

Using comparison judgments to study representations

Steven Langsford
The University of Adelaide

October 31, 2017

Abstract

Three projects are presented, all using comparison data to investigate representations. Processes of comparison are the focus here because of the strong links they create between the abstract representations much of cognition aims to study and unambiguous choice outcomes. The superficial similarities between these projects, that they all use browser based studies to reach relatively large numbers of people, and apply quantitative models to summarize and interpret the results, derive from two things: a common set of concerns with representation structure, and the use of comparison tasks to contrive situations where different representations predict different task behaviors. These basic ideas are applied across different domains to address current questions of representation and measurement in similarity and language.

The first section compares two prominent theories of similarity judgment, transformational similarity and structural alignment, across three studies. The first of these constructs triad stimuli such that the two approaches make opposite predictions, the second measures similarity using an alternative measure of same-different discrimination speed, and a third applies both tasks to a common set of stimuli to clearly resolve their similarities and differences. The results show evidence of a misspecification in the APPLY rule of the transformational account current for geometric shapes, and also show that while same-different discrimination and deliberative comparison measures of similarity judgment are largely consistent, there are differences which appear to arise due to the different time constraints of the two tasks.

The second section investigates a paradigm for testing the impact of transformation learning on similarity and categorization judgments. In this paradigm, a common set of test items follows two different training conditions, such that no test item is present in any training, and the status of each test item as a match, near match, or non-match to the training varies by condition. Responses to identical test items are compared across training conditions to expose the impact of transformation training on similarity and categorization judgment. Across multiple iterations of this basic design I show that the transformations are learned, and that transformation learning does impact similarity and categorization judgment. Change in similarity and categorization ratings due to training are largest in the easiest training conditions where transformations are presented explicitly to participants during training, and less pronounced when transformations are presented implicitly. Some generalization of learning is shown across related transformations, suggesting some similarity structure among transformations.

The third section moves into empirical studies of syntax, comparing different ways of measuring sentence acceptability, the degree to which a sentence appears well-formed to a speaker of that language. This is related to the similarity work in the first and second sections through its use of Thurstonian modeling for structure discovery, which is capable of inferring acceptability scores for each sentence while also avoiding the need to present a rating scale of any kind to participants. This study complements existing work on the Type 1 and Type 2 error rates of the most common measurement techniques with its investigation of within and between participant test-retest reliability. The Likert task is found to be particularly effective. The results presented here show it has particularly good reliability properties and help empirically validate the common practice of interpreting averaged Likert ratings as a fine-grained measure of gradient acceptability.

Dedication

To the Mechanical Turk community. I know the ‘short, fun psychology study’ HITs weren’t always short and fun. But they were all studies: here are the results.

Almost all Mechanical Turk studies implicitly assume that the Mechanical Turk community is reasonably representative of humanity. Working with the Turker community has convinced me that humanity is very smart, very impatient, almost never malicious, and overwhelmingly willing give a good faith effort in doing a task so long as the researcher has made a good faith effort to make the task clear. There’s no particular reason these things had to be true facts about humanity. But they are true, and as a result, all the work presented here was possible. Thank you all.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed:

18/8/2017

Acknowledgements

Science is about people, and the ones at the Adelaide computational cognitive science lab have been great people to try and do science with.

My supervisors for this work, Dani Navarro, Amy Perfors, and Andrew Hendrickson, can't be thanked enough. If this project was a car, they kept an eye on the nuts and bolts to stop the wheels falling off, patiently taught me to drive as I bunny-hopped around thrashing the gears, and gave me a road map with a handy arrow on it marked "You are here." There's no better student experience than that.

Contents

I	General Introduction	8
1	Comparison and representation	9
II	Similarity representation	12
2	Contrasting accounts of similarity	13
2.1	Studying similarity	13
2.1.1	Two approaches to structure	15
2.1.2	When two approaches is one too many	16
3	Transformation or alignment?	18
3.1	Introduction	21
3.1.1	Similarity as structural alignment	22
3.1.2	Similarity as transformation	23
3.2	Experiment 1	23
3.2.1	Method	24
3.2.2	Results	27
3.2.3	Discussion	27
3.3	Experiment 2	28
3.3.1	Method	29
3.3.2	Results	29
3.3.3	Discussion	30
3.4	Experiment 3	31
3.4.1	Method	32
3.4.2	Results	35
3.4.3	Discussion	38
3.5	General Discussion	39
4	Which transformations?	42
5	Transformation learning	45
5.1	Introduction	48
5.2	Experiment 1	48
5.2.1	Method	49
5.2.2	Results	51
5.2.3	Conclusion	52
5.3	Experiment 2	52
5.3.1	Method	52
5.3.2	Results	53
5.4	General Discussion	54
III	Linguistic representation	55
6	From similarity to sentences	56
6.1	The structure of acceptability	56

6.2	Testing the Thurstonian model	58
6.2.1	Simulation tests	58
6.2.2	Interpreting distances on the scale	59
6.3	Reliability studies for sentence acceptability measures	62
7	The reliability of acceptability	65
7.1	Introduction	68
7.1.1	The measures	70
7.1.2	Measure evaluation	71
7.2	Method	72
7.2.1	Sentences	72
7.2.2	Measures	72
7.2.3	General procedure	75
7.2.4	Task-specific procedures	76
7.3	Results	78
7.3.1	Global measures	79
7.3.2	Decision measures	83
7.3.3	Agreement between measures	87
7.4	Summary and Conclusions	88
IV	Conclusions	92
8	Summary and conclusions	93
	References	98
V	Appendices	107
A	Sentence stimuli	108
A.1	Attention check questions	108
A.2	Instruction quiz	108
A.3	Sentence stimuli	108
A.4	Conflict sentences	113
B	Model implementation	114

“Comparison is the death of joy” : Mark Twain

“We can gain intense pleasure only from the contrast”: Freud

“What’s the difference between a zippo and a hippo?”

“One of them is a little lighter.” : Masai Graham.

Part I

General Introduction

Chapter 1

Comparison and representation

This thesis describes three different projects that all use comparison judgment data to try and shed light on the kinds of representations people use. Part II describes a comparison between two prominent theories of similarity judgment, which I tested with ABX triads and same-different comparisons. The results of this study suggested new questions, particularly surrounding the role of transformations in defining categories and similarity, which I tested with another series of comparison judgments, described in Part II Chapter 5. Part III describes a project looking at the psychometric properties of several common sentence acceptability judgment tasks. Although the final results were presented as a methodology paper for experimental linguistics, its original point of departure was a representation question about the structure of sentence acceptability, which motivated a novel adaptation of a well-known procedure in psychophysics based on comparisons, Thurstonian modeling.

So why all the comparisons?

Comparisons are one particularly convenient way to test otherwise unobservable mental representations with simple unambiguous outcomes like clicking a button or tapping a key. Historically, psychology's shift to emphasizing the value of data — as opposed to introspection — came with an explicit refusal to consider complex mental states, not necessarily because they were thought to be false, but because they were unfalsifiable (Watson, 1913; Skinner, 1938). Psychology owes its status as a 'real science' to these early empiricists, but the commitment of the new data-driven psychology to simplicity was challenged by data sets that clearly supported the conclusion that complex mental representations **must** be involved. The most famous of these was in the domain of language, where Chomsky (1959) painstakingly demolished the simple input-output mapping description of language put forward in Skinner's *Verbal Behavior* (Skinner, 1958) in favor of more complex structure: a grammar. What made this such a departure from the prevailing behaviorist philosophy of the time was the fact that these mental structures were not directly observable. However, relatively straightforward facts about language, for instance the apparent need for recursion (Hauser, Chomsky, & Fitch, 2002), has put Occam's razor firmly on the side of the structured representations. Having said that, it's interesting to note that current proponents of the generative grammars first proposed in this early anti-behaviorist salvo are still engaged in an ongoing debate over the status of data in this line of work (Featherston, 2007; Gibson & Fedorenko, 2013; Gibson, Piantadosi, & Fedorenko, 2011), an issue the final section of this thesis engages with directly.

So what was it about language that made it possible to convincingly lift the lid on the behaviorist black box? Is there a way to do the same thing for other cognitive abilities, such as categorization or similarity judgment? Any task that engages the putative representation is a candidate, but not all tasks are equally useful in this regard. As well as engaging the relevant representations in a meaningful way, a useful task should not restrict or direct responses such that the task structure masks representational structure, and must also create situations where different ways of representing the stimuli imply clearly distinct outcomes in responses.

Comparisons have the key properties required. First, comparisons between things necessarily involve some commitment to a particular representation: there is no such thing as a generic comparison, only a comparison *in some respect*, whether this is weight, height, color, likelihood of causing major bodily harm, or some other basis of comparison (Goodman, 1972; Watanabe, 1969; Medin, Goldstone, & Gentner, 1993). Comparisons typically also result in unambiguous decisions. Once people have concluded that something is heavier, taller, redder, or more dangerous, this

conclusion is readily converted unambiguous outcomes such as clicking a button, positioning a slider, or running away. By linking unambiguous observable outcomes with mental representation, comparisons are one way of supplying evidence relevant to the structure of those representation in property induction, categorization, or grammatical acceptability.

Chapter 3 describes a project where comparisons were used to contrast two different approaches to structured similarity: that is, ways of calculating similarity scores that leverage structured representations. People were asked to judge which of two possible alternatives was more similar to a reference item, with the different accounts of similarity making different predictions about which option would be chosen. Follow-up work used another type of comparison, speeded same-difference judgments, to test the same two approaches in a different but related task. On the basis of these results, I identify a particular mis-specification in the implementation of transformational similarity current in the literature for geometric shapes, and discuss how different instantiations of the general framework might accommodate them. This project also showed differences between the deliberative comparison and speeded same-different tasks, adding to existing accounts describing similarity judgment as a process that unfolds over time (Lovett, Gentner, Forbus, & Sagi, 2009; Sagi, Gentner, & Lovett, 2012).

Chapter 5 describes a project that used comparisons to train people on new transformations. In one experiment, people are shown a reference item and asked which of two alternatives belonged to the same category; in another, people are shown two items and asked if they were from the same category or not. All test questions are also in the form of comparisons, asking people how similar two test items were or how likely they were to belong to the same category. The core of this study is the way it manipulates the relationship between the training and the test items: using two different transformation training conditions, test item identity is controlled while manipulating test item status as related or unrelated to training. Although transformation learning is quite general, and relatively little is known about the status of distinctive transformations as features, the results of this work are most relevant to the transformational account of similarity discussed in Chapter 3. The transformational account predicts particularly large changes in similarity when learning a new transformation, and also requires fast transformation learning in new domains. The results regarding these two properties are somewhat inconclusive: people do show changes in their similarity and categorization judgments based on training experience, but are also highly sensitive to task difficulty. One interesting outcome of this work is evidence that in some conditions people generalize learning across related transformations, suggesting the existence of family resemblances between transformations.

Part III moves to a new application domain, sentence acceptability rather than similarity, but applies of many of the same ideas. In it, I take the well established Thurstonian modeling approach from psychophysics, which is based on comparison data and has deep connections with geometric accounts of similarity spaces (Ennis & Johnson, 1993), and adapt it to a linguistics question about the structure of sentence acceptability. The Thurstonian method is particularly well suited to the question of how sentence acceptability is structured, for example whether it is clustered or smoothly varying, because of the way the comparison data it is based on is agnostic as to structure. In this it is unlike popular scale based methods such as Likert scales. A second contribution of this work is the contrast it draws between within and between participant reliability, which allows conclusions to be drawn about the relative contribution of different sources of variability to the overall reliability of sentence acceptability scores. Together, these are complementary lines of investigation that give a detailed quantitative picture of the bias and variance inherent to the different measurement tasks.

Perhaps surprisingly, one striking result of this work was the observation that averages of z -transformed Likert scale ratings give essentially identical acceptability scores to the Thurstonian estimates in this context. This suggests that the standard warnings against over-interpreting Likert scale data based on the ordinal nature of the scale may be overly conservative for the particular case of sentence acceptability, since when a scale-free comparison based method is used, the same results are obtained. Although there is reassuringly high agreement between the acceptability scores assigned to sentences under each method, they are not equally efficient or reliable. A quantitative investigation into the causes of this variation in efficiency and reliability forms the main content of Chapter 7.

It would be fair to say that this is a diverse set of studies. They're related to somewhat different literatures within cognitive science, although they do share a set of common concerns to do with representation structure. Methodologically, the first and third sections can be considered opposites: in the investigation into similarity in Chapter 3, I take the models as given and search

for diagnostic stimuli, whereas when looking at sentence acceptability in Chapter 7, I treat the stimuli as given and try to build a model that can summarize comparisons over these stimuli in an informative way.

However the projects are also related to each other. Their superficial similarities as browser-based studies of people's behavior in contrived experimental scenarios all derive from the heavy use of comparisons. There are other ways to achieve the same thing, but here it is the use of comparisons that gives rise to the key task properties enabling these studies to leverage the power of browser-based studies: comparison tasks engage relatively high level representations not dependent on the fine details of the physical display, produce unambiguous behavioral response outcomes, and can be used to construct situations where different responses distinguish between different representations.

Part II

Similarity representation

Chapter 2

Contrasting accounts of similarity

This section aims to motivate why similarity is worth studying at all, and gives a brief sketch of the main approaches, situating the particular debate that this work aims to contribute to. The motivating logic of the study presented here is outlined, followed by a detailed manuscript presentation of the work.

2.1 Studying similarity

Similarity is a fundamental construct in cognition, forming a critical component in accounts of categorization (Nosofsky, 1986), memory (Chater & Brown, 2008; Forbus, Gentner, & Law, 1995; Hitzman, 1986), property induction, (Smith, Shafir, & Osherson, 1993), problem-solving (Riesbeck & Schank, 2013), and language (Sassoon, 2011).

One simple reason for this broad popularity is that similarity, conceived of as an arbitrary operational construct, is *useful* in accounting for people’s behavior. For example, in category learning, people remain sensitive to the number of shared features between test items and training exemplars even when those shared features are irrelevant or misleading (Brumby & Hahn, 2017; Hahn, Prat-Sala, Pothos, & Brumby, 2010), suggesting ubiquitous and automatic similarity processing. Meanwhile, in the otherwise unrelated field of natural language processing, an analogous process of automatic reference to similarity information is one way of accounting for the way people introduce knowledge about the world when parsing natural language structures that are thought to be hard to distinguish on the basis of linguistic experience alone (Milajevs & Griffiths, 2016).

However observing that ‘similarity’ is a useful term in a diverse range of settings begs the question of *why* this concept is useful, and how best to define it.

One possible grounding point for similarity starts with property induction, which admits a normative treatment (Russell, 1986; Tenenbaum & Griffiths, 2001) formally describing the essential survival ability of being able to make coherent decisions in the face of a changing world that can never be experienced in exactly the same way twice. Cognitive agents need to know things like which temperatures are reasonable for hatching eggs, or what color range indicates a fruit is ripe, without necessarily seeing eggs hatch or fail to hatch across all possible temperatures or tasting fruit of every possible color. Intuitively, it seems obvious that apples similar in color will be similar in ripeness, without ever being perfectly identical in either property: the appeal of normative property-induction accounts of similarity is that they show formally *why* it might be natural to consider this intuitive and obvious. The critical features of the world that lead to a similarity gradient in this sense are just that the property of interest exists in a well behaved consequential region of psychological space, and the agent is trying to infer the extent of the consequential region using one or more samples from the space.

There is some scope for variation in the definition of what constitutes a ‘well behaved’ region: for Shepard (1987), who derived some of the earliest and best known results in this area for the case of a single data point, a well behaved consequential region was continuous, finite, symmetrical around its center, and convex. This account also provisionally assumed that example points were uniformly sampled across the space, and that the space itself was metric. Later work has relaxed some of these assumptions, broadening the possible ways examples could be sampled (Navarro, Dry, & Lee, 2012), allowing for the possibility of complex (possibly disjoint) consequential regions (Navarro, 2006), and re-deriving essentially the same conclusions for data sets consisting of multiple

points in various non-metric spaces (Tenenbaum & Griffiths, 2001; Chater & Vitányi, 2003).

The ability to infer consequential regions of practical equivalence is broadly applicable, and can be interpreted as behind many of the applications cited above to show the popularity of similarity as a construct. For example, applied to the ‘group membership’ property, it becomes an account of categorization, and underpins several successful approaches, notably ALCOVE (Kruschke, 1992), SUSTAIN (Love, Medin, & Gureckis, 2004) and the GCM (Nosofsky, 1986). Applied to the ‘identity’ property it becomes an account of similarity-based recognition memory (Davis, Xue, Love, Preston, & Poldrack, 2014), and so on for the many properties at various levels of abstraction that people might find consequential.

Successful as it has been, normative property induction gives an incomplete account of similarity in at least two major ways (Boroditsky & Ramscar, 2001). One is that property induction is ill-suited for calculating a level of similarity between stimuli from their observed features, where the need for induction does not obviously apply. The other is its fundamental dependence on the structure of the psychological space, which could be interpreted as usurping all the explanatory power, especially if it varies substantially from context to context.

Property induction makes a poor basis for similarity when comparing known common and distinctive features of two things because of the danger of falling into a circular definition. It is clearly vacuous to say that things are similar when they share features and likely to share features when they are similar (Boroditsky & Ramscar, 2001; Hahn & Chater, 1997). As Goodman (1972) argued, when observing rather than inferring features, the need to define of what counts as a feature does all of the explanatory work attributed to similarity. If similarity is interpreted as an estimated probability of sharing additional unobserved features, the question is *which* features add to the count? Weighing more than 75kg may be a critical feature when comparing two boxers, but not two oil tankers, then there’s also the feature of weighing more than 76kg, and 77kg, and so on up to the feature of weighing more than 100,000 tons, which *is* a relevant feature of contrast among oil tankers, but not for nebulae or ants. Even without changing the subject of comparison similarities can vary by context, for example when household items are being grouped in the context of packing for a holiday versus the context of a house fire (Barsalou, 1983).

One possible resolution to this twin challenge of underspecification and heterogeneity of similarity is explored in detail by Medin et al. (1993). This work essentially accepts the main criticism that the basis of comparison is ‘doing all the work’ but considering the basis for comparison itself as the object of study. Under this view, these two concerns are related by the central question of specifying the appropriate representation space, a necessary prerequisite for the property-induction accounts. It’s not that property induction accounts of similarity have failed to consider this issue. Quite the opposite, moving from a physical description of the stimuli to a psychological representation space was a key contribution of Shepard’s systemization of similarity. For example, the attested confusability of red and violet makes little sense in terms of wavelengths, which do not loop, but can be coherently described with an appropriately structured color wheel, which does (Shepard, 1957, 1962). Although Shepard originally worked with metric psychological spaces (Shepard, 1987), this assumption is not strictly necessary (see Tenenbaum & Griffiths, 2001; Chater & Vitányi, 2003). However, allowing such richly structured spaces only makes property-induction accounts of similarity **more** dependent on how that space is specified, and there is no universal normative guide for how such spaces should be constructed (Bellet, Habrard, & Sebban, 2013; Edelman & Shahbazi, 2012; Shahbazi, Raizada, & Edelman, 2016).

One approach to this problem of determining the representation space starts with structure in the world and asks how people might learn representations such that similarity ‘just works’ in the sense of giving well-behaved consequential regions for properties of interest. A relatively extreme version of this kind of approach was explored by C. Kemp, Bernstein, and Tenenbaum (2005), investigating a specific definition of similarity in terms of inference about generative processes. This breaks the potential circularity in inferring similarity from features and features from similarity by appealing to causal processes in the world. Other work grounding representation in the structure of the environment has shown people learning whether dimensions are integral or separable (Austerweil & Griffiths, 2010a), and weighting dimensions according to their importance in a category learning task (Goldstone & Steyvers, 2001), both critical properties for calculating similarity distances.

Chapter 5 of this thesis uses an approach somewhat related to these, exploring a possible relationship between learning and similarity judgment for the case of transformation features. Chapter 3 focuses on the other main avenue of research exploring the structure of the representation

space: evaluating process models accounting for people’s behavior in similarity-related tasks, with representations initially set by theoretical fiat and then justified — or adapted — iteratively with reference to empirical data.

As noted by Gentner (2001), process level theories specifically address the issue of defining a hypothesis space conspicuously absent from arguments based on normative property induction, although the models are not always described in these terms (see eg. Soto, Gershman, & Niv, 2014). The work presented in Chapter 3 is relevant to two such process models, transformational similarity and structural alignment, with a particular focus on the core issue of the representations they use.

The most studied approaches to representation in similarity can be broadly grouped into geometric, feature-based, alignment-based, and transformational accounts (Goldstone, Day, & Son, 2010; Hahn, 2014). Geometric accounts of similarity are often associated with the Shepard’s work on multidimensional scaling (MDS) (Shepard, 1957). Developed alongside the property induction systemization of similarity discussed above, MDS maps similarity judgments to a coherent similarity space. It quickly found applications in a range of structure-discovery problems across phonetics, color, word meaning, memory, and visual discrimination, among others (Shepard, 1974). The technique has proven extremely fruitful and is still in use (see eg. Nosofsky, Sanders, Gerdman, Douglas, & McDaniel, 2017). However, Tversky (1977) pointed out two features of the approach that put it at odds with known properties of human similarity judgment: first, that stimuli are adequately represented as points in some coordinate space, and second, that the similarity spaces were metric. Having shown that the metric properties of symmetry, maximum similarity at identity, and the triangle inequality were violated by human similarity judgment, Tversky (1977) proposed the Contrast Model as an alternative. The Contrast Model, based on feature-set overlap, captured these human-like violations of the metric axioms.

Both of these highly successful ‘classical’ approaches can be considered unstructured theories of similarity, since they do not account for relationships between features, whether these are represented as dimensions of a coordinate space or members of a feature set. People, on the other hand, do seem sensitive to such relationships. For example, a black square and a yellow triangle seems more similar to a black square and a blue circle than a black circle and a blue square, despite the fact that the features ‘black’ and ‘square’ appear in both alternative options: the fact that black and square are bound together is highly salient (Goldstone et al., 2010). Preserving this intuitive effect by allowing conjunctive features such as ‘black-square’ leads to a combinatorial explosion in the number of features (or dimensions) these unstructured models need to consider. The non-independence of features (Goldstone, Medin, & Gentner, 1991) suggests human similarity judgments require structured representations (Biederman, 1985, 1987; Wattenmaker, Nakamura, & Medin, 1988; Likavec & Cena, 2015; Markman, 1999; Yuille & Kersten, 2006).

The work presented here aims to help differentiate two particular approaches to structured representations in similarity judgment, structural alignment and transformational similarity. Although not the only possibilities (see eg. Pothos & Busemeyer, 2011), these two approaches to similarity judgment with structured representations have been particularly prominent. Despite being based on fundamentally different theoretical conceptions of structure, the two approaches have previously proven hard to distinguish due to their highly similar predictions in common test domains.

2.1.1 Two approaches to structure

Structural alignment approaches to similarity (Markman & Gentner, 1993b) build on structure mapping (Gentner, 1983), originally a theory describing how features and their relations to each other could be aligned in analogy. Taking as a starting point prepositional representations and a basic set of desiderata for analogy processing, notably *structural consistency* (requiring matching relations to have matching arguments, and enforcing 1-1 mappings), *relational focus* (privileging relationships over more superficial attributes), and *systematicity* (preferring a connected series of matches over an equal number of disconnected matches), structure mapping provided a model capable of capturing human-like inferences about analogies such as “an electric battery is like a reservoir” or “an atom is like the solar system” (Falkenhainer, Forbus, & Gentner, 1989). The approach has since been extended from analogy to other related domains, including similarity (Gentner, Rattermann, Markman, & Kotovsky, 1995; Gentner & Markman, 1997; Markman & Gentner, 1993a; Taylor & Hummel, 2009), where the ‘goodness of fit’ of an alignment between the

representations may be interpreted as a measure of their similarity. This core idea of structural matching is implemented in a variety of ways by a family of related models, notably SIAM, CAB, and LISA, (Goldstone, 1994; Larkey & Love, 2003; Hummel & Holyoak, 2005) although this list is by no means exhaustive (Guan, Wang, & Wang, 2008).

In contrast, transformational similarity (Hahn, Chater, & Richardson, 2003) takes Kolmogorov complexity as its underlying theoretical basis (Ming & Vitányi, 1997), quantifying the similarity between two representations as the length of the shortest program capable of transforming one into the other. The information theoretic basis of this approach leads to a number of attractive properties, notably linking similarity with a normative need for simplicity assumptions in pattern discovery (Chater, 1999), and allowing a re-derivation of Shepard’s celebrated law of generalization (Shepard, 1987) for arbitrary representations (Chater & Vitányi, 2003). Since finding the shortest program for a given transformation is in general uncomputable (Ming & Vitányi, 1997), attempts to apply this theoretical framework to models of human behavior have focused on specific implementations which fix a set of allowable transformations and compute similarities in terms of the number of basic operations in the shortest route from one representation to the other (Imai, 1977; Chater & Hahn, 1997; Beltran, Liu, Mohanchandra, & Toussaint, 2015; Gershman & Tenenbaum, 2015).

Despite the very different origins of these approaches, they have been surprisingly difficult to differentiate. As detailed below, both enjoy a measure of empirical support, and make many of the same predictions.

2.1.2 When two approaches is one too many

One early effort to compare these different approaches to similarity judgment was carried out by Larkey and Markman (2005). This study contrasted transformational similarity and three different flavors of structure mapping (SME, SIAM, and CAB) on stimuli consisting of pairs of geometric shapes. A single item consisted of two shapes, and each trial consisted of two items, which were rated by participants for similarity on a 1-6 scale. In this study, Larkey and Markman (2005) found that SIAM outperformed the other approaches.

It’s possible, however, that this particular test used a misspecified instantiation of the transformational approach. In particular, the predictions attributed to transformational similarity in Larkey and Markman (2005) were based on the premise that each unique physical relationship between stimulus components would be addressed with a single unique transformation. In response, Hodgetts, Hahn, and Chater (2009b) proposed a particular set of transformations appropriate to geometric shapes –CREATE, APPLY, and SWAP– emphasizing that transformations should be applied at the level of representations rather than physical objects. This instantiation of the transformational approach fit experimental data well. Hodgetts et al. (2009b) showed that models based on a differential weighting of matches in place (MIPs) and matches out of place (MOPs), the basic units of structural alignment approaches, could not match the performance of this particular transformational account on their data.

In the same year, Hahn, Close, and Graf (2009) showed that the transformational account naturally predicted comparison-direction asymmetries in similarity judgment. Asymmetries arise in the transformational approach whenever the number of basic transformations needed to transform A into B is not the same as the number of transformations needed to turn B into A. The transformations used in this case were relatively complex shape morphs, for example a flat shoe gradually changing into a high heel. A ‘preferred’ or ‘more available’ transformation direction was induced by giving participants training experience with the transformation in a particular direction. The observation that participants did indeed give higher similarity ratings to objects related by a transformation in the trained direction is somewhat problematic for alignment accounts. Morph direction was counterbalanced across participants, meaning that either direction of comparison could be considered more similar depending on training experience. Under an alignment account, much like the Contrast Model, similarity asymmetries are most naturally accounted for by systematic differences in the representation of the base and target items: asymmetries are expected when the base representation has a richer or more systematically organized representation, which maximizes the amount of information that can be mapped from base to target (Gentner & Markman, 1997; Medin et al., 1993). It is unclear how experience with a morphing transformation might change the representation of the base item in the required way, inducing a richer or more systematic representation of the base.

Although able to capture these asymmetry effects in a simple and coherent way, the transformational approach faced strong criticisms on other grounds by Grimm, Rein, and Markman (2012). The model comparisons presented in Hodgetts et al. (2009b) were questioned on the grounds of the hidden degrees of freedom involved in specifying an appropriate transformation set, and serious challenges were raised with regard to the kinds of transformations that could be considered feasible. In particular, in a sequence of three related studies with geometric shape stimuli, Grimm et al. (2012) showed the existence of context effects, where differences between stimuli thought to correspond to the same transformation had a larger or smaller impact on similarity depending on whether the transformations were configuration-preserving or configuration-breaking (with reference to Gestalt principles of perceptual organization (Garner & Clement, 1963)). Context sensitivity (and the need for different weightings on different transformations highlighted in their experiment 3) would seriously complicate the calculation of transformation distances. Even without these factors, the computation involved in calculating transformation distances can be non-trivial (Müller, van Rooij, & Wareham, 2009). With them, Grimm et al. (2012) argue that the approach is infeasible.

Hodgetts and Hahn (2012) demonstrated more asymmetry effects consistent with the transformational account of similarity. Given the publication dates, it would be unfair to criticize this work for not directly engaging with the challenges raised by Grimm et al. (2012). Instead, it extended the reach of the transformational similarity account: it used same-different discrimination time as a fine-grained implicit measure of similarity, and found evidence of an asymmetry effect naturally predicted by the transformational account but potentially hard to capture with other approaches. Although it does not engage with the question of context effects, this work does implicitly reply to the charge of hidden degrees of freedom when choosing the transformation set by re-applying the original transformation set from Hodgetts et al. (2009b) without adaptations to new data in a quite different task.

The motivation for the work presented in Chapter 3 was this unresolved status of the ongoing debate between transformational and alignment approaches to similarity. Both approaches enjoyed a measure of empirical support (Beltran et al., 2015; Gershman & Tenenbaum, 2015; Forbus, Ferguson, Lovett, & Gentner, 2016) and both could claim to have been endorsed in direct model comparisons. Transformational similarity's success in direct comparison was slightly more recent, following the introduction of a representation-level transformation set apparently well calibrated to similarity judgments on pairs of geometric shapes, although there seemed to be no published answer to the strong arguments raised by Grimm et al. (2012) regarding how that set was selected, except indirectly via the successful application of it to a new task in Hodgetts and Hahn (2012). The work in Chapter 3 aimed to contribute to this debate by contriving an experimental task which created a direct conflict between the predictions of the two approaches.

To find stimuli where different theories of representation made different predictions, I turned to comparisons. Initially, I constructed a set of triads where the two approaches made different predictions about which of two options was most similar to a reference item. Follow-up work replicated the same-different task used in Hodgetts and Hahn (2012), but using stimuli aimed at providing diagnostic discrimination between the two approaches (rather than testing for asymmetry effects, as in the original). The two different tasks gave seemingly conflicting results. I therefore carried out a third experiment designed to disentangle this apparent contradiction by elements from both previous experiments. The results of this third experiment suggested a particular misspecification in the transformation set current for geometric shapes, and critical differences between the tasks due to their different time profiles.

Chapter 3

Transformation or alignment?

This section reproduces a manuscript currently under review, *Are mental representations aligned or transformed? A comparison between two accounts of similarity-based choice* with authors Steven Langsford, Daniel J Navarro, Amy Perfors, and Andrew Hendrickson.

Statement of Authorship

Title of Paper	Are mental representations aligned or transformed? A comparison between two accounts of similarity-based choice
Publication Status	<input type="checkbox"/> Submitted for Publication
Publication Details	Submitted to Journal of Experimental Psychology, Learning Memory and Cognition June 2017

Principal Author

Name of Principal Author (Candidate)	Steven Langsford		
Contribution to the Paper	Significant contribution to articulation of hypothesis and experimental design, including search for appropriate stimuli. All study implementation and data collection. Carried out analyses, with direction. Initial drafting of associated write-up, plus contribution to subsequent revisions.		
Overall percentage (%)	75%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	18/8/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Amy Perfors		
Contribution to the Paper	I helped design the study, discussed how to analyse the results, and edited the paper.		
Signature		Date	31/08/2017

Name of Co-Author	Daniel Navarro		
-------------------	----------------	--	--

Contribution to the Paper	I helped design the study (especially experiment 3), discussed how to analyse the results, and edited the paper.		
Signature		Date	31/08/2017

Name of Co-Author	Andrew Hendrickson		
Contribution to the Paper	I helped design the study, discussed how to analyse the results, and edited the paper.		
Signature		Date	31/8/2017

Current theories of stimulus similarity propose that people possess structured representations of stimuli, and that comparisons between items are made by processes defined over these representations. Two of the most prominent of these theories are the structural alignment view and the stimulus transformation view, which have proven difficult to discriminate empirically. In this chapter I present three experiments using the simple geometric stimuli commonly used to evaluate these theories, selecting stimuli for which the two competing theories make qualitatively different predictions. Experiment 1 presents evidence from a forced choice task in which participants need to decide which of two alternatives is more similar to a reference item, and appears to find very strong support for the alignment approach. Experiment 2 presents evidence from reaction times in a same-different judgment task, and finds moderate support for the transformation approach. A third experiment aims to reconcile these findings by presenting both tasks to the same participants with a common set of stimuli. I find that (a) the tasks do indeed produce slightly incompatible measures of similarity due to the different demands of a discrimination task and a similarity comparison and (b) the “strong” results in Experiment 1 arise due to a slight misspecification in how the standard version of the transformational model is usually implemented, and requires a modification to the “apply” operation to be consistent with empirical data.

3.1 Introduction

Similarity is an important theoretical construct in cognitive science, and plays a central role in models of categorization (Nosofsky, 1986), memory (Baddeley, 1966; Shulman, 1971), reasoning (Riesbeck & Schank, 2013), problem solving (Novick, 1988), and others. Despite its importance as an explanatory principle, similarity is notoriously difficult to define. It cannot be defined on purely logical grounds (e.g., Goodman, 1972) and the empirical literature makes clear that – among other things – similarity is a flexible quantity that depends on context (e.g., Watanabe, 1985; Barsalou, 1983) and reflects a decision process that unfolds over time (e.g., Goldstone & Medin, 1994; Hendrickson, Navarro, & Donkin, 2015).

As a consequence, there are several theories that seek to explain how similarity is constructed from more primitive mental representations (see Goldstone et al., 2010, for a review). Early work developed models describing similarity as distances in psychological space (Shepard, 1962) or as a function of the number of common and distinctive features (Tversky, 1977). In many situations these simple models work reasonably well (Borg & Groenen, 2005; Tversky & Gati, 1982; Eidenberger & Breiteneder, 2003), but there is now a substantial literature highlighting the systematic ways in which they fail to capture human intuitions about similarity (e.g., Wattenmaker et al., 1988; Fodor, 1975). As a consequence, recent theories have sought to describe similarity in terms of more complex processes defined over structured mental representations. The two most prominent approaches are the *structural alignment* view (Markman & Gentner, 1993b, 1993a; Gentner & Markman, 1997) and the *stimulus transformation* view (Chater & Hahn, 1997; Hahn et al., 2003; Hahn, 2014), and in recent years a number of papers have sought to discriminate between them (Larkey & Markman, 2005; Grimm et al., 2012; Hodgetts et al., 2009b; Hodgetts, Hahn, & Chater, 2009a), with somewhat mixed results.

The work presented here constructs diagnostic tests by focusing on comparisons in which formal models based on the transformational theory make qualitatively different predictions from those made by models based on structural alignment. I conducted three experiments involving two tasks, one a two alternative forced choice task and the other a speeded same-different task. In the forced choice task (Experiment 1), people’s preferences closely matched the predictions of the alignment model, yet a similar approach in a speeded same-different choice task (Experiment 2) suggested a slight advantage for the transformational approach. An attempt to resolve this contradiction by presenting both tasks with a common set of stimuli (Experiment 3) suggests that the apparently-conclusive results from Experiment 1 arise because my “diagnostic” items all exploit a single point of failure in how the transformational *model* (Hodgetts et al., 2009b) instantiates a transformational *theory* of similarity (specifically, how the APPLY operation works). The chapter concludes with a discussion of how different tasks appear to measure slightly different notions of

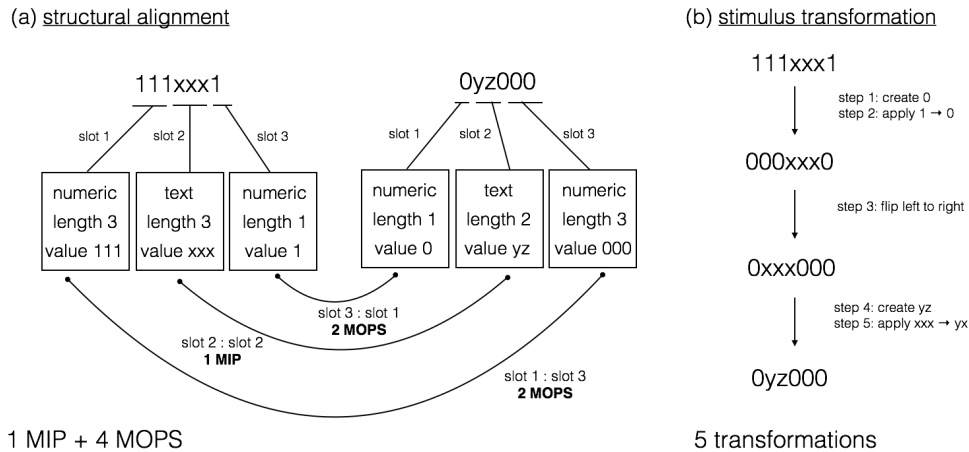


Figure 3.1: Two different mechanisms by which people might judge the similarity of structured stimuli such as 111xxx1 and 0yz000. The structural alignment view (panel a) assumes that features can appear in specific slots, and assesses similarity in terms of features that have the same values in the same slots (MIPs) and features that have the same values but appear in different slots (MOPs). The transformational view (panel b) assumes that the dissimilarity between two items is a function of the number of distinct operations required to transform one stimulus into the other.

similarity, and comment on the constraints that this data set impose on theoretical models of similarity comparison.

