasks). Based on our results, we argue that quantitative data are more reliable than informal data. Critics might say that our data is corrupted by processing factors and that it takes experts to deal with syntactic questions, since syntactic knowledge is tacit. We do not find this defence convincing: We checked our items for testability, which included a check for potential impact of processing effects. Further, processing factors cannot explain why many items that are "bad" according to linguist (marked by a "*") received in-between online ratings. We

think that processing factors might affect a few of the items that syntacticians are concerned with; but this cannot be a reason to not use formal methods at all (as this affects the adequacy of syntactic data at large).

This is not to say that all syntacticians have to conduct judgement tasks all the time. Sometimes it is not necessary (one does not need a judgement task to back the claim that "Pete went home" is grammatical) and sometimes it is just not practicable, e.g. when the researcher is doing fieldwork (for which established methods already exist). But for the investigation of many phenomena, a judgement task is beneficial (the same applies to other formal methods, e.g. corpus analyses).

Chapter 2 and Chapter 3 were looking at questions of *how* to best conduct acceptability judgement tasks. In Chapter 2, we looked at syntactic constructions from different modalities (i.e. primarily occurring in spoken vs written language) and asked whether the mode of presentation or the degree of formality had an influence on the outcome of a judgement task. Differences in modality are of interest, because they can reveal aspects of how syntactic phenomena are processed by the human brain. We analysed our experimental results with a linear mixed effect model and found no difference between auditory and textual questionnaires and no difference between informal and formal questionnaires. With respect to these factors, acceptability ratings are very robust. The downside of our findings is that when it comes to investigating questions of modality, syntacticians have to resort to other methods, in most cases corpus analyses. This also means that a textual questionnaire should suffice for most judgement tasks.

Chapter 3 is motivated by the fact that linguists rarely make use of Z-scores or ordinal data for their judgement data. Instead, they stick to standard data. While there are theoretical arguments for standard data, it is also well known from psychology and other fields that Z-scores and ordinal data are typically advantageous. To look

into this matter, we randomly sampled 36 sentences from the syntactic literature and tested how well they can be distinguished using standard data vs Z-scores vs ordinal data. Effectively, we tested for statistical power. We found that when it comes to statistical power, Z-scores and ordinal data are superior to standard data. This is not to say that using Z-scores or ordinal data is a must, but linguists should at least seriously consider their use. And if a researcher decides to not use them, this decision should be justified. In general, though, we would recommend to report standard data (and ideally also the Z-scores or the ordinal results) and to use Z-scores or ordinal data for the statistical analyses.

In Chapter 4, we validate the TOST, a common similarity test. Initially, we intended to do this for syntactic judgement data, but then realised that our formula applies in general, i.e. to any kind of (normally distributed) data. It has long been unknown how to set the TOST's critical parameter δ. Through data simulations, we provided guidelines on how to do this. Caution should be exercised, though, with low and high N's.

If we had to provide a take-home message, it would be this: There is nothing special about linguistic data that justifies the rejection of quantitative methods to the extent that we currently see. Staunch proponents of researcher introspection claim that it takes an expert to judge and explain such data, because syntactic knowledge is tacit. We disagree: In most cases, such judgements can also be provided by laymen (but: it takes a linguist to explain the judgements in depth). Crucially, relying on layman judgements has the advantage that quantitative methods can be used, which increases data reliability and allows the application of standard statistical tests.

There are a few issues that we would like to follow up on in the future. First, we observed a surprising amount of gradience in Chapter 1. We think that the extent of gradience has to be explained, as it was not caused by processing factors. Second,

the question arises as to how syntax and semantics compare with respect to the validity of informal data, as many semanticists rely on researcher introspection, as well. Thirdly, throughout our experiments, we observed a worrisome amount of non-cooperative behaviour. Appendix 6 touches on this, but it needs to be addressed in more depth.

Chapter 6

Appendix: Detecting Non-Cooperative Participants in Acceptability Judgement Tasks

6.1 Introduction

In recent years, several studies validated Amazon Mechanical Turk (AMT) as a means to recruit participants and to run experiments (e.g. Snow et al., 2008, Buhrmester, Kwang, and Gosling, 2011, Mason and Suri, 2012, Crump et al., 2013, Berinsky, Huber, and Lenz, 2012, Paolacci et al., 2010; syntax specific: Gibson et al., 2011, and Sprouse, 2011). At the same time, it has been noted that a considerable percentage of AMT participants are non-cooperative (e.g. Kittur et al., 2008, Downs, Holbrook, Sheng, et al., 2010, Kazai et al., 2011, Eickhoff and de Vries, 2013).

However, to the best of our knowledge, few judgement studies, syntactic or otherwise, employ adequate techniques to prevent, discourage, and detect non-cooperative behaviour. In this appendix, we show that failing to exclude non-cooperative par-

ticipants has potentially damaging effects on the reliability of the results. Further, we outline strategies to discourage and to detect non-cooperative behaviour. Some of these strategies are generic, some of them are specific to judgement tasks.

OVERVIEW In Section 6.2.1, we examine the core concepts used in this appendix. We then look at the prevalence of non-cooperative behaviour, the demographics of non-cooperative participants, and the impact on the results (Section 6.2.2). In Section 6.2.3, we look at different types of non-cooperative behaviour and in Section 6.3, we discuss strategies to prevent, discourage, and detect non-cooperative behaviour. Section 6.4 concludes this appendix.

6.2 Further Background

6.2.1 Core Concepts

Judgement Studies In a judgement study, participants rate certain stimuli with respect to a certain percept or concept. Such percepts and concepts can include perceived brightness, perceived harmony, the likability of brands, syntactic acceptability, etc. Judgements are given on some kind of scale (e.g. a 7-point scale). In Chapter 1, Chapter 2, and Chapter 3, we conducted several syntactic acceptability judgement tasks.

Non-Cooperative Behaviour We consider any participant who does not comply with the task non-cooperative. For judgement studies, not submitting your true judgement is the most serious way of not complying with the task (others include: being inattentive, trying to accommodate the researcher, etc.).