3.1.1 Similarity as structural alignment

The structural alignment approach to similarity takes its inspiration from theories of analogy (Gentner, 1983; Gentner & Markman, 1997). In these theories a stimulus is not merely represented in terms of a set of features: features are bound to specific “slots” or “roles”. As is the case for simpler feature-matching models, the similarity between two stimuli is driven by correspondences between their features. However, unlike feature-matching models, the contribution of a particular shared feature is dependent on the context in which they appear. If the shared features both serve the same role (or appear in the same slot) they will typically make a larger contribution to similarity. This kind of feature match is referred to as a *match in place* (MIP). In contrast, when a shared feature appears in different slots, it constitutes a *match out of place* (MOP) and typically makes a much smaller contribution to stimulus similarity. Illustrating this idea, Figure 3.1a shows how a very simple structural alignment model might assess the similarity between the strings 111xxx1 and 0yz000. A natural way to represent these stimuli is to break them into three slots, yielding representation that looks like [111] [xxx] [1] and [0] [yz] [000]. The contents of each slot could be characterized in terms of a set of features (e.g., *numeric*, *length 3*, etc.), and a natural way of aligning these structures might be to map slot 1 of the first item onto slot 3 of the second one. Because 111 and 000 are both *numeric* and *length 3*, this mapping yields two matches, and since these matches occur in the context of mapping two different slots (i.e., slot 1 to slot 3), this correspondence yields two MOPs. In contrast, if slot 2 of the first item is mapped onto slot 2 of the second item, the one feature shared by xxx and yz yields one MIP.

The structure mapping process shown in Figure 3.1 highlights the fact that there are a number of substantive choices that need to be made when developing a concrete model. In the first instance, the structure needs to be defined. A representation in which 111xxx1 is divided into three slots [111] [xxx] [1] can yield different results to one in which each character is a separate slot (i.e., [1] [1] [1] [x] [x] [x] [1]). Similarly, the mapping process needs to be specified. In Figure 3.1, each slot from the first stimulus is mapped to exactly one slot in the second stimulus, producing a strictly consistent map. However it is entirely possible that similarity judgments are less than strict, and might allow (for instance) the 111 in slot 1 of 111xxx1 to be partially matched (in place) to the 0 in slot 1 of 0yz000 in addition to partially matching (out of place) the 000 in slot 3. Because of these issues, there are several different ways in which the structural alignment view has been instantiated as a formal model of similarity, including SIAM (Goldstone, 1994) and CAB (Larkey & Love, 2003). A variety of models are compared side-by-side in Larkey and Markman

(2005), but here it will suffice to consider a very simple approach that counts the number of MIPs and MOPs, insofar as most of these models make similar predictions for the stimuli considered here.

3.1.2 Similarity as transformation

The transformational approach (Hahn et al., 2003; Hahn, 2014) defines the dissimilarity between stimuli in terms of the minimum number of operations that are required to transform one stimulus into the other (Chater & Hahn, 1997). The more steps required to mentally “transform” one object into another, the less similar those two objects are. For example, in order to convert 111xxx1 into 000xxx0, all the 1s need to be replaced with 0s. In contrast, turning 111xxx1 into 000yz0 would be a more complicated operation, since that would also require converting the xxx substring into a yz string. In its most general form, transformation distance can be characterized as a measure of information distance (Bennett, Gács, Li, Vitányi, & Zurek, 1998), and has a number of desirable properties. For example, it admits an alternative derivation of Shepard’s much studied universal law of generalization (Shepard, 1987), but it applies over arbitrarily structured representations rather than a Euclidean psychological space (Chater & Vitányi, 2003).

As with the structural alignment view, the transformational perspective requires the researcher to make substantive choices in order to produce a specific similarity model. In particular, the modeler must make choices as to what counts as a single “transformation”. Earlier, when the example 111xxx1 was turned into 000xxx0 by replacing all the 1s with 0s, how many operations are involved? If each replacement is a single operation, then a total of four operations are involved, but if several replacements can be applied as a single action then only one is required. Similarly, does the transformational process get to introduce a new feature (i.e., the 0s) for free, or does it require an operation to do so?

A number of versions of the transformational approach have been considered in the literature and applied to different domains. Early work considered transformations that are most applicable to binary strings (Imai, 1977), but other models have been defined and tested for simple geometric shapes (Hodgetts et al., 2009b), complex real-world visual objects (Hahn et al., 2009), sounds (Beltran et al., 2015), and spoken words (Hahn & Bailey, 2005). Since the stimuli will consist of pairs of geometric shapes, I relied on the transformation model described by Hodgetts et al. (2009b) (see also Cheries, Newman, Santos, & Scholl, 2006; Káldy & Leslie, 2003; Hodgetts & Hahn, 2012). According to this model, there are three primitive operations that are relevant to these stimuli: SWAP, CREATE, and APPLY. These operations are defined in the following way. A SWAP operation will reverse the positions of two objects or features. When a particular feature is not already present, a CREATE operation is required to introduce it. Finally, if a feature is present, it can be APPLIED to as many locations as is necessary using only a single operation. The free application of subsequent APPLY transformations for a single feature is necessary to explain previous results in the literature (Hodgetts & Hahn, 2012) and was the basis for the design of Experiment 1.

The transformational view has been applied successfully in many domains including sounds (Beltran et al., 2015) and phrases (Gershman & Tenenbaum, 2015), but the evidence has been rather more mixed in other situations. The impact of any particular transformation on similarity has been shown to be context-dependent (Grimm et al., 2012), complicating the computation of transformation distances. It has also been argued that transformational similarity models perform more poorly than structural alignment models in domains where the two are directly comparable (Larkey & Markman, 2005), but this claim has been disputed and evidence to the contrary offered (Hodgetts et al., 2009a).

3.2 Experiment 1

Although structural alignment and stimulus transformation are very different theories, they are not simple to separate empirically (Larkey & Markman, 2005; Hodgetts et al., 2009a): models based on both theories tend to produce good quantitative fits to the data. To motivate the first experiment, consider the three stimuli shown in Figure 3.2. These stimuli are defined in terms of two slots (left and right), and the objects that fill each slot have two features (shape and color). The objects shown at the top comprise the base item x , and the two alternatives a and b are shown

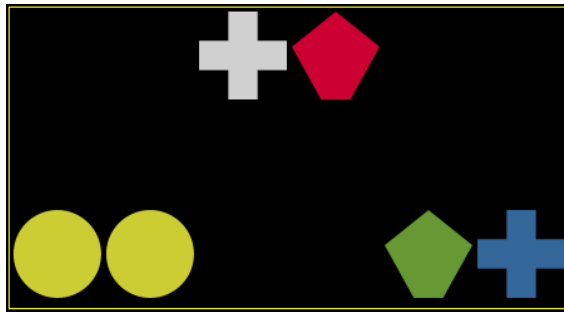


Figure 3.2: An example trial from Experiment 1, using a two alternative forced choice design. Participants were asked “Which option is most similar to the pair on top?”

below. In a forced choice task, participants are shown all three stimuli and asked to indicate whether the base item x is more similar to the option on the left (item a) or to the option on the right (item b). Let $S(a, x)$ denote how similar a is to x , and – in cases when the similarity relation is asymmetric – let $S(x, a)$ denote how similar x is to a .

What do the models predict about these items? First consider the predictions made by a structural alignment model. A MIP occurs when a particular feature (e.g., a circle, a square, the color yellow) appears in the same location. If a feature appears in different location it counts as a MOP. As illustrated in Figure 3.3, when comparing option a to the base x there are no feature matches of any kind and so these items are in no sense similar. In contrast, items b and x both contain crosses and both contain pentagons, but these occur in different positions (e.g., the cross is on the left in x and on the right in b). These stimuli share 0 MIPs and 2 MOPs. Throughout the chapter I consider two variations of this alignment model, a STRICT model that requires every feature in one stimulus to map to exactly one feature in the other, and a LOOSE model that considers every possible way of mapping features. Subtle differences in the different instantiations of the structural alignment view notwithstanding, it seems uncontroversial to claim that $S(a, x) < S(b, x)$ is a natural prediction for the stimuli in Figure 3.3, as it is for all stimuli in the experiment.

In contrast, consider the predictions of the transformational model. Assuming that the base item x is presented to participants first, and that participants assess similarity by attempting to transform x into the two alternatives a and b , it is clear from inspection of Figure 3.3b that the transformational view makes the opposite prediction to the alignment view. It takes fewer operations to transform the base item into the stimulus on the left than to the one on the right, and thus $S(b, x) < S(a, x)$. That being said – to foreshadow later experiments – it is important to note that the *direction* of the transformation is important. Figure 3.3b assumes the transformation is from base to target, and shows that $S(a, x) > S(b, x)$. However, because the operations are not symmetric, the similarities often reverse if the transformations are made from target to base, yielding $S(x, a) < S(x, b)$. Accordingly, there are two versions of the transformational model to consider depending on assumed direction (FORWARD or REVERSE). Nevertheless, as this example illustrates, it is possible to construct stimulus triads (a, b, x) that produce a qualitative reversal in which the transformational view predicts $S(a, x) > S(b, x)$ and the structural alignment view predicts $S(a, x) < S(b, x)$. The experiments presented here are based on stimuli that possess this property.

3.2.1 Method

Participants

50 participants were recruited via Amazon Mechanical Turk. Of these 48 were included in the final analysis, with two exclusions for making two or more errors on attention-check items (see below). Ages ranged from 19 to 69 with a mean of 33.96, 58% were male. Participants were from the United States, with one from India. Participants were each paid \$1.85 US for an average of 13 minutes work, an effective hourly rate of \$8.43.

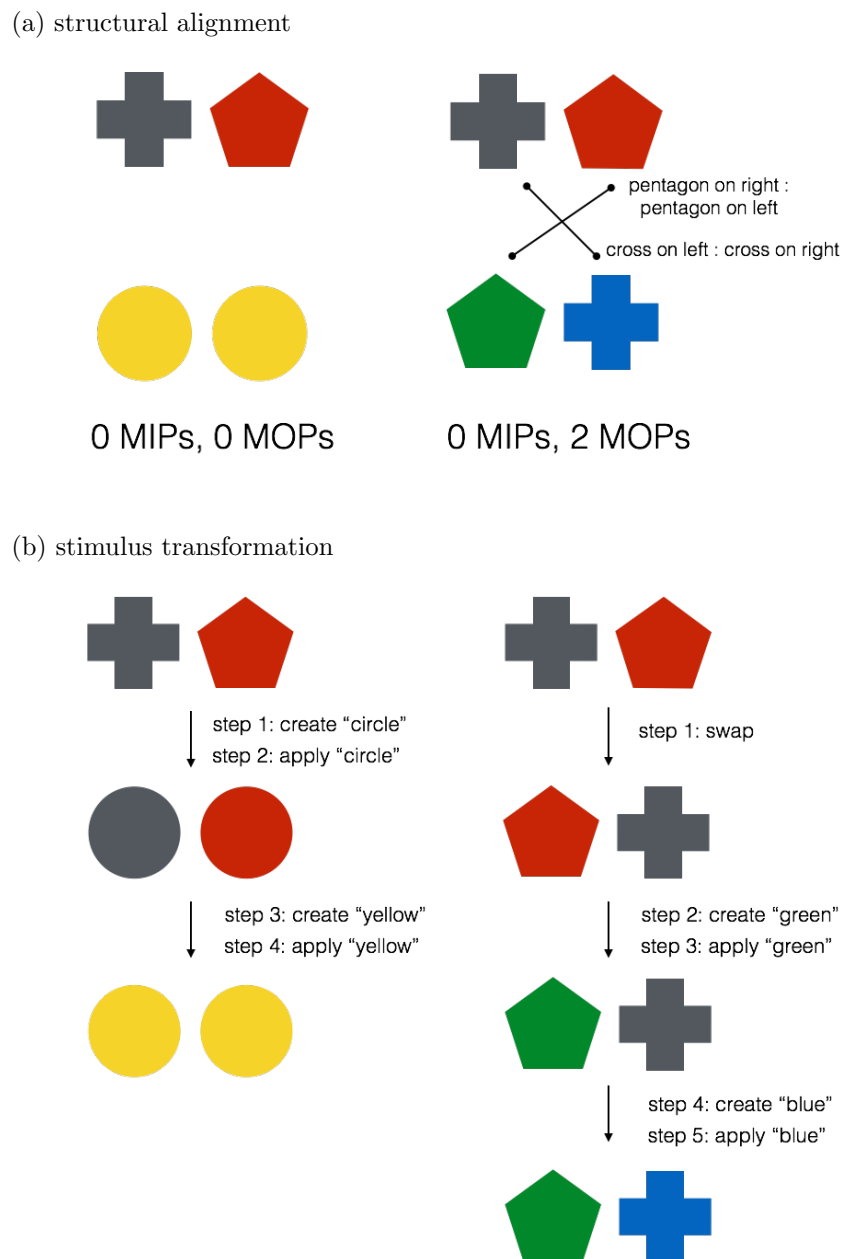


Figure 3.3: The structural alignment view predicts that the stimulus pair on the right are more similar than the pair on the left. In contrast, the transformational view predicts that the pair of items on the left are more similar than the pair shown on the right.

Table 3.1: The 11 distinct test triads (see main text for details). In all triads there is a *base* item (x) and two possible comparison options a and b . In both cases the transformational model judges the first similarity $S(a, x)$ to be higher than the second one $S(b, x)$. The structural alignment view always predicts the opposite effect. When describing the transformations below, \leftrightarrow denotes a swap operation \oplus denotes a feature create operation, and \triangleright denotes a feature application. For clarity, $\triangleright\triangleright$ denotes a “double apply” operation in which a feature is applied to two locations.

	Base	Option	Alignment (MIPS-MOPS)		#	Transformation		
			Strict	Loose		#	Reverse	
1a	AwBw	AxBx	2-0	2-0	2	$[\oplus x] [\triangleright\triangleright x]$	2	$[\oplus w] [\triangleright\triangleright w]$
1b		CwAw	2-1	2-3	3	$[\oplus C] [\triangleright C] [\leftrightarrow]$	3	$[\leftrightarrow] [\oplus B] [\triangleright B]$
2a	AwBw	CwCw	2-0	2-2	2	$[\oplus C] [\triangleright\triangleright C]$	4	$[\oplus A] [\triangleright A] [\oplus B] [\triangleright B]$
2b		BwDw	2-1	2-3	3	$[\oplus D] [\triangleright D] [\leftrightarrow]$	3	$[\leftrightarrow] [\oplus A] [\triangleright A]$
3a	AwBw	AxBx	2-0	2-0	2	$[\oplus x] [\triangleright\triangleright x]$	2	$[\oplus w] [\triangleright\triangleright w]$
3b		ByAw	1-2	1-3	3	$[\oplus y] [\triangleright y] [\leftrightarrow]$	2	$[\leftrightarrow] [\triangleright w]$
4a	AwBw	BxBx	1-0	1-1	3	$[\triangleright B] [\oplus x] [\triangleright\triangleright x]$	4	$[\oplus A] [\triangleright A] [\oplus w] [\triangleright\triangleright w]$
4b		CyBw	2-0	2-1	4	$[\oplus C] [\triangleright C] [\oplus y] [\triangleright y]$	3	$[\triangleright w] [\oplus A] [\triangleright A]$
5a	AwBw	BxBx	1-0	1-1	3	$[\triangleright B] [\oplus x] [\triangleright\triangleright x]$	4	$[\oplus A] [\triangleright A] [\oplus w] [\triangleright\triangleright w]$
5b		CwDw	2-0	2-2	4	$[\oplus C] [\triangleright C] [\oplus D] [\triangleright D]$	4	$[\oplus A] [\triangleright A] [\oplus B] [\triangleright B]$
6a	AwBw	CxBx	1-0	1-0	4	$[\oplus C] [\triangleright C] [\oplus x] [\triangleright\triangleright x]$	4	$[\oplus A] [\triangleright A] [\oplus w] [\triangleright\triangleright w]$
6b		ByDw	1-1	1-2	5	$[\oplus D] [\triangleright D] [\oplus y] [\triangleright y] [\leftrightarrow]$	4	$[\leftrightarrow] [\triangleright w] [\oplus A] [\triangleright A]$
7a	AwAx	ByAy	1-0	1-1	4	$[\oplus B] [\triangleright B] [\oplus y] [\triangleright\triangleright y]$	5	$[\triangleright A] [\oplus w] [\triangleright w] [\oplus x] [\triangleright x]$
7b		CzAw	1-1	1-2	5	$[\oplus C] [\triangleright C] [\oplus z] [\triangleright z] [\leftrightarrow]$	4	$[\leftrightarrow] [\triangleright A] [\oplus x] [\triangleright x]$
8a	AwBx	CyCy	0-0	0-0	4	$[\oplus C] [\triangleright\triangleright C] [\oplus y] [\triangleright\triangleright y]$	8	$[\oplus A] [\triangleright A] [\oplus B] [\triangleright B] [\oplus w] [\triangleright w] [\oplus x] [\triangleright x]$
8b		DxEw	0-2	0-2	5	$[\oplus D] [\triangleright D] [\oplus E] [\triangleright E] [\leftrightarrow]$	5	$[\leftrightarrow] [\oplus A] [\triangleright A] [\oplus B] [\triangleright B]$
9a	AwBw	CxCx	0-0	0-0	4	$[\oplus C] [\triangleright\triangleright C] [\oplus x] [\triangleright\triangleright x]$	6	$[\oplus A] [\triangleright A] [\oplus B] [\triangleright B] [\oplus w] [\triangleright\triangleright w]$
9b		ByDw	1-1	1-2	5	$[\oplus D] [\triangleright D] [\oplus y] [\triangleright y] [\leftrightarrow]$	5	$[\leftrightarrow] [\oplus A] [\triangleright A] [\oplus x] [\triangleright x]$
10a	AwBx	CyCy	0-0	0-0	4	$[\oplus C] [\triangleright\triangleright C] [\oplus y] [\triangleright\triangleright y]$	8	$[\oplus A] [\triangleright A] [\oplus B] [\triangleright B] [\oplus w] [\triangleright w] [\oplus x] [\triangleright x]$
10b		DwEw	1-0	1-1	5	$[\oplus D] [\triangleright D] [\oplus E] [\triangleright E] [\triangleright w]$	6	$[\oplus A] [\triangleright A] [\oplus B] [\triangleright B] [\oplus x] [\triangleright x]$
11a	AwBw	CxCx	0-0	0-0	4	$[\oplus C] [\triangleright\triangleright C] [\oplus x] [\triangleright\triangleright x]$	6	$[\oplus A] [\triangleright A] [\oplus B] [\triangleright B] [\oplus w] [\triangleright\triangleright w]$
11b		DyEw	1-0	1-1	6	$[\oplus D] [\triangleright D] [\oplus E] [\triangleright E] [\oplus y] [\triangleright y]$	5	$[\oplus A] [\triangleright A] [\oplus B] [\triangleright B] [\triangleright w]$

Materials

As illustrated in Figure 3.2, each stimulus consisted of a pair of shapes, where the shapes in question could be squares, triangles, circles, pentagons or crosses. Each shape could take on one of five possible colors: red, green, blue, yellow, and white. On any given trial participants would be shown three such stimuli, the base item x and the two alternatives a and b . An automated search was used to construct a set of 11 triads for which the two approaches make opposing predictions, with the transformational model always predicting that alternative a is more similar, and the alignment model always predicting b is more similar. The search focused on triads in which the prediction of the alignment model was identical regardless of if matches were STRICT and each feature could only match one feature in the other item or LOOSE and a feature could match other features. Thus the relative weighting of MIPs and MOPs in the alignment model was irrelevant as long as both contributed positively to similarity.

The 11 triads are listed in Table 3.1: the logical representation of a particular stimulus is written Cx-Dy, where the Cx part indicates that the object on the left has shape C and color x, and Dy indicates that the object on the right has shape D and color y. The assignment of actual feature values (e.g., square, circle) to logical values (e.g., C, D) was randomized, as was the left-right position on the screen of the a and b options. Test items were augmented with 4 attention check items in which one of the response options was identical to the base. Ninety trials were presented to each participant, with direct repetition avoided by randomly assigning feature values to logical values on each presentation, and by counterbalancing the roles played by shape and color dimensions in each pattern.¹

Procedure

The experiment was conducted online and delivered through the browser. After reading the experimental instructions (and passing a short quiz checking that the participants had understood the task), participants proceeded to the sequence of 90 trials, presented in a random order. On each trial, the base stimulus was displayed at the top of the screen. Initially *only* the base item was shown in order to ensure that participants focused on that item first, and only after the participant pressed the space bar were the two alternatives displayed. The text on screen asked participants to assess “Which option is most similar to the pair on top?” Responses were given using the keyboard: participants used the letter Q to select the left option and the letter P to indicate the right one. The display is illustrated in Figure 3.2. At the end of each trial all stimuli were removed, and a screen was shown containing a ‘next’ button, and an indication of how many trials remained.

3.2.2 Results

All participants chose the alignment-consistent option more often than the transformation-consistent option, with the proportion of such choices ranging from 68% to 97% across subjects. Moreover, as Figure 3.4 illustrates, a similar pattern emerges when the data are broken down by item: across items, the proportion of alignment-consistent choices ranged from 58% to 98%. In all but one case the corresponding 95% confidence interval excludes 50%.

3.2.3 Discussion

The results from Experiment 1 are apparently unambiguous, with the alignment approach performing much better. However, one concern with these results from the transformational perspective is that the direction of comparison matters. In constructing the triads, I assumed that participants would start with the base item x and transform it into one or both of the comparison items a or b . However, if participants ran the transformations in the other direction, the predictions made by the transformational model sometimes (though not always) reverse. Loosely inspired by the approach taken by Hodgetts and Hahn (2012), I attempted to control this by ensuring that the base item was displayed first in each trial, thereby encouraging them to attend to and process the base item first. Nevertheless, given that methodology here does not precisely mirror theirs, there

¹The minimum number of repetitions for each pattern was 6, but patterns 2 and 8 were presented 12 times due to a coding error which mistakenly identified equivalent forms as distinct.

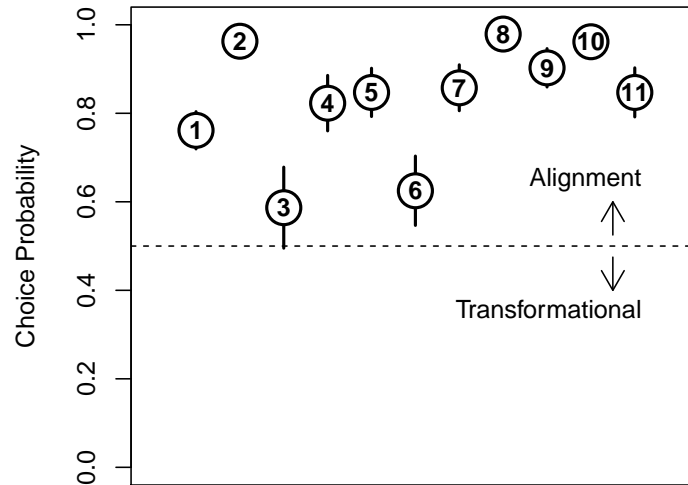


Figure 3.4: Choice proportions for all 11 stimulus triad conditions in Experiment 1. Bars represent 95% credible intervals around the proportion of “alignment-consistent” choices for each participant. Overall, the results appear to strongly favor the structural alignment model.

is some uncertainty about whether the direction of transformation was suitably constrained. This concern is directly addressed in Experiment 2.

3.3 Experiment 2

The asymmetric nature of the transformational similarity model was discussed and explored by Hodgetts and Hahn (2012), who employed a speeded same-different judgment task, and the dependent measure was the response time (RT) taken to make correct decisions. The more time required to correctly discriminate between two stimuli, the more similar they are assumed to be. The key design feature was stimulus presentation asynchrony: in their task Hodgetts and Hahn (2012) presented the stimuli in succession with a short gap between them. The asymmetry of presentation enforces an order of comparison, and participants’ judgments in their experiments were consistent with the assumption that people make the comparison by taking the first-presented item and transforming it into the second one.

Since stimuli in Experiment 2 were pairs rather than triads, the constraints on the stimuli were also different, and new stimuli were selected to meet them. In particular, the requirements for uniqueness were more restrictive for pairs than triads simply because pairs involve fewer features. This extra constraint meant that pooling the individual pairwise options from the triads used in Experiment 1 would lead to an unbalanced design with trial diversity confounded with condition, since in Experiment 1 individual base-target comparison pairs were allowed to repeat among triads if the contrasting pair was different. However, by not restricting the pairs to be components of the triads more possible pairs could be considered because there was no requirement to share a base item between pairs. As in Experiment 1, a search over the space of possible stimuli was used in an attempt to find highly informative items. This search focused on finding a set of stimuli that had a high negative correlation between similarity computed based on either transformation distance or the number of alignable features. This resulted in a collection of 14 distinct pairs, covering 12 distinct intersections of alignment and transformation similarity scores (with the two collisions considered distinct because they arrived at the same alignment score using different combinations

of MIPs and MOPs).

3.3.1 Method

Participants

102 adults were recruited via Amazon Mechanical Turk. Of these 97 were included in the final analysis. Exclusions were for low accuracy in the same-different task indicating lack of attention or misunderstanding of the task (accuracy less than 80%, 3 participants), or for self-reported color-blindness (2 participants). Ages ranged from 20 to 72 with a mean of 34.39, 63% were male. Participants were from the United States, with one from India. Participants were paid \$1.85 for an average of 11 minutes work, an effective hourly rate of \$9.94.

Materials

The stimulus set consisted of the 14 logically distinct pairs in Table 3.2, which also lists the predictions of the (FORWARD) transformational model as well as the number of MIPs and MOPs according to the STRICT and LOOSE mapping rules. Additionally, the table lists the total number of possible feature matches (equivalent to the unweighted sum of MIPs and MOPs under the LOOSE mapping rule).

Each participant viewed 84 pairs, 42 ‘same’ and 42 ‘different’ trials in shuffled order. ‘Different’ trials consisted of a set of 14 distinct pairs presented three times in different configurations: a base configuration, a reversal of the features within an item assigned to the left or the right, and a reversal of the role played by the color and shape dimensions. ‘Same’ trials used the same set of initial items as the ‘different’ trials. The specific feature values assigned to roles in a pattern was randomized, with possible shape and color values the same as in Experiment 1.

Procedure

Participants were presented with a black page with a central display area delineated with a yellow border, and instruction text “[z]=different” and “[m]=same” displayed in bottom left and right corners of the screen. The size of this display area scaled to 70% of the available screen size, or until increasing the width would distort the aspect ratio of the stimuli, which was fixed. On each trial participants were shown a pair of geometric shapes, which remained on screen for 900ms. The location of the pair was variable, falling in a randomly selected location drawn from a 5x5 grid of possibilities. The screen was then cleared for 150ms, before drawing a second pair of shapes to a new location. This second pair remained visible until participants responded by pressing either Z or M to indicate if the second pair was identical to the first. In all cases, the key indicating the ‘same’ response was on the participant’s dominant side (M for the 83 right handed participants, Z for the 14 left handed). Participants were given feedback on their responses before starting the next trial. Correct responses were followed by a blank screen for 500ms, while incorrect responses caused the message “Wrong response” to be displayed for 2000ms.

Preprocessing

The fastest and slowest 5% of trials were excluded from analysis to screen for anticipatory guessing and lapses of attention, removing responses faster than 423ms or slower than 1457ms. In addition, trials where the browser recorded a loss of focus event were also dropped (8 trials from 5 different participants). Only correct responses to “different” pairs were analyzed. To mitigate the effects of individual variability in response speed, I normalized the response time within-subject by subtracting the mean RT for each participant.

3.3.2 Results

The median normalized response times are plotted in Figure 3.5 as a function of transformation distance (left) and number of feature matches (right), with each logically distinct stimulus pair plotted as a single point, and error bars corresponding to 95% bootstrapped confidence intervals. Viewed solely from the perspective of transformation distance, the result is largely in agreement

Table 3.2: Stimulus pairs used in Experiment 2, counts of the numbers of MIPs and MOPs using both strict and loose mapping rules, total number of possible feature matches (counted in the loose sense that does not enforce 1:1 mappings), and the operations required to transform the base item into the target item. As before \leftrightarrow denotes a swap operation \oplus denotes a feature create operation, and \triangleright denotes a feature application. For clarity, $\triangleright\triangleright$ denotes a “double apply” operation in which a feature is applied to two locations. Note that in this design the number of MOPs under a loose matching rule is always equal to the sum of MIPs and MOPs under the strict matching rule; but this is not true in general.

	Base	Target	Matches	MIPs-MOPs		Transformation	
				Strict	Loose	#	Operations
1	AvBw	CwCx	1	0-1	0-1	5	$[w \leftrightarrow v]$ $[\oplus C]$ $[\triangleright \triangleright C]$ $[\oplus x]$ $[\triangleright x]$
2	AvAv	AwBv	4	2-0	2-2	4	$[\oplus B]$ $[\triangleright B]$ $[\oplus w]$ $[\triangleright w]$
3	AvAw	BxAx	2	1-0	1-1	4	$[\oplus B]$ $[\triangleright B]$ $[\oplus x]$ $[\triangleright \triangleright x]$
4	AvAv	AvBv	6	3-0	3-3	2	$[\oplus B]$ $[\triangleright B]$
5	AvBv	BwCx	1	0-1	0-1	7	$[B \leftrightarrow A]$ $[\oplus C]$ $[\triangleright C]$ $[\oplus w]$ $[\triangleright w]$ $[\oplus x]$ $[\triangleright x]$
6	AvBv	AvAw	4	2-0	2-2	3	$[\triangleright A]$ $[\oplus w]$ $[\triangleright w]$
7	AvBw	BwCx	2	0-2	0-2	5	$[Bw \leftrightarrow Av]$ $[\oplus C]$ $[\triangleright C]$ $[\oplus x]$ $[\triangleright x]$
8	AvAw	BxBv	0	0-0	0-0	6	$[\oplus B]$ $[\triangleright \triangleright B]$ $[\oplus x]$ $[\triangleright x]$ $[\oplus y]$ $[\triangleright y]$
9	AvAw	BxBx	0	0-0	0-0	4	$[\oplus B]$ $[\triangleright \triangleright B]$ $[\oplus x]$ $[\triangleright x]$
10	AvAw	BxAx	3	1-1	1-2	5	$[Av \leftrightarrow Aw]$ $[\oplus B]$ $[\triangleright B]$ $[\oplus x]$ $[\triangleright x]$
11	AvAv	BvBv	4	2-0	2-2	2	$[\oplus B]$ $[\triangleright \triangleright B]$
12	AvBw	CwCw	2	1-0	1-1	3	$[\triangleright w]$ $[\oplus C]$ $[\triangleright \triangleright C]$
13	AvBw	BxAx	2	0-2	0-2	3	$[A \leftrightarrow B]$ $[\oplus x]$ $[\triangleright \triangleright x]$
14	AvBv	AvAv	6	3-0	3-3	1	$[\triangleright A]$

with Hodgetts and Hahn (2012). The more transformations required, the shorter the RT, with the correlation of $r = -.84$ corresponding to a Bayes factor of 129:1 against the null hypothesis. However, the results are also somewhat in agreement with the predictions of a simple feature matching model, and the correlation of $r = .76$ provides a Bayes factor of 21:1 against the null. Taken together, these findings suggest modest evidence (Bayes factor of 6:1) favoring the transformational model over the feature matching model.

Turning to the alignment model, the constraints imposed by the stimulus design outlined in Table 3.2 ensure that the weighted MIPs and MOPs model makes the same predictions regardless of whether one assumes a strict matching rule or whether a loose matching rule is applied, because the number of MOPs under the loose matching rule happens to be a linear combination of the number of MIPs and MOPs under the strict matching rule. As such one needs to consider one of the two alignment models. However, even at the best fitting weighting value (in which 1 MOP = 0.70 MIP, under the strict rule) the alignment model does not fit the data any better than the simple feature matching model, correlating at $r = 0.77$. A Bayes factor analysis suggests that the alignment model is slightly dispreferred to the feature matching model (BF = 2.6) and moderately dispreferred to the transformational model (BF = 15.6).

3.3.3 Discussion

The general pattern of results in Experiment 2 can be summarized in terms of two key findings. Firstly, these results replicate the finding by Hodgetts and Hahn (2012) that reaction time to a same-difference tasks shows a strong relationship with transformation distance. In fact they also incidentally replicate the finding that asymmetric similarities can be predicted by the transformational model: Pairs 4 and 14 are logically equivalent except for the order in which the two items are presented, and there is a suggestion of a small RT difference (about 30ms) between these pairs that mirrors the prediction of the transformational model (though it should be noted that this study was not powered to detect this difference reliably and statistical tests on this comparison gave different answers depending on what assumptions were made - this evidence is at best suggestive). Secondly, while I do find clear evidence that the transformational model outperforms a simple feature matching model on these items, I did not find any clear evidence that the distinction between MIPs and MOPs made any substantial contribution to RT: that is, the alignment model “fits” the data only insofar as the feature matching account is a special case of a weighted MIPs and MOPs model.

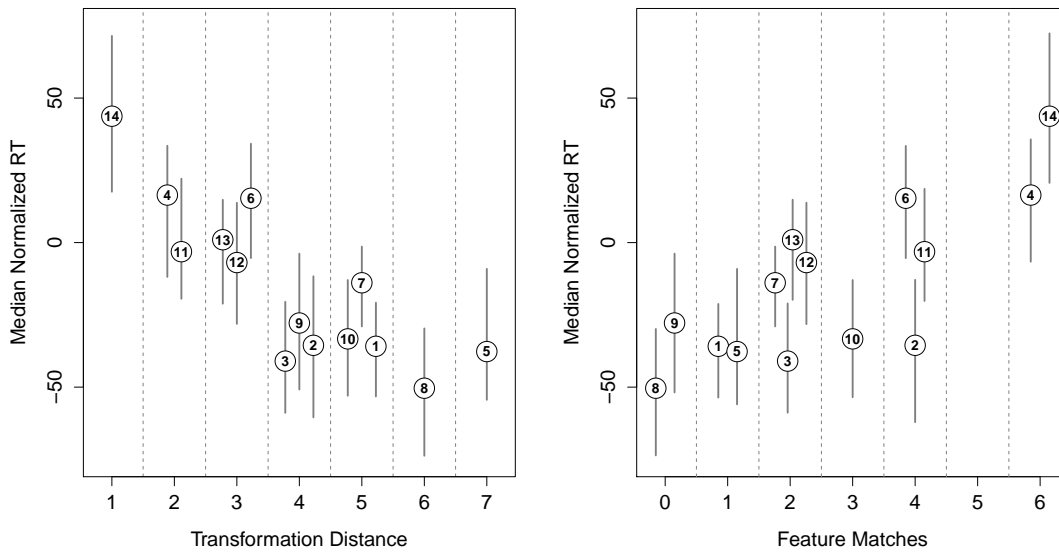


Figure 3.5: Median response times for each item plotted as a function of transformation distance (left) and total number of feature matches (right). As one might expect, both models make sensible predictions, in which stimuli predicted to be more similar produce slower responses.

3.4 Experiment 3

The results of Experiments 1 and 2 together constitute a puzzle. Experiment 1 finds very strong evidence for alignment, whereas Experiment 2 finds modest evidence favoring transformation. Why does this happen? Perhaps the transformational model failed in Experiment 1 because the direction of transformation could not be controlled? Perhaps – noting that every one of the transformation-preferred options in Experiment 1 involves the use of the APPLY transformation to apply a feature to multiple slots – there is something important about how the transformation model specifies the APPLY operation? Perhaps there is a fundamental difference in how people behave in similarity choice tasks and same-difference judgments? To test these possibilities, Experiment 3 presented the tasks from Experiment 1 and Experiment 2 to the same participants using a common set of stimuli. In order to meet the constraints imposed by the two different tasks simultaneously, the stimuli were simpler than those used in the previous experiments. They varied on only one dimension rather than two, and never involved the SWAP operation in their transformation scores. In this simplified domain it was possible to construct stimuli that induce different patterns of predictions under six similarity models relevant to the apparent conflict in Experiments 1 and 2. I included four variations of the transformational model:

- *Forward transformation, free apply.* (FT-FA) The original transformational similarity scheme, referred to as “free apply” below for the way it assigns a cost of one when applying multiple instances of a single feature. This was the model used for designing the stimuli and analysis of Experiments 1 and 2.
- *Forward transformation, costly apply.* (FT-CA) A modified transformation scheme where each application of a feature value incurs a cost of one.
- *Reverse transformation, free apply.* (RT-FA) The similarity scores under the free apply scheme when comparing in a target-to-base direction.
- *Reverse transformation, costly apply.* (RT-CA) The similarity scores under the costly apply scheme when comparing in a target-to-base direction.

By including all four of these models, it should be possible to differentiate between the “direction of transformation” explanation of Experiment 1 and the “cost to apply” explanation. Similarly I included two simple alignment based models:

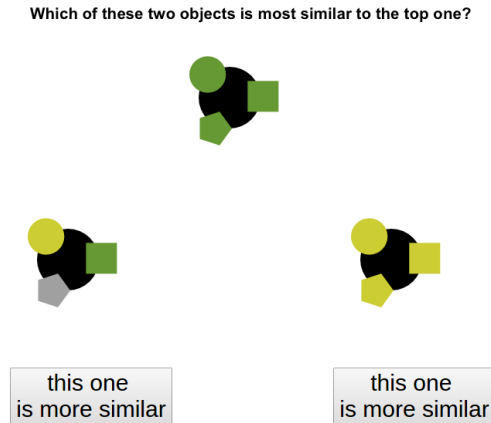


Figure 3.6: Screenshot of the triads task as presented in Experiment 3. Unlike Experiments 1 and 2, these stimuli varied on a single dimension, the colors assigned to the three embedded shapes. The same-different task (not shown) used the same stimuli.