However, how do we know whether a participant is "telling the truth"? In some cases, it is obvious that a participant is being non-cooperative (for instance, if a participant rates a textual stimuli like "The horse raced past the barn fell" as acceptable within 320 ms, it is clear that he/she could not have even read the sentence). In other cases, it is harder to tell whether a participant is behaving truthfully (e.g. if he/she judges the same sentence as acceptable after 30 s).[1]

Amazon Mechanical Turk With its Mechanical Turk, Amazon offers a crowdsourcing platform to recruit participants ("workers") for various tasks ("HITs"), among them judgement studies. AMT also offers a tool to design studies so that they can be run directly on AMT, but also allows for running studies on external websites. (Many of the references in Section 6.1 also offer excellent introductions to AMT.)

From our own experience, there will be more non-cooperative participants when recruiting participants via AMT than with "offline" recruiting methods. When we ran a conventional pen and paper study and recruited our participants on campus, we did not notice any issues with non-cooperative behaviour (Juzek, unpublished manuscript). Similar observations hold for the study in Chapter 2, Section 2.2.1, in which we did not notice a problem with non-cooperative participants. However, when we ran our first pilots on AMT, we immediately noticed that there were several participants whose ratings could not have been genuine.

6.2.2 Prevalence, Demographics, and Impact

PREVALENCE Downs, Holbrook, Sheng, et al., 2010, estimated that 38.9% of their AMT participants were non-cooperative. Kazai et al., 2011, reported a

[1] N.B.: We do not consider incomplete results from stopping mid-questionnaire and dropping out as non-cooperative, because doing this is a participant's right.

number between 23.9% and 56.8% (depending on whether one counts "incompetent" and "sloppy" participants as non-cooperative).

The numbers that we encountered throughout this thesis are considerably lower, because we already employed techniques to prevent and discourage non-cooperative behaviour (see below for details). 7% of our participants in Chapter 1 were non-cooperative. This number stands at 11% for Chapter 2 and at 8% for Chapter 3. With adequate prevention but no discouragement, we think this number will be around 15% (for details on how this number is derived, see Section 6.3).

DEMOGRAPHICS The demographics of the participants that we deemed cooperative are as follows: We recruited slightly more male participants than female participants (178 female participants vs 209 male participants; 46.0% vs 54.0%) and their mean age is 34.8 years (standard deviation: 10.3). The picture is quite different for the participants that we deemed non-cooperative: 28 of the 36 non-cooperative participants are male (77.8%) and their mean age is 26.9 (4.7). Thus, while our AMT participants are quite diverse in general, the non-cooperative participants (as defined below) have a specific demographic: They tend to be men in their mid- to late twenties.

There is an obvious question: If we think these participants were not truthful in their acceptability judgements, how can we be sure that their answers to our demographic questions are correct? The answer is simple: We cannot be sure. However, there are two reasons why we think these demographics are based on correct answers. First, when it comes to their ratings, participants were not aware that we were screening their ratings to the extent we did. We suspect that many of them genuinely thought that they would go undetected (see below for details on our detection techniques). But when filling in personal information, they will have been aware that giving nonsensical answers (e.g. home country: "Pluto") will be detected

quickly. Alternatively, they would have had to make up answers that do not look suspicious (e.g. instead of giving their real home country, say the USA, they answer "Canada"). However, in our view, the motivation of most non-cooperative participants is that they want to get through the survey quickly, at minimal effort. Making up ingenious answers that do not look suspicious will take more effort than typing the real, readily available answer. Thus, we think these demographics are, by and large, correct.[2]

IMPACT ON THE RESULTS All this would not matter if non-cooperative behaviour did not affect the reliability of the results in any way. Unfortunately, it does. This has been documented for AMT data in general (e.g. Downs, Holbrook, Sheng, et al., 2010, Kittur et al., 2008, Kazai et al., 2011, Eickhoff and de Vries, 2013) and we think it also holds for judgement data.

Throughout this thesis, we used AMT to recruit participants for three of our experiments (for the judgement task, though, we used an external website). Each experiment comprises four sessions so that there are twelve sessions in total (see Table 6.1). For each session, we check how the overall mean ratings by the cooperative participants compare to the overall mean ratings by the non-cooperative participants. We test for differences, using Wilcoxon Signed Rank Tests (as mentioned in Chapter 3, a good deal of our data is not normally distributed). The overall mean ratings and the results of the Wilcoxon Signed Rank Tests are also given in Table 6.1.

The mean ratings by cooperative and non-cooperative participants differ considerably and in eight of eleven sessions, the mean ratings were significantly different. This is an indication that data by cooperative and non-cooperative participants

[2]There is one exception to this: Anyone under the age of 18 will have changed their age to appear over 18 for legal reasons. This issue, though, concerns the demographics of the cooperative participants, as well.

Session	Measurement Method	Overall Mean: Cooperative Participants	Overall Mean: Non-Cooperative Participants	Significance W.S.R.T. p-value
Chapter 1 Session 1 "Binary Part 1"	2-Point S.	0.54 (0.33)	0.63 (0.32)	$<$ **0.05**
Chapter 1 Session 2 "Binary Part 2"	2-Point S.	0.49 (0.33)	NA	NA
Chapter 1 Session 3 "Gradient Part 1"	7-Point S.	4.41 (2.11)	4.46 (1.71)	$>$ 0.05
Chapter 1 Session 4 "Gradient Part 2"	7-Point S.	4.28 (2.12)	3.77 (1.50)	$<$ **0.05**
Chapter 2 Session 1 "Informal Auditory"	5-Point S.	3.61 (1.47)	3.74 (0.85)	$>$ 0.05
Chapter 2 Session 2 "Informal Textual"	5-Point S.	3.54 (1.48)	3.24 (1.23)	$<$ **0.05**
Chapter 2 Session 3 "Formal Auditory"	5-Point S.	3.63 (1.47)	3.34 (0.81)	$<$ **0.05**
Chapter 2 Session 4 "Formal Textual"	5-Point S.	3.64 (1.46)	3.54 (1.17)	$<$ **0.05**
Chapter 3 Session 1 "Binary Likert Sc."	2-Point S.	0.49 (0.50)	0.67 (0.48)	$<$ **0.05**
Chapter 3 Session 2 "Gradient Likert Sc."	7-Point S.	4.07 (2.19)	3.76 (1.40)	$>$ 0.05
Chapter 3 Session 3 "Thermometer M."	Therm. M.	0.48 (0.34)	0.61 (0.16)	$<$ **0.05**
Chapter 3 Session 4 "Magnitude Est."	Magn. Est.	1.33 (0.79)	1.10 (0.75)	$<$ **0.05**

Table 6.1: The experimental sessions of the three experiments for which we used AMT, the overall mean ratings both by the cooperative participants and by the non-cooperative participants (including the population standard deviations in parentheses), and the outcomes for a difference test between the two groups. (N.B.: We did not detect any non-cooperative participants in Session 2 of Chapter 1. W.S.R.T. = Wilcoxon Signed Rank Tests.)

differ in quality. Crucially, the ratings by the cooperative participants vary considerably. We would expect this, because these ratings include a wide range of syntactic constructions.