- *Strict alignment.* (SA) A count of MIPs+MOPs where any individual feature can be involved in at most one match.
- *Loose alignment.* (LA) A count of MIPs+MOPs where features can be involved in any number of matches simultaneously.

By including these two models it can be determined whether the data place substantive constraints on alignment models, and to investigate the extent to which it remains possible to distinguish between alignment and transformation in the general cases.

3.4.1 Method

Participants

246 participants were recruited via Amazon Mechanical Turk. Of these, 213 were included in the final analysis, with two participants excluded for reporting some degree of color blindness, four for incorrect responses on attention-check triads for which one option was identical to the base, and 27 for attaining accuracy rates of less than 80% in the same-different task. Ages ranged between 18 and 72 with a mean of 36.3, and 56% of participants were male. Included participants were predominantly from the United States, with two from India and one each from Canada and Ireland. Participants were paid \$1.50 for an average of 11 minutes work, an effective hourly rate of \$7.67.

Materials

Individual items consisted of a black circle of radius 40px with three geometric shapes embedded in the perimeter, each one with an area approximately 1/3 that of the black circle. These shapes were evenly spaced around the perimeter and consisted of a circle in the upper left, a square on the extreme right, and a pentagon in the lower left (see Figure 3.6). Items differed from each other only in the colors assigned to these shapes, which could be red, green, blue, yellow, or grey. In the triad task participants were shown three stimuli, whereas in the same-different task two items were shown. The logical design of stimuli and predictions of each model are listed in Tables 3.3 and 3.4 (see below for details). Assignment of colors to roles indicated by the letter codes for each stimulus was randomized on each trial. The left/right positioning of options A and B was also randomized, and the order of trials shuffled.

Procedure

Each participant completed the triad task and the same-different task, with tasks presented in a random order. The procedure in the triad task mirrored the procedure in Experiment 1, and the procedure in the same-different task mirrored Experiment 2. For the triad task participants

Table 3.3: Predictions made by six similarity models for the triad task in Experiment 3. Letter codes represent a pattern of matching/distinct colors, with the particular color assigned to each letter randomized on each trial. The triads were constructed to ensure that models would make different predictions for a subset of items, while balancing nuisance variation in properties of the base and option items. All possible pairs of models differ on at least four triads, and for calibration purposes there are four triads for which all models make the same prediction.

	Base	A	B	SA	LA	FT-FA	FT-CA	RT-FA	RT-CA
1	aaa	aab	abb	A	A	=	A	=	A
2	aaa	abc	bbb	A	A	B	=	A	A
3	aaa	abc	bbc	A	A	=	A	A	A
4	aab	aaa	aac	=	A	A	A	=	=
5	aab	aaa	abb	=	A	=	=	B	B
6	aab	aaa	acb	=	A	A	A	B	B
7	aab	aaa	acc	A	A	A	A	A	A
8	aab	aaa	bbb	A	A	=	A	=	A
9	aab	aac	abb	=	=	B	B	B	B
10	aab	aac	acb	=	A	=	=	B	B
11	aab	aac	bbb	A	A	B	=	=	A
12	aab	aac	ccb	A	A	=	A	=	A
13	aab	abb	acc	A	A	A	A	A	A
14	aab	acc	bbb	=	B	B	B	B	=
15	aab	acc	ccb	=	A	=	=	B	=
16	abc	aaa	abd	B	A	A	=	B	B
17	abc	aba	add	A	A	A	A	A	A
18	abc	aba	ddd	A	A	A	A	A	A

Table 3.4: Model predictions about stimulus similarity for each pair in the same-different task in Experiment 3: similarity values are given for the two alignment models (SA and LA) and dissimilarity values are given for the two transformational models (FA and CA). The final column in the table denotes the relative frequency of each stimulus pair: same trials and different trials are equally likely overall, and are equally likely for all three distinct base stimulus types (i.e. **aaa**, **aab** or **abc**), ensuring that the presentation of the base item carries no information about the correct response.

	Base	Target	SA	LA	FA	CA	Freq.		Base	Target	SA	LA	FA	CA	Freq.
1	aaa	aaa	3	9	0	0	21	12	aab	bbb	1	3	1	2	3
2	aaa	aab	2	6	2	2	3	13	aab	ccb	1	1	2	3	3
3	aaa	abb	1	3	2	3	3	14	aab	ccc	0	0	2	4	3
4	aaa	bbb	0	0	2	4	3	15	abc	aaa	1	3	1	2	3
5	aaa	bcd	0	0	6	6	12	16	abc	aba	2	3	1	1	3
6	aab	aaa	2	6	1	1	3	17	abc	abc	3	3	0	0	21
7	aab	aab	3	5	0	0	24	18	abc	abd	2	2	2	2	3
8	aab	aac	2	4	2	2	3	19	abc	add	1	1	2	3	3
9	aab	abb	2	4	1	1	3	20	abc	ade	1	1	4	4	6
10	aab	acb	2	3	2	2	3	21	abc	ddd	0	0	2	4	3
11	aab	acc	1	2	2	3	3								

provided judgments for the 18 triads listed in Table 3.3 and two check trials in which one target was identical to the base item. As illustrated in Figure 3.6, people were presented with a set of three items in a triangle configuration on a white background. A reference item appeared at the top of the screen under the title “Which of these two objects is most similar to the top one?” with left and right options below it. Each of the left and right options had underneath it a button labeled “This one is more similar”. Participants responded by clicking on either of these buttons. A blank screen was displayed for 1000ms between each trial, with response buttons disabled for a further 600ms after the new triad was drawn to the screen.

For the same-different task, participants were presented with the 132 pairs from Table 3.4 in random order. In the same-different task, participants were presented with a white page with the title “Is the second object the same or different from the first one?” The center of the page was taken up by a central display area with a black border, with fixed aspect ratio but a size depending on the viewing device as in Experiment 2. The instruction text “Type z for same” and “Type m for different” was displayed in the upper left and right corners, with all left-right conventions and response mappings being reversed for left handed participants. At the beginning of each trial, an item was presented in one of the 5x5 possible locations within the display field, visible for a randomly selected duration between 500ms and 1000ms. At that point, this item was erased and a second item drawn at a new location not previously occupied by the first one. The second item remained visible until a response was made. Participants were given feedback on their responses. Correct responses progressed to the next trial after clearing the screen for 500ms, while incorrect responses caused the message “Incorrect response” to be displayed for 2000ms.

Design: Triad task

The stimuli were constructed to maximize differences between the predictions of the six similarity scoring schemes (See Table 3.3 for details). The maximum agreement between any two models over the eighteen triads was fourteen shared predictions (between SA and RT-CA), with a minimum of six (between FT-FA and RT-CA) and an average of eleven. The same-different task strictly requires that the base item not carry information about the response type, and I also attempted to make each item type roughly equally frequent. As a calibration check – to make sure that the task produces sensible results for “easy” cases – there are four triads (7, 13, 17, 18) for which all six items make the same prediction. Of the remaining 14 triads, seven produce two different predictions across all models (no-preference under some models and a preference under the others), while seven have every possible result endorsed by at least one of the similarity models (some preferring neither option, some preferring option A, and some B).

Design: Same-different task

The design of the same-different task was structured to control for several possible confounds while still allowing a meaningful comparison between the two tasks, and is outlined in Table 3.4. The number of “same” trials (pairs 1, 7 and 17) was identical to the number of “different” trials (all other pairs). The base item (i.e., the stimulus presented first) was roughly identical, with 32% of trials displaying an **aaa** item first, 36% presenting **aab** and 32% of trials displaying an **abc** base item. Critically, for all three base items, 50% of trials required a “same” response and 50% required a “different” response, ensuring that participants could not guess what response would be required until both items had appeared on screen. I also aimed to balance the number of distinct colors that appear in a trial (e.g., an **aaa:bcd** trial has 4 distinct colors), in those cases where it was possible to do so. For the “same” trials there are, 8 instances in Table 3.4 in those instances with two colors (pair 7) and 7 with three colors (pair 17). There are 6 instances with 2 distinct colors (pairs 2, 3, 4, 6, 9 and 12) and 7 instances with 3 distinct colors (pairs 8, 10, 11, 13, 14, 15, 16), ensuring that for those cases where it was logically possible for a same and different trial to be equally colorful (i.e., 2 and 3 distinct colors), the base rate of “same” and “different” responses was roughly matched. However, because it is impossible for a “different” trial to contain only a single color, and impossible for a “same” trial to have more than three, the overall colorfulness of the display did differ between same and different trials: on average a “different” trial contained 3.1 distinct colors, whereas on average a “same” trial contained only 2 distinct colors.

The same-different task is strongly directional, and so do I not consider the possibility of a “reverse” transformational model among the model predictions. This leaves only four models to

Table 3.5: Choice proportions for all conditions in the triad task in Experiment 3. In this table, the CI denotes a 95% credible interval. The right side of the table indicates the conditions for which each model fails to make the correct prediction (see Figure 3.7 for a visual depiction).

	Triad (X → A:B)	%A [CI]	A:B	SA	LA	FT-FA	RT-FA	FT-CA	RT-CA
1	aaa → aab:abb	96% [93,98]	205:8			✗	✗		
2	aaa → abc:bbb	49% [42, 56]	104:109						
3	aaa → abc:bbc	79% [73, 84]	169:44						
4	aab → aaa:aac	43% [37, 50]	92:121						
5	aab → aaa:abb	70% [63, 76]	149:64				✗		✗
6	aab → aaa:acb	84% [78, 88]	178:35	✗					✗
7	aab → aaa:acc	92% [87, 95]	195:18						
8	aab → aaa:bbb	95% [92, 97]	203:10			✗	✗		
9	aab → aac:abb	72% [66, 78]	154:59			✗	✗	✗	✗
10	aab → aac:acb	83% [77, 87]	176:37	✗		✗	✗	✗	✗
11	aab → aac:bbb	96% [92, 98]	204:9			✗	✗	✗	
12	aab → aac:ccb	95% [91, 97]	202:11			✗	✗		
13	aab → abb:acc	89% [84, 92]	189:24						
14	aab → acc:bbb	53% [46, 60]	113:100						
15	aab → acc:ccb	46% [39, 53]	98:115						
16	abc → aaa:abd	5% [3, 9]	11:202		✗	✗	✗	✗	
17	abc → aba:add	92% [88, 95]	196:17						
18	abc → aba:ddd	97% [93, 98]	206:7						
Failures: (✗)				2	1	7	8	4	4

consider: strict alignment (SA), loose alignment (LA), forward free apply (FA), and forward costly apply (CA). Within this design the two alignment based models make similar predictions to one another ($r = 0.86$), as do the two transformational models ($r = .91$), and as such the design does not easily distinguish between these possibilities. Looking at correlations between alignment and transformational models, the costly apply model is fairly closely correlated with both the strict ($r = -.87$) and loose ($r = -.82$) alignment models. However, the free apply model is quite distinct from both versions of the alignment model ($r = -.61$ in both cases), and accordingly the design has most power to discriminate between those models.

3.4.2 Results

The data from the triad task are summarized in Table 3.5: for each triad the leftmost columns list the proportion of participants choosing option A along with a 95% credible interval and counts of the absolute number of participants making each choice. On the right hand side, the table highlights the qualitative failures for all six similarity models. To determine what counts as a failure, I took a conservative approach: a prediction of option A in Table 3.3 is considered consistent with the model if the true probability of choosing option A lies between 0.5 and 1, and the data are deemed inconsistent with the model only if the entirety of the 95% credible interval falls outside that range. A model prediction of option B is evaluated in the same way, using the range 0 to 0.5, and – again, conservatively – a model that predicts indifference to the two options is only labeled as a qualitative failure if the credible interval falls outside the range 0.25 to 0.75. Using this approach, a very clear pattern is observed. Consistent with the findings from Experiment 1 both alignment models perform well: the SA model fails on 2 of the 18 triads, and the LA model fails on 1. Moreover, the transformational models with free APPLY perform poorly (FT-FA fails on 7 triads and RT-FA fails in 8 cases), again mirroring the findings from Experiment 1. However, when switching to a costly APPLY model, the performance of the transformational models improves considerably (FT-CA and RT-CA both have 4 failures), though not quite to the level of performance that the alignment-based models produce. This pattern of results is depicted visually in Figure 3.7 – on the triad task the alignment models retain some advantage over transformation models, but the size of the advantage is greatly attenuated when the costly APPLY operation is used.

For the same-different task, I preprocessed the data using the same procedure as in Experiment 2: the fastest and slowest 5% of response times were excluded from analysis to screen for anticipatory guessing and lapses of attention, removing responses faster than 463ms or slower than 1584ms. To mitigate the effects of individual variability in response speed, response times were normalized by subtracting the participant’s overall mean response time from each trial. Over all participants, mean accuracy was .88 and median accuracy .91 (before any exclusion criteria were applied). Only correct responses to different trials were analyzed. The pattern of results

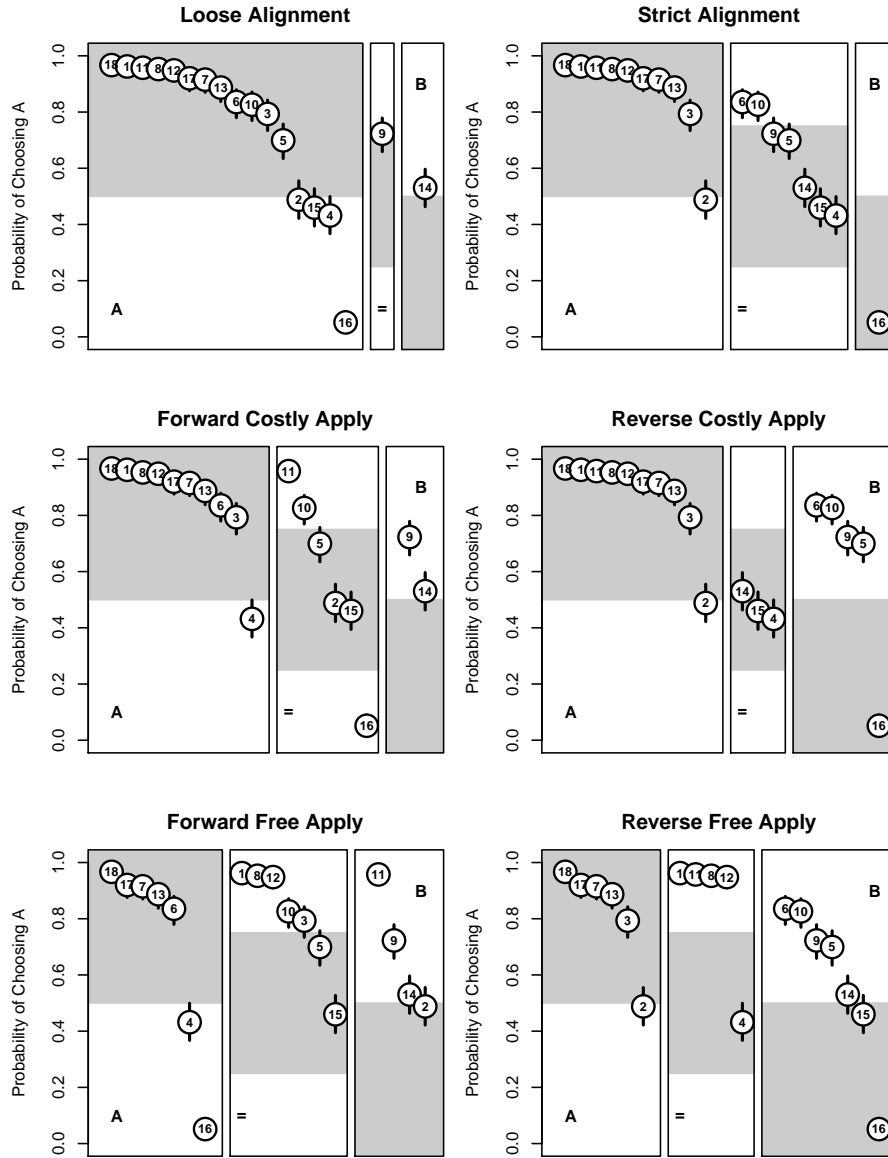


Figure 3.7: Performance of the six similarity models when applied to the triad task in Experiment 3. Each panel is divided into three sub-panels, corresponding to those experimental conditions in which the model in question predicts option A in Table 3.5 (left), option B (right), or indifference (middle). Grey shaded areas correspond to response probabilities that are deemed “consistent” with the model prediction. Each marker plots the proportion of participants choosing option A, and error bars show 95% credible intervals. Conditions for which the entire credible interval lies outside the shaded area are deemed inconsistent with the model prediction.

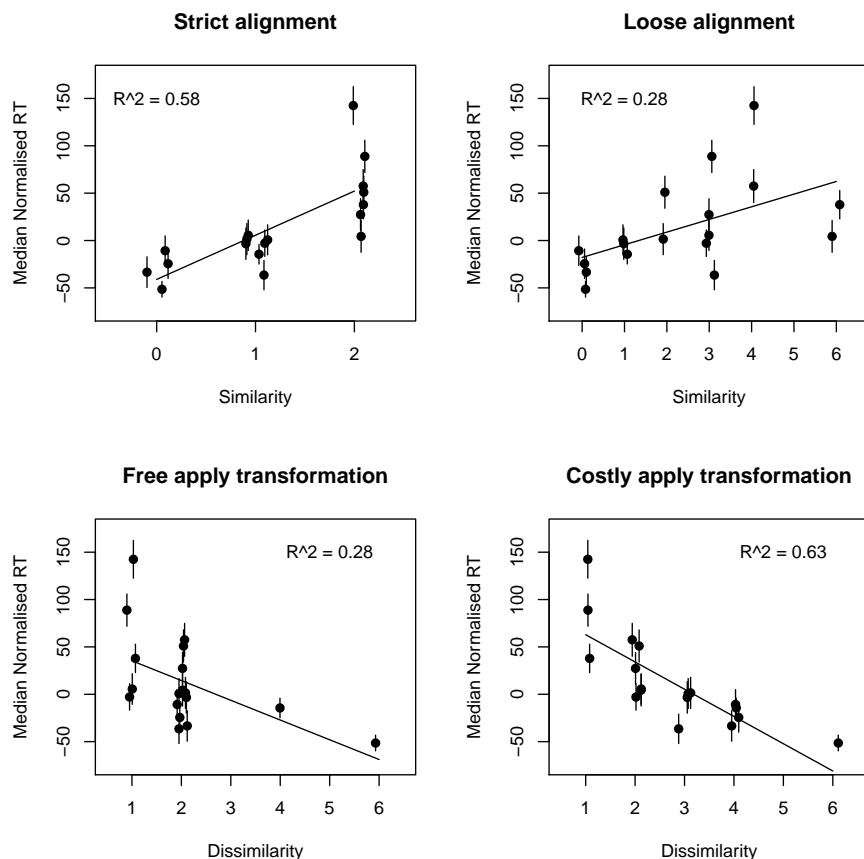


Figure 3.8: Median normalized RT data for the same-difference task, plotted against the predicted similarity (or dissimilarity) values under each of the four models.

are plotted in Figure 3.8 and again mirror my previous findings. Consistent with Experiment 2, the transformational approach shows a slight advantage. When calculated using the costly APPLY method, the transformational distance explains 63% of the variance in the median RT across stimulus pairs. The best performing alignment model uses the strict method, and explains 58% of the variance in the data. Neither of the other two models performs especially well: the loose alignment model and the free-APPLY transformation model both explain only 28% of the variance. In short, while the same-different results from Experiment 3 afford a more nuanced interpretation than those from Experiment 2 (only some versions of alignment and transformation models are successful) the core result is the same: transformational and alignment models perform comparably on the same-different task, with perhaps a slight advantage to the transformational model.

Comparing the two tasks to one another, there is a slight mismatch between the two, again consistent with Experiments 1 and 2: the alignment model performs slightly better in both triad tasks, and the transformational approach fares slightly better in both same-different tasks. By including both tasks in the one study using the same stimuli, the data from Experiment 3 allow us to investigate this. To compare the triad task and the same-different task, note that each triad (ABX) in the forced choice task maps onto two pairwise similarities (AX and BX) in the same-different task. To assess the degree of agreement between the two tasks, Figure 3.9 plots the difference in RT in the same-different task against the choice probabilities in the triad task. As is immediately clear from inspection, the data for all but three triads (5, 9 and 13) fall on or near a linear function. If these three triads are ignored, the two tasks are very closely related, yielding a correlation of $r = 0.86$. To the extent that any systematic differences between tasks exist, these three triads seem most likely to be the source of the discrepancy.

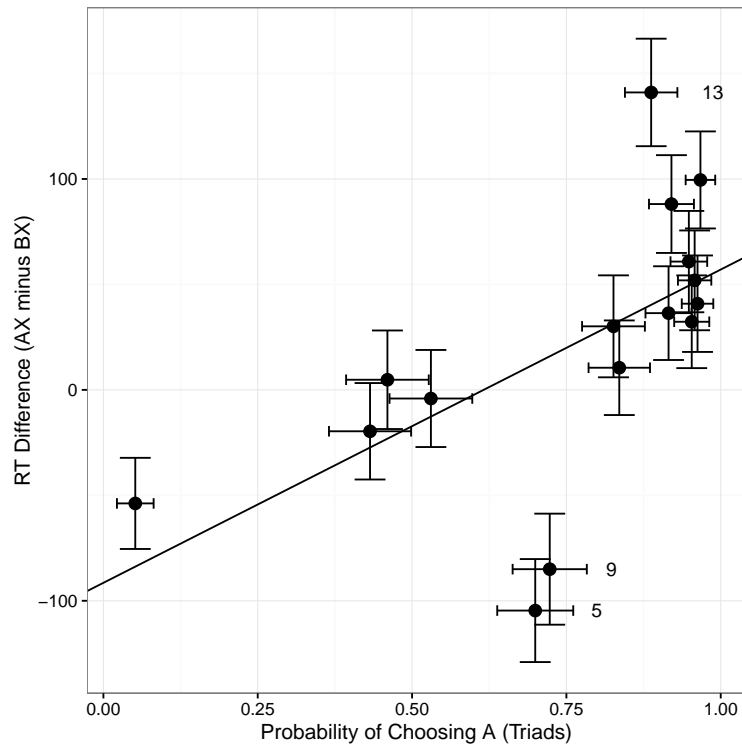


Figure 3.9: Degree of agreement in the same-different task and the triad task. If the three “special cases” (triads 5, 9 and 13) are ignored, the agreement is quite close, producing a correlation of $r = 0.86$. Note that this plot does not include triads 2 or 3 because these require the `aaa:abc` pair and (for balance reasons) this pair was not included in the same-different task.

3.4.3 Discussion

The results from Experiment 3 replicate the key findings from both Experiments 1 and 2, and help resolve a number of questions. First, the results from the triad task suggest that most (but not all) of the alignment advantage in Experiment 1 can be attributed to the fact that the design exploited the free APPLY transformation, which performs considerably worse than the costly APPLY. Second, the results from the same-different task suggest that a strict alignment approach is somewhat superior to a loose alignment approach, and both tasks suggest that a costly APPLY operation produces a better performing transformational model. Third, Experiment 3 replicates the finding that alignment models do slightly better on the triad task and transformation models slightly better on the same-different task, and provides some suggestion that – while the two tasks are very closely related – the origin of the discrepancy may lie in a small number of comparisons (triads 5, 9 and 13) for which human responses are inconsistent across the two tasks. Indeed, setting these triads to one side, the alignment and transformation models perform almost indistinguishably on the triad task: the strict alignment model has one fewer failure than the best transformation model (the forward transformation with costly apply model), but one would hardly want to draw strong conclusions on the basis of a single triad stimulus.

If triads 5, 9 and 13 are indeed the source of the discrepancies between tasks, what makes them special? Inspection of Table 3.3 shows that these triads are the only triads in which one pair (AX or BX) contains a feature mismatch and the other does not. For instance, triad 9 has `aab` as the base item and has `aac` and `abb` as the two potential choices. One stimulus pair (`aab` and `aac`) contains a feature mismatch, with the `b` feature being unique to one item and `c` being unique to the other. This difference produces much faster reaction times for same-different judgments of the pair containing the mismatched feature that does not produce strong (or even consistent) preference for or against choosing that pair in the triad task.

In a same-different judgment, the mismatching features are likely to produce a visual pop-out effect, quickly revealing that the two items are not identical. In contrast, the identification problem for the `aab-abb` pair cannot be solved in the same way as there are no distinctive features: to correctly discern that these are distinct items the visual system must solve the feature binding

problem (Treisman & Gelade, 1980) before responding. As this is a slow, attention-dependent process, reaction times will necessarily be slower to the **aab-abb** pair than the **aab-aac** pair. Although it is not straightforward to determine how such task characteristics relate to model differences, a speculative possibility is as follows. Whenever a feature mismatch exists, the transformational model requires that a **CREATE** operation be employed, which makes the two items more distinctive than they otherwise might be. The simple alignment-based models I consider here do not: the number of MIPs is identical in both pairs, as is the number of MOPs (regardless of whether the alignments are loose or strict). The ability to emphasize feature mismatches via the **CREATE** operation may allow the transformational approach to perform well on same-different judgment tasks.

In the triad task, the pattern of responding is quite different for the same items. People have a fairly strong preference (72%) to endorse **aab** and **aac** as more similar to one another than **aab** and **abb**. This preference makes considerable sense: the fact that the first pair both share the **aa?** structure seems more compelling as a *pattern* (and hence a basis for similarity) than the fact that the second pair share the **a?b** structure. At the very least, there does not seem to be good reason to treat **a?b** as a *more* compelling basis for similarity than **aa?**. In this context, the alignment model seems to be in closer agreement with human behavior. More broadly, this mismatch highlights the fact that identification decisions (as in the same-different task) do have a different structure than similarity judgments (as in the triad task) and while the two are closely related they can differ in non-trivial ways. In fact, the transformational model is designed around a kind of identity-mapping problem: what operations are required to make two objects identical? The alignment model is more focused on structure-mapping, and as such is arguably more closely linked to a pattern matching problem. Given this, it is perhaps rather sensible that the performance of the two models is different across the tasks in precisely the way observed empirically here. Nevertheless, it should be noted that this account is somewhat speculative, and some caution is warranted.

3.5 General Discussion

My goal in this project was to try to distinguish between transformational and structure alignment approaches to assessing stimulus similarity – a goal that met with only partial success. Perhaps unsurprisingly, I did not find a perfectly diagnostic test that allows these two frameworks to be cleanly separated, but across three experiments I was able to place some constraints on both frameworks. For transformational models, the experiments presented here suggest that the original specification of the **APPLY** operation is not quite accurate: the cost for “applying” a feature does seem to increase with the number of slots to which it is applied. Without such a change it is difficult to see how the transformational account can capture the results in Experiment 1 and 3. However, this is difficult to align with previous results that rely on a “free” apply to capture reaction time differences (Hodgetts & Hahn, 2012). For the alignment framework, I find evidence that a “loose” alignment that allows a single feature in object A to be mapped to multiple features in object B performs poorly when applied to a same-different task (in Experiment 3), and a “strict” alignment mechanism that forces a set of consistent alignments produces a better fit. That said, some caution is required: even though the distinction between loose and strict mapping is mirrored in the literature on structure alignment, the particular alignment models I used are greatly simplified when compared to SIAM (Goldstone, 1994), CAB (Larkey & Love, 2003), or LISA (Taylor & Hummel, 2007).²

Taking the three experiments together, the evidence suggests that the strict alignment model and the costly-apply transformation model perform comparably well to one another, but their performance is not consistent across tasks. The alignment model performs somewhat better in a triad task (Experiments 1 and 3), whereas the transformational approach performs better in a same-different task (Experiments 2 and 3), noting that this may in fact stem from the subtle differences between the two tasks. A triad task completed without time pressure involves an explicit similarity comparison, and in that sense is kind of pattern matching problem. In contrast,

²I did consider using the more complicated models but had some concerns about model flexibility. Either I would have had to run SIAM with parameters fixed by previous research, which may not generalize well across the various task changes considered here, or I would need to estimate a number of free parameters. As such, I chose to restrict analysis to a simplified MIPs-versus-MOPs model, though I acknowledge this limits the scope of conclusions somewhat.

the speeded same-different task is a discrimination problem and while similarity is relevant, it is possible to solve a discrimination task by detecting a single discrepant feature without necessarily processing the complete object. This finding is reminiscent of other work (Hendrickson et al., 2015; Hendrickson, Navarro, & Donkin, n.d.) investigating the time course of similarity. For instance, using an evidence accumulation model of response time in a speeded same-different task, Hendrickson et al. (n.d.) found that evidence for time-inhomogeneous information accrual processes: during a single feature match information becomes available before the MIP/MOP distinction exists and can trigger discrimination decisions even before any structural information about the stimulus is available. With this in mind, it is not entirely clear that the triad task and same-different tasks necessarily measure the “same” similarity construct.

On a different note, it should be pointed out that there is an additional source of ambiguity in these results, pertaining to the mental representation of stimuli. As noted in connection to Figure 3.1, in order to employ a structure mapping model to assess the similarity between 111xxx1 and 0yz000 it is critical to determine what counts as a “slot” and what counts as a “feature”. Different ways of describing the stimuli can produce different similarities. As much as possible I tried to design stimuli for which this concern does not arise (on the assumption that in each case there is one “obvious” way to describe the stimuli), but it may well be that there are exceptions. For example, consider triad 10 from Experiment 3, denoted as $aab \rightarrow aac : acb$ in Table 3.5. Participants showed a strong preference to rate aab and aac as more similar to one another than aab and acb , an effect that was predicted by only one of the six models (loose alignment: one of the worse performing models in general). However, the model failures may not be due to an inability to describe similarity comparison *processes*, but might instead be caused by an inappropriately specified stimulus representation. As shown in Figure 3.6 the stimuli were designed with three visually distinct shapes, intended to prevent people from mentally rotating the stimuli or treating the three locations as exchangeable. For a stimulus like acb it seems entirely reasonable to assume a three-slot representation $[a] [c] [b]$. However when *two* of the three stimuli in the triad happen to have the same features in the same two locations, it is quite possible that people might chunk those two locations into a single slot, yielding stimulus representations $[aa] [b]$ and $[aa] [c]$ for the other two items. When the stimuli are recoded this way, all models will produce the right effect. From the alignment perspective, the $[aa] [b]$ and $[aa] [c]$ items share one MIP (i.e., aa), whereas $[aa] [b]$ and $[a] [c] [b]$ have a single MIP (i.e., b) and a non-alignable difference caused by the fact that one object has two slots and the other has three (Goldstone et al., 1991). From the transformational perspective, to transform $[aa] [b]$ into $[aa] [c]$ requires two operations (CREATE C and APPLY C), but to transform $[aa] [b]$ into $[a] [c] [b]$ presumably requires three operations. First the $[aa]$ chunk needs to be SPLIT into $[a][a]$, and then the CREATE C and APPLY C operations follow. If there is a possibility that people make use of chunked representations when processing stimuli, then both theories can be adapted to accommodate these results.

The transformational and alignment approaches to similarity are both broad frameworks, each supporting a diverse menagerie of possible implementations. Since all models are wrong (Box, 1979), it is perhaps unsurprising that it was possible to find stimuli that caused problems for the particular implementation of transformational similarity current for geometric shapes, or a simple MIP+MOP lowest-common-denominator implementation of the alignment approach. However, I would argue these ‘wrong’ models are extremely useful in the way they inform thinking on the bigger issues of what representations people use and how they can be compared. At this more abstract level, these results do not offer strong constraints: but they do sharpen the questions, in the spirit of Samuel Karlin’s assertion that “the purpose of models is not to fit the data but to sharpen the questions”.³

Specifically, at the implementation level, requiring a costly apply operation improves the agreement between the data presented here and the transformational theory of similarity. The original multiple-apply transformation exposes the model to a set of special cases where people consistently make choices incompatible with the theory. Although this could be considered a relatively minor model misspecification, the result raises the more general question of how to appropriately determine the set of transformations used in the transformational approach. Along the same lines, these results highlight the dependence of alignment approaches on the definition of a “slot”. Perhaps surprisingly given the extremely simple nature of these stimuli, under an alignment interpretation of the results it seems that participants flexibly adapted the nature of the chunking structures they were using depending on apparent relations between features.

³11th R. A. Fisher Memorial Lecture, Royal Society 20, April 1983.

At the more abstract framework level, these results sharpen the old questions of “How do you know you have the right representations?” and “How do you know you have the right task?” The short answer to the first question is simply that you don’t, but here I sharpen the question with a concrete demonstration of the problem in the domain of simple geometric stimuli: is [aa] [b] different from [a] [a] [b]? The second question also admits a short answer, that there’s no such thing as a wrong task, only different tasks. Here again it’s possible to sharpen the question, in this case with a demonstration of specific timing-dependent differences between two tasks both aimed at the same general construct of ‘similarity judgment’.

The comparison between transformational and alignment-based approaches to similarity presented above does not answer these sharpened questions. But I suggest these tasks highlight promising places to look, in the way they link higher-level questions of representation to more readily testable claims about the relationships between these abstractions and perceptual similarity, pattern matching, discrimination, and visual pop-out.

Chapter 4

Which transformations?

One of the main results in Chapter 3 was a question: how can researchers (and people in general) arrive at a ‘good’ set of transformations? The question was raised in the specific context of a transformational account of similarity, but it’s also true that regardless of how people process similarity, the world is full of transformations that need to be learned. For example, people tend to take transformations due to aging into account when identifying faces (Mark, Todd, & Shaw, 1981), and track object identities by inferring motion from a sequence of still images (Freyd, 1983). When action stimuli are stripped of as much visual information as possible and rendered as a minimal collection of moving dots, people find activities such as walking or running instantly recognizable and base detailed inferences on just this motion information (Troje & Basbaum, 2008). Whether these things are ‘just ordinary features’ in the same sense as FEATHERS or RED is debatable, but they are clearly involved in natural category structures in the world.

So how do people actually process and use this transformation-like structure in the world? Since transformations are so ubiquitous, one reasonable position might be that in fact there is nothing particularly special about them. Why not simply give the feature FLIES the same status as the feature FEATHERS for the category BIRD? Transformations may or may not be directly observable, but in this sense they are no more strange than latent features such as CARNIVORE or HAS-SESAMOID-BONES routinely invoked in inference and property induction (Navarro & Griffiths, 2008). Transformations themselves could admit category structures just like anything else, using a more abstract representation space that allows for the way transformations unfold over time (Lamberts, 2003; Pollick & Paterson, 2008). There are however a few different ways in which transformations could potentially be distinct from other types of feature.

One view is that transformations are higher-order features composed of other, more basic, perceptual features. Under this view, a transformation is a series of more primitive features chained together over time (Sadanand & Corso, 2012). However it’s also possible that transformations are in fact *more* fundamental than other types of features, in the sense that a having a set of identity-preserving transformations defines the invariants that let us identify perceptual features in the first place (Austerweil & Griffiths, 2013) For example, recognizing that shape features are invariant under rotation might precede their use in object recognition (Ullman, 1996; Graf, 2006).

Given these different accounts it’s probably fair to say that it’s not currently clear what role transformations might play in similarity and categorization. However the theory of transformational similarity discussed in the first section of this thesis makes some quite strong predictions in this regard. Firstly, it describes a world where transformations are ubiquitous and involved early in perceptual processing. Less obviously, it also commits to a world where new transformations are learned quickly and are strongly associated with a specific context. These additional properties arise as a consequence of the need for similarity judgments to be computationally tractable (Müller et al., 2009).

Efficiency is important to all accounts of similarity, which is known to be highly efficient. The speeded same-different responses presented in Chapter 3 require similarity to be processed in at most hundreds of milliseconds, see Hendrickson et al. (2015); Lovett et al. (2009) for tighter more detailed bounds. However efficiency is particularly relevant to transformational similarity, where the Kolmogorov complexity measure at the heart of the theory is not just hard to compute, but provably uncomputable (Ming & Vitányi, 1997). As a result, all practical implementations of information-distance measures must sacrifice the full generality of the theory (where the choice

of programming language contributes only an additive logarithmic factor to the distance between representations) for a specific instantiation where particular transformation distances and the time needed to compute them depend on the specifics of the program doing the calculation (see M. Li, Chen, Li, Ma, & Vitányi, 2004; Bennett et al., 1998). Since different instantiations render different comparisons efficiently, the known need for high efficiency can be used to provide constraints on the specific program involved. In the context of the similarity judgments considered here, these constraints are naturally expressed in terms of limits on plausible transformation sets.

Müller et al. (2009) investigated these constraints at a highly general level, working from a broadly inclusive definition of transformational similarity and making minimal assumptions about the nature of the representations. Even abstracting away these specific details, they were able to identify a set of bottleneck parameters which must be controlled in any plausible implementation of transformational similarity, with plausibility here defined as ‘polynomial running time.’ These critical parameters include constraints on the longest transformation distances, and on the size of all intermediate representations on the transformation path, but the constraints most relevant to the project described below are the ones on the transformations sets: the set of transformations considered *must* be small and context-specific. The full details appear in Müller et al. (2009), but intuitively, the search for shortest paths is only feasible if there are a limited number of choices at each potential crossroads and the target is not too far away.

The constraint that transformations must be few and context-specific implies that under the transformational account people should readily learn new transformations for new domains. Life-long chopstick users won’t be able to use generic ‘object similarity’ schemes to make similarity judgments between various Western eating utensils, while newcomers to Asian calligraphy will need new transformations to judge the similarities that unite different *seal* scripts and distinguish them from *running* scripts, and so on for each domain. Moreover, each domain will only admit a small number of allowable transformations at any one time. So is it possible to observe this learning process in a controlled setting? Are the observed learning rates and any subsequent changes in similarity judgment compatible with the predictions of the transformational account of similarity? The work presented in Chapter 5 aims to find out.