Typically, ratings by non-cooperative participants also vary less, as less variation is a sign of non-discerning, non-genuine ratings. To look into this matter further, we calculate the individual variances: That is, we determine how much the ratings vary within each participant. Table 6.2 gives the averaged individual variances. We also add F-tests to test whether the variances between cooperative and non-cooperative participants are different. In eight of the eleven tests, the variances differ significantly. This is a clear indicator that ratings by non-cooperative participants are less informative and inferior in quality.

6.2.3 Types of Non-Cooperative Behaviour

During our experiments, we observed the following patterns in behaviour: cooperative behaviour, non-cooperative behaviour, and (probably) inattentive behaviour.

The vast majority of participants gave plausible ratings at plausible reaction times (i.e. the time it takes a participant to rate a stimulus).[3] We deem such participants cooperative.

However, there are a few participants whose ratings are implausible (i.e. non-discerning) and whose reaction times are unrealistically fast. We deem such participants non-cooperative. Following Kazai et al. (2011), we call such participants

[3]As to plausible ratings: Most of our experiments included extremely bad and extremely good reference sentences. We established their badness/goodness in pilots and previous experiments. If a participant gave higher ratings to the "bad" items than to the "good" items, then we consider such ratings implausible (details in Section 6.3).
As to plausible reaction times: Building on a formula in Bader and Häussler (2010:289) and our experience from previous experiments, we are able to define how fast is too fast (for details, see Section 6.3).

Session	Measurement Method	Mean Individual Variances: Cooperative Participants	Mean Individual Variance: Non-Cooperative Participants	Significance F-Test p-value
Chapter 1 Session 1 "Binary Part 1"	2-Point S.	0.33	0.31	> 0.05
Chapter 1 Session 2 "Binary Part 2"	2-Point S.	0.32	NA	NA
Chapter 1 Session 3 "Gradient Part 1"	7-Point S.	2.03	1.64	$<$ **0.05**
Chapter 1 Session 4 "Gradient Part 2"	7-Point S.	2.04	1.44	$<$ **0.05**
Chapter 2 Session 1 "Informal Auditory"	5-Point S.	1.41	0.83	$<$ **0.05**
Chapter 2 Session 2 "Informal Textual"	5-Point S.	1.44	1.03	$<$ **0.05**
Chapter 2 Session 3 "Formal Auditory"	5-Point S.	1.40	0.80	$<$ **0.05**
Chapter 2 Session 4 "Formal Textual"	5-Point S.	1.38	1.15	$<$ **0.05**
Chapter 3 Session 1 "Binary Likert Sc."	2-Point S.	0.49	0.48	> 0.05
Chapter 3 Session 2 "Gradient Likert Sc."	7-Point S.	2.01	1.19	$<$ **0.05**
Chapter 3 Session 3 "Thermometer M."	Therm. M.	0.30	0.14	$<$ **0.05**
Chapter 3 Session 4 "Magnitude Est."	Magn. Est.	0.63	0.51	> 0.05

Table 6.2: The means of the individual variances for the experimental sessions of the three experiments. We compare the variances with F-tests, for which we give the significance levels. (N.B.: We did not detect any non-cooperative participants in Session 4 of Chapter 1.)

spammers. We observed two types of spammers, though: Some spammers rush through the entire survey and both their overall times and their mean reaction times are implausibly fast. These are easy to detect by their overall times. However, we observed that a majority of spammers are more clever: They wait on one or two items in order to make their overall times look plausible (for details, see Section 6.3).

We also observed a few participants whose ratings are implausible, but whose reaction times are fine. Following our definition, this is also non-cooperative behaviour. We think the most likely explanation for this kind of behaviour is that those participants are *inattentive*.

6.3 Prevention, Discouragement, and Detection

There are three major ways of dealing with non-cooperative behaviour: prevent it, discourage it, detect it.

PREVENTION Among others, Eickhoff and de Vries (2013) recommended some pre-study prevention techniques, including filtering by prior performance (ibid.:129). In all AMT studies throughout this thesis, we required that potential participants had to have an Amazon Mechanical Turk approval rate of at least 98% and to have finished at least 5000 approved tasks. (These criteria are less strict than what it takes to become an AMT "master worker"). Obviously, there is a trade-off at this point: One could have more lenient recruitment criteria, resulting in more non-cooperative participants. Or one could set really strict criteria, which, however, might result in a rather homogenous pool of participants: That is, only participants who "do AMT at a (semi-)professional level" will be admitted.

DISCOURAGEMENT Prevention is about not letting potentially non-cooperative participants take part in one's study. However, some *potentially* non-cooperative participants might have still gotten in. We wish to discourage them from being non-cooperative. To do so, we included an on-line warning mechanism in two of our studies. As described in Chapter 1 and Chapter 3, the mechanism works as follows: If a participant's reaction time for a stimulus is below what is physically possible, then we count this as a violation. After a certain number of violations (e.g. four in Chapter 1), a pop-up window appears, warning the participant (see Figure 6.1). A second pop-up window appears after even more violations. For Likert Scales, we consider anything below 400 ms as unrealistically fast, as this is less than half of the expected reading time of typical three-word-sentences (we built on a formula in Bader and Häussler, 2010:289, in determining a sentence's expected reading time). Using the Thermometer Method and Magnitude Estimation, we consider anything below 1200 ms as unrealistically fast (in these methods, the participant has to click on a text field, type in a rating, and then press the "rate" button in order to submit the rating; we estimate that these additional actions take about an extra 800 ms).