Similarity and categorization judgments are known to change with domain knowledge (Medin, Lynch, Coley, & Atran, 1997; Honoré-Chedozeau, Lelièvre-Desmas, Ballester, Chollet, & Valentin, 2017; Shafto & Coley, 2003). Moreover, Hahn et al. (2009) specifically showed similarity judgments changing with transformation learning over short time-frames. Chapter 5 describes a series of studies looking at the feasibility of training people on new transformations in a simplified artificial setting over short time scales and relatively low numbers of trials, and explores whether or not such learning had a detectable impact on similarity or categorization judgment. It builds on the work of Hahn et al. (2009) in a number of ways: by considering a new set of transformations in a simple and flexible stimulus space, by separating measures of learning from measures of similarity change, and by considering both similarity and categorization judgment.

People are known to be able to learn transformations, and alternative approaches to similarity that simply consider transformations as latent features also predict that similarity judgments would be impacted by learning about a new feature that two stimuli share. So evidence of learning and applying transformations to similarity do not in themselves constitute a strong test of the transformational approach. However establishing a paradigm for transformation learning and similarity judgment would be especially useful for testing transformational approaches to similarity. Given evidence of transformation learning, the problem of determining which transformations were being used by people would be heavily constrained by the requirement that transformation sets be small and context specific.

Putting the particular concerns of transformational similarity to one side, transformations as features also have unusual properties relevant to other approaches to similarity. In particular, their nature as features extended over time presents unique challenges for the broad family of geometric approaches to similarity (Lamberts, 2003). From a research methodology point of view, unique opportunities are offered by the way experience with a transformation can be manipulated to give variable degrees of exposure to a feature for physically identical test stimuli. However to date there has been relatively little work in this area (although see Austerweil, Griffiths, and Palmer (2016); Austerweil and Griffiths (2010b) and references therein).

To investigate a potential paradigm for studying transformation learning, I turned again to comparisons. The key properties of the task were a separation of measures of transformation learning from measures of similarity, and a crossed design with two different training conditions

such that physically identical test stimuli had a different relationship to the trained stimuli for different people. Isolating this difference allowed me to separate out the effects of the training manipulation from the inherent properties of the stimuli. I also separated questions about category membership from questions about similarity. I found that transformation learning was associated with changes in similarity and categorization, but also that people's learning and patterns of generalization were highly sensitive to the details of the presentation format. These results give somewhat ambiguous support for the transformational account of similarity. The core requirement of a relationship between learned transformations and similarity supported, but there is also evidence of family relationships between transformations that suggest something like graded availability of transformations might be required to give a full account of the data, complicating the calculation of transformation distances. For other theories of similarity, the main contribution of these results is their description of how differences in presentation format impact task difficulty.

Chapter 5

Transformation learning

The work in this section was published as: Langsford, Hendrickson, Perfors, Navarro (2017) *When do learned transformations influence similarity and categorization?* In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.) Proceedings of the 39th Annual Conference of the Cognitive Science Society. (pp. 2530-2535)

Statement of Authorship

Title of Paper	When do learned transformations influence similarity and categorization?
Publication Status	<input type="checkbox"/> Published
Publication Details	Langsford, Hendrickson, Perfors, Navarro (2017) <i>When do learned transformations influence similarity and categorization?</i> In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.) Proceedings of the 39th Annual Conference of the Cognitive Science Society.

Principal Author

Name of Principal Author (Candidate)	Steven Langsford		
Contribution to the Paper	Significant contribution to articulation of experimental hypothesis and design. All implementation and data collection. Performed all analyses, with direction. Initial drafting of the associated write-up.		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	18/8/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Amy Perfors		
Contribution to the Paper	I helped to design the study, articulate the core hypotheses, frame the paper, and edit the manuscript.		
Signature		Date	31/08/2017

Name of Co-Author	Daniel Navarro		
Contribution to the Paper	I helped with the design of the study, commented on data analysis and assisted with the writing of the manuscript.		

Signature		Date	31/08/2017
-----------	--	------	------------

Name of Co-Author	Andrew Hendrickson		
Contribution to the Paper	I helped design the study, discussed how to analyse the results, and made a significant contribution to editing the manuscript.		
Signature		Date	31/08/2017

Langsford, S., Hendrickson, A., Perfors, A. and Navarro, D. (2017) When do learned transformations influence similarity and categorization? In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.) *Proceedings of the 39th Annual Conference of the Cognitive Science Society (CogSci 2017), held 26-29 July 2017*, Cognitive Science Society, London, UK, pp. 2530-2535.

NOTE: This publication is included in the print copy of the thesis held in the University of Adelaide Library.

Part III

Linguistic representation

Chapter 6

From similarity to sentences

This section moves into a new application domain within cognitive science, examining the empirical toolkit for studying syntax. Two aspects of this thesis in particular are novel contributions to this area: one is the adaptation of the well-known Thurstonian modeling approach to the sentence acceptability setting, and the other is an investigation into the test-retest reliability of various measurement techniques.

The main connection with the themes of this thesis appears in the first of these, the Thurstonian modeling of sentence acceptability. As discussed in more detail below, this technique shares common roots with the structure-discovery methods discussed earlier in the context of similarity spaces. The Thurstonian method relies on comparison data to make inferences about the latent structure of the representations involved. Here, the technique is useful in the context of an ongoing debate over the graded-vs-clustered structure of acceptability, and is also motivated by the desire to combine the best features of forced choice judgments with the advantages of scale-like data.

In the process of testing and validating the Thurstonian model, I found that the existing literature did not fully address the test-retest reliability of the most popular elicitation tools for sentence acceptability. Although other important aspects of reliability have been studied in this context, notably type one and two error rates (Sprouse & Almeida, 2017), and sensitivity to effect size (Weskott & Fanselow, 2011), test-retest reliability is distinct from both of these. It has the advantage of being interpretable in the absence of a reference ground truth. Moreover, it can also offer additional information about the source of variability by contrasting within-participant with between-participant test-retest reliability; this comparison has no analogue in error rate.

The following section first motivates the question of the structure of acceptability, then presents evidence that the particular Thurstonian model I implemented is capable of addressing this question. After motivating the adaptation of Thurstonian modeling to this context, it gives a brief introduction to work on the properties of popular methods used in this area, looking in particular at their reliability and sensitivity. This is followed by the main chapter, which combines all of these elements in a comparison between the Thurstonian model and five other formal measures of acceptability. Through studies of their test-retest reliability and cross measure agreement, this work provides a quantitative picture of the different sources of variability these different elicitation tasks are vulnerable to, shows their relative efficiency, and tests for the presence of task-induced biases.

6.1 The structure of acceptability

The structure of sentence acceptability is one of those questions where the answer is obvious to everyone, but it's not necessarily the *same* answer (Aarts, 2007; Aarts, Denison, Keizer, & Popova, 2004; Fanselow, 2006; Sorace & Keller, 2005). For some, acceptability is clearly a statistical phenomenon related to the likelihood of a sentence under some appropriate language model, best described by a real number (eg Manning, 2003; Halevy, Norvig, & Pereira, 2009). Others prefer the discrete formalisms associated with Chomsky (1965), and highlight the possibility that acceptability ratings combine several different factors, arguing that attested measures of gradient acceptability (eg Lau, Clark, & Lappin, 2016) conflate a strongly categorical grammaticality component with gradient performance factors (Hofmeister, Jaeger, Arnon, Sag, & Snider, 2013).

The Thurstonian model presented here is relevant to this debate because of its ability to fit a

range of different latent acceptability structures. In this it is unlike other common acceptability judgment tasks, which admit either interpretation. For example, categorical acceptability judgments (particularly binary ones) intuitively support a categorical view of acceptability matching a literal interpretation of the discrete response format. However such responses can also be aggregated to give a real-valued proportion of endorsements (Myers, 2009). The process used to analyze the data depends on a theoretical commitment to a particular interpretation of the responses, something that is not directly answerable by the responses themselves.

Thurstonian modeling is one way to address this question. The method was developed by Louis Thurstone to describe comparison judgments in terms of a latent representation space (Thurstone, 1927), and is a descriptive modeling technique that applies whenever things can be compared on a subjective continuum. Thurstone’s 1927 paper gives degrees of greyness, perceived weight, and handwriting quality as examples, the many subsequent uses of the approach have included movie ratings and presidential popularity (Selker, Lee, & Iyer, 2017), various kinds of food and other product preferences (O’Mahony, 2003) and many many more (see Ennis (2016) for a review).

The core of the Thurstonian technique is the construction of a latent scale capturing people’s “impressions” of the stimuli. Assuming only that people’s impressions are symmetrically variable around some mean value characteristic of the stimulus, the fact that the units of the scale are arbitrary allow it to be structured such that the variability of impressions is well described by a normal distribution. Specifying a mapping between impressions on this scale and choice responses allows the construction of an appropriate scale from data, inferring the position on the scale representing the mean “impression” of each stimulus. The particular implementation presented here uses a Gibbs sampler, JAGS, to do this (Plummer, 2003). The inferred means and variability values on the latent scale are interpretable as an efficient summary of the comparison data used to construct the scale. Given an assumption of transitivity, they can be used to infer probable outcomes for comparisons that may not have been presented to participants, allowing a full n^2 matrix of possible contrasts to be inferred from considerably fewer than $n(n - 1)$ comparisons (Thurstone, 1931). Although violations of transitivity are certainly possible (Tversky, 1969), in many applications it is a reasonable assumption (see Cavagnaro & Davis-Stober, 2014), and in the particular project described here, the use of multiple elicitation tasks including rating scales on individual items provide a cross-check against the possible influence of contrast items on perceived acceptability.

This ability to predict comparison outcomes is not just about leveraging an assumption of transitivity to efficiently estimate a subjective ordering: to the extent that scores on the inferred latent scale allow predictions about comparison outcomes, distances on the scale capture something about the representation space (Ashby & Lee, 1991). Interest in the structure of the representation space has given Thurstonian modeling an important place in the similarity literature: multidimensional scaling approaches, a core formalism for geometric approaches to similarity (Borg & Groenen, 2005), can be considered a multidimensional generalization of the same basic idea (Nosofsky, 1992).

For the particular question of sentence acceptability, the Thurstonian method does not offer a complete solution since it does not distinguish between different factors that might independently contribute to overall acceptability. However, currently the most common approaches to eliciting sentence acceptability judgments do not directly address the question of representation. The most basic methodology for constructing linguistic arguments – simply presenting acceptable and pathological examples validated by introspection – either makes no claim as to the structure of acceptability or at best can be interpreted as offering a three-way categorical outcome distinguishing between perfectly acceptable, questionable, and uncontroversially unacceptable (Gibbs, 2006). While non-introspective approaches such as Likert scale ratings by groups of native speakers can claim greater objectivity, which comes with its own substantial benefits, it is not clear that such scales are any more informative about the structure of acceptability (Schütze, 1996; Geeraerts, 2006). Likert scales in particular are *ordinal* judgments, with the psychological distance between the various response option levels unknown and quite likely to be different between different pairs of adjacent options: the difference in acceptability causing someone to change their rating from a ‘3’ to a ‘4’ need not be the same as that required to shift a rating from a ‘4’ to a ‘5’. Moreover, both the level of acceptability each response option is taken to represent and the sizes of distances between the options are likely to be different for different people (Gonzalez-Marquez, 2007).

One possible response to the limitations of Likert scale data is magnitude estimation judgment. Originally adapted from its psychophysics origins (Stevens, 1956) to linguistics by Bard, Robertson, and Sorace (1996), magnitude estimation asks participants to assign a numerical acceptability score

to a series of test sentences using a reference item/score pair to calibrate their responses. Magnitude estimation was originally intended to provide truly interval data, in which score differences map directly onto acceptability differences (Stevens, 1956; Cowart, 1997). Interpreted literally in line with the instructions given to participants, such responses should reflect representation structure, and data of this type has been presented as evidence in claims about the structure of acceptability, for example that it is categorical rather than gradient (Sprouse, 2007). However more recent work (some by the same authors) shows that a literal interpretation of magnitude estimation scores as interval data directly reflecting psychological distances is untenable. Magnitude estimation scores violate commutativity (Sprouse, 2011a) and are insensitive to the choice of reference in ways that are incompatible with the literal interpretation (Sprouse, 2008).

One possible response might be to modify the magnitude estimation task into something that people can do more easily. One such modification appears in Featherston (2009b). The proposed method approximates a continuous scale with a large number of response options (more than 20) and supplies two references rather than one, placed near the top and near the bottom of the expected response range (although allowing responses outside this range, if a test item goes above or below the range spanned by the references).

The Thurstonian approach I investigated here presents an alternative possibility: it shifts the responsibility for quantifying acceptability scores away from participants and onto to a model of choice behavior, presenting participants with the easier task of just making comparisons. One motivation for this approach was the hope that the relative simplicity of the comparison task for participants would translate into higher reliability and clearer data. The other was due to the similarity literature, which suggests that scale-free comparison judgments are particularly well suited to structure discovery in domains like sentence acceptability where the underlying structure of the space is unclear.

6.2 Testing the Thurstonian model

The Thurstonian model describes a simplified process for selecting either the more acceptable sentence or a ‘no difference’ option from trials comparing pairs of sentences. The decision process for each trial was modeled as follows: for each comparison, acceptability ‘impressions’ were drawn for each sentence from a distribution with a mean at the group consensus for the acceptability of that sentence and a variance that differed between participants but was constant for each participant across sentences. The difference between the sampled values was then compared to a participant’s criterion. If the absolute value of the difference was greater than the criterion, the model predicted that the more acceptable sentence should be endorsed; otherwise, it made an ‘equal’ response. Implementation details of the model can be found in Appendix B.

I tested the structure-discovery capability of this particular instantiation of the Thurstonian approach in two main ways. One was in simulation tests, checking that the model could successfully recover a variety of simulated acceptability structures. The other was with real-world data, where I examined whether distances on the inferred scale could be used to derive well-calibrated estimates of the choices people would make when facing novel combinations of these sentences.

6.2.1 Simulation tests

This section reports the performance of the Thurstonian model on simulated datasets. All simulated participants followed the model’s decision process, but various experimental design decisions and facts about participant’s decision parameters and the simulated sentence acceptability were allowed to change. For each simulated decision, acceptability impressions for each of the two sentences were drawn from an underlying distribution with a community consensus mean and a participant-specific degree of variability, with participants deciding which sentence to endorse based on whether the differences in perceived acceptability were larger than their subjective criterion.

Simulations testing experimental design decisions such as sample size and the size of the item set were used to check the feasibility of the the experimental approach but are not reported here. As expected, these simulations showed that performance depends on the ratio of participants to items, and gave an estimate of the feasible design space for this modeling approach given a range of possible participant variability levels. These simulations were used to select the number of items and participants used in the actual study. Given that the model was found to be feasible for plausible experimental scenarios, the main question of interest here is its performance in recovering

different underlying distributions of ‘true’ sentence acceptability. For the particular simulations presented here, the other variables were fixed at: 300 sentences, 150 simulated participants making 40 responses each, with participant variability drawn from $|N(10, 10)| + 10$, and a criterion drawn from $|N(10, 10)| + 5$. These settings are chosen to be (roughly) consistent with the real-word data. In particular, these variability and criterion settings result in a similar proportion of ‘equal’ endorsements to that observed in the real data set ($\sim 25\%$).

The issue of interest for the representation debate is whether the model can recover different acceptability structures. If it can, this suggests it would be informative about the structure of human representations if applied to human data. To test this, I evaluated the model on a variety of acceptability structures (captured as different distributions). The candidate distributions tested were UNIFORM (with sentence acceptabilities ranging uniformly from 0-100), CLUSTERED (with acceptability scores of only 10, 50 or 90, with roughly equal numbers in each group), CUBIC (with acceptability scores following a cubic equation), and SIGMOID (i.e., with acceptability scores drawn from a logistic function scaled to fall in a 0-100 range).¹

Results of these simulations are plotted in Figure 6.1. The first column shows the different ‘true’ acceptability structures (in order: CUBIC, UNIFORM, SIGMOID, and CLUSTERED). In each, the plot shows the acceptability of each sentence when arranged by rank. The second column demonstrates that model is capable of recovering the structure in each case: when the model estimates are arranged according to the true acceptability rank of each sentence, the distribution shape matches the true underlying distribution. The third column demonstrates the model’s success more precisely by plotting model acceptability estimates on the x -axis and true simulated acceptability on the y -axis. This presentation makes it easier to see differences between the model estimates and the simulation truth: if the model is successfully recovering the acceptability structure, these plots should be linear. The only case showing any kind of systematic difference from the simulated truth is the cubic distribution, where the model gives less extreme ratings than the simulation truth warrants for the best and worst sentences. This is a relatively minor effect and probably emerged because of the model priors, which were chosen with the aim of preventing the ends of the scale from becoming too extreme. Details appear in Appendix B.

6.2.2 Interpreting distances on the scale

Simulated data is not the only way to test whether the THURSTONE model can accurately capture the structure of acceptability in the stimulus set. The THURSTONE acceptability scores can be interpreted as predictions about future acceptability judgments, which admits tests of predictive validity. The decision rule used to map acceptability scores onto comparison responses can be applied to any two acceptability scores to produce a prediction of the probability of endorsement for each possible response. Chapter 7, which investigates the test-retest reliability of sentence judgments, involves datasets supporting exactly this kind of test. The chapter contains the full details, but the key property here is that it conducted a BETWEEN PARTICIPANTS replication study. In it, each new participant saw a different subset of the overall set of 300 sentences of interest. As a result, the majority (4472 of 5000, 89%) of the comparisons made in the replication were novel combinations that had not been presented to any participants in the INITIAL data set. This gave us a natural test of the THURSTONE model: the fit to the INITIAL data set was used to generate predictions for the new comparisons that appeared in the BETWEEN PARTICIPANTS data set.

One simple measure of predictive success is to count the number of times the response made was the one predicted as having the highest probability on the basis of the fit to the INITIAL data. This yields a predictive accuracy of 66.62% (3331 correct out of 5000 responses). This is, however, a crude measure: it largely reflects the sign of the difference in acceptability scores while ignoring the actual distance. To test if the distances between acceptability scores support interpretation as a psychological distance, the level of confidence in each prediction is also important.

Figure 6.2 visualizes one possible test of this prediction calibration property. Outcome probabilities predicted by the model were rounded to one decimal place, producing 11 bins from 0 to 1, and all possible responses to comparisons appearing in the BETWEEN PARTICIPANTS data set were grouped by predicted probability of that response occurring. If the predictions of the model are well-calibrated, the proportion of responses in each bin which were actually observed should closely match the probability associated with that bin. Figure 6.2 shows the probability of a response occurring based on the model fit to INITIAL data on the y -axis, and the proportion of

¹The equation used to generate the data was $fn(x) = \frac{1}{1+e^{-.1*(x-50)}} \times 100$.

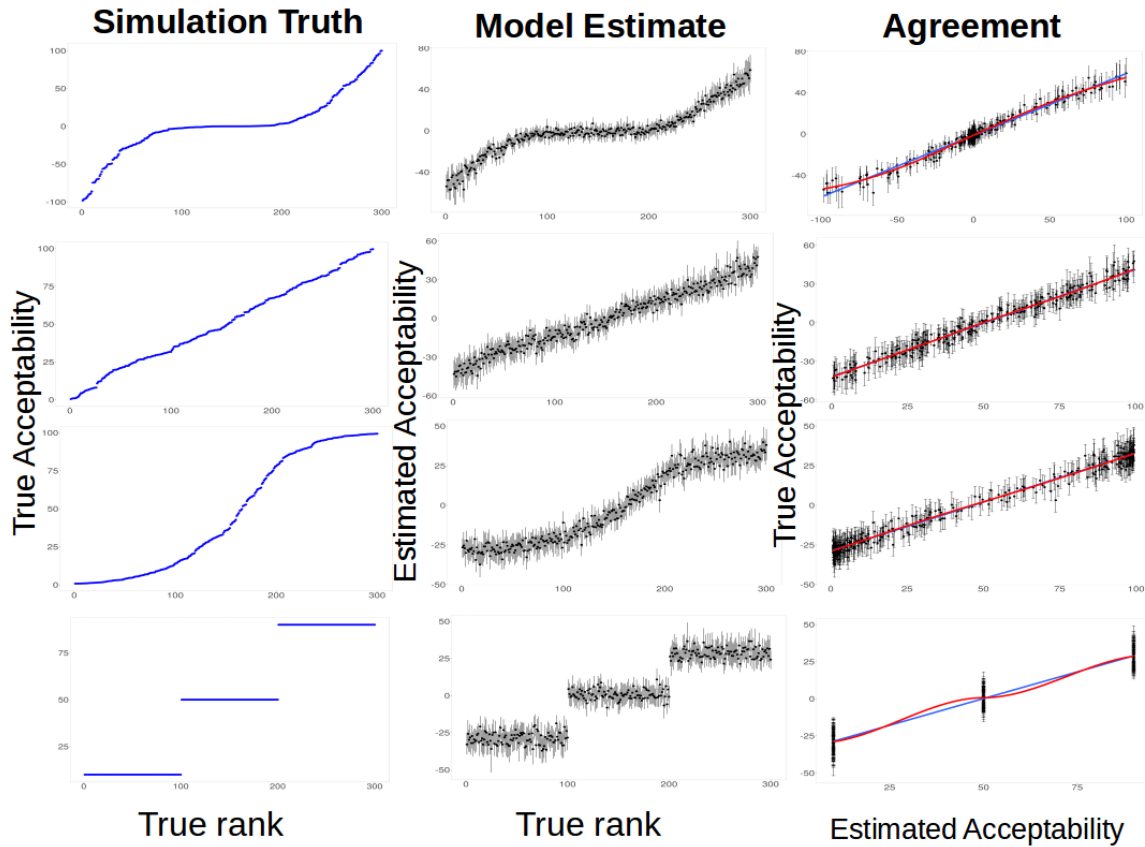


Figure 6.1: Testing the Thurstonian model’s ability to recover various acceptability structures. The leftmost column shows the ‘ground truth’ of the simulated acceptability values on the y axis, and rank order on the x-axis. The distributions were (in order) CUBIC, UNIFORM, SIGMOID, and CLUSTERED. The middle column shows the acceptability scores estimated by the model, positioned on the same x-axis of true simulation rank. The third column shows the agreement between the two, with estimated acceptability on the x-axis and simulation truth on the y-axis. The blue fit line shows the best fitting linear model: perfect correlation would entail all points falling exactly on this line. The red line shows a loess-smoothed local fit: differences between the linear and smoothed fits visually highlight any differences in structure between the simulation truth and the local model estimates. In most cases, the two lines coincide, indicating that the model successfully recovers the simulated distribution regardless of the shape of the distribution. The only exception is that when the underlying distribution is cubic, the smoothed line departs slightly from the linear fit at the ends of the scale by giving less extreme estimates than the simulation truth warrants.

these responses which were actually observed on the x-axis. The model appears well calibrated, in that the prediction probabilities closely match the observed proportion at all levels of confidence. One widely-used metric for quantifying prediction calibration is the Brier score (Brier, 1950), essentially the mean squared error of the prediction, with lower values indicating better predictions. The model fit to INITIAL responses achieves a Brier score of 0.27 when predicting the BETWEEN PARTICIPANTS responses. Brier scores are mainly useful for comparing competing predictors in a particular domain, but since the other acceptability measures considered in Chapter 7 do not produce such predictions, they cannot be compared on this measure. Still, the relatively low Brier score is somewhat reassuring. To the extent that the THURSTONE acceptability estimates express meaningful distances – and these distances embody valid predictions about people’s acceptability judgments – the structure described by the estimated score distances can be interpreted as expressing something about the structure of sentence acceptability.

So what *was* the structure that the THURSTONE model recovered based on the data presented in Chapter 7? This data, described in detail in that chapter, consisted of participant sentence acceptability judgments of 300 sentences of varying acceptability. Figure 6.3 shows the structure of acceptability that the THURSTONE model recovered based on it. Although all intermediate

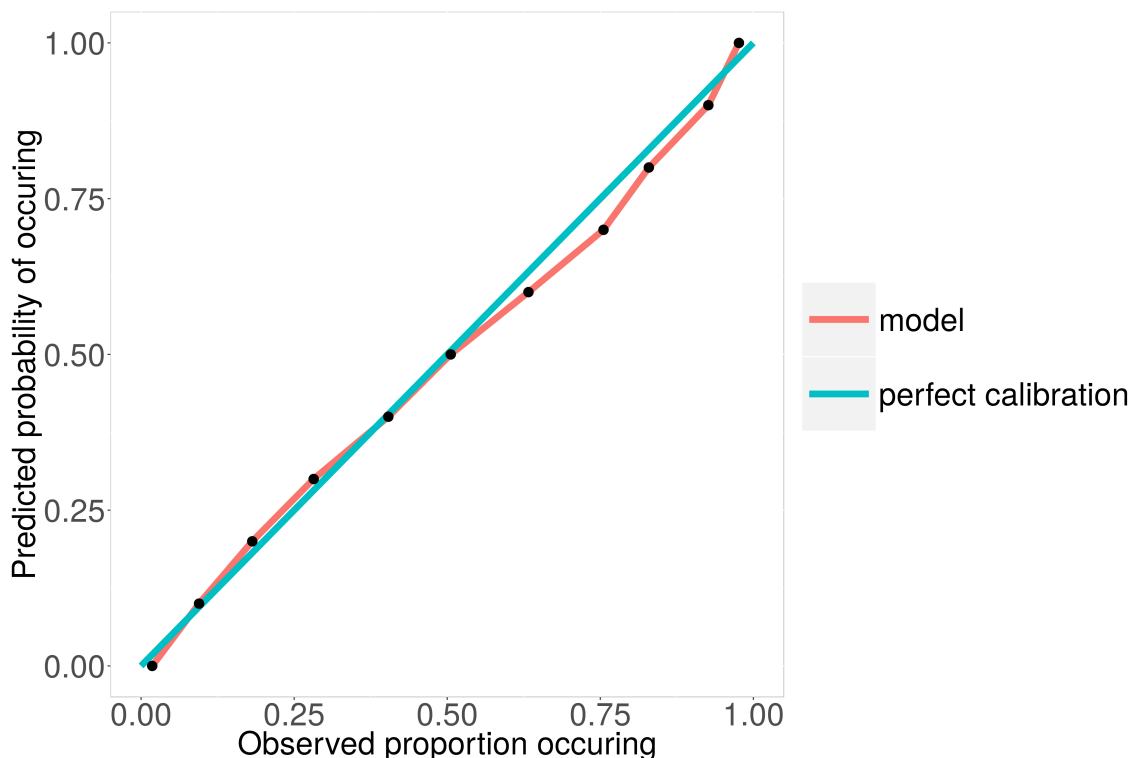


Figure 6.2: Calibration of predictions made by the Thurstonian model. Each possible outcome for each judgment appearing in the BETWEEN PARTICIPANTS replication study was assigned a probability of occurring by the model on the basis of its fit to the INITIAL data set, which involved the same sentences but different individual comparisons. These predictions were binned by rounding them to a single decimal place, and for each bin, the proportion of those outcomes which were actually observed in the BETWEEN PARTICIPANTS data set was calculated. Perfect calibration (blue) would involve every bin showing a proportion of observed outcomes exactly identical to the bin probability. The model’s predictions lie quite close to this line. Good calibration at this level of detail implies that distances between THURSTONE acceptability scores have a valid interpretation in terms of people’s ability to discriminate their acceptability.

values between the highest and lowest acceptability scores are represented, a histogram of the acceptability scores shows a bimodal distribution, suggesting an underlying latent structure with two clusters.

Of course, the analysis here and the sentences used are more suitable for validating the structure-recovery ability of the model than they are for making any comment about the structure of sentence acceptability more generally. In this case, the recovered structure reflects a property of the sentence set that was known in advance: they were constructed by linguists in acceptable/unacceptable pairs, each pair illustrating an argument put forward in *Linguistic Inquiry* (Sprouse, Schütze, & Almeida, 2013). This advance knowledge about the structure of the test sentences is a nice sanity check on the plausibility of the structure recovered, but also strongly limits the generality of any conclusions drawn about the ‘true’ distribution of sentence acceptability.

I leave the survey of acceptability distributions given different sentence sampling or construction schemes to future work, but note that to the extent that the debate is over the appropriate interpretation of discrete response data, the results presented in Chapter 7 *do* suggest a resolution. This is evident thanks to the strong agreement reported between the structure-agnostic THURSTONE acceptability scores and the scale-based LIKERT scores. As detailed in that chapter, means of z -transformed Likert responses can be interpreted as indicating gradient levels of acceptability.

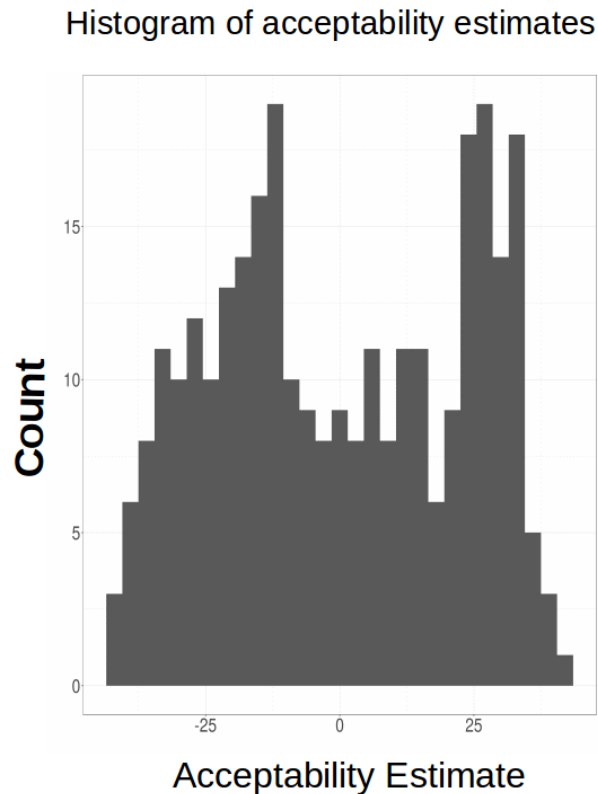


Figure 6.3: Histogram showing distribution of acceptability scores inferred by the Thurstonian model. Scores are bimodal, with one peak around an acceptability of -15 and another around 30. These units are arbitrary: they roughly correspond to ranks 100 and 230 out of 300.

6.3 Reliability studies for sentence acceptability measures

The structure of acceptability discussed in detail above is only one aspect of a larger discussion about measurement methods in linguistics research, and in particular the status of acceptability judgments as evidence.

Early work in this area was prompted by concerns about the practice of ‘armchair linguistics’, which considered phrases or sentences as the primary unit of evidence on which linguistic theories were built, taking for granted that the acceptability status of these sentences would be immediately obvious to a native speaker (Phillips, 2009). With reference to previously discredited introspective approaches in psychology (see Danziger, 1980), critics pointed out that the intuitions of a linguist about a sentence they constructed themselves to demonstrate a particular point of syntax might not be the same as those of the broader language community (Spencer, 1973; Schütze, 1996; Wasow & Arnold, 2005; Dabrowska, 2010).

Proponents of informal approaches argued in response that linguists were mainly concerned with phenomena that gave very large effect sizes, making multiple opinions on a particular acceptability difference redundant (Phillips & Lasnik, 2003; Phillips, 2009; Featherston, 2009a). Arguing for the acceptance of informal approaches on ‘obvious’ cases moved to defend the legitimacy of a large literature built on such informal tests, but left open the question of how to decide what counts as an obvious case (Linzen & Oseki, 2015).

Recent systematic work comparing expert and naive judgments has largely supported the argument that the majority of claims published in the linguistics literature are consistent with the results of formal tests against the judgments of large numbers of naive native speakers (Culbertson & Gross, 2009; Sprouse & Almeida, 2012; Sprouse et al., 2013). However the same program of research has shown that even for contrasts with large effect sizes, formal tests offer more information than informal ones. As well as giving an objective measure of whether a test sentence is more or less acceptable than a control to a language community, a formal test can also give an indication of the size of the difference, and the relative acceptability of both sentences on a global acceptability scale (Sprouse & Schütze, 2017). It has also been argued that as a result of much productive work

on large effects, smaller effects have become increasingly important to further progress (Gibson & Fedorenko, 2013; Gibson, Piantadosi, & Fedorenko, 2013). Rather than focusing on questions of strict *veracity* of claims with regard to the presence or absence of an effect, as in the original framing of the formal/informal measures debate, it may be more useful to focus on questions of *informativeness* (Featherston, 2007).

The different types of information: veracity of the presence of an effect, effect size, and global acceptability status of test items, all make different demands on the measurement tool. In particular, each comes with some cost in time and participant-hours. Myers (2009) describes a range of possible experimental designs, varying in numbers of participants and the types of responses offered, which could be expected to cover different trade-offs between the information gained and cost of running the experiment. The discussion given by Myers (2009) makes the key point that objectivity and methodological rigor need not be prohibitively expensive or complicated: subsequent work in the same vein (to which the project presented here also belongs) has looked at providing more detailed quantification of the reliability of results for different testing methods and result types.

At one end of this scale, Mahowald, Graff, Hartman, and Gibson (2016) examine the minimal experiment needed to give high confidence in the veracity of an ‘obvious’ effect, the original question in the formal/informal methods debate. They find that five unanimous judgments by independent raters in a forced-choice task gives 95% confidence that endorsement rates for the preferred sentence would be over 75% in a large-n formal study, assuming that the distribution of effect sizes being tested is well represented by a sample drawn from 10 years of *Linguistic Inquiry* articles appearing in Sprouse et al. (2013).

However if the target of a study is the extra information unique to formal tests, the estimates of effect size and overall acceptability status of test items, more responses are needed. These finer-grained questions are addressed in Sprouse and Almeida (2017) specifically focusing on statistical power in hypothesis tests for forced-choice, Likert scale, magnitude estimation, and binary acceptability tasks. In brief, this study found that the forced choice task was the most sensitive to contrasts between conditions, but noted that it did not give global acceptability information, as Likert scales and Magnitude Estimation do. The latter two measures were found to have similar sensitivity.

As described in the previous section, one contribution of the project presented in Chapter 7 was the application of Thurstonian modeling to the sentence acceptability context, which could reasonably be expected to benefit from the high sensitivity of the forced choice task while also supplying global acceptability information. A second contribution was the use of between and within participant replications to test reliability.

Test-retest reliability is strongly related to the Type 1 and Type 2 error rates that are the main focus of Sprouse and Almeida (2017). To the extent that true effects are stable in the world, they drive stable patterns of responses resulting in both low error rates and high test-retest reliability. Test-retest reliability is however distinct from error rates in not appealing to a ground truth, and in the way the contrast across between and within participant test-retest reliability separate out different sources of variability (Bland & Altman, 2007, 1999).

Type 1 and Type 2 error rates are defined with reference to the true state of the world: a Type 1 error is the spurious endorsement of an effect where none truly exists, and a Type 2 error is the rejection of a true effect. As a result, studying error rates is very resource intensive, since validating high confidence ground truths for the small effect sizes of greatest interest for sensitivity studies requires large numbers of responses. For example, the tests presented in Sprouse and Almeida (2017) rely on reference effects established in (Sprouse et al., 2013) and independently replicated by Häussler, Juzek, and Wasow (2016); Mahowald et al. (2016). Sprouse et al. (2013) argues that there is no way to settle the superiority of a measure in a simple comparison study, since inconsistencies between measures offer no information about accuracy. From this perspective, assessments of measure accuracy require a deep prior understanding of the particular effects under study. While this is true of inconsistencies between different measures, the same argument does not apply to tests of self-consistency. In that case, it is not necessary to identify which if either of two different results is accurate to determine that at least one of them is in error. The possibility of consistent errors (for example, consistent failures to detect a small effect) mean that these inconsistency rates do not give the same information as error rates, and do not directly address statistical power. However in combination with null decision rates, test-retest reliability measures offer a broadly applicable metric of measure quality without necessarily requiring large scale studies

validating ground truths for a range of effect sizes.

Another advantage of the test-retest reliability method is that contrasting between participant test-retest reliability and within-participant test-retest reliability offers some information about the source of variability, a contrast which has no analogue in error rates. There are many potential sources of variability when measuring psychological constructs such as a sentence's acceptability, even when the measures are within-participants and are not widely separated in time. The underlying construct may itself be unstable and subject to some inherent variability (Vul & Pashler, 2008). Even for stable constructs, lapses of attention or misunderstanding of task instructions may randomly introduce responses that do not accurately reflect the person's perception of the test sentence. Between-participant replication attempts are subject to all of these sources of variability, plus individual differences in ability or interpretation of the task, and potentially also presentation differences such as item neighborhood effects. By measuring the extent to which between-participant replications vary more than within participant replications with other factors controlled, it's possible to estimate the extent to which results may be impacted by the extra factors specific to between-participant replications.

The next chapter presents the performance of six different sentence acceptability elicitation tasks on these tests of measure quality. This work is related to the previous chapters in its use of comparisons to examine representation structure, but also engages with the primary methodological concerns of linguists working in this area, which center around reliability and efficiency. The tasks include Likert scales, proportion endorsement in targeted forced choice tasks, proportion endorsement in randomized forced choice tasks, Thurstonian modelling of randomized forced choice tasks, and magnitude estimation with or without a z -transformation applied to the resulting scores. Results suggest that all tasks have surprisingly high test-retest reliability, although Likert scales and the Thurstonian model do especially well. Moreover, more fine-grained analyses allow a detailed quantitative description of the various sources of variation impacting acceptability judgments.

Chapter 7

The reliability of acceptability

This chapter reproduces a manuscript currently under review *Quantifying sentence acceptability measures: Reliability, bias, and variability* with authors Steven Langford, Amy Perfors, Andrew Hendrickson, Lauren Kennedy, and Daniel Navarro.