This warning mechanism has a clear effect. The experiments in Chapter 1 and Chapter 3 made use of this warning mechanism. In Chapter 1, only five out of 143 participants (4%) went so fast that we excluded them for being too fast. In Chapter 3, this number stood at six out of 137 participants (4%). Chapter 2 did not make use of this warning mechanism and we had to exclude fifteen out of 142 participants (11%) for being too fast. Thus, including this warning mechanism cut this kind of non-cooperative behaviour by more than 50%. It seems as if once participants realise that (clear) non-cooperative behaviour can be detected, they refrain from being non-cooperative.

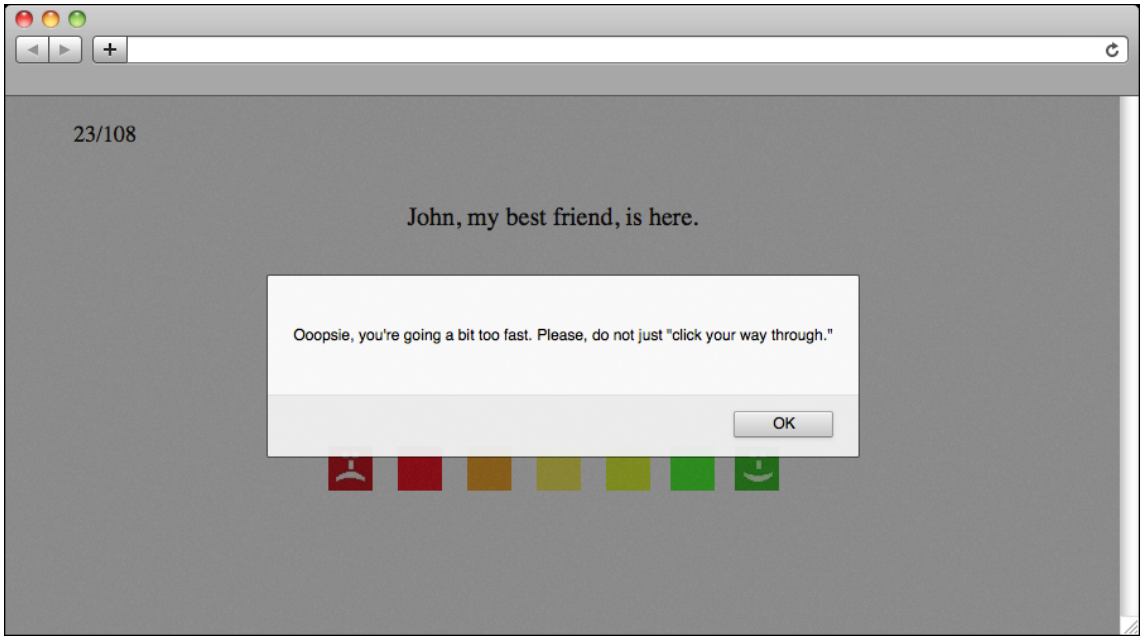


Figure 6.1: A screen capture of our on-line warning mechanism. This is the first warning message.

(The second message is less friendly and read: "Sorry, you're going too fast and you might not get approved. If you are getting this message although you're doing the task properly, please continue as before.")

DETECTION In this subsection, we outline two mechanisms to detect non-cooperative behaviour: First, analysing a participant's median reaction time and second, setting up booby trap items.

To the best of our knowledge, most judgement studies do not use any mechanisms to detect non-cooperative behaviour. There are some studies that exclude participants based on their reaction times, but this was determined by a participant's mean reaction time ($\bar{\mu}$ in the formula below). To do this, one would take the average of all participants' mean reaction times ($ave(\bar{\mu}_{1-N})$) and exclude participants whose mean reaction times differ by e.g. one and a half standard deviations (σ) from the average of all participants' mean reaction times. θ_{mean} defines the lower and upper thresholds for this criterion.

Subject Type	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Mean RT	Median RT
Cooperative 1	4520 ms	2370 ms	3810 ms	3900 ms	4180 ms	6130 ms	4152 ms	4040 ms
Cooperative 2	3760 ms	2780 ms	3200 ms	4375 ms	4250 ms	5320 ms	3948 ms	4005 ms
Cooperative 3	3530 ms	1840 ms	3080 ms	3160 ms	3270 ms	5320 ms	3367 ms	3215 ms
...	...	...	...	...	...	...	...	...
Cooperative 10	4520 ms	2370 ms	3810 ms	3900 ms	4180 ms	6130 ms	4152 ms	4040 ms
"Simple" Spammer	360 ms	730 ms	260 ms	510 ms	390 ms	1180 ms	572 ms	450 ms
"Clever" Spammer	540 ms	660 ms	490 ms	320 ms	24790 ms	510 ms	4552 ms	525 ms
						Average: (Stdev):	3746 (1060)	3378 (1406)

Table 6.3: An illustration of why a mean-based exclusion criterion is not sufficient to detect "clever" spammers. "Clever" spammers are aware that researchers can see overall times of completion on AMT. Thus, they make their overall time look OK, typically by waiting on one or two items (here Item 5). A better way to detect spammers is to calculate median reaction times. This way, both the "simple" and the "clever" spammers are detected easily. While these data are fabricated, they are based on real data from our twelve sessions (cf. Table 6.1).

$$\theta_{mean} = ave(\bar{\mu}_{1-N}) \pm 1.5\sigma \tag{6.1}$$

$$\text{where } ave(\bar{\mu}) = \frac{\bar{\mu}_1 + \bar{\mu}_2 + ... + \bar{\mu}_N}{N} \text{ and } \sigma = stdev(\bar{\mu}_{1-N})$$

In our view, this mean based approach is an inadequate measure, because it would only detect the "simple" spammers (i.e. those who rush through the entire questionnaire), but not the "clever" ones (i.e. those who wait on one or two items). To illustrate this point, consider the reaction times in Table 6.3. The data in Table 6.3 are fabricated, but they are representative of real data from the twelve experimental sessions mentioned in Table 6.1.