Statement of Authorship

Title of Paper	Quantifying sentence acceptability measures: Reliability, bias, and variability
Publication Status	<input type="checkbox"/> Submitted for Publication
Publication Details	This publication was first submitted to <i>Glossa</i> in April 2017, and a revised version re-submitted in August 2017.

Principal Author

Name of Principal Author (Candidate)	Steven Langsford		
Contribution to the Paper	<p>Articulated the initial experimental hypothesis and designed the initial experiment.</p> <p>Implemented the study and collected all data.</p> <p>Performed all analyses, with direction, including implementation and testing of the model used.</p> <p>Initial drafting of the associated write-up and contributed to subsequent revisions.</p>		
Overall percentage (%)	75%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	18/8/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Amy Perfors		
Contribution to the Paper	I helped design the study (including making suggestions part of the way through that significantly changed its form) and contributed extensively to writing the paper, although Steve did all initial drafts and most of the literature search and review.		
Signature		Date	31/08/2017

Name of Co-Author	Daniel Navarro		
Contribution to the Paper	I helped with the design of the study and made a small contribution to the writing of the paper.		
Signature		Date	31/08/2017

Name of Co-Author	Andrew Hendrickson		
Contribution to the Paper	I helped design the study, discussed how to analyse the results, and made small contributions to editing the paper.		
Signature		Date	31/08/2017

Name of Co-Author	Lauren Kennedy		
Contribution to the Paper	I contributed to the design of the model, contributed to discussion surrounding desirable traits of measurement and made small contributions to editing earlier stages of the paper.		
Signature		Date	1/09/2017

Understanding and measuring sentence acceptability is of fundamental importance for linguists, but although many measures for doing so have been developed, relatively little is known about their variability and bias. In this chapter, I quantify the contribution of different sources of variability by contrasting within- and between- participant test-retest reliability, which allows us to estimate the contribution of individual differences to the total variability of the consensus scores. By contrasting results with and without response-style mitigation in analyses, I further isolate the impact of response styles. By contrasting acceptability scores arrived at using different elicitation tasks, I test potential sources of bias due to properties of the task. The measures examined include Likert scales, two versions of forced-choice judgments, magnitude estimation, and a novel measure based on Thurstonian approaches in psychophysics. I reproduce previous findings of high reliability for effects, and extend these results to the reliability of acceptability differences between individual items. I find that Likert scales have particularly good reliability, and their agreement with the Thurstonian method suggests the limits of the discrete scale does not impose structure on results.

7.1 Introduction

Acceptability judgments have formed a large part of the study of language since at least Chomsky (1965). They are one of many sources of evidence, alongside corpus linguistics (Sampson, 2007), psychological experiments (Noveck & Reboul, 2008), and neuroscience techniques (Shalom & Poeppel, 2007), that each offer distinct and complementary information about language (Arppe & Järvikivi, 2007). One major factor in the popularity of acceptability judgments is the way they allow theories to be tested against artificial constructions that passive observation would rarely or never provide (Schütze, 1996). For instance, acceptability judgments can differentiate between constructions that are ungrammatical and those that are rare or missing but still grammatical.

Acceptability judgments come in a number of possible forms, each with their own advantages and disadvantages. The main differences between different forms are in the kind of response required from the participant. People can be offered a discrete rating scale, a real-valued scale, or be asked to make a relative comparison between items. The choice of what response options to offer is critical in two important respects: it determines the statistical tests available to researchers, and it may also significantly influence people's interpretation of the task. For these reasons, the characteristics of different kinds of acceptability measures are well studied. It is well known that acceptability judgment data are influenced by details such as the selection of participants (Dabrowska, 2010), sample size (Mahowald et al., 2016), task structure (Featherston, 2008), participant engagement (Häussler & Juzek, 2017), and data processing decisions (Juzek, 2015).

Most of the existing literature focuses on the question of to what extent acceptability judgment data can be used to adjudicate about individual phenomena or effects of linguistic interest (e.g., by presenting pairs of sentences that capture a specific contrast relevant to a particular theoretical claim). However, one might be interested in evaluating the range of acceptability measures along other dimensions as well. To what extent do acceptability judgments from different elicitation tasks support claims about the global structure of acceptability across different sentences and phenomena? To what extent do different measures of acceptability agree with each other about specific items or sentences? To what extent is each measure robust to differences within individuals at different time points? This chapter focuses on exploring these questions.

In the work presented here, I attempt to quantify the extent to which acceptability judgment data from a variety of different elicitation tasks supports different kinds of claims: claims about the global structure of acceptability across a large set of diverse sentences, claims based on the magnitude of acceptability differences, and claims made at the level of single items or sentences. This is done by quantifying the relative contribution of multiple factors – individual participant differences, sample size, task structure, and response-style mitigation in data processing – to the empirical reliability of acceptability scores over specific items (rather than over specific effects) for different measures. I chose to focus on reliability because reliability places a ceiling on how appropriate acceptability judgments are as a test of linguistic theories. Just as replicable effects are the foundation of theory, repeatable measurements are the foundation of effects. Moreover, understanding what factors influence the reliability of a measure can be informative about exactly what that measure reflects.

The approach taken here aims to differentiate between possible sources of bias and variance.

It is currently unclear what proportion of the variability seen in acceptability judgment data is due to lapses of attention, idiolect differences between participants, differences in interpretation of acceptability scales, or interference from simultaneously presented items. A standard response to the diversity of potential sources of variability is to give them all equal status as *noise independent of the linguistic effect* and ask what can be concluded about true linguistic effects (focused on specific phenomena) in the presence of this noise, regardless of its source. An extensive literature explores this question, looking at the chance of identifying an effect where none exists (Sprouse & Almeida, 2011; Sprouse et al., 2013), the chance of failing to identify an effect that is truly present (Sprouse & Almeida, 2017), and differences in sensitivity of different measures compared on a particular known effect (Weskott & Fanselow, 2011). The consensus of such studies is that acceptability judgments are highly reliable across replications (Sprouse & Almeida, in press).

As this literature shows, differentiation between different sources of bias and variance is not strictly necessary in order to test specific linguistic **effects**, which are the primary currency of linguistic research. Many measures of sentence acceptability have good psychometric properties when they are used for such a purpose (e.g., testing whether a set of sentences licensed under some linguistic theory have different acceptability than a set of sentences that are not licensed). If such differentiation is not necessary, why attempt to do so here?

The first reason is that such differentiation is important if acceptability judgments are to be used to explore questions that are not focused on hypothesis testing about specific linguistic effects. For instance, it is quite possible that the nature of the elicitation task may impose structure on the *overall distribution* of acceptability scores across multiple kinds of sentences. Thus, understanding to what extent different tasks do this is important for investigations of the global structure of acceptability in language. Such investigations would include issues like whether acceptability is gradient or strongly clustered (Lau et al., 2016; Hofmeister et al., 2013), whether there are dialect or language differences in global acceptability structure, or whether low acceptability sentences show greater variability than high acceptability ones. Indeed, global acceptability judgments (if they are reliable) may even provide a means to differentiate between dialects or evaluate the knowledge or fluency of individual speakers.

The second reason it might be useful to distinguish between different sources of variability is the expectation that some of these sources fall under an experimenter's control and can be minimized. Different elicitation tasks may vary in their vulnerability to particular sources of variability, which affects their relative quality as scientific instruments. In general, a task that is more difficult might be expected to incur greater variability due to distraction or mistaken responding. Tasks with a small number of unambiguous response options, such as forced choice tasks, may be less vulnerable to response style variability than tasks with flexible free response options that are open to differences of interpretation, such as magnitude estimation. Conversely, forced choice tasks may be more vulnerable to item neighborhood effects, with sentences potentially processed differently in the context of a contrast rather than in isolation. How much do these tasks vary and how large are these different sources of variation? My goal is to provide a quantitative answer to this question.

The many possible sources of bias and variability cannot be completely disentangled, since they are generally all present in some unknown degree in every response. I give quantitative bounds for the distinct contribution of certain sources of variability in two different ways.

First, I contrast between and within participant test-retest reliability. Between-participant test-retest reliability is an important metric of measure quality in its own right, since no strong conclusions can be drawn from the results of a measure if it is liable to give different answers to the same question on different occasions (Kline, 2013; DeVellis, 2016; Porte, 2013; Brandt et al., 2014). While distinct in the way it avoids appealing to a ground truth, between-participant test-retest reliability is closely related to error-rate reliability, if the underlying truth is considered stable over the time scales involved. As such, it is widely reported in existing work on the reliability of acceptability judgment data (Sprouse et al., 2013; Sprouse & Almeida, 2017). However test-retest reliability within the same participant can offer additional information, especially when contrasted with between-participant reliability. This contrast, which has no analogue in error rates, is informative about the composition of the variability: variability inherent to the construct itself and random noise due to inattention or other error can be expected in both, while individual differences in response style and subjective acceptability only contribute to the variability of between-participant replications. As a result, between-participant replications are expected to be less reliable, and the size of the reliability gap quantifies the combined impact of these particular

sources of variability. Even further decomposition into the source of this within/between reliability gap is possible as well. For instance, the variability due to response style differences can be estimated by examining the effect of data pre-processing steps (e.g. z -transformation of scores) known to mitigate this particular source of variability.

Second, I contrast these within and between participant test-retest reliability results for measures based on different tasks. The tasks differ primarily in the kind of response options offered, which could potentially impose structure on results. For example, asking people to give responses on a discrete Likert scale might force them to collapse distinct acceptabilities onto one response if there are too few options or encourage them to make spurious distinctions if there are too many (Carifio & Perla, 2008; Schütze, 1996, 2011). The comparisons involved in forced choice judgments could also direct people’s attention to specific syntactic details, particularly when the two sentences are related, as is typical of a well-controlled test pair. This might lead to different acceptability ratings than if each sentence was considered in isolation (Cornips & Poletto, 2005). Contrasts between measures are therefore useful both in identifying the best-performing measures (Sprouse et al., 2013; Sprouse & Almeida, 2017) and to test the degree of agreement between them (Weskott & Fanselow, 2011; Schütze, 2011; Sprouse & Almeida, 2012). However, from the perspective of decomposing sources of bias and variance, distinct tasks may also be differently vulnerable to different sources of variability. As a result, they can be used to cross-check against each other’s potential biases.

The structure of this chapter is as follows. I first give a detailed introduction to the measures considered in this chapter, the processing steps and statistical tests associated with each, and the series of experiments that provide the data. When reporting the results the primary focus is on test-retest reliability; it is first evaluated in terms of raw score correlation of all sentences in a dataset, then in terms of the decisions yielded by each measure on particular contrasts of interest. For each of these I compare within and between participant reliability and examine the impact of sample size. I conclude by examining the mutual agreement between the measures, with reference to expert judgments in the published literature. The discussion explores some limitations of this work, gives recommendations for researchers interested in measuring sentence acceptability, and describes some possible future directions.

7.1.1 The measures

Early work on the reliability of formal measures was prompted by concerns about the practice of ‘armchair linguistics’, which considered phrases or sentences as the primary unit of evidence on which linguistic theories were built, taking for granted that the acceptability status of these sentences would be immediately obvious to a native speaker. With reference to previously discredited introspective approaches in psychology (Danziger, 1980), critics pointed out that the intuitions of a linguist about a sentence they constructed themselves to demonstrate a particular point of syntax might not be the same as those of the broader language community (Spencer, 1973; Schütze, 1996; Wasow & Arnold, 2005; Dabrowska, 2010). Proponents of informal approaches argued in response that linguists were mainly concerned with phenomena that gave very large effect sizes, making multiple opinions on a particular acceptability difference redundant (Phillips & Lasnik, 2003; Phillips, 2009; Featherston, 2009a). This approach defended the legitimacy of the large literature built on such informal tests, but left open the question of how to decide what counts as an obvious case (Linzen & Oseki, 2015).

Recent systematic work comparing expert and naive judgments has largely supported the argument that the majority of claims published in the linguistics literature are consistent with the results of formal tests against the judgments of large numbers of naive native speakers (Culbertson & Gross, 2009; Sprouse & Almeida, 2012; Sprouse et al., 2013). However the same program of research has shown that even for contrasts with large effect sizes, formal tests offer more information than informal ones. As well as giving an objective measure of whether a test sentence is more or less acceptable than a control to a language community, a formal test can also give an indication of the size of the difference, and the relative acceptability of both sentences on a global acceptability scale (Sprouse & Schütze, 2017). It has also been argued that as a result of much productive work on large effects, smaller effects have become increasingly important to further progress (Gibson & Fedorenko, 2013; Gibson et al., 2013).

One potential drawback of formal methods is their higher cost in time and participant-hours. However, as Myers (2009) points out, objectivity and methodological rigor need not be prohibitively

expensive or complicated. Moreover, cost depends in part on the measurement task as well as the question being asked. For instance, many fewer judgments are required for a forced-choice task on an ‘obvious’ effect (Mahowald et al., 2016) than for answering finer-grained questions about statistical power or sensitivity (Sprouse & Almeida, 2017).

My goal in this work was to evaluate all of the most commonly used formal measures of sentence acceptability, as well as variants on them, in order to isolate and expose the impact of task-specific assumptions. The primary distinction between existing measures is whether they ask participants to give each sentence a rating on a scale of some sort (a rating task) or make a choice between two sentences (a choice task). The two rating tasks considered here are LIKERT scales and Magnitude Estimation (ME), while the two choice tasks involve either deciding between two related sentences (TARGET PAIRS) or two random sentences (RANDOM PAIRS). This yields four separate tasks, but for two I separately evaluate alternative statistical methods for transforming the raw results, giving six distinct measures. One task for which I consider multiple analyses is magnitude estimation, where scores can be log transformed (ME(LOG)) or both log and z -transformed (ME(z -SCORE)). The other is the judgments involving random sentence pairs, which can either be used directly or input into a THURSTONE model based on a standard measurement approach in psychophysics.

The six measures, ME(z -SCORE), ME(LOG), LIKERT, THURSTONE, TARGET PAIRS, and RANDOM PAIRS are described in more detail in the Method section. One reason for this choice of tasks is to reflect current practice: LIKERT, TARGET PAIRS, and ME are probably the most common instruments for eliciting acceptability judgments (Podesva & Sharma, 2014). However another consideration is their diversity of assumptions. In particular, LIKERT and ME each supply a particular rating scale, while the choice tasks do not. A key contribution of this project is the presentation of the THURSTONE model, which allows comparisons between these perspectives by inferring scale structure from choice data (Thurstone, 1927). The THURSTONE model is capable of representing a wide range of latent acceptability structures: the degree of consistency between the structure inferred from choice task data and rating task data gives an indication of the extent to which the researcher-supplied scales impose structure on people’s responses.

7.1.2 Measure evaluation

In this section I systematically investigate three criteria for evaluating each of the six measures: test-retest reliability, agreement, and robustness to sample size. Measure agreement is an important check of validity for diverse measures claiming to reflect the same underlying construct. Here I am also interested in the vulnerability of different measures to different sources of noise, with the goal of allowing researchers to minimize the variability in results that are due to controllable properties of the elicitation task rather than the linguistic construct of interest. Although robustness to sample size is not directly related to the decomposition of measure variability and bias that is the main focus of this project, I include it as important information for readers interested in the implications of this work for study design.

Test-retest **reliability** can be defined at various levels from responses (when repeating questions within-participants) to items (an aggregation of many responses) to effects (which aggregate over many theoretically-related items). Here I am primarily concerned with the item level, for several reasons. First, effect-level reliability is already well studied. Second, including only one item per effect (as here) allows us to maximize variability across items and thus creates a much more stronger test of each *measure*. If a measure is highly reliable even across an extremely varied sentence set, this is more informative than finding that it is reliable along a more narrow set of stimuli. Finally, item-level reliability is not itself well-studied, yet is theoretically important: if people’s judgments about specific items are reliable for a given measure, a much wider range of theoretical claims about language are open to study with this data type.

The assessment of reliability depends in part on the nature of the hypothesis being tested. Some researchers might be particularly interested in a *decision* problem: determining whether people make different judgments for two different sentences or kinds of sentences. Others might be interested in an *estimation* problem, being able to accurately position sentences relative to each other on an acceptability scale. Here I evaluate reliability using both kinds of assessment. For a decision problem, I rely on statistical significance testing of the difference between acceptability scores produced by a particular measure for the two sentences. This allows us to precisely characterize uncertainty in the estimate of the difference for each pair of sentences, and compare that degree of uncertainty across measures in a principled way. For estimation problems, I calculate

correlations between scores from different time periods or people. Reliability at this level of detail is relevant to claims about the overall structure of acceptability, for example whether or not it exhibits strong clustering (Sprouse, 2007).

A secondary factor explored here is **sensitivity to sample size**. I do this by systematically repeating the reliability analyses with the judgments derived from different sample sizes of participants and comparing this to the results from the full sample. This is directly useful in estimating the sample size required for a target level of reliability in studies using these measures. It also gives an indication of how efficiently these measures are able to extract information from responses; this is useful because different methods might take different numbers of trials to produce reliable answers (L. Li, Malave, Song, & Yu, 2016).

The final factor of interest is the **agreement** between measures. This is of interest not only because substantial agreement suggests that the measures reflect genuine acceptability judgments rather than superficial measure-specific behavior, but also because such agreement provides converging evidence about the nature of those judgments. Cross-measure agreement is better studied than reliability (Weskott & Fanselow, 2011; Schütze, 2011; Sprouse & Almeida, 2012), but still has not been investigated within the full array of measures considered here. It is therefore valuable as a replication and extension of previous work.

7.2 Method

7.2.1 Sentences

In order for the comparisons to be fair, all of the measures are evaluated on the same set of sentences. Sprouse et al. (2013) selected these sentences from a subset of English data points published in *Linguistic Inquiry* between 2001 and 2010. Sprouse et al. (2013) subdivide these sentences into 148 distinct linguistic phenomena, roughly corresponding to 150 distinct sources (with two instances where different sources discussed the same construction). Each linguistic phenomenon was then represented by multiple items (eight instances). Since the focus here is not on the content of any particular linguistic claim, I selected one matched pair of acceptable/unacceptable items at random from the 150 distinct sources to create a set of 300 sentences. This decision limits my ability to make claims about the status of any particular phenomenon, since each is represented by a single item. However, my focus is on the reliability and variability inherent to specific *measures*, and for this the diversity of sentences is a significant advantage: it is important to evaluate them over the full range of sentence acceptability levels and effect sizes. In addition, with this data it is also possible to estimate the variability associated with individual items. The full list of sentences appears in Appendix A.

7.2.2 Measures

The reliability and sample size analyses involve comparing the six different measures of sentence acceptability described above. When analyzing agreement, I additionally include informal expert judgments from the published literature (INFORMAL). The procedures for deriving scores and significance tests for each measure are given below, followed by the details of data collection. Table 7.1 summarizes this information.

Informal

The INFORMAL measure captures the binary judgments presented in the *Linguistic Inquiry* journal for each of the sentences in question. For each of the 150 pairs, one sentence was judged to be acceptable and one was unacceptable (as noted with a judgment diacritic like * or ? in the journal). I include this measure because of the intense interest in comparing informal and formal methods (Sprouse et al., 2013; Gibson & Fedorenko, 2013; Munro et al., 2010; Myers, 2012; Featherston, 2007; Sprouse & Almeida, 2012), although the main focus is on evaluating the test-retest reliability and mutual consistency of the formal methods. One important caveat for the interpretation of the comparison with INFORMAL results presented here is the fact that each phenomenon is represented by a single example sentence, rather than the multiple items as is the usual practice for formal studies (Myers, 2009). For my purpose here (i.e., investigating item-level reliability and especially the extent to which acceptability judgment data supports tests of global structure), this feature of

the item set is an advantage: to the extent that different instances of the same phenomenon have similar acceptability, using one item per phenomenon gives the maximum variability over the item set and maximum coverage over the acceptability space. However it also means that there is some risk that any specific phenomenon in question will be represented by an atypical example. Since INFORMAL measures are single judgments, test-retest reliability measures do not apply, so they are assessed only in terms of cross-measure agreement.

Likert

In a typical LIKERT task, each sentence is presented with a series of possible acceptability rating options. This task is widely used in the psychological literature (Likert, 1932; Hartley, 2014) and is generally considered fairly intuitive. LIKERT scales are one of the most widely-used formal measures of linguistic acceptability (Schütze & Sprouse, 2014) and have been shown to substantially agree with informal judgments (Sprouse et al., 2013), with experts and non-linguists coming to largely the same conclusions (Culbertson & Gross, 2009). In these experiments, following Sprouse et al. (2013) and Mahowald et al. (2016), I aggregated LIKERT scores by first converting each individual participants' responses to z -scores. The acceptability score for each sentence was thus the average of all z -scores associated with it. This normalization scheme mitigates the impact of individual differences in response style.

Magnitude Estimation

ME tasks were developed to estimate the relative magnitude of differences between items by supplying interval data (Bard et al., 1996). In this procedure, adapted from psychophysics (Stevens, 1956), participants are given an initial reference item to calibrate their judgments and then asked to compare other items by assigning them any positive real number. Although unable to provide true ratio data as initially claimed (Weskott & Fanselow, 2008; Sprouse, 2008, 2011a), ME is still commonly used (Keller, 2003; K. Johnson, 2011; Featherston, 2005; Schütze, 2011; Cowart, 1997; Murphy, Vogel, & Opitz, 2006; Erlewine & Kotek, 2016). Interpreted as a linear scaling task rather than a direct recording of people's mental representations, it is distinct from other measures in the extreme freedom it gives for arbitrarily precise responses, although whether that extra variability actually encodes information about linguistic effects has been questioned (Weskott & Fanselow, 2011). ME has been shown to agree with other forms of acceptability judgment (Keller & Asudeh, 2001; Weskott & Fanselow, 2009).

The typical aggregation scheme for ME data in linguistics, following Bard et al. (1996), is to average the log of the raw scores associated with each item (Sorace, 2010; Weskott & Fanselow, 2011). Originally, this was because the log transform is natural for ratio data, which is the form requested in the instructions to participants. Recent work has shown that participants are in general unable to produce responses conforming to the properties of true ratio scales (Sprouse, 2011a), and it may in fact be impossible to do so since acceptability does not have a clearly defined zero point. I adopt the log transformation here primarily because it has historically been a standard approach for reducing the impact of the outliers typical of ME data. Moreover, other possibilities, such as trimming the data or Winsorizing, would remove information.

In order to evaluate the role played by response style differences, I additionally investigate the impact of also applying a z -transformation, which is sometimes recommended for ME scores for that purpose (Fukuda, Goodall, Michel, & Beecher, 2012; Sprouse & Almeida, 2011; Featherston, 2005). The z -transformation mitigates response style differences in two ways. First, participant ratings are scored relative to their mean rating (which compensates for individual differences in which part of the scale people use) and distances are expressed in standard deviation units (which compensates for individual differences in the range of the scale that they use). By contrasting the test-retest reliability of ME data both with and without the z -transform applied, it is possible to see how effective it is in mitigating response style differences. Specifically, the z -transformation is predicted to improve reliability in the between-participant replication to a much greater extent than the within-participant replication, since response style differences are a between-participant source of variability. To the extent that it is effective, this contrast gives an indication of the degree to which variation in ME scores can be attributed to variability in people's usage of the scale.¹

¹I also ran all of these analyses with raw judgments (no transformations at all), judgments receiving only the z -transform (rather than z and log), or judgments that were converted to ranks. None had superior reliability than LIKERT or THURSTONE, and judgments that did not incorporate some way of taming outliers did not produce

Target pairs

The TARGET PAIRS judgment task asks people to select the more acceptable sentence of two candidates specifically chosen to isolate a particular contrast of theoretical interest. This is perhaps the simplest measure. By focusing only on the differences that are of theoretical interest, this measure increases the statistical power for determining the differences within those targeted pairs, but sacrifices the ability to compare pairs to one another. The TARGET PAIRS comparison is widely used (Rosenbach, 2003; Myers, 2009) and has been shown to substantially agree with informal judgments (Sprouse & Almeida, 2011; Sprouse et al., 2013).

Acceptability scores in the TARGET PAIRS task are considered to be the proportion of times the preferred option was chosen, without distinguishing between responses indicating equal acceptability or the alternative option. Unlike the other aggregate scores, this measure does not capture global structure, since decisions regarding each pair are isolated by design. The primary outcome of interest for this measure is the outcome of the significance test of the estimated proportion (\hat{P}) with respect to the number of judgments (N). This was calculated by determining if the 95% confidence interval around the estimated proportion included random guessing (0.5). If the interval did not include 0.5, the null hypothesis that people did not prefer one sentence over the other was rejected. The standard formula for calculating a confidence interval around a proportion was used:

$$\hat{p} \pm Z_{crit} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \text{ where } Z_{crit} = 1.96.$$

Thurstone

The THURSTONE measure, which has a long history in psychophysics (Thurstone, 1927; Roberts, Laughlin, & Wedell, 1999; Fabrigar & Paik, 2007), is used to make inferences about the subjective perception of stimuli based on forced-choice comparison data. The basic idea is to ask people to make acceptability judgments about a random subset of pairs drawn from a large set of stimuli (for us, this corresponds to asking people to give forced-choice judgments on two sentences sampled at random from the full set of 300 sentences). It is important that the pairs are random rather than the theoretically-motivated pairs as in the TARGET PAIRS task because comparing each sentence to many others imposes strong statistical constraints on the set of possible orderings of all of the sentences (Thurstone, 1927). Distances on the inferred acceptability scale are given meaning by the model's mapping between acceptability differences and probability of endorsing a particular response. The observed responses constrain the plausible outcomes for responses in unobserved comparisons, assuming transitivity of acceptability. As a result, only a small subset of all possible pairs is necessary to make inferences about the acceptability of all of the sentences relative to each other. Technical details for the THURSTONE measure are described thoroughly in Appendix B.

The THURSTONE model represents the acceptability of each sentence as a distribution over its inferred acceptability scale. To derive an overall acceptability score from this measure, I simply take the mean of each distribution as representing the acceptability score for that sentence. For a decision rule corresponding to the significance test in other measures, I constructed a credible interval over the difference between sentences. The distribution of credible differences was generated by repeatedly sampling from each posterior and taking the difference of those samples. The result was considered inconclusive if the range between 0.025 and 0.975 quantiles included 0, otherwise the observed difference was considered significant.

The THURSTONE model has a well-established record of performance in other domains that require inferring latent acceptability orderings, such as product preferences in marketing research (O'Mahony, 2003; Ennis, 2016). It is also a prominent tool in the 'wisdom of crowds' literature, where it is used to define a meaningful consensus aggregating over individual judgments that cannot be simply be averaged together (Miller, Hemmer, Steyvers, & Lee, 2009; Selker et al., 2017). Previous work on experimental syntax methodology has identified forced choice tasks as a particularly sensitive and reliable method of eliciting acceptability judgments (Sprouse et al., 2013; Schütze & Sprouse, 2014), while noting that they are restricted in the way they give limited ordinal information about only the particular sentences involved in the contrast at hand. The THURSTONE method retains main benefits of this task type, which are the simple unambiguous response options and the way individual items can target arbitrarily small acceptability differences, while also aggregating information over all responses to derive a real-valued acceptability score that is directly comparable over all items. By providing real-valued data on a psychologically

meaningful results.

Task	Measure	Sentence Score	Hypothesis test
Targeted contrasts	TARGET PAIRS	Proportion endorsements	Difference of proportions
Random contrasts	RANDOM PAIRS	Proportion endorsements	Difference of proportions
Random contrasts	THURSTONE	Mean posterior acceptability	Credible interval
Magnitude estimation	ME(LOG)	Mean of log responses	t-test
Magnitude estimation	ME(z -SCORE)	Mean of z -transformed log responses	t-test
Likert rating	LIKERT	Mean of z -transformed ratings	t-test

Table 7.1: Method summary: I examined four different tasks, two choice tasks and two rating tasks, analyzing two of these in two different ways for a total of six distinct measures. For each of these measures, I evaluate the set of acceptability scores for all sentences (which supports comparisons using Pearson correlations) as well as decisions made on pairs of sentences (which allows us to focus on targeted contrasts between two particular sentences in a hypothesis-testing framework)

meaningful scale (Borg & Groenen, 2005; Nosofsky, 1992), THURSTONE modeling draws on much of the same motivation that originally drove the adoption of ME (Schütze, 2011). By shifting the responsibility for quantifying acceptability from participants to a measurement model, it avoids problems associated with the difficulty people have using the number line in the requested way (Sprouse, 2008, 2011a).

Random Pairs

The THURSTONE model requires choice task data over random pairs rather than the theoretically related pairs that are usually compared in a choice task. This means that there is a dataset – the raw scores on RANDOM PAIRS – which can provide a baseline against which to compare the THURSTONE and the TARGET PAIRS measures. Analyzing the RANDOM PAIRS measure may be helpful in both determining how much of the performance of the THURSTONE measure depends on the model, as well as in quantifying the impact the choice of contrast sentence has in the TARGET PAIRS task.

As in the TARGET PAIRS measure, the proportion of trials in which a sentence was endorsed over the alternative or the both-equal option was taken as its overall acceptability score. Unlike TARGET PAIRS, this is an estimate of global acceptability across the whole set of sentences considered, albeit a noisy measure that depends on the randomly sampled set of alternative sentences each sentence appeared with. The significance test applied was the same test of proportion equality applied to TARGET PAIRS.

7.2.3 General procedure

To examine within and between participant reliability, three data sets were needed, an INITIAL reference set, followed by a WITHIN PARTICIPANTS replication and a BETWEEN PARTICIPANTS replication. Participants involved in the WITHIN PARTICIPANTS replication gave the series of acceptability judgments used in the INITIAL dataset. They then performed a short distractor task designed to interfere with their ability to remember their answers to particular items, after which they repeated the same set of acceptability judgments (in a different random order) to create the WITHIN PARTICIPANTS data set. A second group of participants was recruited to supply the BETWEEN PARTICIPANTS data set: the same procedure was used, except that these participants did not see the distractor task or give a second set of judgments.

In order to keep the time commitment per participant under approximately 30 minutes, I divided the four tasks into two groups that were presented to the same set of participants, with the RANDOM PAIRS task grouped with the LIKERT rating task, and the TARGET PAIRS task grouped with the ME task. With these groupings each participant saw one choice task and one rating task, which minimized possible fatigue due to always making the same type of judgment or interference between similar task types. Furthermore, requiring participants to complete more than one task increases the time and attention expended between responses to identical items. This decreases the chance that responses reflect an explicit memory of the first judgment for the WITHIN PARTICIPANTS replication.

In the first set of trials (the first half of the experiment for the INITIAL/WITHIN PARTICIPANTS group, the entirety of the study for the BETWEEN PARTICIPANTS group) participants saw two blocks (one rating task and one choice task) of 42 trials each. The order of tasks within a block was randomized for each participant, and the order of items within each task was randomized on

each presentation of a block. Each block contributed 40 trials to the data analysis. The additional two questions were attention checks designed to have a clear correct answer, used only to exclude participants whose incorrect responses indicated either inattention or misunderstanding of the task (see Appendix A). Participants involved in the BETWEEN PARTICIPANTS study completed at this point, while those involved in the WITHIN PARTICIPANTS study then did the distractor task, followed by a repetition of the exact same trials, with the same task order as the initial presentation but a re-drawn random order of items within each task. No sentences were repeated in different items for any one participant. Each participant thus saw only a random subset of the 300 sentences, but across participants all sentences were seen a similar number of times.

The distractor task was based on a change blindness demonstration (Simons & Rensink, 2005). It was chosen because it is non-linguistic and known to be a very attention-grabbing task (Rensink, O'Regan, & Clark, 1997). During it, people were shown two images that were identical except for one difficult-to-identify discrepancy: for instance, one showed a city street in which the window of one of the buildings was present in one image and absent in the other. The images were presented sequentially and repeatedly for 800ms each with an 800ms white mask in between. Participants were asked to identify the discrepancy and click on it. Once they had done so or thirty seconds had elapsed, they were shown another pair of alternating images. There were six such images. Because the point of this task was just to provide a break between the acceptability judgment tasks, performance was not analyzed.

In all conditions participants saw the same general set of instructions, shown below:

This study will ask you some questions about the acceptability of sentences. There's no objective standard for what makes a phrase feel 'more acceptable', but we're confident that you'll know it when you see it. Some phrases are natural while others are clumsy or just plain wrong, and we expect you'll find it pretty easy to judge how acceptable a phrase is, even across very different topics. There are two different types of question. Some of the questions will ask you to give a sentence an acceptability rating. Others will ask you to compare two sentences and say which one is more acceptable.

All participants were asked to answer two multiple choice questions to make sure they understood the instructions (see Appendix A) before beginning the experiment. Those who did not answer both questions correctly were returned to the instructions page and could not begin until both were answered correctly.

7.2.4 Task-specific procedures

Random pairs blocks

In these blocks, people were presented with three vertically arranged options. Each was surrounded by a blue border under the title "Which of the two sentences is most acceptable?". The first two options were sentences randomly drawn from the full pool of 300. The third option read "These two sentences are equally acceptable." Participants clicked on a sentence to choose it, as shown in Figure 7.1(b). As in the LIKERT blocks, a progress marker indicating the item and block number was displayed, and no feedback was given.

Target Pairs blocks

These trials were exactly the same as the RANDOM PAIRS trials in the other version of the experiment. The only difference is that the sentences were both in the pair of theoretical interest rather than randomly selected from the entire set; an example is shown in Figure 7.1(d).

Likert blocks

In these blocks, on each trial people saw a single sentence surrounded by a blue border under the title "Please rate the acceptability of this sentence." Under the sentence was a row of five unmarked buttons labeled "Bad" at the far left and "Good" on the far right, as shown in Figure 7.1(a). Below this was a progress marker giving the trial and block number. Clicking any of the response buttons disabled them for 500ms and displayed the next sentence to be judged. No feedback was given.

<p style="text-align: center;">(a) Likert example</p> <p>Please rate the acceptability of this sentence.</p> <p style="border: 1px solid black; padding: 2px;">Which coworker did George yawn before insulting?</p> <p style="text-align: center;">Bad <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Good</p>	<p style="text-align: center;">(b) Random pairs example</p> <p>Which of the two sentences is most acceptable?</p> <p style="border: 1px solid black; padding: 2px;"><input type="radio"/> Larry cooked her husband the meal</p> <p style="border: 1px solid black; padding: 2px;"><input type="radio"/> Who brought what?</p> <p style="border: 1px solid black; padding: 2px;"><input type="radio"/> These two sentences are equally acceptable.</p>
<p style="text-align: center;">(c) Magnitude estimation example</p> <p>If this sentence gets an acceptability rating of one hundred...</p> <p style="border: 1px solid black; padding: 2px;">Who said that my brother was kept tabs on by the FBI? 100</p> <p style="text-align: center;">...what should these get?</p> <p style="border: 1px solid black; padding: 2px;">The book is long and the essay is short. ?</p> <p style="border: 1px solid black; padding: 2px;">The virtuoso practices any pieces only rarely. ?</p>	<p style="text-align: center;">(d) Target pairs example</p> <p>Which of the two sentences is most acceptable?</p> <p style="border: 1px solid black; padding: 2px;"><input type="radio"/> What did John wonder what he bought?</p> <p style="border: 1px solid black; padding: 2px;"><input type="radio"/> John wondered what he bought.</p> <p style="border: 1px solid black; padding: 2px;"><input type="radio"/> These two sentences are equally acceptable.</p>

Figure 7.1: **Example trials for each of the four question types.** In one version of the experiment ((a) and (b)), participants saw blocks of sentences presented in random order in a LIKERT task and a choice task in which the sentences were randomly drawn from the entire sentence pool (RANDOM PAIRS). In the other version ((c) and (d)), the blocks were either in a magnitude estimation (ME) or typical choice task in which the sentence pairs were theoretically motivated (TARGET PAIRS). For each measure, the associated panel reflects the appearance of a typical trial

ME blocks

In these blocks, people saw six pages of seven sentences each. On each page the top of the screen contained a fixed title banner that remained in position when the page was scrolled. It consisted of some reminder instructions (“If this sentence gets an acceptability rating of one hundred...[reference sentence] ... what should these get?”). The reference sentence, following Sprouse (2011b), was “Who said that my brother was kept tabs on by the FBI?”), and was surrounded by a black border that also contained a non-editable text box in the lower right corner that was pre-filled with the reference value 100. This was followed by the test items, which were surrounded by a blue border and contained an editable text box in the lower right corner initially containing a question mark. An example is shown in Figure 7.1(c).

Test items were arranged vertically with seven to a page with approximately two or three test sentences visible at once on the screen and the remaining sentences visible by scrolling. Each set of seven sentences was followed by progress marker and a next button which presented a fresh set of seven sentences, with no option to return to a previously rated set. Input was restricted to positive numbers, and no feedback was given, other than a prompt to give positive number ratings in order to continue if an unparsable or empty input was detected when the next button was clicked.

In order to ensure that people understood the ME task, before they rated any sentences each participant practiced the task on line lengths. They were required to give ratings for six different test lines (relative to a reference line length of 100). There were five test lines presented in random order, with lengths ranging between $\times 0.01$ and $\times 2.5$ of the reference line. Although the exact lengths of test lines were randomized to avoid encouraging participants to only use round numbers, there was one example each of very short (length $\sim 25\%$ of the reference line), short ($\sim 75\%$), roughly equal ($\sim 125\%$), long ($\sim 175\%$), and very long ($\sim 225\%$) lines². During these

²The ME specific instructions were:

Some of the questions will ask you to compare the acceptability of each sentence to a standard reference sentence and tell us the result as a number. The standard reference sentence always has an acceptability rating of 100. A sentence that is twice as good should get a rating that is twice as large, a sentence that is half as good should get a rating that is half as large, and so on. Any positive number is a valid rating, please do try to use a wide range of numbers. More detailed responses carry more information about how acceptable you feel the sentences are, and that’s really what we’re interested in. Having

practice trials there was feedback on every response, and people did not continue to the next trial until their estimates were within 30 of the correct answer. Participants successfully completing this practice were considered to have understood the process of ME.