If one excludes participants by their mean, one would only exclude "simple" spam-

mers, but not "clever" ones. But this is not good enough: Of the 26 non-cooperative participants that we labelled as spammers, only eleven are "simple" spammers. That is, fifteen out of 26 spammers (58%) were clever enough to make their overall time look normal. However, with a median based approach ($\mu_{1/2}$), they can be detected, too. The formulas for the lower and upper thresholds (θ_{lower} and θ_{upper}, respectively) are:

$$\theta_{lower} = \overline{\mu_{1/2_{1-N}}} - 1.5\sigma \tag{6.2}$$

$$\theta_{upper} = \overline{\mu_{1/2_{1-N}}} + 4\sigma \tag{6.3}$$

$$\text{where } \overline{\mu_{1/2_{1-N}}} = \frac{\mu_{1/2_1} + \mu_{1/2_2} + \ldots + \mu_{1/2_N}}{N} \text{ and } \sigma = stdev(\mu_{1/2_{1-N}})$$

There is another point that needs to be addressed. Those studies that do exclude participants based on their reaction times typically set symmetric thresholds (e.g. the mean of all mean reaction times ±1.5 standard deviations). Often, this works for the lower threshold. However, we think that it is overly strict on the upper end. For instance, for our study in Chapter 1, a symmetric criterion (of ±1.5 standard deviations) would have excluded participants with a median reaction time of about six seconds. Six seconds per sentence is not overly slow and in our view, there are no reasons to exclude a participant with such reaction times. This is why we set our lower criterion at median reaction time −1.5 standard deviations (about two seconds for our experimental stimuli), but our upper criterion at median reaction time +4 standard deviations (about ten seconds).

The second strategy is to include what we call "booby trap items" (in the literature, they are sometimes referred to as "screening questions" or "filtering items"). Booby

	Chapter 1	Chapter 2	Chapter 3
Prevention AMT Filtering	yes	yes	yes
Discouragement On-line Warning	yes	no	yes
Detection Median-based Reaction Times	yes (concerned 4/143 subjects)	yes (15/142 subjects)	yes (6/137 subjects)
Detection Booby Trap Items	yes (5/143 subjects)	no	partly* (5/137 subjects)

Table 6.4: An overview of the different techniques of preventing non-cooperative behaviour, discouraging it, and detecting it, as discussed above. (*In Chapter 3, we used the initial calibration items as booby trap items.)

trap items are stimuli that are clear cut cases (i.e. the researcher expects that these items should receive extreme ratings) and if a participant does not get them right, then that participant is non-cooperative (probably because he/she is not paying attention). What does "get them right" mean? We set the following criterion: If, for any given participant, the average of the bad booby trap items is higher than the average of the good booby trap items, then this means that the participant was not getting it right. Also, in Chapter 1, we randomly interspersed the booby trap items in the final two thirds of the questionnaires. Putting them at the beginning at the questionnaire, like we did in Chapter 3, does not help detect participants who "doze off" mid-questionnaire.

Table 6.4 gives an overview of the different strategies that we use throughout the thesis to counter non-cooperative behaviour.

With Table 6.4 in mind, we can briefly come back to the prevalence of non-cooperative behaviour. Using the outlined techniques to prevent non-cooperative behaviour, we expect that about 11% of the participants are non-cooperative with respect to

their reaction times (cf. Chapter 2). This number is reduced to almost a third by including the on-line warning mechanism (which we employed in Chapter 1 and Chapter 3, but not in Chapter 2). Further, the booby trap items detected a few more non-cooperative participants (4%). Thus, with adequate prevention, but without discouragement, we expect about 15% of the participants on AMT to be non-cooperative (Chapter 2's 11% + Chapter 1 and Chapter 3's 4%).[4]

6.4 Concluding Remarks

Even when using strict recruitment criteria, i.e. preventing non-cooperative participants to take part in the first place, we estimate that about 15% of the participants recruited through AMT are non-cooperative. We presented a technique to discourage non-cooperative behaviour, viz. an on-line warning mechanism that tells participants when they are clearly behaving in a non-cooperative manner. As to detecting non-cooperative behaviour, we recommend analysing participants' median reaction times (instead of their mean reaction times) and including booby trap items. We base our suggestions on the syntactic acceptability judgement studies from Chapter 1, Chapter 2, and Chapter 3. However, we expect that they are also applicable to judgement studies in other fields, including in economics (e.g. consumer studies), sociology (e.g. studies on the perceived friendliness of certain stereotypes), or psychology (e.g. studies on the perceived brightness of certain stimuli).

[4]This estimate leads to another question: How come about 15% of our participants were non-cooperative, despite the fact that each participant had to have an approval rate of at least 98% and had to have at least 5000 tasks finished? We think the following explains this discrepancy. We *would have* failed to approve only the most blatant violators, which only concerns a handful of participants per experiment. But, we say "would have", because we approved all our participants, including the non-cooperative ones. This is for a simple reason: Many researchers are just as concerned about their reputation as many of their participants. If a researcher frequently turns down participants, then he or she acquires a bad reputation. Once a researcher gets to the point of having a bad reputation, participants are less inclined to take part in his/her studies. For this reason, we think, many researchers, including ourselves, hesitate to fail to approve non-cooperative participants.

References

Adger, D., Ramchand, G., 2003. Predication and equation. Linguistic Inquiry 34 (3): 325-359.

Altman, D. G., Bland, J. M., 1995. Absence of evidence is not evidence of absence. British Medical Journal 311: 485.

Ané, C. 2010. Testing Random Effects. Retrieved 16 July 2015 from: http://www.-stat.wisc.edu/∼ane/st572/notes/lec21.pdf

Asudeh, A., 2011a Local grammaticality in syntactic production. In E. M. Bender and J. E. Arnold (eds.), Language from a Cognitive Perspective, 5179. CSLI Publications.

Bader, M., Häussler, J., 2010. Toward a model of grammaticality judgements. Journal of Linguistics 46: 273-330.

Bader, M., unpublished manuscript. Processing ambiguous and unambiguous case features - does modality matter?

Bakker, M., van Dijk, A., Wicherts, J. M., 2012. The rules of the game called psy-

chological science. Perspectives on Psychological Science 7: 543-554.

Bard, E. G., Robertson, D., Sorace, A., 1996. Magnitude Estimation of linguistic acceptability. Language, 72 (1): 32-68.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G., 2014. Package "lme4" (Version 1.1-7). lme4: Linear mixed-effects models using "Eigen" and S4. http://CRAN.R-project.org/package=-lme4.

Berinsky, A. J., Huber, G. A., Lenz, G. S., 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. Political Analysis 20: 351-368.

Berman, M. G., Jonides, J., Lewis, R. L., 2009. In search of decay in verbal short-term memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 35: 317-333.

Bever, T. G., 1970. The cognitive basis for linguistic structures. In J. R. Hayes (ed.), Cognition and the Development of Language. New York, NY: John Wiley and Sons.