Participants

There were four rounds of recruitment to cover the two pairs of tasks (LIKERT/RANDOM PAIRS and ME/TARGET PAIRS) in two presentation formats (a two-session format giving INITIAL and WITHIN PARTICIPANTS data, and a single-session format giving BETWEEN PARTICIPANTS data).

Two-session LIKERT and RANDOM PAIRS 150 adults were recruited via Amazon Mechanical Turk. Participants were paid US\$3.00 for an average of 33 minutes work. They ranged in age from 20 to 65 (mean: 34.6) and 81 of them (55%) were male. Fifteen people were excluded from the analysis: three had non-compatible browsers so their data failed to save, one reported being a non-native English speaker, and 11 gave at least one incorrect response to the attention check questions. Of the 135 remaining participants, 133 were from the US and two were from India. Three reported speaking additional languages other than English but all 135 included participants reported being English native speakers.

Two-session ME and TARGET PAIRS 160 adults were recruited via Amazon Mechanical Turk. Participants were paid US\$4.00 for an average of 38 minutes work. They ranged in age from 19 to 66 (mean: 34.0) and 91 of them (57%) were male. Twenty-five people were excluded from the analysis: one reported being a non-native English speaker, two were found to have participated in the previous round, and 22 gave at least one incorrect response to the attention check questions. Of the 135 remaining participants, 132 were from the US, with one each from India, Chile, and Ireland. One reported speaking an additional language other than English but all 135 included participants reported being English native speakers.

Single-session LIKERT and RANDOM PAIRS 150 adults were recruited via Amazon Mechanical Turk. Participants were paid US\$1.60 for an average of 21 minutes work. They ranged in age from 22 to 69 (mean: 34.5) and 93 of them (62%) were male. Twenty-three people were excluded from the analysis: two had participated in a previous round, four reported being non-native English speakers, and 17 gave at least one incorrect response to the attention check questions. Of the 127 remaining participants, 125 were from the US, one was from Dominica, and one was from India. Two reported speaking additional languages other than English but all 127 included participants reported being English native speakers.

Single-session ME and TARGET PAIRS 151 adults were recruited via Amazon Mechanical Turk. Participants were paid US\$3.00 for an average of 31 minutes work. They ranged in age from 18 to 70 (mean: 34.8) and 89 of them (59%) were male. Fourteen people were excluded from the analysis: four reported being non-native English speakers, and 10 gave at least one incorrect response to the attention check questions. Of the 137 remaining participants, 135 were from the US with one participant from Canada and one from the United Kingdom. Three reported speaking additional languages other than English but all 137 included participants reported being English native speakers.

7.3 Results

I begin by examining the test-retest reliability of the scores derived from each measure. For these analyses, I use the Pearson correlation between scores drawn from the relevant data sets: INITIAL and WITHIN PARTICIPANTS for **within participant reliability** or INITIAL and BETWEEN PARTICIPANTS for **between participant reliability**. Reliability at this level of detail may be required to test claims involving comparisons over more than two items, such as whether or not acceptability exhibits strong clustering, or claims expressed in terms of the degree of difference between items rather than the binary presence or absence of a difference (Gibson et al., 2013; Sorace & Keller, 2005).

said that, you don't need to spend a lot of time doing a deep analysis of every little detail, we're much more interested in your first impressions.

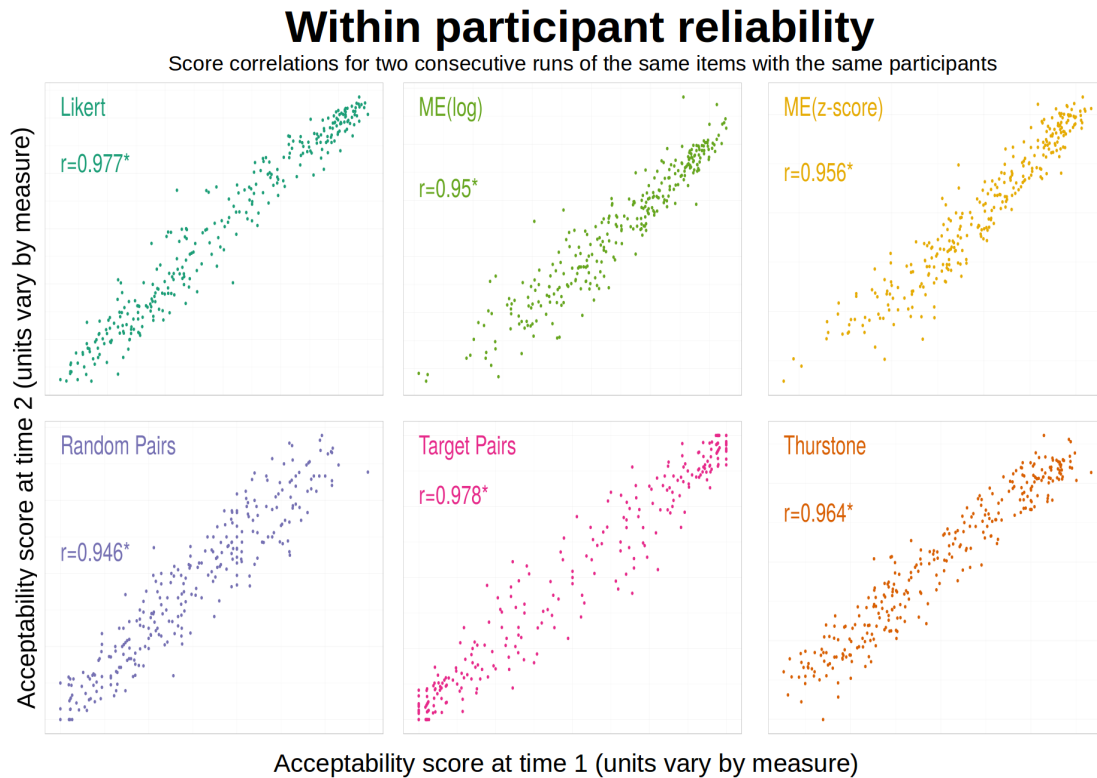


Figure 7.2: **Within-participant reliability measured by correlations between sentence acceptability rankings.** All of the formal measures aggregate responses into an acceptability score for each sentence. For each, the x -axis reflects the score using that measure in the INITIAL data, while the y -axis reflects the score from that measure in the WITHIN PARTICIPANTS data. The r values indicate Pearson's correlation coefficient, and the stars (*) indicate significance at $p < 0.001$. All measures are both highly linear and highly significant, suggesting that all these measures have good within-participant reliability

7.3.1 Global measures

Reliability

I quantify the global reliability of a measure across different data sets using the Pearson correlation between acceptability estimates. Correlations between scores obtained between scores based on the INITIAL dataset and those based on the WITHIN PARTICIPANTS replication data are shown in Figure 7.2, with the score based on INITIAL responses on the x -axis and scores based on WITHIN PARTICIPANTS replication on the y -axis. The strong linear relationships obtained show that all measures were highly reliable. Test-retest correlations were large and statistically significant for every measure. LIKERT scores and TARGET PAIRS were the most reliable measures.

Correlations between scores obtained in INITIAL data and BETWEEN PARTICIPANTS data are shown in Figure 7.3, with scores obtained from the INITIAL data on the x -axis and scores obtained from the BETWEEN PARTICIPANTS replication on the y -axis. As in the WITHIN PARTICIPANTS case, all measures were highly reliable, with all correlations large and statistically significant. However, each correlation is somewhat lower than the within-participant counterpart. This extra variation must be driven by those factors unique to the BETWEEN PARTICIPANTS case: either individual differences among the participants in the two participant pools or item effects due to the re-drawing of the items shown to participants (within-participants tests used identical items each time).

Given that all the measures seem to be relatively reliable, it is natural to test whether the relative differences in reliability can be considered significant. One way to test the significance of the differences observed between these correlations is to bootstrap 95% intervals around them. I used the R package *boot* (Davison & Hinkley, 1997) to generate adjusted bootstrap percentile intervals (BCa) around the r^2 estimate of variance explained in re-test scores given only scores from the INITIAL data set, assuming linearity.

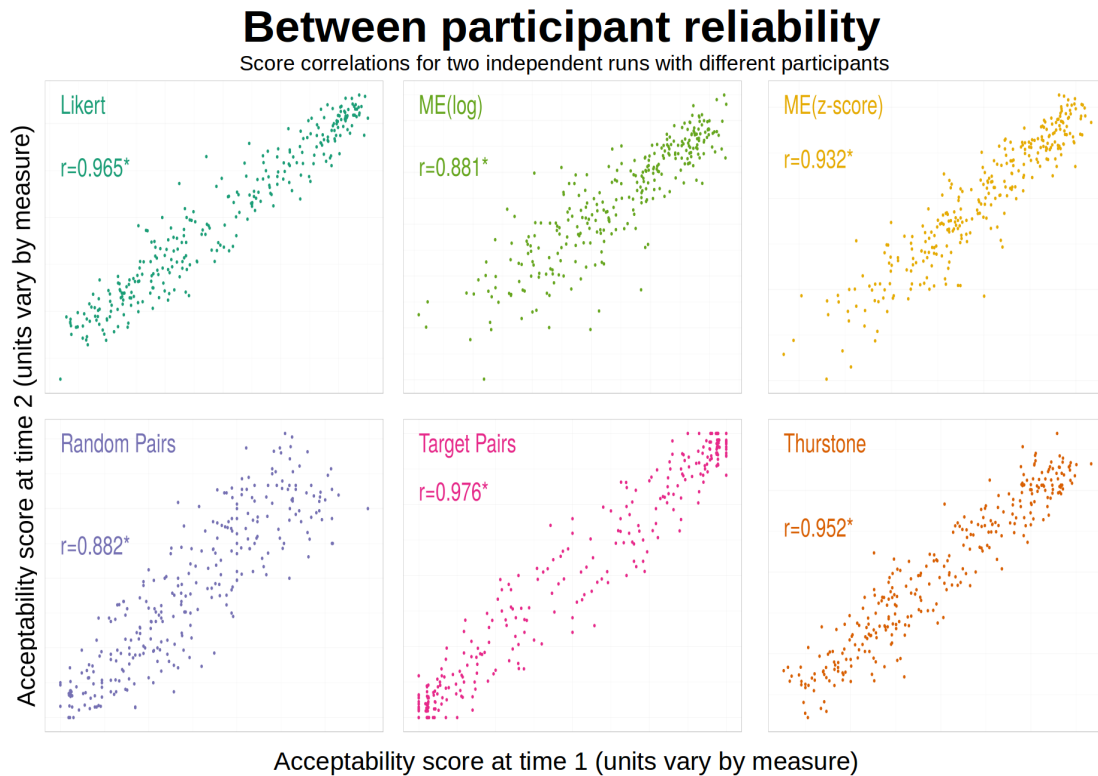


Figure 7.3: **Between-participant reliability measured by correlations between sentence acceptability rankings.** All of the formal measures aggregate responses into an acceptability score for each sentence. For each, the x -axis reflects the ranking derived using that measure in the INITIAL data. The y -axis reflects the score from that measure in the BETWEEN PARTICIPANTS data. The r values indicate Pearson’s correlation coefficient, and the stars (***) indicate significance at $p < 0.001$. Between-participant reliability is naturally lower than within-participant reliability for all measures, but the relationship between scores derived from the two data sets are still linear and highly significant. All these measures show good between-participant reliability

Results are shown in Figure 7.4. There is a significant difference in correlations between within-participant and between-participant reliability for the ME and RANDOM PAIRS measures but not the other ones. The relatively large drop in correlation when moving to from within to between-participant data for the ME scores is most likely driven by individual differences in response styles, as shown by the large reduction in this gap when mitigating response style differences using ME(z-score). In contrast, there is limited scope for response style differences in the RANDOM PAIRS task, so the large drop in reliability when moving from within to between-participants data is likely to reflect the fact that new items were drawn, which gave each sentence a different set of comparison sentences.

What is most noteworthy about these effects is the fact that the LIKERT and THURSTONE scores do not appear to suffer from them. Despite LIKERT ratings being vulnerable in theory to response style differences, these results suggest they do not appear to be a major source of variation in practice. Although the THURSTONE acceptability estimates are based on exactly the same responses the RANDOM PAIRS endorsement proportions derive from, they do not show strong item effects, which is a testament to the robust nature of the THURSTONE approach.

Sample size dependence

The analyses so far yield estimates of between- and within-participant reliability of global sentence acceptability judgments for each measure, but all involve the full sample of judgments derived from all included participants. Although even this quantity of judgments is relatively cheap and straightforward using platforms such as Amazon Mechanical Turk, it is important to understand how robust reliability is when sample sizes are lower. By repeatedly dropping some subset of

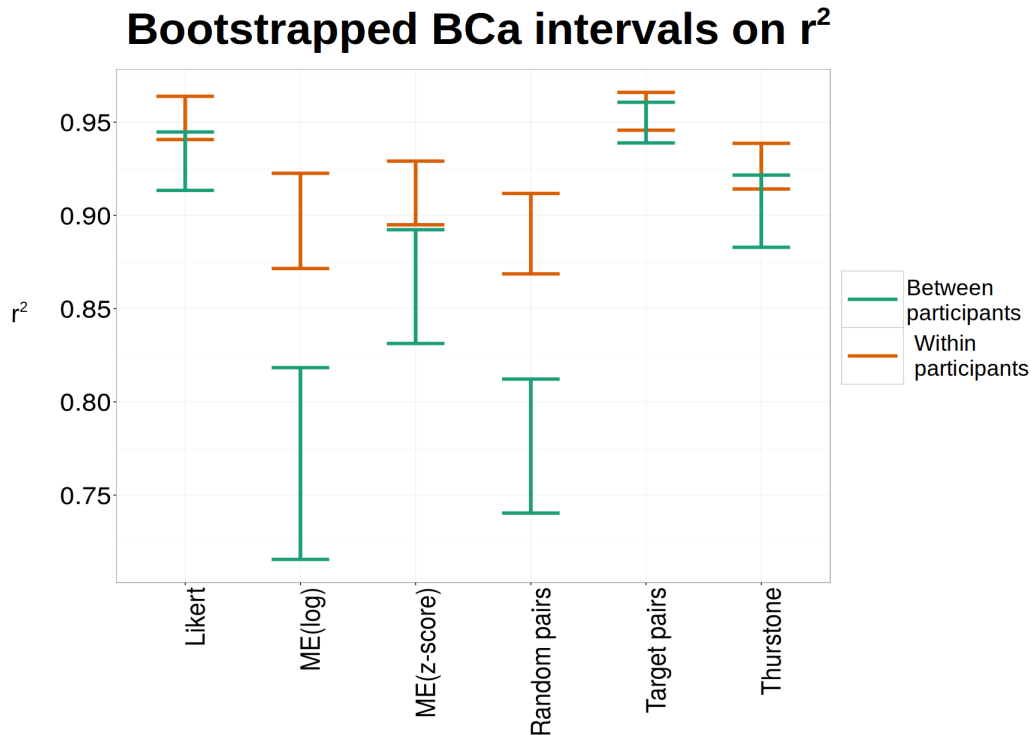


Figure 7.4: **Comparing reliability correlations with bootstrapped r^2 .** The correlations indicating score reliability were compared across scores and across between/within contrasts by bootstrapping a 95% BCa interval. These intervals are of the r^2 for the linear model predicting scores in the second data set (WITHIN PARTICIPANTS or BETWEEN PARTICIPANTS) from scores derived from the INITIAL dataset, and are based on 1000 samples. The results show that ME scores and RANDOM PAIRS scores are significantly impacted by participant and item effects where the other scores are not. TARGET PAIRS is the single most reliable measure. Of the measures allowing global comparisons, the most reliable is LIKERT

participants at random from the full sample and re-running all analyses on the retained participants only, I obtained estimates for the number of participants required for a given level of reliability up to the level achieved in the full sample. These required-sample-size estimates are directly useful for researchers planning future studies, and also give an indication of how efficiently each measure extracts information from its input. All measures can be expected to asymptote to some maximum level of reliability given the underlying variability of responses, with more efficient measures approaching this maximum more quickly.

I explored sample size by performing a sub-sampling procedure in which only a subset of participants were drawn without replacement³ from the total population (of around 150) at sample sizes ranging from 30 to 120 in increments of 10. Only the subset of judgments was used to derive the reliability estimates. I carried out 30 repetitions of the sub-sampling procedure at each sample size and averaged them to estimate the reliability measures at that sample size. Although this smooths out variation associated with the random choice of participants retained, it does not fully reflect the variability expected at each sample size because the repetitions cannot be totally independent. Especially when the sub-sample is a large proportion of the full sample, there is extensive overlap in the data retained across iterations. Sub-sampling was also constrained to only allow samples where every item appeared at least once so that an acceptability score was always computable for each sentence and the targeted comparisons were guaranteed to be feasible.

As Figure 7.5 shows, reliability decreased for every measure with decreasing sample size, but less reliable measures also showed larger decreases and the drops were higher for between-participant

³Other work (Sprouse & Almeida, 2017) draws samples with replacement for similar analyses, but unlike the work presented here, their items were organized into lists, preserving an even distribution across people. Because in the current study participants all rated different sets of sentences, sampling people multiple times greatly distorts the distribution of items within the dataset in a way that they would never be distorted had that been the target sample size.

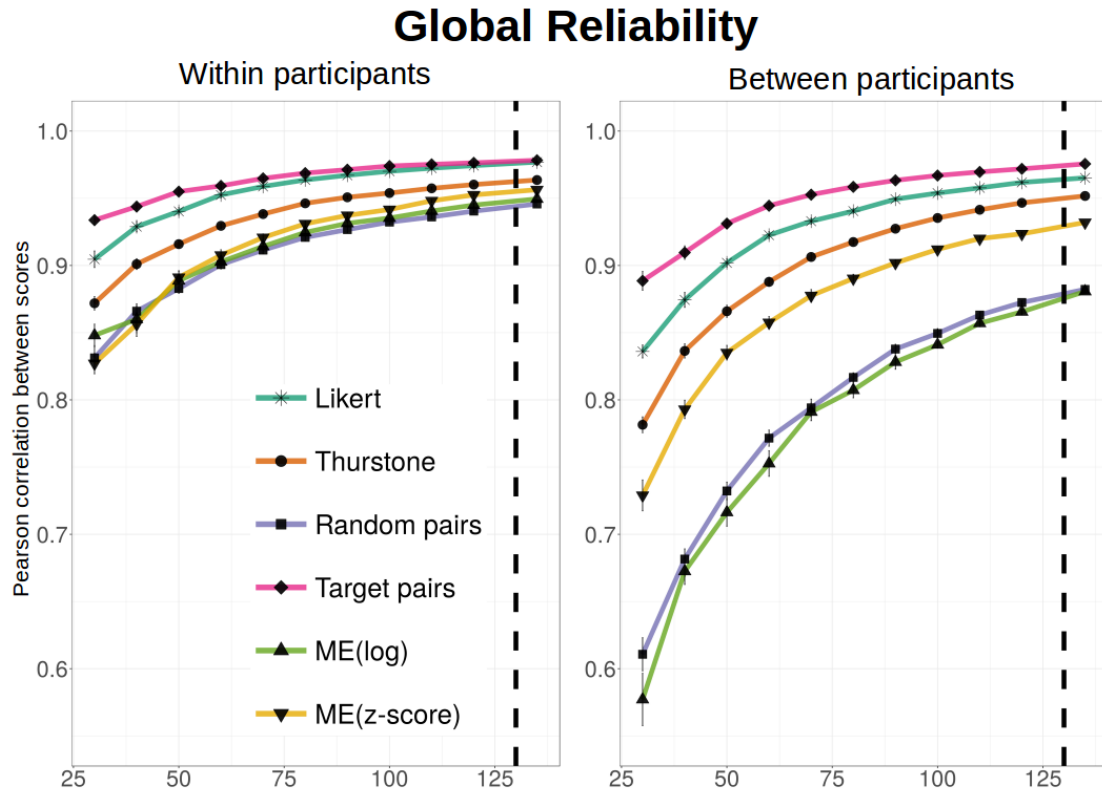


Figure 7.5: **Global reliability measured by Pearson correlation.** Pearson correlations (y-axis) between acceptability estimates based on different data sets were used to quantify the reliability of each measure over different sample sizes (x-axis). Results to the right of the dotted vertical bar are based on the full sample, results to the left are averages of 30 samples of the designated size drawn from the full sample. For all measures, smaller sample sizes are less reliable, but gains in reliability from increasing sample size become progressively smaller. **Within-participant reliability** (left panel) shows variability in estimates based on responses from the same people to the same items, and can be interpreted as variability deriving from the difficulty of the task and the inherently probabilistic nature of people’s responses. **Between-participant reliability** (right panel) is subject to the same sources of noise plus individual differences and variability introduced when re-drawing the items presented, so contrasting within and between participant reliability indicates the vulnerability of each measure to these extra factors. These results show that RANDOM PAIRS and ME scores are particularly vulnerable to participant and item differences, and that TARGET PAIRS and LIKERT ratings are consistently most reliable

reliability. Reassuringly, the relative ordering of measures did not change and most became reasonably close to their performance on the full dataset at samples between 50 and 100 people. These results also suggest that the most reliable measures are most efficient, as they approach their maximum reliability more quickly in the number of responses.

Discussion

Overall, all of the measures have high test-retest reliability, especially LIKERT, THURSTONE, and TARGET PAIRS; the most reliable judgments are obtained by TARGET PAIRS. This task is unusual in not offering acceptability scores that are comparable across all sentences: of the measures that do offer global comparisons, LIKERT scores are most reliable. RANDOM PAIRS and the two ME scores were the least reliable. Contrasting the within and between-participant r^2 values suggests one possible reason: these scores are particularly vulnerable to individual response style or item effects. Of these two possibilities, individual differences in response style is likely to be the major contributing factor for ME, as shown by the way the z -transformation improves reliability and reduces the gap between within and between participant reliability. The RANDOM PAIRS measure is more likely to be showing item effects. There is little scope for response style differences in a choice task, but the measure is clearly sensitive to the changing identity of the alternative choices,

which were re-drawn for the new participants.

In principle, LIKERT scales are also vulnerable to response style differences, and since the THURSTONE scores are based on the same input as random-pairs they are exposed to the same item effects. However both measures include protection against these influences: z -transformation in the case of LIKERT scores and the inferred latent scale for the THURSTONE scores. These results suggest that in practice these protections are effective.

Examining subsets of participants shows that the relative reliability of the different measure types does not change with sample size, and that the most reliable measures were also the least impacted by the number of participants. For the sentences considered here, reliability scores approached their maximum values at approximately 100 participants, which with 40 trials per participant and 300 items corresponds to an average of 13.3 trials per item. The degree of variability in responses might be expected to vary with the particular sentences used, so this relation between reliability and number of trials per item holds only to the extent that the sentences considered here represent a typical range of acceptability for research targets.

7.3.2 Decision measures

Global reliability is useful when testing claims applicable to diverse collections of sentences, but some hypotheses are most naturally tested with targeted contrasts between particular pairs of sentences. Does each measure yield the same *decision* about which item of a pair is more acceptable? This sort of targeted comparison can expose changes in acceptability due to a particular syntactic manipulation while controlling for other factors like length, plausibility, and complexity. The global scores discussed above do allow pair-wise contrasts based simply on the difference between two acceptability scores, but for targeted contrasts researchers would typically conduct a measure-specific significance test instead. These are preferable because they take full advantage of a researcher's knowledge of the test structure to appropriately characterize the variability associated with the acceptability estimates, which in turn offers control of the long-run Type 1 error rate.

Of course, if a researcher's primary goal was to evaluate a particular theoretical claim, they would present participants with multiple item pairs that all instantiate the syntactic manipulation of interest, rather than the one-item-per-effect that I have evaluated here. It is nevertheless interesting for us to evaluate the decision reliability of *items* as shown here, for several reasons. First, if items are highly reliable across tests or people, that is both noteworthy and highly reassuring about whether *effects* might also be reliable. Second, looking at item-level decision reliability is still informative about the overall reliability of each measure, and can tell us about the sources of variability within each measure.

These analyses investigate the reliability of each of the measures with regard to the decisions a researcher would draw based on a significance test for a contrast of interest. The particular significance tests I used differ for each measure as described in the *Measures* section in Table 7.1: some involve t-tests while for others, because of the structure of the data, more complex analyses are necessary. I consider only the 150 targeted contrasts used in the targeted pairs task, reflecting the particular linguistic phenomena under investigation in the original *Linguistic Inquiry* articles. Since the main focus here is the reliability of decisions rather than the content of any particular decision, I did not control for multiple comparisons in any of these tests, mimicking the situation that would obtain if each contrast was being studied independently. As in the previous analysis, I contrast within and between-participant reliability.

There are three outcomes relevant to the test-retest reliability and sensitivity of such decisions: the number of *inconsistent* decisions across time points, the number of those inconsistencies which involve decision *reversals*, and the *null decision count*. Each individual decision admits three possible outcomes: option A is more acceptable, option B is more acceptable, or the null hypothesis of no difference cannot be rejected. For each measure, I evaluate the number of decisions (of 150 pairs) which were **inconsistent** (i.e., at one time option A was selected by the measure but at the other time either option B or the null was). An inconsistent measure indicates that an error of some sort (either Type 1 or Type 2) was made at some point, but it in most cases it is impossible to determine what kind of error it was. Flipping from option A to option B, a **reversal** indicates a Type 1 error, and is quite rare: no measure produces a reversal on the full dataset. Flipping from a null to a non-null result could be either a Type 1 error (if the non-null result was incorrect) or a Type 2 error (if the null result was incorrect).

An indication of the sensitivity of a measure is given by the number of **null decisions** (i.e.,

	LIKERT	ME(z-SCORE)	ME(LOG)	THURSTONE	RANDOM PAIRS	TARGET PAIRS
Inconsistent decisions						
Within	11	15	18	12	26	13
Between	17	21	28	18	27	15
Null decisions						
Initial	24	35	44	31	49	11

Table 7.2: **Decision reliability measured by agreement on targeted contrasts.** The reliability of each measure was quantified based on the number of decisions, out of a 150 total, that suggested different conclusions at different time points. All inconsistent decisions here were significant at one time point and null in the other: sign reversals appeared only at smaller sample sizes. To test if high reliability was based on insufficient power resulting in consistent null decisions, the total number of null results is also shown. There was no significant difference in the number of inconsistent decisions across measures for within- or between- participant datasets, however there was a significant difference in the number of null decisions, with the TARGET PAIRS measure showing the fewest null decisions

the measure was unable to reject the null hypothesis in the INITIAL data set): it would be possible for a measure to never produce an inconsistent decision, but only because it was unable to ever reject the null hypothesis, which would not be a very interesting measure.

The raw numbers of inconsistent and null decisions are shown in Table 7.2. No significant differences in the number of inconsistent measures was found either within participants ($\chi^2(5) = 10.933, p = 0.0527$) or between participants ($\chi^2(5) = 8.0842, p = 0.15$), but there were significant differences in the number of null decisions ($\chi^2(5) = 37.35, p < 0.001$). In particular, TARGET PAIRS had notably fewer null responses: it was a more sensitive measure. There were no reversals across data sets for any measure in the full sample, but some did occur at smaller sample sizes.

Sample size dependence

As in the previous analysis, I examined the impact of sample size on decision reliability by analyzing 30 random subsets of the full sample for each increment of 10 participants between 30 and 120. Results are shown in Figure 7.6. As expected, within-participant reliability is generally higher than between-participant reliability. TARGET PAIRS is highly reliable and less sensitive to sample size, although at sample sizes of 75 and above it enjoys no particular advantage over the most reliable global scores, LIKERT and THURSTONE. A similar sample size analysis for the null results is shown in Figure 7.7, whose left panel indicates the number of sentence pairs for which each measure concluded the evidence was insufficient to reject the null in the INITIAL data set.

Overall, then, TARGET PAIRS appears to be both highly reliable and extremely sensitive, yielding relatively few inconsistent decisions (Figure 7.6) even with very low numbers of null decisions. That said, the main drawback of the measure is evident upon comparing the effect of sample size on the number of decision **reversals** across measures: the number of decisions on which the measure indicates one option was significantly more acceptable at one time but the same measure indicates that the *other* option was significantly more acceptable at the other time. Reversals are one kind of inconsistency, but are singled out here because they reflect a larger and more consequential difference than inconsistencies that involve being unable to reject the null in one sample but not another. As the right panel of Figure 7.7 indicates, only TARGET PAIRS yields decision reversals. Especially at small sample sizes, it may show a statistically significant preference for one item of a pair only to prefer the *other* item with more data. Thus, its sensitivity comes at a cost – being more likely to completely flip in the direction of a statistically significant judgment.

Discussion

The relative reliability of the targeted contrast decisions derived from each measure is consistent with the ordering observed in the score correlation analysis, with TARGET PAIRS and LIKERT the most reliable measures. The consistency of decisions is however distinct from the consistency of the underlying scores because of the way it depends on the threshold for rejecting the null. When the null hypothesis is true, the number of agreements across replications is completely determined by the alpha level of the test, but in the presence of a real effect it also depends on the sensitivity of the measure and the true effect size. In this study the effect sizes are constant across measures, so differences between measures reflect their sensitivity. A measure’s sensitivity

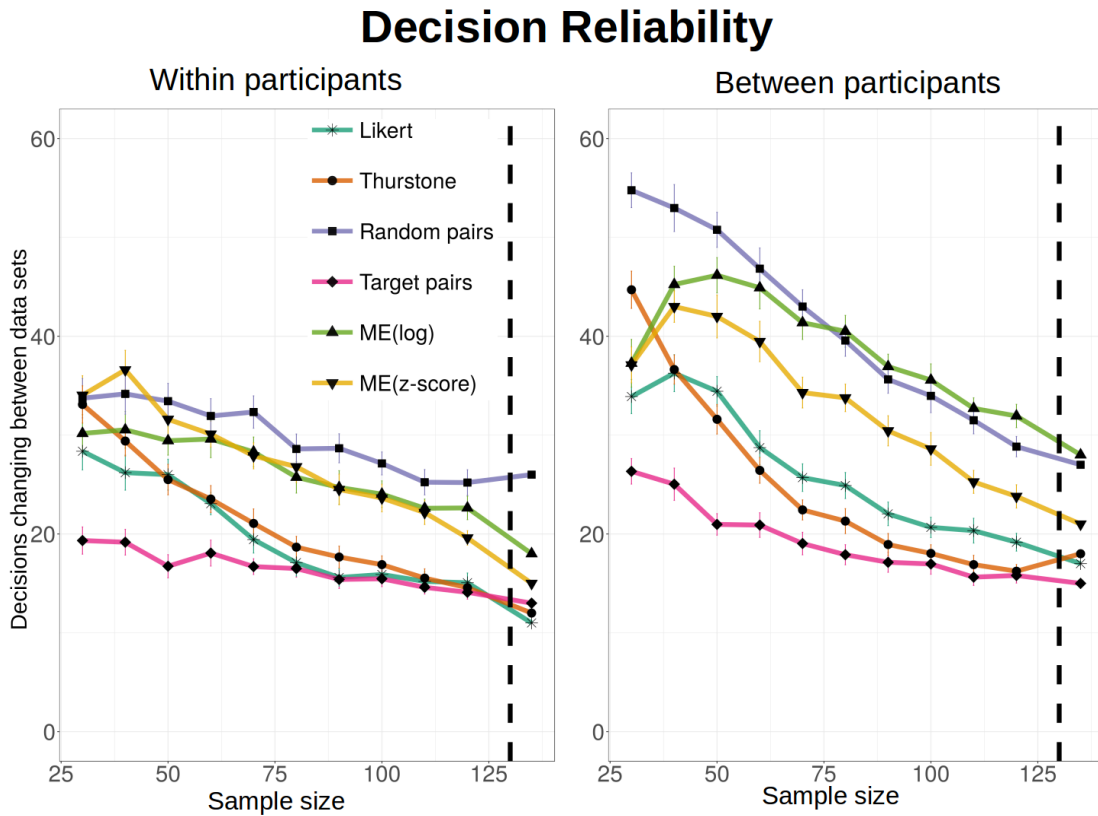


Figure 7.6: **Decision reliability measured by agreement on targeted contrasts.** The reliability of each measure was quantified as the number of decisions resulting in inconsistent outcomes (y-axis) tested across various sample sizes (x-axis). Outcomes were considered inconsistent if a contrast was considered null at one time point but significant at another, or the two results were significant in opposite directions: either case guarantees an error, although the first scenario is ambiguous as to error type. Results to the right of the dotted vertical bar are based on the full sample, results to the left are averages of 30 samples of the designated size drawn from the full sample. For all measures, smaller sample sizes are less reliable. Within-participant reliability (left panel) shows variability in estimates based on responses from the same people to the same items, and can be interpreted as variability deriving from the difficulty of the task and the inherently probabilistic nature of people’s responses. Between-participant reliability (right panel) is subject to the same sources of noise plus individual differences and variability introduced when re-drawing the selection of items presented. Contrasting within and between participant reliability indicates the vulnerability of each measure to these extra factors. Between-participant reliability also gives a direct measure of how well acceptability scores might be expected to replicate. These results show that RANDOM PAIRS and ME scores are particularly vulnerable to participant and item differences, and that THURSTONE and LIKERT scores are consistently reliable and approach the reliability of the TARGET PAIRS method

ultimately depends on the information content of responses, and the extent to which information is lost by the process of aggregating responses to produce an acceptability score. Unlike the alpha level, this is not a property of the decision rule and can only be estimated empirically. It is determined by the allowable range of variability in responses and the extent to which observed variability is systematic. An ideal measure would be high on both, but the two properties conflict in the sentence acceptability context to the extent that increasing the flexibility of response options makes the task more difficult. The different tasks considered here represent different trade-offs in intuitive ease-of-use (helping participants respond systematically) and expressiveness (widening the range of response options). As previous authors have noted (Fukuda et al., 2012; Weskott & Fanselow, 2008), the greater expressiveness of ME’s free responses appears to be offset by increases in unsystematic variation. As in the score correlations discussed above, contrasting between and within participant reliability suggests that this extra noise is introduced by individual participants’ idiosyncratic use of the scale, and can be mitigated by z -transformation. LIKERT appears to be effective in the compromise it achieves between allowing variability in responding and constraining unsystematic variation.

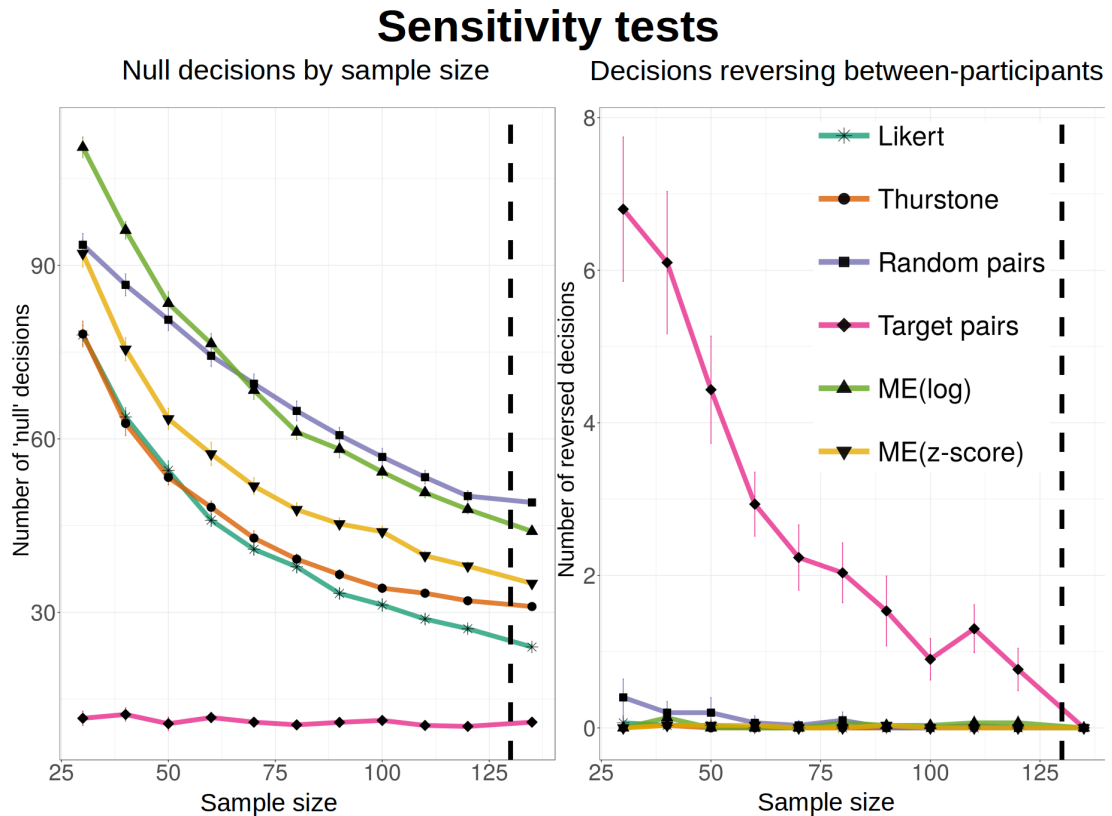


Figure 7.7: **Sensitivity tests by sample size.** The number of non-significant differences declared by each decision rule (left panel) and the number of times significant effects appear to reverse between samples (right panel). Each measure evaluated the same 150 contrasts between target-pairs of interest. Results to the right of the dotted vertical bar are based on the full sample, results to the left are averages of 30 samples of the designated sample size drawn from the full sample. TARGET PAIRS stands out as rejecting the null most often, and at a similar rate across a range of sample sizes. Taken together, these results highlight the unique properties of the TARGET PAIRS measure, which almost always arrives at a decision even at small sample sizes, but with decisions that may not be stable under repeated measurement. Decision reversals are shown for the between-participants test, the within-participants analogue shows the same qualitative pattern with roughly half the number of decision reversals at each sample size for TARGET PAIRS. The other measures are less sensitive in the sense of producing more null decisions, but more conservative in the way repeated samples only ever disagree on the magnitude of an effect, never its sign. Among these conservative measures, LIKERT ratings are the most likely to detect acceptability differences

TARGET PAIRS was found to achieve a very high power on the limited contrasts it considers, in that it arrived at fewer null decisions and was relatively insensitive to sample size. The risk of producing significant results in the wrong direction, as shown by the decision reversals, is a consequence of this high power along with the fact that controlling Type 1 error rates does not entail controlling error magnitudes. The design of the decision rule used allows for the possibility of significant findings in the wrong direction so long as the rate of such outcomes obeys the specified limits (Cumming & Maillardet, 2006; Gelman & Tuerlinckx, 2000). The practice of testing several instances of any one phenomenon of interest (Schütze, 2011) provides protection against these potential sign errors, since measurement error will be randomly distributed across individual items. The cost is inflation of the item set size, which may be a relatively minor burden compared to increasing the number of participants to the levels required for comparable power with a more conservative measure such as LIKERT or THURSTONE.

Although item effects may be driving the higher level of sign errors for TARGET PAIRS, it is unlikely that they are responsible for the relatively greater number of null decisions yielded by the LIKERT measure here than were reported in other work (Sprouse et al., 2013; Mahowald et al., 2016; Häussler et al., 2016). On closer examination, this difference probably emerges because the dataset presented here involved fewer responses per item. For instance, the Sprouse et al. (2013) results are based on 12 or 13 responses per item, with 8 items per phenomenon giving

roughly 100 responses per effect. The data presented here involve approximately 20 responses per item (with some variability due to the random draw of items), but since there is only one item per phenomenon this is also 20 responses for each effect. With the smaller number of responses there are naturally more null decisions. Despite the much smaller Ns per item, the fact that there were still relatively few null decisions is reassuring, especially in light of the fact that a researcher investigating a specific effect would test multiple items.

7.3.3 Agreement between measures

Having investigated the within and between-participant reliability of all of the measures, a natural question is whether they give the same answers as *each other*. Indeed, the question of how well different measures compare to informal judgments (as published in scholarly journals or textbooks) is much of the focus of related research in the literature (Sprouse et al., 2013; Gibson & Fedorenko, 2013; Munro et al., 2010; Myers, 2012; Featherston, 2007; Sprouse & Almeida, 2012).

The work presented here extends this body of existing work by also incorporating comparisons to the INFORMAL measure of acceptability, though that is not the main focus. Instead, I explore a much wider set of comparison measures between all of the formal measures. I examine both global structure and decision agreement, using the same tests of agreement between measures as the reliability analyses, but for agreement between measures rather than across data sets.

Figure 7.8 summarizes these results, presenting both correlations between acceptability estimates and the percentage of decisions (out of 150 total) where the measures arrived at the same conclusions. The scores used in this analysis were from the INITIAL dataset for LIKERT and TARGET PAIRS measures, and from the BETWEEN PARTICIPANTS data for THURSTONE, RANDOM PAIRS, and ME scores. This avoids comparing unrelated measures on data derived from the same participants, potentially inflating their agreement. Related measures (the two versions of ME, or the two analyses of the random pairs choice task) are derived from the same data sets, so any disagreements between them are consequences of the different analysis only.

Discussion

There was substantial agreement between measures. Measures based on the same responses (THURSTONE/RANDOM PAIRS and ME(LOG)/ME(z-SCORE)) were highly correlated ($r \approx 0.96$). Between measures based on different responses, correlations ranged between 0.74 and 0.95, with the highest agreement appearing between the LIKERT and THURSTONE measures. This highest level of agreement is comparable to the reliability between participants for these measures, indicating that switching from one measure to the other introduces no more variation than using the same measure twice, despite substantial differences in the presentation of the task.

All measures are largely consistent with the informal judgments. Almost all differences were observed in the expected direction, although on average 23% of these differences were deemed too small to be considered statistically significant. This rate of null decisions is higher than that previously reported for these measures in Sprouse et al. (2013) primarily because these decisions are based on more responses per item but many fewer responses per effect, due to the way each effect was represented by a single sentence pair. Although these design decisions limit conclusions about any particular effect, because they were constant over measures, contrasts can still be drawn between the measures. In particular, this comparison highlights a striking difference between TARGET PAIRS and the other measures, as it identified 14 contrasts in the opposite of the predicted direction, a disagreement rate of 9.3%. There are several possible explanations for these inconsistencies. It is possible that they represent measurement errors, although the sub-sampling analysis suggests that it is unlikely that re-measurement at the sample size considered here would ever be expected to reverse more than one decision. More likely is that in most cases these are genuine effects counter to the predicted direction, but they are extremely small and TARGET PAIRS is the only measure with enough power to identify them. This interpretation is supported by the fact that most of these contrary decisions by TARGET PAIRS are identified as null results by the other measures.

Measurement error seems unlikely for cases where two different measures agree with each other and informal judgments, so there may also be instances where contrasting highly similar sentences directs people's attention to features of the sentences that are less salient when the two items are presented separately. One example of such an effect is the pair of sentences *There are leaves burnt* and *There are leaves green*. This pair of sentences, constructed by Sprouse et al. (2013),









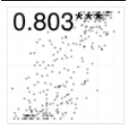
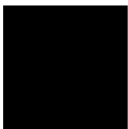




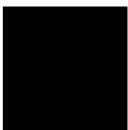
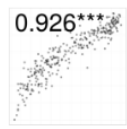
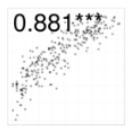


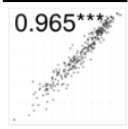

	Likert	Random pairs	Thurstone	Target pairs	ME(log)	ME(z-score)	Informal
Likert		80%	88%	84%	79%	84%	Consistent:125 Inconclusive:24 Inconsistent:1
Random pairs	0.921*** 		86%	74%	76%	80%	Consistent:100 Inconclusive:48 Inconsistent:2
Thurstone	0.951*** 	0.963*** 		81%	78%	82%	Consistent:118 Inconclusive:31 Inconsistent:1
Target pairs	0.811*** 	0.778*** 	0.803*** 		66%	73%	Consistent:125 Inconclusive:11 Inconsistent:14
ME(log)	0.894*** 	0.846*** 	0.88*** 	0.74*** 		89%	Consistent:98 Inconclusive:52 Inconsistent:0
ME(z-score)	0.926*** 	0.881*** 	0.916*** 	0.758*** 	0.965*** 		Consistent:110 Inconclusive:40 Inconsistent:0

Figure 7.8: **Agreement between different measures.** The lower left of this table presents Pearson correlation as a measure of global agreement between scores. The upper right presents the percentage of targeted contrasts for which different measures arrived at the same conclusion (A is more acceptable, B is more acceptable, or the evidence is inconclusive). The exception to this format is the informal expert judgment measure of acceptability, which presents only the number of decisions where informal judgments were consistent with the formal measure, the number that were inconsistent, and the number of decisions where the formal measure found the evidence inconclusive. Overall inter-measure agreement is strong. Among the formal measures, the highest agreement obtains between the LIKERT and THURSTONE scores. Agreement with the informal judgments is also high, with TARGET PAIRS notable as the only measure indicating an appreciable number of contrary conclusions

follows a pattern designed to demonstrate that (English) passives have event arguments (Basilico, 2003). The option endorsed in the INITIAL TARGET PAIRS data is *There are leaves green*.⁴ One possible explanation for this counterintuitive result is that it arises from people recognizing that the sentences are identical except for the words *burnt* and *green*, then responding to a strong association between *green* and *leaves*. The interpretation that the choice is somehow induced by the particular contrast of these two sentences is supported by the agreement of the measures that do not presenting the two sentences together that *There are leaves burnt* is the more acceptable option. The practice of using multiple items targeting each phenomenon under study would be an effective defense against this kind of potentially misleading result, since it depends heavily on the *green/leaves* association, which would be unlikely to have an analogue in other sentences targeting this particular passive construction. Across all three data sets collected, only 19 contrasts are involved in a decision result conflicting with expert judgment at any point. These are presented in Appendix A.

7.4 Summary and Conclusions

The main focus in this work is the test-retest reliability survey of the most common tasks used to measure sentence acceptability. All tasks considered here showed high reliability, with even the least reliable measure, RANDOM PAIRS, producing large positive correlations across re-test

⁴This result re-appears in the WITHIN PARTICIPANTS replication, but not the BETWEEN PARTICIPANTS replication, which is null.

data sets. By contrasting within-participant reliability with between-participant reliability on the same sentences with the same measures, I estimated what proportion of the variability observed can be attributed to factors unique to the between-participant replication. In all cases between-participant reliability was lower, and this reliability drop was particularly pronounced for ME and RANDOM PAIRS, suggesting these measures are particularly vulnerable to variability across people or how items are paired together. The TARGET PAIRS and LIKERT ratings showed not only the highest within-participant reliability but also had the least amount of decrease in reliability when comparing between- to within-participant correlations. This pattern is a hallmark of well-calibrated measurement instruments.

Secondly, I ask to what extent acceptability estimates depend on the particular assumptions of each measurement tool, and whether the conclusions a researcher would reach would change based on the measurement task they used. Here I find high consistency between measures, including near-uniform agreement with expert judgment. The least accurate global score (RANDOM PAIRS) was still highly correlated ($r \approx .9$) with the most accurate global score (LIKERT). Where disagreements occurred between the measures, it was usually in the magnitude rather than the direction of the difference, with the less reliable scores more likely to not reject the null for closely matched pairs.

This overall consistency is striking given the structural differences between these tasks, especially between the LIKERT and THURSTONE tasks. Both these measurement tasks incorporate strong assumptions, and in different domains have not always agreed with each other (Roberts et al., 1999; Drasgow, Chernyshenko, & Stark, 2010). Specifically, the assumptions made by the LIKERT task center around people’s interpretation of the scale, which may impose structure on responses (Carifio & Perla, 2008; Schütze, 1996) or be vulnerable to differences in response style (T. R. Johnson, 2003; Lee, Jones, Mineyama, & Zhang, 2002). The THURSTONE measure avoids these issues by removing the researcher-supplied scale and forcing a discrete choice, but instead assumes transitivity of acceptability, which is known to be violated in similar preference-ranking tasks (Tversky, 1969). Such violations have been observed in sentence acceptability judgments (Hindle & Ivan, 1975; Danks & Glucksberg, 1970)

A core contribution of this work is that these measures provide converging evidence in the domain of sentence acceptability: theoretically motivated concerns about the restrictions a fixed LIKERT response scale imposes on participants turn out not to matter in practice, with the scale-free THURSTONE measure based on choice task data arriving at essentially identical acceptability estimates. Although the LIKERT and THURSTONE acceptability scores agree, LIKERT scores are marginally more reliable and have the advantage of more easily accepting additional sentences into an existing set of comparisons.

Despite the close agreement between measures, TARGET PAIRS stands out as having noteworthy decision reliability. It showed the highest power, yielding very few null results, but as a result was also the only measure vulnerable to complete reversals of a significant decision. This pattern is characteristic of high-powered tests, where significant differences observed under high-noise/low information conditions tend to entail exaggerated estimates of effect size (Loken & Gelman, 2017). While TARGET PAIRS is the highest performing measure in terms of test-retest consistency, and maintains this performance at small sample sizes, the relatively few errors it produces at low sample sizes can be of a qualitatively different and potentially much more misleading kind. Relatedly, the TARGET PAIRS measure had by far the highest disagreement with the informal expert ratings of any measure, endorsing the informally dispreferred sentence on 14 of the items (9.3%) while the other measures endorsed at most two. When using the TARGET PAIRS measure it is critical for researchers to include multiple pairs of target sentences within the same construct to increase decision reliability.

These studies show that ME tasks produce acceptability scores that are consistent with the other measures but somewhat less reliable. Contrasting the within and between participant test-retest reliability shows that this greater variability is likely to be due to variation in participant response styles, which appears as noise in the final measure. This source of variability can be mitigated somewhat by processing the scores using a transformation sensitive to response style, such as the z -transform. However, this is less effective than offering restricted responses in the task itself, as the LIKERT and THURSTONE measures do. In general, although ME measures performed overall better than expected, they were still consistently inferior to most of the alternatives.

Although these results can be expected to be indicative of the relative test-retest reliability of these measures, the particular reliability results observed here depend to some extent on factors such as the specific sentences and the number of trials per participant, which were controlled across

measures to ensure the comparisons were fair. For the rating tasks, reliability can be expected to be a function of the number of trials per item, so the analysis over participant sample sizes gives some indication of how reliability might be expected to change with different sentence set sizes. The situation is less clear for the THURSTONE and RANDOM PAIRS measures, which may be sensitive to the diversity of contrasts presented as well as the average number of presentations per sentence. By choosing to hold the set of sentences constant I ensured that each measure was tested on the same range of effect sizes, but this does limit the generalizability of the reliability results presented here. They hold to the extent that these 150 sentences are representative of the kinds of sentences commonly used for sentence grammaticality judgments, which does not seem unreasonable given the way they were sampled from a prominent linguistics journal (Sprouse et al., 2013).

Although individually these measures make a range of assumptions that could be considered strong limitations, the high agreement between them suggests that these measure-specific assumptions do not have a strong impact on acceptability judgments. The results presented here indicate that if multiple items targeting the same contrast are used, none of the methods considered here have an appreciable chance of giving a strongly misleading result (although there are differences in efficiency, with ME measures requiring more trials for any given level of reliability).

While these studies find that the most common measurement tasks are all reasonably effective, the LIKERT task performed especially well. In addition to achieving relatively high test-retest reliability, the results also suggest that the LIKERT measure admits a stronger interpretation of sentence acceptability scores than is usually attributed to it. These findings suggest that the interpretation of LIKERT data need not be constrained by concerns that the limited response scale may impose structure on the data, or that the subjective distance between response options is unknown and may vary between people. The structure suggested by the LIKERT data is in high agreement with the structure suggested by the THURSTONE measure. Since the latter is both agnostic about the underlying structure of acceptability and capable of recovering various clustered or gradient but non-linear distributions of acceptability, this high agreement suggests that the nature of the LIKERT scale is not significantly shaping the structure of acceptability judgments it yields. The minimal difference between within-participant test-retest reliability and between-participant test-retest reliability suggests that the z -transformation offers effective protection against potential differences in the interpretation of the scale.

One interesting aspect of the results hinges on the fact that the dataset examined here involved only one item per effect. This was intentional since it thus made the item set maximally variable and offered a stronger test of each measure. The result indicating that many of these measures can reliably reflect *global* acceptability, rather than just effect-level acceptability, is gratifying and reassuring. It is also interesting that the *item*-level reliability is so high, differing from other work measuring effect-level reliability primarily in yielding slightly higher numbers of null decisions at lower sample sizes (Sprouse et al., 2013; Mahowald et al., 2016; Häussler et al., 2016). Aside from this, item-level reliability was in this case nearly as good as effect-level reliability incorporating multiple items. Taken together with the high item-level variability observed around effects in other work (Sprouse et al., 2013), this may suggest that people are surprisingly consistent on specific items but that the effect-level phenomena within any given item can at least sometimes be obscured by lexical choices or other superficial differences between sentences.

In terms of design recommendations for researchers interested in efficiently obtaining results that replicate with high confidence, the studies presented here replicate previous results pertaining to the reliability of effects defined as ordinal relationships between sentence classes and extend them to include recommendations for ensuring the reliability of distances between individual items. These results reproduce both the general finding that acceptability judgments are highly reliable in between-participant replications (Sprouse & Almeida, 2012; Sprouse et al., 2013), and also more detailed claims such as the high power of TARGET PAIRS (Schütze & Sprouse, 2014), the lack of extra information in the extra variability of ME ratings (Weskott & Fanselow, 2011), and the qualitative relationship between decision reliability and sample size (Mahowald et al., 2016). The results presented here further show that these reliability results extend to estimation analyses, with a high correlation in the acceptability scores assigned by different tasks to different sentences.

Overall, this work demonstrates that formal acceptability results are even more informative than previously realized. They agree substantially with each other (as well as informal measures) across the global structure of acceptability, not just individual targeted sentence pairs. Moreover, the best-performing measures (like LIKERT and THURSTONE) appear not to impose substantial

structure of their own onto the pattern of acceptability responses. This licenses researchers to use acceptability judgments to address a wider variety of questions than previously – from identifying dialectal or language differences (or possibly even individual fluency) using acceptability judgments, to investigating the global structure of grammatical knowledge (e.g., is it all-or-none or multi-dimensional?). Not all of these questions may pan out, but the investigation into the properties of the formal tools available to study them is an important prerequisite of such work. The positive results presented here suggest that acceptability judgment data may be more informative about these questions than previously thought.

Part IV

Conclusions

Chapter 8

Summary and conclusions

This thesis has presented three related projects linked by their use of comparison data to help shed light on representations. This final chapter briefly reviews the main findings and adds some further discussion relating the results to the existing literature, outlining limitations, and giving possible future directions.

The first of these projects compared two prominent theories of similarity judgment, the transformational account and a simple alignment approach. In the first study, triad stimuli were constructed such that the two accounts gave different predictions as to which of two options was most similar to a reference ‘base’. Participants overwhelmingly endorsed the options indicated as more similar by the alignment account. In a follow-up study, a same-different discrimination task gave somewhat different results. In this task, the speed of correct ‘different’ responses is thought to reflect similarity, with more similar pairs more difficult to distinguish and therefore slower. The transformational approach better captured the timing of ‘different’ judgments. A third study examining both choice and same-different discrimination tasks with a common set of stimuli concluded that the two tasks gave largely consistent measures of similarity across stimuli, with three exceptions possibly reflecting differences due to the differing time demands. The third study also compared a number of closely related variations on the models being considered. As a result of these contrasts, an explanation for the differences between studies one and two based on comparison direction was ruled out. Instead, the balance of evidence suggested the the main factors contributing to the observed pattern of results were the different time demands over the two tasks and a particular misspecification of the APPLY operation in the transformational account current for geometric shapes.

The original goal of Experiment 1 was to find a set of diagnostic items for which alignment-based models would typically make qualitatively different predictions to the transformation model, using simple perceptual stimuli that have been used in previous work (Larkey & Markman, 2005; Hodgetts et al., 2009a; Hodgetts & Hahn, 2012). However, because both approaches typically provide realistic (and therefore similar) accounts of human similarity judgment, the search for diagnostic tests tends to focus on special cases. From a statistical perspective this would usually be considered good practice, insofar as it maximizes the power of the experimental design. However, what actually happened in Experiment 1 is that the diagnostic tests all exploited a “single point of failure” in the transformational model, leading to minor modifications to the model but not yielding a decisive result. Similarly, the results in Experiment 3 impose some constraints on alignment-based models of perceptual similarity. However, it is less than clear that either of these advances is entirely informative regarding the more substantive question of whether transformation or alignment provides the better *theory* of similarity. Given the ease with which both modeling frameworks can be modified to accommodate discrepant findings – particularly when we allow the stimulus representation to be changed – these results do not strongly support one framework over the other.

The original intention of the study was to distinguish between these frameworks, but this may have been a somewhat quixotic goal. If the transformations or feature representations are allowed to vary arbitrarily, both frameworks are unfalsifiable. The only achievable version of this goal would be to restrict the viable set of transformations or feature representations to an implausibly narrow set, ideally one incapable of generalizing across tasks. Such a result could be taken as evidence at the framework level, but the constraints imposed by the data collected here are simply

not that strong.

Although the results presented here do not require implausible gymnastics from either theory, it is also true that some accommodation is required, and the fact that the basic building blocks of each framework admit variation is theoretically important. This is especially true of the transformational approach, where the results presented here add to arguments presented by Grimm et al. (2012) questioning the specification of the transformation set used. The work presented here also implicitly suggests a solution to the problem of how to specify a transformation set – incrementally, in response to data, as is done here by working through a number of closely related possibilities. Admittedly this is a very slow process, especially given the fact that transformation sets must be domain specific (Müller et al., 2009). The analogous comparison between the variations on basic structural alignment models is somewhat limited by the way I sought a lowest-common-denominator for several much more sophisticated alignment based schemes such as SIAM, LISA, or CAB (Goldstone, 1994; Taylor & Hummel, 2009; Larkey & Love, 2003). However a search process evaluating stimuli on the basis of this simplified measure did help eliminate ‘easy cases’ and returned stimuli that exposed a fundamental question of chunking in representation: is [aa] [b] different from [a] [a] [b]? From this point of view, the most interesting aspect of these results would be the follow up work they suggest exploring candidate explanations for why triads 5, 9 and 13 were exceptional.

Is it useful to go down this path of pursuing ever finer distinctions in people’s similarity judgments on such heavily simplified stimuli? The whole program behind process accounts of similarity seems open to question in a world where all models are wrong, and the state of the art in closely related AI visual processing is instead built on an unashamedly uninterpretable bottom up process using very different representations (eg Wang et al., 2017). I would argue that it is useful, because the primary value of the project is in the human responses rather than the specific theories, the data, not the debate. In order to test the more sophisticated theories of the future, cognitive science should be able to produce stimuli on demand that are both simple enough to be interpretable and rich enough to capture a meaningful picture of human similarity processing. Debates like the one explored here between proponents of transformational and structural alignment accounts of similarity make that possible. For example, it may be useful to distinguish stimuli where human responding systematically differs from the strict alignment account, and seek to include such examples in tests of future theories, despite their relative rarity when sampling stimuli at random in this domain of pairs of geometric shapes. No commitment to the literal veracity of the strict alignment account is necessary for this to be useful. The work presented here is itself an example of this kind of logic. Although it constructed stimuli with reference to an oversimplified version of structural alignment and a misspecified version of transformational similarity, by avoiding ‘easy cases’ it ended up highlighting stimuli that exposed differences between the speeded same-different task and deliberative comparisons. The distinction between these tasks was not apparent in earlier work using stimuli from the same domain. The admittedly imperfect models examined here show that the tasks are distinct, and between them give a procedure for identifying those stimuli where the distinction is particularly obvious. The models that succeed them (for example the adjusted instantiation of transformational similarity with costly apply) will almost certainly still be wrong, but can be expected to enable improved searches for stimuli highlighting further properties of human similarity judgment.

The second of the three projects presented here looked at how learning a transformation might influence similarity and categorization judgment. Participants were shown one of two different training conditions before proceeding to a common set of test items. The different training conditions meant that identical test items had different relation-to-training status depending on training condition. The study focused mainly on whether relation-to-training status was associated with differences in ratings for similarity or category membership questions. Study one, which gave a particularly easy form of training where participants were shown the transformation explicitly at each step, established that such training could cause shifts in people’s similarity and categorization judgments for novel stimuli instantiating the trained transformation. Follow-up work moving to implicit rather than explicit presentation of transformations showed people’s response to the increased difficulty of the task. Transformation learning was still possible, but this lower-confidence learning appeared to have somewhat less influence on categorization and similarity judgment, particularly for similarity. Across the two studies, this work also showed transfer of learning across transformations, suggesting a family resemblance structure for transformations. The extent of transfer also depended on the difficulty of the learning task as manipulated by the explicit/implicit

presentation format of the transformations.

This work was primarily exploratory. The many minor changes to the experimental procedure documented in the method section were helpful to hone in on a feasible experimental paradigm, but also introduce variability that precludes strong conclusions from comparisons across the different data sets. For reasons of space, the procedural variations are reported without much discussion in the proceedings paper, but they were quite critical to this work. For example, the paper notes that pilot testing indicated the grid stimuli could not be too visually complex, so the number of colors in each grid was limited. This fact is not particularly relevant to the main question of interest, the relation between transformation learning and *changes* in similarity judgment, and is therefore mentioned without further elaboration, but it was critical prerequisite knowledge for designing the main test. There are many such details, another is the inclusion and subsequent removal of the ‘identity’ transformation condition, which was primarily useful as a test for item effects: if for example the test stimuli related by one or the other of the main transformations was at ceiling or floor when neither of the relevant transformations was seen, the main comparison between the two could potentially be misleading. In the event, the stimuli related by the color transformation were found to be a little more similar than those related by the movement transformation when neither transformation was seen, but neither approached ceiling or floor similarity.

Although many of these learnability results were highly specific to the particular task explored here, some are more general, such as the distinction between observed and inferred transformations that formed the main point of contrast between experiments one and two. This is particularly relevant to the theory of transformational similarity, as many of the transformations involved in comparing representations are unobservable. For example, the transformations involved in comparing a glass with a coffee mug would need to account for a change in material. There is no barrier to constructing transformations that do this, it could be as simple as a CREATE operation for the appropriate feature followed by APPLY, but regardless of its exact form, the resulting transformation whatever it is cannot exist in the natural world. If transformations like these are to be learned, they must be inferred not observed. The results presented here do suggest that such learning is feasible on the short timescales needed for consistency with transformational similarity, but also that the resulting difference in task difficulty is not trivial.

With these learnability results in hand, future work could use this paradigm in more cleanly confirmatory studies of the relationship between transformation learning and similarity judgment. From this perspective the most interesting result of this work is probably the evidence of generalization suggesting family relationships between transformations. Transformation features offer strong challenges to both approaches to similarity discussed in Chapter 2. It’s true that the learnability of these simple transformations suggest that the computational complexity constraints put forward by Müller et al. (2009) do not present insurmountable barriers to transformational accounts of similarity, since people showed some evidence of the required quick learning of domain specific transformations. However, any graded availability of transformations such as that implied by family resemblances between transformations would seriously complicate the calculation of transformation distances. Considering the same results from a structural alignment perspective is also challenging, since it’s not immediately clear what the appropriate representation is for transformations, making it difficult to determine what to align on. General theories of similarity do need to be able to account for transformations and actions, which have a number of unique properties (Austerweil & Griffiths, 2010b; Lamberts, 2003; Pollick & Paterson, 2008). The work presented in this thesis was only partially successful in providing a simple lab-friendly task capturing the relevant phenomena, but did establish the feasibility of such an approach and give some preliminary results.

The third project compared different ways of measuring sentence acceptability. Although applied in a different area of cognition, this work drew on ideas from the the similarity work in projects one and two, in particular in its use of a Thurstonian model for measuring acceptability. This approach, which has deep connections to the similarity literature, uses comparison data to make inferences about representation. In total six different measures of sentence acceptability were examined with test-retest reliability as the main measure of performance, complementing existing work on the Type 1 and Type 2 error rates of these measures. The test-retest reliability analysis indicated that Likert scales gave particularly high reliability, and in particular small differences when contrasting between and within participant reliability suggested that individual differences in interpretation of the scale were effectively controlled by using a z -transformation on raw responses. High agreement with the scale-free Thurstonian measure suggested that providing a discrete scale

did not impose structure on responses. That averages of z -transformed Likert scale responses give identical answers to a scale-free estimate of acceptability from comparison data supports the legitimacy of interpreting such data as representing a fine-grained estimate of gradient acceptability. While magnitude estimation was found to give similar acceptability estimates to other measures, a greater variability across replications suggested this is a less efficient measure, and contrasting between and within participant reliability with and without a z -score transformation of the data showed that this higher variability is most likely to be due to individual differences in the interpretation of the scale. Targeted forced choice tasks were found to have extremely high power, with a very low null decision rate but a corresponding vulnerability to a small proportion of sign errors for significant differences based on smaller sample sizes.

This work contributes to the existing literature in this area by informing the interpretation of acceptability judgment data and by offering information relevant to study design. In one sense the most important results here are the least surprising ones. Since the primary justification for taking averages of Likert scale data is that it just works, a result licensing the interpretation of averaged Likert scale data is the opposite of revolutionary. It does in fact just work. However this result was not a foregone conclusion. The objections to interpreting averages of Likert scale data are theoretically sound: the Likert scale is a discrete scale, discrete scales provide ordinal data, ordinal data cannot simply be averaged (Yusoff & Mohd Janor, 2014; Chimi & Russell, 2009; Harwell & Gatti, 2001). The opposing ‘just works’ argument is based to a large extent on simulation studies showing that popular parametric tests invoking an interval interpretation of these averages are robust to the violations of their assumptions imposed by Likert scale data (see for example Carifio and Perla (2007) and references therein). Although this does provide a principled basis for a kind of ‘just works’ argument in specific cases such as the ANOVA F -statistic, it does not address the core theoretical concern around the validity of interpreting averaged ordinal data (Jamieson, 2004; Cliff, 2014). As is often the case, it seems that both sides of this debate are right: the ‘just works’ argument is very often valid but it is conditional, depending in part on properties of the thing being measured (S. Kemp & Grace, 2010). To the best of my knowledge, the work presented in this thesis provides the first empirical evidence that these conditions are met for the specific case of sentence acceptability judgment.

The results of this work also have implications for study design. If the range of acceptability differences represented in the test sentences collected by Sprouse et al. (2013) can be taken as representative, then the reliability results here provide quite specific guides: for estimation questions, the Likert elicitation task is preferable to the other methods examined here, and ideally should use at least 30 responses per participant (for the z -transformation to be effective) and approximately 13 to 15 responses per item, beyond which only limited improvements in reliability is possible. Where the hypothesis takes the form of a directed contrast between two sentence classes, the targeted forced choice task is preferable, but the use of multiple items to represent each sentence class is important, and sample sizes higher than the minimum necessary to achieve statistically significant results are desirable to avoid sign errors. In this regard, the recommendations presented here overlap substantially with existing work (Sprouse & Almeida, 2011, 2017; Mahowald et al., 2016; Erlewine & Kotek, 2016; Gibson et al., 2011). As well as functioning as an independent replication of these existing results, the work presented here adds some detail showing *why* the different elicitation tasks have these different performance profiles, using the decomposition of variability afforded by the within/between participant reliability contrast. The clearest example of this is the identification of how response styles contribute to the attested high variability of the magnitude estimation task. Since the same tests are applied to all measures, this information is also available for the other tasks, it’s just that for the other measures potential concerns such as response style or item neighborhood effects are mainly notable by their absence.

As noted in the discussion section of Chapter 7 this work is subject to a number of limitations. Some important possible sources of bias were controlled here in order to test the different measures under comparable conditions, precluding any conclusions about the size of their impact. For example, participants may be quite sensitive to the instructions, particularly the way ‘acceptability’ is introduced. This may be an important effect, especially when interpreting the results of diverse studies across the literature, but since this factor was held uniform across all experiments here no comment on it is possible on the basis of these results. Similarly, it’s possible that people are sensitive to properties of the stimulus set, such as calibrating their responses to the range of variation exhibited by test items, but this too was controlled in these studies by applying all measures to the same sentences. As a consequence it’s not possible to tell how important these

effects might be and whether they impact the different measures differently on the basis of these results.

As well as excluding a few controlled sources of bias and variability from consideration, the narrow focus of this work on properties of the elicitation tasks meant that it also did not consider any issues of syntax itself. It's worth asking what kinds of research question this reliability work could support, and whether they go beyond the questions that are currently being asked in any useful way.

Two anonymous reviewers of this work independently pointed out that the presentation of item-level reliability results here stands in contrast with the way such acceptability data is usually used, which is to identify statistically significant differences between syntactic classes. My hope is that this difference is part of a recent trend in this area towards finer levels of analytic detail. A recent case study articulating a rallying call for this kind of approach is Vasisht, Nicenboim, Chopin, and Ryder (2017). In brief, this study shows a syntactic violation previously identified by the longer average reading times it produces is better modeled as a mixture of fast successful parses and occasional but dramatically slower failures to parse. Although this example is in reading times not acceptability judgment, the main message of the paper still applies: the shape of the whole distribution of responses to individual items is informative. Further, identifying differences in terms of the whole distribution can motivate quite different working theories of the underlying process. Although the initial identification of the effect via mean reading times was sound, asking 'what raises failure-to-parse rates' is a very different question from 'what slows down reading speeds'. In this case the first question makes closer contact with the data, the second is subtly misleading. The item level reliability results presented here suggest that an analogous argument may apply to acceptability differences between syntactic classes, in that response reliability is high enough to motivate modeling response distributions as mixtures with item and individual level contributions. Doing so might finally answer the most common broad objection to acceptability judgments as a source of linguistic data, that they contain confounds such as plausibility and parsing effort unrelated to the putative syntactic target of a particular contrast (Bornkessel-Schlesewsky & Schlewsky, 2007). I would argue that these different factors most definitely exist, but that they are best considered as important factors in natural language use rather than just experimental confounds. One way to address such factors would be to examine a large and diverse set of sentences, maximizing independent variation along these different dimensions. The reliability work presented in this thesis would support research of this kind in two ways. Studies looking large diverse item sets could leverage the efficiency and between-participant reliability results presented here in their design. In addition, the Thurstonian model represents a small first step towards describing full response distributions with a response model, and could be elaborated to capture structure beyond the simple one dimensional acceptability scale I consider here, such as hierarchical clustering of speakers or items.

Final overview

Although these three projects are each different and related to somewhat different literatures in cognitive science, they are related by their use of comparisons to investigate questions of representation in similarity, categorization, and language. Through asking people to make comparisons, this work gives some constraints on the transformational and alignment approaches to similarity, raises questions about the specification of representations in structured accounts of similarity, and gives an example of time course effects making theoretically consequential differences in different measures of similarity. It explores transformations as latent features, showing the impact of some simple presentation manipulations on a relationship between transformation learning, similarity, and categorization judgment. Finally, it also investigates the measurement of sentence acceptability, adding to the toolkit of experimental linguistics. These are diverse results relevant to different areas, but they all draw on the way comparisons link easily observed behaviors like button presses with the true object of interest, complex structured mental representations.

References

- Aarts, B. (2007). *Syntactic gradience: The nature of grammatical indeterminacy*. Oxford University Press.
- Aarts, B., Denison, D., Keizer, E., & Popova, G. (2004). *Fuzzy grammar: the nature of grammatical categories and their representation*. Oxford University Press.
- Andrews, J., Livingston, K., Auerbach, J., Altiero, E., & Neumeier, K. (2014). Does learning to categorize visual stimuli based on motion features produce learned categorical perception effects? In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 3170–3170).
- Arppe, A., & Järvikivi, J. (2007). Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory*, 3(2), 131–159.
- Ashby, F. G., & Lee, W. W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General*, 120(2), 150.
- Austerweil, J. L., & Griffiths, T. L. (2010a). Learning hypothesis spaces and dimensions through concept learning. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 73–78).
- Austerweil, J. L., & Griffiths, T. L. (2010b). Learning invariant features using the transformed Indian buffet process. In *Advances in neural information processing systems* (pp. 82–90).
- Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric bayesian framework for constructing flexible feature representations. *Psychological review*, 120(4), 817.
- Austerweil, J. L., Griffiths, T. L., & Palmer, S. E. (2016). Learning to be (in) variant: Combining prior knowledge and experience to infer orientation invariance in object recognition. *Cognitive Science*.
- Baddeley, A. D. (1966). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *The Quarterly Journal of Experimental Psychology*, 18(4), 362–365.
- Bader, M., & Häussler, J. (2010). Toward a model of grammaticality judgments. *Journal of Linguistics*, 46(02), 273–330.
- Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 32–68.
- Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, 11(3), 211–227.
- Basilico, D. (2003). The topic of small clauses. *Linguistic Inquiry*, 34(1), 1–35.
- Bellet, A., Habrard, A., & Sebban, M. (2013). A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*.
- Beltran, J. F., Liu, X., Mohanchandra, N., & Toussaint, G. T. (2015). Measuring musical rhythm similarity: Statistical features versus transformation methods. *International Journal of Pattern Recognition and Artificial Intelligence*, 29(2).
- Bennett, C. H., Gács, P., Li, M., Vitányi, P. M., & Zurek, W. H. (1998). Information distance. *IEEE Transactions on Information Theory*, 44(4), 1407–1423.
- Biederman, I. (1985). Human image understanding: Recent research and a theory. *Computer vision, graphics, and image processing*, 32(1), 29–73.
- Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2), 115.
- Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical methods in medical research*, 8(2), 135–160.
- Bland, J. M., & Altman, D. G. (2007). Agreement between methods of measurement with multiple observations per individual. *Journal of biopharmaceutical statistics*, 17(4), 571–582.
- Borg, I., & Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

- Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2007). The wolf in sheep's clothing: Against a new judgement-driven imperialism. *Theoretical Linguistics*, 33(3), 319–333.
- Boroditsky, L., & Ramscar, M. (2001). First, we assume a spherical cow... *Behavioral and Brain Sciences*, 24(04), 656–657.
- Box, G. E. (1979). Robustness in the strategy of scientific model building. *Robustness in statistics*, 1, 201–236.
- Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, 50, 217–224.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1), 1–3.
- Brumby, D. P., & Hahn, U. (2017). Ignore similarity if you can: A computational exploration of exemplar similarity effects on rule application. *Frontiers in Psychology*, 8, 424–438.
- Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education*, 42(12), 1150–1152.
- Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. *Journal of Social Sciences*, 3(3), 106–116.
- Cavagnaro, D. R., & Davis-Stober, C. P. (2014). Transitive in our preferences, but transitive in different ways: An analysis of choice variability. *Decision*, 1(2), 102.
- Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *The Quarterly Journal of Experimental Psychology: Section A*, 52(2), 273–302.
- Chater, N., & Brown, G. D. (2008). From universal laws of cognition to specific cognitive models. *Cognitive Science*, 32(1), 36–67.
- Chater, N., & Hahn, U. (1997). Representational distortion, similarity and the universal law of generalization. In *Simcat97: Proceedings of the interdisciplinary workshop on similarity and categorization* (pp. 31–36).
- Chater, N., & Vitányi, P. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, 47(3), 346–369.
- Cherries, E. W., Newman, G. E., Santos, L. R., & Scholl, B. J. (2006). Units of visual individuation in rhesus macaques: objects or unbound features? *Perception*, 35(8), 1057.
- Chimi, C. J., & Russell, D. L. (2009). The likert scale: A proposal for improvement using quasi-continuous variables. In *Information systems education conference, washington, dc*.
- Chomsky, N. (1959). A review of B. F. Skinner's Verbal Behavior. *Language*, 35(1), 26–58.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.
- Cliff, N. (2014). *Ordinal methods for behavioral data analysis*. Psychology Press.
- Cornips, L., & Poletto, C. (2005). On standardising syntactic elicitation techniques (part 1). *Lingua*, 115(7), 939–957.
- Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments*. Sage Publications.
- Culbertson, J., & Gross, S. (2009). Are linguists better subjects? *The British Journal for the Philosophy of Science*, 60, 721–736.
- Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: where will the next mean fall? *Psychological Methods*, 11(3), 217.
- Dabrowska, E. (2010). Naive v. expert intuitions: An empirical study of acceptability judgments. *The linguistic review*, 27(1), 1–23.
- Danks, J. H., & Glucksberg, S. (1970). Psychological scaling of linguistic properties. *Language and Speech*, 13(2), 118–138.
- Danziger, K. (1980). The history of introspection reconsidered. *Journal of the History of the Behavioral Sciences*, 16(3), 241–262.
- Davis, T., Xue, G., Love, B. C., Preston, A. R., & Poldrack, R. A. (2014). Global neural pattern similarity as a common basis for categorization and recognition memory. *Journal of Neuroscience*, 34(22), 7472–7484.
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge: Cambridge University Press.
- DeVellis, R. F. (2016). *Scale development: Theory and applications* (Vol. 26). Sage publications.
- Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology*, 3(4), 465–476.