Biber, D., Gray, B., 2011. Is conversation more grammatically complex than academic writing? In M. Konopka, J. Kubczak, C. Mair, F. Štícha, U. H. Waner (eds.),

Grammar and Corpora 2009: 47-62. Tübingen: Narr Verlag.

Boring, E., 1953. A history of introspection. Psychological Bulletin 50: 169-189.

Bornkessel-Schlesewsky, I., Schlesewsky, M., 2007. The wolf in sheep's clothing: against a new judgment-driven imperialism. Theoretical Linguistics 33: 319-333.

Borsley, R. D., 2005. Introduction. Lingua 115 (11): 1475-1480.

Bowers, H., 1973. Grammatical relations. MIT: Ph.D. Dissertation.

Box, G. E. P., Hunter, W. G., Hunter, J. S., 1978. Statistics for experimenters. New York, NY: John Wiley.

Brians, P., 2015. Common errors in English usage. Retrieved 20 Dec 2014 from: http://itre.cis.upenn.edu/∼myl/languagelog/archives/000122.html

Buhrmester, M., Kwang, T., Gosling, S. D., 2011. Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? Psychological Science 6 (1): 3-5.

Chafe, W., 1992. Writing vs. speech. In W. Bright (ed.), Oxford International Encyclopedia of Linguistics 4: 257-259. Oxford, UK: Oxford University Press.

Chomsky, N., 1965. Aspects of the theory of syntax. Cambridge, MA: MIT Press.

Churchland, P. M., 1985. Reductionism, qualia and the direct introspection of brain

states. Journal of Philosophy 82: 8-28.

Cieri, C., Miller, D., Walker, K., 2004. The Fisher Corpus: a resource for the next generations of speech-to-text. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004).

Clark, M., 2009. Equivalence testing. Retrieved 16 Dec 2013 from: www.unt.edu/-rss/class/mike/5700/Equivalence%20testing.ppt

Cohen, J., 1988. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum (2nd ed.).

Comrie, B., 1984. Subject and object control: syntax, semantics, pragmatics. Proceedings of the Tenth Annual Meeting of the Berkeley Linguistics Society: 450-464.

Comrie, B., 1985. Reflections on subject and object control. Journal of Semantics 4: 47-65.

Cowart, W., 1997. Experimental syntax: applying objective methods to sentence judgments. London, UK: Sage Publications.

Cox, E. P., 1980. The optimal number of response alternatives for a scale: a review. Journal of Marketing Research 17 (4): 407-422.

Crump, J. C., McDonnell, J. V., Gureckis, T. M., 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PLoS ONE, 8(3). Retrieved 14 July 2015 from: http://www.plosone.org/article/info%3Adoi%2F10-

.1371%2Fjournal.pone.0057410

Culbertson, J., Gross, S., 2009. Are linguists better subjects? British Journal for the Philosophy of Science 60: 721-736.

Culicover, P.W., Jackendoff, R., 2001. Control is not movement. Linguistic Inquiry 32 (3): 493-512.

Culicover, P.W., Jackendoff, R., 2010. Quantitative methods alone are not enough: response to Gibson and Fedorenko. Trends in Cognitive Sciences 14: 234-235.

Davies, M., 2008-present. The Corpus of Contemporary American English. http://-corpus.byu.edu/coca/.

den Dikken, M., et al., 2007. Data and grammar: means and individuals. Theoretical Linguistics 33: 335-352.

Dodge, R., 1912. The theory and limitations of introspection. American Journal of Psychology 23: 214-229.

Downs, J. S., Holbrook, M. B., Sheng, S., Cranor, L. F.. Are your participants gaming the system?: Screening Mechanical Turk Workers. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 10): 2399-2402.

Edelman, S., Christiansen, M., 2003. How seriously should we take minimalist syntax? Trends in Cognitive Sciences 7: 60-61.

Eickhoff and de Vries, 2013. Increasing cheat robustness of crowdsourcing tasks.

Information Retrieval Journal 16: 121-137.

Elliott, A.C., Woodward W.A., 2007. Statistical analysis quick reference guidebook with SPSS examples. London, UK: Sage Publications.

Erteschik-Shir, N., 1973. On the nature of island constraints. MIT: Ph.D. Dissertation.

Erteschik-Shir, N., 2005. Bridge phenomena. In M. Everaert, H. van Riemsdijk (eds.), The Blackwell Companion to Syntax (1): 284-294. Oxford, UK: Blackwell.

Fanselow, G., Frisch, S., 2006. Effects of processing difficulty on judgments of acceptability. In G. Fanselow, C. Féry, R. Vogel, M. Schlesewsky (eds.), Gradience in Grammar: 291-316. Oxford: Oxford University Press.

Fanselow, G., Häussler, J., Weskott, T., 2013. Are you certain? On confidence affecting acceptability ratings. Poster Presented at the International Conference on Linguistic Evidence, February 13-15, 2014, Tübingen, Germany.

Featherston, S., 2007. Data in generative grammar: the stick and the carrot. Theoretical Linguistics 33: 269-318.

Featherston, S., 2008. Thermometer judgments as linguistic evidence. In C. M. Riehl, A. Rothe (eds.), Was ist Linguistische Evidenz?: 69-89. Aachen: Shaker Verlag.

Featherston, S., 2009. Relax, lean back, and be a linguist. Zeitschrift für Sprach-

wissenschaft 28: 127-132.

Ferreira, F., Swets, B., 2005. The production and comprehension of resumptive pronouns in relative clause "island" contexts. In A. Cutler (ed.), Twenty-first Century Psycholinguistics: Four Cornerstones: 263-278. Mahwah, NJ: Lawrence Erlbaum Associates.

Fodor, J. D., 1998. Learning to parse? In D. Swinney (ed.), Anniversary Issue of Journal of Psycholinguistic Research 27 (2): 285-318.

Foster, J., Parker, I., 1995. Carrying out investigations in psychology: methods and statistics. London, UK: Wiley-Blackwell.

Frazier, L., 2012. Two interpretive systems for natural language? Proceedings of the Twenty-Fifth Annual CUNY Conference on Human Sentence Processing.

Ghiselli, E. E., 1939. All or none versus graded response questionnaires. Journal of Applied Psychology 23: 405-415.

Gibson, E., Fedorenko, E., 2010. Weak quantitative standards in linguistics research. Trends in Cognitive Sciences 14: 233-234.

Gibson, E., Fedorenko, E., 2013. The need for quantitative methods in syntax and semantics research. Language and Cognitive Processes 28 (1-2): 88-124.

Gibson, E., Piantadosi, S., Fedorenko, E., 2013. Quantitative methods in syntax / semantics research: a response to Sprouse & Almeida (in press). Language and

Cognitive Processes 28 (3): 229-240.

Gibson, E., Piantadosi, S., Fedorenko, K., 2011. Using Mechanical Turk to obtain and analyze English acceptability judgments. Language and Linguistics Compass 5: 509-524.

Gonick, L., Smith, W., 1993. The cartoon guide to statistics. New York, NY: Harper Perennial.

Grewendorf, G., 2007. Empirical evidence and theoretical reasoning in generative grammar. Theoretical Linguistics 33: 369-381.

Guy, G. R., 1988. Language and social class. In. F. J. Newmeyer (ed.), Linguistics: The Cambridge Survey: Volume 4, Language: The Socio-Cultural Context: 37-63. Cambridge, UK: Cambridge University Press.

Hansen, K., Carls, U., Lucko, P., 1996. Die Differenzierung des Englischen in Nationale Varianten. Eine Einführung. Berlin: Erich Schmidt Verlag.

Haumann, D., 1997. The syntax of subordination. Berlin: Mouton de Gruyter.

Hemphill, J. F., 2003. Interpreting the magnitudes of correlation coefficients. American Psychologist 58 (1): 78-79.

Hofmeister, P., Sag, I., 2010. Cognitive constraints and island effects. Language 86 (2): 366-415.

Hong, J. I., Li, F. C., Lin, J., Landay, J. A., 2001. End-user perceptions of formal

and informal representations of web sites. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 01): 385-386.

Jackson, F., 1986: What Mary didn't know. Journal of Philosophy 83: 291-295.

Jackson, S., 2011. Research methods and statistics: a critical thinking approach. Wadsworth, UK: Cengage.

James, W., 1890. The principles of psychology. New York, NY: Holt.

Juzek, T. S., unpublished manuscript. Measuring acceptability: comparing conventional and alternative normalisation methods.

Juzek, T. S., unpublished manuscript. Textual and auditory stimuli in acceptability judgement tasks.

Kazai, G., Kamps, J., Milic-Frayling, N., 2011. Worker types and personality traits in crowdsourcing relevance labels. Proceedings of Twentieth International Conference on Information and Knowledge Management (ACM CIKM): 1941-1944.

Keller, F., 2000. Gradience in grammar: experimental and computational aspects of degrees of grammaticality. University of Edinburgh: Ph.D. Dissertation.

Kelman, H. C., 1967. Human use of human subjects - the problem of deception in social psychological experiments. Psychological Bulletin 67 (1): 1-11.

Kilpatrick, F. P., Cantril, H., 1960. Self-anchoring scaling: a measure of individuals'

unique reality worlds. Washington DC: The Brookings Institution.

Kittur, A., Chi, E. H., Suh, B., 2008. Crowdsourcing user studies with Mechanical Turk. Proceeding of the Twenty-Sixth Annual Sigchi Conference on Human Factors in Computing Systems: 453-456. New York: ACM.

Kluender, R., 1992. Deriving islands constraints from principles of predication. In H. Goodluck, M. Rochemont (eds.), Island Constraints: Theory, Acquisition and Processing: 223-58. Dordrecht: Kluwer.

Koch, P., Oesterreicher, W., 1985. Sprache der Nähe - Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. Romanistisches Jahrbuch 36: 15-43.

Kochanski, G. P., 2005. Brute force as a statistical tool. Retrieved 15 July 2015 from: http://kochanski.org/gpk/teaching/0401Oxford/bfi.pdf

Korn, J. H., 1997. Illusions of reality: a history of deception in social psychology. Albany, NY: State University of New York Press.

Kövecses, Z., 2000. American English - an introduction. Peterborough, Canada: Broadview Press.

Leech, L., 2000. Grammars of spoken English: new outcomes of corpus-oriented research. Language Learning 50: 675-724.

Legate, J. A., 2010. On how how is used instead of that. Natural Language and

Linguistic Theory 28: 121-134.

Likert, R., 1932. A technique for the measurement of attitudes. Archives of Psychology 140: 1-55.

Liu, X. S., 2014. Statistical power analysis for the social and behavioral sciences. New York, NY: Routledge.

Lyons, W., 1986. The disappearance of introspection. Cambridge, MA: MIT Press.

Mahowald, K., Graff, P., Hartman, J., Gibson, E., submitted. SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments.

Martin, R., 2001. Null case and the distribution of PRO. Linguistic Inquiry 32 (1): 141-166.

Mason and Suri, 2012. Conducting behavioral research on Amazon's Mechanical Turk. Behavioral Research Methods 44: 1-23.

Matthews, P. H., 2007. Oxford concise dictionary of linguistics. Oxford, UK: Oxford University Press (2nd ed.).

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., Reed, G. M., 2001. Psychological testing and psychological assessment: a review of evidence and issues. American Psychologist 56

(2): 128-165.

Meyerhoff, M., 2006. Introducing sociolinguistics. London: Routledge.

Miller, J., Weinert, R., 1998. Spontaneous spoken language. Syntax and Discourse. Oxford, UK: Clarendon Press.

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., Schnoebelen, T., Tily, H., 2010. Crowdsourcing and language studies: the new generation of linguistic data. Proceedings of Workshop At NAACL 2010 Creating Speech and Text Language Data with Amazon's Mechanical Turk.

Myers, J. L., 1972. Fundamentals of experimental design. Boston, MA: Allyn and Bacon.

Nagel, T., 1974: What is it like to be a bat? Philosophical Review 83: 435-450.

Nugent, W. R., 2004. A validity study of scores from self-anchored-type scales for measuring depression and self-esteem. Research on Social Work Practice 14 (3): 171-179.

O'Donnell, R. C., 1974. Syntactic differences between speech and writing. American Speech 49: 102-110.

Paolacci, G., Chandler, J., Ipeirotis, P. G., 2010. Running experiments on Amazon Mechanical Turk. Judgment and Decision Making 5: 411-419.

Phillips, C., 2010. Should we impeach armchair linguists? In A. Iwasaki, H. Hoji,

P. Clancy, S.-O. Sohn (eds.), Japanese-Korean Linguistics: 49-64. Stanford, CA: CSLI Publications.

Phillips, C., Lasnik, H., 2003. Linguistics and empirical evidence: reply to Edelman and Christiansen. Trends in Cognitive Sciences 7: 61-62.

Pollard, C., Sag, I. A., 1994. Head-Driven Phrase Structure Grammar. Chicago, IL: University of Chicago Press.

Postal, M., Pullum, G. K., 1978. Traces and the description of English complementizer contraction. Linguistic Inquiry 9 (1): 1-29.

Poulton, E. C., 1989. Bias in quantifying judgements. Hove, UK: Erlbaum.

Prince, E. F., 1990. Syntax and discourse: a look at resumptive pronouns. Proceedings of the 16th Annual Meeting of the Berkeley Linguistics Society: 482-497.

Pullum, G. K., 2003. Corpus fetishism. Retrieved 27 Apr 2009 from: http://itre.cis.upenn.edu/∼myl/languagelog/archives/000122.html

R Core Team, 2015. R: A language and environment for statistical computing. Retrieved 31 May 2015 from: http://www.R-project.org

Richter, S. J., Richter, C., 2002. A method for determining equivalence in industrial applications. Quality Engineering 14 (3): 375-380.

Riddiough, R., Thomas, D., 1998. Statistics for higher mathematics. Surrey, UK:

Thomas Nelson & Sons Ltd.

Rosenbaum, P., 1967. The grammar of English predicate complement constructions. Cambridge, MA: MIT Press.

Ross, J., 1967. Constraints on variables in syntax. MIT: Ph.D. Dissertation.

Schütze, C. T., 1996. The empirical base of linguistics: grammaticality judgments and linguistic methodology. Chicago, IL: University of Chicago Press.

Schuirmann, D. J., 1981. On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. Biometrics 37: 617.

Simmons, J.P., Nelson L.D., Simonsohn U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science 22: 1359-1366.

Snow, R., O'Connor, B., Jurafsky, D., Ng, A. Y., 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. Proceedings of the Conference on Empirical Methods in Natural Language Processing: 254-263.

Sorace, A., Keller, F., 2005. Gradience in linguistic data. Lingua 115: 1497-1524.

Sprouse, J., 2009. Magnitude Estimation and the non-linearity of acceptability judgments. Proceedings of the West Coast Conference on Formal Linguistics 27.

Sprouse, J., 2011. A validation of Amazon Mechanical Turk for the collection of ac-

ceptability judgments in linguistic theory. Behavior Research Methods 43: 155-167.

Sprouse, J., 2013. Experimental syntax in 2013: triumphs and challenges. Workshop: Understanding Acceptability Judgments at the University of Potsdam.

Sprouse, J., Almeida, D., 2012. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. Journal of Linguistics 48: 609-652.

Sprouse, J., Almeida, D., 2013. The role of experimental syntax in an integrated cognitive science of language. In K. Grohmann, C. Boeckx (eds.), The Cambridge Handbook of Biolinguistics: 181-202. Cambridge, UK: Cambridge University Press.

Sprouse, J., Schütze, C. T., Almeida, D., 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. Lingua 134: 219-248.

Stevens, S. S., 1946. On the theory of scales of measurement. Science 103: 667-688.

Szabolcsi, A., 2006. Strong and weak islands. In M. Everaert, H. van Riemsdijk (eds.), The Blackwell Companion to Syntax (4): 479-532. Oxford: Blackwell.

Takahashi, S., 2008. The manga guide to statistics. San Francisco, CA: No Starch Press.

Tannen, D. (ed.), 1982. Spoken and written language: exploring orality and literacy. Norwood, NJ: Ablex.

Titchener, E. B., 1912. Prolegomena to a study of introspection. American Journal

of Psychology 23: 427-448.

Vandergriff, I., 2005. Weil der schmeckt so gut! The Learner as Linguist. Die Unterrichtspraxis Teaching German 38 (1): 61-73.

Walker, J., Almond, P., 2010. Interpreting statistical findings. Maidenhead, UK: Open University Press.

Wallendorf, M., Brucks, M., 1993. Introspection in consumer research: implementation and implications. Journal of Consumer Research 20 (3): 339-359.

Wardhaugh, R., Fuller, J. M., 1986. An introduction to sociolinguistics. Oxford, UK: Blackwell.

Wasow, T., Arnold, J., 2005. Intuitions in linguistic argumentation. Lingua 115 (11): 1481-1496.

Wasow, T., Levy, R., Melnick, R., Zhu., H., Juzek, T., in press. Processing, prosody, and optional to. In L. Frazier, E. Gibson (eds.), Papers from the Workshop on Prosody and Processing, Amherst 2013.

Weijters, B., Cabooter, E., Schillewaert, N., 2010. The effect of rating scale format on response styles: the number of response categories and response category labels. International Journal of Research in Marketing 27: 236-247.

Wellek, S., 2003. Testing statistical hypotheses of equivalence. Boca Raton, FL:

CRC Press.

Weskott, T., Fanselow, G., 2008. Variance and informativity in different measures of linguistic acceptability. Proceedings of the West Coast Conference on Formal Linguistics 27 (WCCFL 27): 431-439.

Weskott, T., Fanselow, G., 2009. Scaling issues in the measurement of linguistic acceptability. In S. Featherston, S. Winkler (eds.), The Fruits of Empirical Linguistics (1): 231-245. Berlin: Mouton de Gruyter.

Weskott, T., Fanselow, G., 2011. On the informativity of different measures of linguistic acceptability. Language, 87 (2): 249-273.

Westlake, W. J., 1976. Symmetric confidence intervals for bioequivalence trials. Biometrics 32: 741-744.

Winer, B. J., Brown, D. R., Michels, K. M., 1971. Statistical principles in experimental design. New York, NY: Mcgraw-Hill (2nd ed.).

Winter, B., 2013. Linear models and linear mixed effects models in r with linguistic applications. Retrieved 24 Apr 2015 from: http://arxiv.org/pdf/1308.5499.pdf

Wolf, C., 1985. Status. In A. Kuper, J. Kuper (eds., The Social Science Encyclopedia: 842-843. London: Routledge and Kegan.