- Edelman, S., & Shahbazi, R. (2012). Renewing the respect for similarity. *Frontiers in computational neuroscience*, 6.
- Eidenberger, H., & Breiteneder, C. (2003). Visual similarity measurement with the feature contrast model. In *Electronic imaging 2003* (pp. 64–76).
- Ennis, D. M. (2016). *Thurstonian models: Categorical decision making in the presence of noise*. Institute for Perception.
- Ennis, D. M., & Johnson, N. L. (1993). Thurstone-Shepard similarity models as special cases of moment generating functions. *Journal of Mathematical Psychology*, 37(1), 104–110.
- Erlewine, M. Y., & Kotek, H. (2016). A streamlined approach to online linguistic surveys. *Natural Language & Linguistic Theory*, 34(2), 481–495.
- Fabrigar, L. R., & Paik, J. S. (2007). Thurstone scales. In N. Salkind (Ed.), *Encyclopedia of Measurement and Statistics* (p. 1003-1005). SAGE publications.
- Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial intelligence*, 41(1), 1–63.
- Fanselow, G. (2006). *Gradiance in grammar: Generative perspectives*. Oxford University Press.
- Featherston, S. (2005). Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua*, 115(11), 1525–1550.
- Featherston, S. (2007). Data in generative grammar: The stick and the carrot. *Theoretical linguistics*, 33(3), 269–318.
- Featherston, S. (2008). Thermometer judgments as linguistic evidence. *Was ist linguistische Evidenz*, 69–89.
- Featherston, S. (2009a). Relax, lean back, and be a linguist. *Zeitschrift für Sprachwissenschaft*, 28(1), 127–132.
- Featherston, S. (2009b). A scale for measuring well-formedness: Why syntax needs boiling and freezing points. *The fruits of empirical linguistics*, 1, 47–73.
- Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard University Press.
- Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2016). Extending SME to handle large-scale cognitive modeling. *Cognitive Science*.
- Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive science*, 19(2), 141–205.
- Freyd, J. J. (1983). The mental representation of movement when static stimuli are viewed. *Perception & Psychophysics*, 33(6), 575–581.
- Fukuda, S., Goodall, G., Michel, D., & Beecher, H. (2012). Is Magnitude Estimation worth the trouble? In *Proceedings of the 29th West Coast Conference on Formal Linguistics* (pp. 328–336).
- Garner, W. R., & Clement, D. E. (1963). Goodness of pattern and pattern uncertainty. *Journal of Verbal Learning and Verbal Behavior*, 2(5-6), 446–452.
- Geeraerts, D. (2006). Methodology in Cognitive Linguistics. In G. Kristiansen, M. Achard, R. Dirven, & F. Ruiz de Mendoza Ibañez (Eds.), *Cognitive linguistics: Current applications and future perspectives* (pp. 21–49). Berlin/New York: Mouton de Gruyter. doi: 10.1515/9783110197761.1.21
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3), 515–534.
- Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3), 373–390.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2), 155–170.
- Gentner, D. (2001). Exhuming similarity. *Behavioral and Brain Sciences*, 24(04), 669–669.
- Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American psychologist*, 52(1), 45.
- Gentner, D., Rattermann, M. J., Markman, A., & Kotovsky, L. (1995). Two forces in the development of relational similarity. *Developing cognitive competence: New approaches to process modeling*, 263–313.
- Gershman, S., & Tenenbaum, J. (2015). Phrase similarity in humans and machines. In *Proceedings of the 37th annual conference of the cognitive science society*.
- Gibbs, R. W. (2006). Introspection and cognitive linguistics: should we trust our own intuitions? *Annual Review of Cognitive Linguistics*, 4(1), 135–151.

- 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, K. (2011). Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass*, 5(8), 509–524.
- Gibson, E., Piantadosi, S. T., & Fedorenko, E. (2013). Quantitative methods in syntax/semantics research: A response to sprouse and almeida (2013). *Language and Cognitive Processes*, 28(3), 229–240.
- Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(1), 3-28.
- Goldstone, R. L., Day, S., & Son, J. Y. (2010). Comparison. In *Towards a theory of thinking* (pp. 103–121). Springer.
- Goldstone, R. L., & Medin, D. L. (1994). Time course of comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(1), 29-50.
- Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive psychology*, 23(2), 222–262.
- Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, 130(1), 116.
- Gonzalez-Marquez, M. (2007). *Methods in cognitive linguistics* (Vol. 18). John Benjamins Publishing.
- Goodman, N. (1972). *Problems and projects*. Bobbs Merrill.
- Graf, M. (2006). Coordinate transformations in object recognition. *Psychological Bulletin*, 132(6), 920.
- Grimm, L. R., Rein, J. R., & Markman, A. B. (2012). Determining transformation distance in similarity: Considerations for assessing representational changes a priori. *Thinking & Reasoning*, 18(1), 59–80.
- Guan, Y., Wang, X., & Wang, Q. (2008). A new measurement of systematic similarity. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(4), 743–758.
- Hahn, U. (2014). Similarity. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(3), 271–280.
- Hahn, U., & Bailey, T. M. (2005). What makes words sound similar? *Cognition*, 97(3), 227–267.
- Hahn, U., & Chater, N. (1997). Concepts and similarity. *Knowledge, concepts and categories*, 43–92.
- Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition*, 87(1), 1–32.
- Hahn, U., Close, J., & Graf, M. (2009). Transformation direction influences shape-similarity judgments. *Psychological Science*, 20(4), 447–454.
- Hahn, U., Prat-Sala, M., Pothos, E. M., & Brumby, D. P. (2010). Exemplar similarity and rule application. *Cognition*, 114(1), 1–18.
- Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2), 8–12.
- Hartley, J. (2014). Some thoughts on Likert-type scales. *International Journal of Clinical and Health Psychology*, 14(1), 83–86.
- Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, 71(1), 105–131.
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598), 1569–1579.
- Häussler, J., & Juzek, T. (2017). Detecting and discouraging non-cooperative behavior in online experiments using an acceptability judgment task. In *A blend of MaLT: Selected contributions from the methods and linguistic theories symposium 2015* (Vol. 15, p. 73).
- Häussler, J., Juzek, T., & Wasow, T. (2016). To be grammatical or not to be grammatical – Is that the question?.
- Hendrickson, A., Navarro, D. J., & Donkin, C. (n.d.). Evidence accumulation in same-different judgments: Integrating featural similarity with structural knowledge using a linear ballistic accumulator.
- Hendrickson, A., Navarro, D. J., & Donkin, C. (2015). Quantifying the time course of similarity. In D. Noelle et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society* (pp. 908–913).

- Hindle, D., & Ivan, S. (1975). Some more on anymore. In *Analyzing variation in language: Papers from the second colloquium on new ways of analyzing variation* (p. 89-110).
- Hitzman, D. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review*, *93*(4), 411–428.
- Hodgetts, C. J., & Hahn, U. (2012). Similarity-based asymmetries in perceptual matching. *Acta psychologica*, *139*(2), 291–299.
- Hodgetts, C. J., Hahn, U., & Chater, N. (2009a). The role of transformations and structure in the same-different paradigm. In *Proceedings of the 31st annual conference of the cognitive science society*.
- Hodgetts, C. J., Hahn, U., & Chater, N. (2009b). Transformation and alignment in similarity. *Cognition*, *113*(1), 62–79.
- Hofmeister, P., Jaeger, T. F., Arnon, I., Sag, I. A., & Snider, N. (2013). The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments. *Language and Cognitive Processes*, *28*(1-2), 48–87.
- Honoré-Chedozeau, C., Lelièvre-Desmas, M., Ballester, J., Chollet, S., & Valentin, D. (2017). Knowledge representation among assessors through free hierarchical sorting and a semi-directed interview: Exploring Beaujolais wines. *Food Quality and Preference*, *57*, 17–31.
- Hummel, J. E., & Holyoak, K. J. (2005). Relational reasoning in a neurally plausible cognitive architecture an overview of the LISA project. *Current Directions in Psychological Science*, *14*(3), 153–157.
- Imai, S. (1977). Pattern similarity and cognitive transformations. *Acta Psychologica*, *41*(6), 433–447.
- Jamieson, S. (2004). Likert scales: how to (ab) use them. *Medical education*, *38*(12), 1217–1218.
- Johnson, K. (2011). *Quantitative methods in linguistics*. John Wiley & Sons.
- Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika*, *68*(4), 563–583.
- Juzek, T. S. (2015). *Acceptability judgement tasks and grammatical theory* (Unpublished doctoral dissertation). University of Oxford.
- Káldy, Z., & Leslie, A. M. (2003). Identification of objects in 9-month-old infants: integrating ‘what’ and ‘where’ information. *Developmental Science*, *6*(3), 360–373.
- Keller, F. (2003). A psychophysical law for linguistic judgments. In *Proceedings of the 25th annual conference of the cognitive science society* (pp. 652–657).
- Keller, F., & Asudeh, A. (2001). Constraints on linguistic coreference: structural vs. pragmatic factors. In *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 483–488).
- Kemp, C., Bernstein, A., & Tenenbaum, J. B. (2005). A generative theory of similarity. In *Proceedings of the 27th annual conference of the cognitive science society* (pp. 1132–1137).
- Kemp, S., & Grace, R. C. (2010). When can information from ordinal scale variables be integrated? *Psychological methods*, *15*(4), 398.
- Kline, P. (2013). *Handbook of psychological testing*. Routledge.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.
- Lamberts, K. (2003). An exemplar model for perceptual categorization of events. *Psychology of Learning and Motivation*, *44*, 227–260.
- Larkey, L. B., & Love, B. C. (2003). CAB: Connectionist analogy builder. *Cognitive Science*, *27*(5), 781–794.
- Larkey, L. B., & Markman, A. B. (2005). Processes of similarity judgment. *Cognitive Science*, *29*(6), 1061–1076.
- Lau, J. H., Clark, A., & Lappin, S. (2016). Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cognitive Science*.
- Lee, J., Jones, P., Mineyama, Y., & Zhang, X. E. (2002). Cultural differences in responses to a Likert scale. *Research in nursing & health*, *25*(4), 295–306.
- Li, L., Malave, V., Song, A., & Yu, A. (2016). Extracting human face similarity judgments: Pairs or triplets? *Journal of Vision*, *16*(12), 719–719.
- Li, M., Chen, X., Li, X., Ma, B., & Vitányi, P. M. (2004). The similarity metric. *IEEE Transactions on Information Theory*, *50*(12), 3250–3264.
- Likavec, S., & Cena, F. (2015). Property-based semantic similarity: What counts? In *Proceedings of the 3rd international workshop on artificial intelligence and cognition* (pp. 116–124).

- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140:44–60.
- Linzen, T., & Oseki, Y. (2015). The reliability of acceptability judgments across languages. *New York: New York University*.
- Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, 355(6325), 584–585.
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, 111(2), 309–332.
- Lovett, A., Gentner, D., Forbus, K., & Sagi, E. (2009). Using analogical mapping to simulate time-course phenomena in perceptual similarity. *Cognitive Systems Research*, 10(3), 216–228.
- Mahowald, K., Graff, P., Hartman, J., & Gibson, E. (2016). SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language*, 92(3), 619–635.
- Manning, C. D. (2003). Probabilistic syntax. *Probabilistic linguistics*, 289–341.
- Mark, L. S., Todd, J. T., & Shaw, R. E. (1981). Perception of growth: A geometric analysis of how different styles of change are distinguished. *Journal of Experimental Psychology: Human Perception and Performance*, 7(4), 855.
- Markman, A. B. (1999). *Knowledge representation*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Markman, A. B., & Gentner, D. (1993a). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language*, 32(4), 517–535.
- Markman, A. B., & Gentner, D. (1993b). Structural alignment during similarity comparisons. *Cognitive psychology*, 25(4), 431–467.
- McNicol, D. (2005). *A primer of signal detection theory*. Psychology Press.
- Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological review*, 100(2), 254.
- Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive psychology*, 32(1), 49–96.
- Milajevs, D., & Griffiths, S. (2016). Treating similarity with respect: How to evaluate models of meaning. *lingbuzz/002984*.
- Miller, B., Hemmer, P., Steyvers, M., & Lee, M. D. (2009). The wisdom of crowds in rank ordering problems. In *9th International Conference on Cognitive Modeling*.
- Ming, L., & Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its applications*. Springer Heidelberg.
- Morey, R. D., Rouder, J. N., & Jamil, T. (2014). BayesFactor: Computation of Bayes factors for common designs [Computer software manual]. Retrieved from <http://CRAN.R-project.org/package=BayesFactor> (R package version 0.9.8)
- Müller, M., van Rooij, I., & Wareham, T. (2009). Similarity as tractable transformation. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 50–55).
- Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., . . . Tily, H. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk* (pp. 122–130).
- Murphy, B., Vogel, C., & Opitz, C. (2006). Cross-linguistic empirical analysis of constraints on passive. In *Presentation to the Symposium on Interdisciplinary Themes in Cognitive Language Research*.
- Myers, J. (2009). The design and analysis of small-scale syntactic judgment experiments. *Lingua*, 119(3), 425–444.
- Myers, J. (2012). Testing adjunct and conjunct island constraints in Chinese. *Language and Linguistics*, 13(3), 437.
- Navarro, D. J. (2006). From natural kinds to complex categories. In *Proceedings of the 28th annual conference of the cognitive science society* (pp. 621–626).
- Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, 36(2), 187–223.
- Navarro, D. J., & Griffiths, T. L. (2008). Latent features in similarity judgments: A nonparametric bayesian approach. *Neural computation*, 20(11), 2597–2628.
- Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, memory, and cognition*, 10(1), 104.

- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.
- Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual review of Psychology*, *43*(1), 25–53.
- Nosofsky, R. M., Sanders, C. A., Gerdman, A., Douglas, B. J., & McDaniel, M. A. (2017). On learning natural-science categories that violate the family-resemblance principle. *Psychological Science*.
- Noveck, I., & Reboul, A. (2008). Experimental pragmatics: A Gricean turn in the study of language. *Trends in Cognitive Sciences*, *12*(11), 425–431.
- Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 510.
- O’Mahony, M. (2003). Discrimination testing: a few ideas, old and new. *Food Quality and Preference*, *14*(2), 157–164.
- Phillips, C. (2009). Should we impeach armchair linguists? *Japanese/Korean Linguistics*, *17*, 49–64.
- Phillips, C., & Lasnik, H. (2003). Linguistics and empirical evidence: Reply to Edelman and Christiansen. *Trends in Cognitive Sciences*, *7*(2), 61–62.
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (pp. 20–22).
- Podesva, R., & Sharma, D. (2014). *Research methods in linguistics*. Cambridge University Press.
- Pollick, F. E., & Paterson, H. (2008). Movement style, movement features, and the recognition of affect from human movement. *Understanding events: From perception to action*, 286–308.
- Porte, G. (2013). Who needs replication? *CALICO Journal*, *30*(1), 10–15.
- Pothos, E., & Busemeyer, J. (2011). A quantum probability explanation for violations of symmetry in similarity judgments. In *Proceedings of the cognitive science society* (Vol. 33).
- Rensink, R. A., O’Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, *8*(5), 368–373.
- Riesbeck, C. K., & Schank, R. C. (2013). *Inside case-based reasoning*. Psychology Press.
- Roberts, J., Laughlin, J., & Wedell, D. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement*, *59*(2), 211–233.
- Rosenbach, A. (2003). Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive in English. *Topics in English Linguistics*, *43*, 379–412.
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225–237.
- Russell, S. J. (1986). A quantitative analysis of analogy by similarity. In *Proceedings of the national conference on artificial intelligence* (pp. 284–288).
- Sadanand, S., & Corso, J. J. (2012). Action bank: A high-level representation of activity in video. In *Computer vision and pattern recognition (CVPR)* (pp. 1234–1241).
- Sagi, E., Gentner, D., & Lovett, A. (2012). What difference reveals about similarity. *Cognitive Science*, *36*(6), 1019–1050.
- Sampson, G. (2007). Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory*, *3*(1), 1–32.
- Sassoon, G. W. (2011). Adjectival vs. nominal categorization processes: The rule vs. similarity hypothesis. *Belgian Journal of Linguistics*, *25*(1), 104–147.
- Schütze, C. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. University of Chicago Press.
- Schütze, C. (2011). Linguistic evidence and grammatical theory. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(2), 206–221.
- Schütze, C., & Sprouse, J. (2014). Judgment data. In R. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (p. 27–51). Cambridge University Press.
- Selker, R., Lee, M. D., & Iyer, R. (2017). Thurstonian cognitive models for aggregating top-n lists. *Decision*, *4*(2), 87.
- Shafto, P., & Coley, J. D. (2003). Development of categorization and reasoning in the natural world: Novices to experts, naive similarity to ecological knowledge. *Journal of Experimental Psychology-Learning Memory and Cognition*, *29*(4), 641–648.

- Shahbazi, R., Raizada, R., & Edelman, S. (2016). Similarity, kernels, and the fundamental constraints on cognition. *Journal of Mathematical Psychology*, *70*, 21–34.
- Shalom, D. B., & Poeppel, D. (2007). Functional anatomic models of language: assembling the pieces. *The Neuroscientist*.
- Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.
- Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika*, *27*(2), 125–140.
- Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika*, *39*(4), 373–421.
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.
- Shulman, H. G. (1971). Similarity effects in short-term memory. *Psychological Bulletin*, *75*(6), 399.
- Simons, D., & Rensink, R. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences*, *9*(1), 16–20.
- Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York: Appleton-Century-Crofts.
- Skinner, B. F. (1958). *Verbal behavior*. Acton, MA: Copley Publishing Group.
- Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition*, *49*(1), 67–96.
- Sorace, A. (2010). Using magnitude estimation in developmental linguistic research. *Experimental methods in language acquisition research*, 57–72.
- Sorace, A., & Keller, F. (2005). Gradiance in linguistic data. *Lingua*, *115*(11), 1497–1524.
- Soto, F. A., Gershman, S. J., & Niv, Y. (2014). Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychological review*, *121*(3), 526.
- Spencer, N. J. (1973). Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of psycholinguistic research*, *2*(2), 83–98.
- Sprouse, J. (2007). Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics*, *1*, 123–134.
- Sprouse, J. (2008). Magnitude estimation and the non-linearity of acceptability judgments. *Proceedings of the West Coast Conference on Formal Linguistics*, *27*, 397–403.
- Sprouse, J. (2011a). A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language*, *87*(2), 274–288.
- Sprouse, J. (2011b). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167.
- Sprouse, J., & Almeida, D. (2011). Power in acceptability judgment experiments and the reliability of data in syntax. *Master's Thesis, University of California, Irvine & Michigan State University*.
- Sprouse, J., & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*, *48*(03), 609–652.
- Sprouse, J., & Almeida, D. (2017). Design sensitivity and statistical power in acceptability judgment experiments. *Glossa*, *2*(1), 1.
- Sprouse, J., & Almeida, D. (in press). Setting the empirical record straight: Acceptability judgments appear to be reliable, robust, and replicable. *Behavioral and Brain Sciences*.
- Sprouse, J., & Schütze, C. T. (2017). Grammar and the use of data. In *The Oxford Handbook of English Grammar Location*. Oxford 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. (1956). The direct estimation of sensory magnitudes: Loudness. *The American journal of psychology*, *69*(1), 1–25.
- Taylor, E. G., & Hummel, J. E. (2007). Perspectives on similarity from the LISA model. *Analogies: Integrating Multiple Cognitive Abilities*, *5*, 21.
- Taylor, E. G., & Hummel, J. E. (2009). Finding similarity in a model of relational reasoning. *Cognitive Systems Research*, *10*(3), 229–239.

- Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, 24(04), 629–640.
- Thurstone, L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273.
- Thurstone, L. (1931). Rank order as a psycho-physical method. *Journal of Experimental Psychology*, 14(3), 187.
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136.
- Troje, N. F., & Basbaum, A. (2008). Biological motion perception. *The senses: A comprehensive reference*, 2, 231–238.
- Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, 76(1), 31.
- Tversky, A. (1977). Features of similarity. *Psychological review*, 84(4), 327.
- Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological review*, 89(2), 123.
- Ullman, S. (1996). *High-level vision: object recognition and visual cognition*. Bradford/MIT Press.
- Vasishth, S., Nicenboim, B., Chopin, N., & Ryder, R. (2017). Bayesian hierarchical finite mixture models of reading times: A case study.
- Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19(7), 645–647.
- Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., ... Tang, X. (2017). Residual attention network for image classification. *ILSVRC2017 winners: method described in arXiv:1704.06904 preprint*.
- Wasow, T., & Arnold, J. (2005). Intuitions in linguistic argumentation. *Lingua*, 115(11), 1481–1496.
- Watanabe, S. (1969). *Knowing and guessing: a quantitative study of inference and information*. Wiley, New York.
- Watanabe, S. (1985). *Pattern recognition: human and mechanical*. John Wiley & Sons, Inc.
- Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological review*, 20(2), 158.
- Wattenmaker, W. D., Nakamura, G. V., & Medin, D. L. (1988). Relationships between similarity-based and explanation-based categorization. In *Contemporary science and natural explanation: Commonsense conceptions of causality* (p. 205-241). New York University Press.
- Weskott, T., & Fanselow, G. (2008). Variance and informativity in different measures of linguistic acceptability. In *Proceedings of the 27th West Coast Conference on Formal Linguistics (WCCFL)* (pp. 431–439).
- Weskott, T., & Fanselow, G. (2009). Scaling issues in the measurement of linguistic acceptability. *The Fruits of Empirical Linguistics*, 1, 229–245.
- Weskott, T., & Fanselow, G. (2011). On the informativity of different measures of linguistic acceptability. *Language*, 87(2), 249–273.
- Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7), 301–308.
- Yusoff, R., & Mohd Janor, R. (2014). Generation of an interval metric scale to measure attitude. *SAGE Open*, 4(1), 2158244013516768.

Part V

Appendices

A Sentence stimuli

A.1 Attention check questions

In addition to the 300 sentences below, which formed the basis of the analyses in this paper, each experiment incorporated some attention-check questions. These were intended to be used to identify participants who were not paying attention to the task or did not understand it. As such, they were two pairs of sentences with a clear answer, as shown below (the first sentence of the pair is the better one).

- Mike read things quickly. / Read things, Mike did quickly.
- Eva was killed by John. / Eva was killed from John.

In the INITIAL / WITHIN PARTICIPANTS two-session experiment format, in TARGET PAIRS and RANDOM PAIRS choice tasks, both pairs of sentences were presented in both halves of the experiment. In the LIKERT and ME rating tasks, one sentence from each pair was presented in the first half and the other sentence from each was presented in the second half. A participant was considered to have failed the attention check if they chose the incorrect sentence over the correct one in a choice task or ranked an incorrect sentence more highly than a correct one in a rating task.

The BETWEEN PARTICIPANTSSingle-session format was identical except that no sentences appeared in the second half (since there was no second half). Participant exclusions were determined before any other analyses and the attention check questions were not included in the main analysis for any measure.

A.2 Instruction quiz

Participants in all conditions had to pass a two-question instruction quiz testing their understanding of acceptability as defined in the instructions (reported in the main *Method* section. These may have influenced their interpretation of the task. The questions asked were:

Which of these best describes 'sentence acceptability', the thing you'll be asked to make judgments about?

- How likely the sentence is to offend someone.
- Whether the sentence expresses positive or negative sentiment.
- How 'well formed' the sentence is, whether it sounds natural.
- How likely the sentence is to be true.

How can you tell us about the acceptability of these sentences?

- You'll be asked to give sentences an acceptability rating.
- You'll be asked to compare two sentences and say which one is more acceptable
- You'll be asked to describe what is wrong with some badly worded sentences.
- Both of options 1 and 2, but not 3.

Expected answers were options 3 and 4 respectively.

A.3 Sentence stimuli

The sentences used here were a randomly selected subset of those collected by Sprouse et al. (2013). Although drawing randomly from the instances presented for each phenomenon, we did sample uniformly across the 150 sources present in the full collection when drawing the set of 150 pairs of sentences used here. Sprouse et al. (2013) identify repeated phenomena among this set, one example appearing in our subset is the pair of sentences *I seem eating sushi* / *I like eating sushi* and the pair *I tend taking vacations* / *I like taking vacations*, which Sprouse et al. (2013) consider equivalent. In addition to direct repeats, we expect further theoretical connections to be

present among related but non-identical phenomena, but for the purposes of the analyses presented above the critical property of this set of sentences is just that they cover a range of acceptability values representative of linguistics research. As noted in the main text, multiple items testing a single phenomenon are necessary to make reliable claims regarding its status: occasional repeats notwithstanding, in general we chose to sacrifice the ability to make claims about the status of individual phenomena in order to increase the diversity of the contrasts considered, since the latter is more directly relevant to the test-retest measure of reliability.

Pair Id	Proposed as acceptable	Proposed less acceptable
1	Brittany attempted to touch the porcupine.	Brittany touched plenty of.
2	It seems to him that Kim solved the problem.	He seems to that Kim solved the problem.
3	Scott intended to run for class president.	Scott intended to have run for class president.
4	Debra convinced Elliot that he would wash the dishes.	Debra convinced Elliot to have washed the dishes.
5	Celia convinced Brad that he will have eaten dinner by the time she gets home.	Celia convinced Brad that he would have eaten dinner by the time she gets home.
6	Charlene believed Shawn to be helpful.	Charlene believed Shawn to write the paper.
7	Thomas read in the paper that the stock market crashed.	Thomas read in the paper the stock market crashed.
8	It is important for one to sleep regularly.	It is important one to sleep regularly.
9	The expectation that Lauren will graduate is reasonable.	The expectation Lauren will graduate is reasonable.
10	How likely to sell the house is Diane?	How likely to be a stock market crash is there?
11	Carla appears to have graduated high school.	Carla was hoped to graduate high school.
12	'Stop bullying me!' shouted the overweight child fearfully.	'Stop bullying me!' shouted fearfully the overweight child.
13	Jason drove his car, and Tara rode her bike.	Jason drove his car, and Tara her bike rode.
14	Annie was insulted.	Was insulted Annie.
15	The money was stolen.	The money was stolen the money.
16	Beth hitchhiked to Los Angeles and Robert drove to San Diego.	Beth hitchhiked Los Angeles and Robert drove to San Diego.
17	Richard may have been hiding, but Blake may have been doing so too.	Richard may have been hiding, but Blake may have done so too.
18	They all have eaten and they have all done so quickly.	They all have eaten and they have done all so quickly.
19	Arty is sick, and Mark is too.	Arty is sick, and Mark does so too.
20	John tried to win.	John tried himself to win.
21	Dale found Brooke after frightening himself.	Dale found Brooke after frightening herself.
22	Olivia told Gregory when to exonerate himself.	Olivia told Gregory when to exonerate herself.
23	Last month there was a plan to promote me.	Last month there was a plan to promote oneself.
24	Edward created a website in order for us to promote ourselves.	Edward created a website in order for us to promote himself.
25	Melissa's pledge to Dan to take care of herself.	Melissa's pledge to Dan to take care of himself.
26	Eric shouted to Maria to believe in herself.	Eric shouted to Maria to believe in himself.
27	Jack asked Sally to be allowed to take care of himself.	Jack asked Sally to be allowed to take care of herself.
28	She called Victor and yelled at him.	She called Victor and yelled at.
29	There has been an announcement made in the newspaper.	There has been made an announcement in the newspaper.
30	There has been a man considered violent.	There has been considered a man violent.
31	She wrote her boyfriend a poem.	She wrote a poem her boyfriend.
32	We believed Ben with all our hearts to be innocent.	We believed with all our hearts Ben to be innocent.
33	We consider there to be three problems.	We consider there three problems.
34	I told Mandy who would win the election.	I told Mandy who the hell would win the election.
35	During no storm should people leave their computers on.	What during no storm should people leave on?
36	I don't think that I will take any musicians to the office.	I don't think that any musicians, I will take to the office.
37	Who the hell asked who out?	Who asked who the hell out?
38	Clare didn't promise Rita a red cent.	Clare didn't promise every employee a red cent.
39	I visited a city yesterday near the city that John did.	I visited a city near the city yesterday that John did.
40	I told you when we met that Bill will come to the party.	I told you that Bill when we met will come to the party.
41	I suggested when we dated that Laura should cut her hair.	I suggested that Laura when we dated should cut her hair.
42	Melanie prefers for everyone to get a raise that you do.	Melanie prefers for everyone you do to get a raise.
43	It rained torrentially.	It torrentially rained.
44	Mike believes Phil to be a genius.	Mike believes to be a genius.
45	There are likely to be students in the library.	There are likely students to be in the library.
46	There might seem to be fossils several miles underground.	There might fossils seem to be several miles underground.
47	The flowers were planted to attract hummingbirds.	The flowers plant seasonally to attract hummingbirds.
48	The bureaucrat was bribed deliberately.	The bureaucrat bribes deliberately.
49	The book was written truthfully.	The was written book truthfully.

50 Picking which suit to wear in the morning makes Helen late
to work.
51 Someone better leave town.
52 John wondered what he bought.
53 Which coworker did George compliment before insulting?
54 It appears that tonight Marjorie is staying over.
55 Who did I see hug Natalie?
56 There are leaves burnt.
57 The doctors are almost all wealthy.
58 What do you complain that the neighbor turns on at night?
59 It appears that a certain player will leave the team, but which
player is still a secret.
60 The professor said that the score of one the assignments is
going to be posted online by the end of the day, but I don't
remember which.
61 Melissa said she read about one of Shakespeare's plays, but
I don't know which play.
62 Richard and Christine hate each other.
63 Sabrina gave her elderly mother a beautiful new coffee mug
at breakfast time.
64 The poor were overlooked by the politicians and the bankers.
65 The government gives no help to any poor people.
66 Julian learned to drive, and Rita learned to compose music.
67 The instructor put more solutions than the TA did on the
board.
68 John intended to give the children something nice to eat,
and give the children a generous handful of candy he did.
69 The table was being set by the waiter.
70 We supporters of democrats are just as worried about the
economy as you supporters of republicans.
71 He was the assistant.
72 The brother and sister that were playing all the time had to
be sent to bed.
73 The hammer with the black handle and the screwdriver with
the square tip are in your toolbox.
74 The book is long and the essay is short.
75 This is a shelf.
76 The lobby of the movie theater with the fantastic sound sys-
tem was empty.
77 There seems to be a new deal in the works.
78 One interpreter tried to be assigned to every visiting diplo-
mat.
79 Nobody has gone anywhere.
80 The virtuoso only rarely practices any pieces.
81 It seemed at that time that Robert had confessed.
82 What they children believe is that they will get some candy.
83 They denied and we suspected that Sean would buy the car.
84 Matt believed that Ben read a book and Lilly that Ben
watched a movie.
85 That Addison bit the boy, Jena didn't believe.
86 At that time, what did they believe that Peter fixed?
87 This is the child who I think will walk your dog.
88 Eric wondered how Lisa learned to dance a certain dance,
but it's not clear what dance.
89 They consider a teacher of Chris geeky.
90 Cassie is more talented than intelligent.
91 They anticipate that everybody will contact Fred that you
do.
92 Fran searched the web for Danny.
93 The eerie sound frightened Seth.
94 I find it annoying that usually this bus is late.
95 A dog bigger than my corgi started snarling.
96 Jessica shouted at a girl as nervous as her daughter.

Picking who to wear that new suit in the morning makes late
to work.
Anyone better leave town.
What did John wonder what he bought?
Which coworker did George yawn before insulting?
Tonight appears that Marjorie is staying over.
Who was seen hug Natalie?
There are leaves green.
The doctors almost all are wealthy.
What do you notice if the neighbor turns on at night?
It appears that a certain player will leave the team, but which
player it does is still a secret.
The professor said that the score of one the assignments is
going to be posted online by the end of the day, but I don't
remember which he did.
Melissa said she read about one of Shakespeare's plays, but
I don't know which play she did.
Each other hate Richard and Christine.
Sabrina gave at breakfast time her elderly mother a beautiful
new coffee mug.
The poor were overlooked and the bankers by the politicians.
The government gives any help to no poor people.
Julian learned to drive, and Rita did to compose music.
The instructor put more solutions on the board than the TA
did on the handout.
John intended to give the children something nice to eat,
and give the children he did a generous handful of candy.
The table was being heavy.
We supporters of democrats are just as worried about the
economy as you of republicans.
He was assistant.
Brother and sister that were playing all the time had to be
sent to bed.
Hammer with the black handle and screwdriver with the
square tip are in your toolbox.
Book is long and essay is short.
This is shelf.
The movie theater with the fantastic sound system's lobby
was empty.
There desires to be a new deal in the works.
One interpreter each tried to be assigned to every visiting
diplomat.
Anywhere has nobody gone.
The virtuoso practices any pieces only rarely.
It seemed at that time Robert had confessed.
What they children believe is they will get some candy.
They denied and we suspected Sean would buy the car.
Matt believed Ben read a book and Lilly Ben watched a
movie.
Addison bit the boy, Jena didn't believe.
What did they believe at that time that Peter fixed?
This is the child who I think that will walk your dog.
Eric wondered how Lisa learned to dance, but it's not clear
what dance.
Who do they consider a teacher of geeky?
Cassie is more than intelligent talented.
They anticipate that everybody you do will contact Fred.
Fran searched Danny the web.
The eerie sound frightened Seth at the ghost.
I find it annoying for usually this bus to be late.
A bigger dog than my corgi started snarling.
Jessica shouted at as nervous a girl as her daughter.

97 The heat turned the meat rotten.
98 It seems a woman is in the yard.
99 There is likely to appear a man.
100 It is unimaginable for Mary to arrive on time.
101 What did you contribute to whom?
102 Some frogs and a fish are in the pond.
103 The more you give, the happier you will be.
104 Brian asked who sent what.
105 He met Bush, about whom he wrote several stories.
106 He would have been fired.
107 There tend to be storms at this time of year.
108 I like eating sushi.
109 I like taking vacations.
110 The tennis players watched the ball bounce a crazy bounce
off the line.
111 How much money is there in your account?
112 Into which room walked three men?
113 Onto which floor did the wine glass fall?
114 Mr. Ted said we should write five pages each night, but the
actual doing of it turned out to be impossible.
115 From the Atlantic to the Pacific there are people who oppose
immigration reform.
116 It will take five to seven days for your cat to feel better.
117 The newlyweds travelled a few days and then settled down
at a motel.
118 Larry cooked her husband the meal
119 There seems to be a box on your doorstep.
120 Dale was suspended by the principal.
121 The freezer was defrosted to remove the ice build-up.
122 The chamber was flooded intentionally.
123 John broke a cup, and Mary did so too.
124 Linda graduated high school before her brother.
125 Leonard talks about the things Sally likes.
126 Eric began a romantic relationship with an employee who he
later found himself superior to.
127 Who brought what?
128 He envied me my success after the promotion.
129 Maggie is difficult to compliment without embarrassing.

130 That Mary was going out with Luke bothered you.
131 Angela wonders which story about Elaine is online.
132 The teacher denied George his extra credit.
133 In that room anyone who stays long enough is given horrible
headaches.
134 The prom queen was picked some flowers before the award
ceremony.
135 The child that the clown gave the creeps to at the party is
still upset.
136 The children are certain to have all been picked up.
137 Shannon stole food for her family to enjoy.
138 The question was answered feeling nervous.
139 I informed Stan that I wanted to wash the car together.
140 The gardener and handyman painted the fence together.
141 There have arisen in these negotiations all of the issues the
lawyers warned us about.
142 Caroline likes cupcakes and Lisa cookies.
143 Kelsey speaks Japanese more fluently than Jason does.
144 The girls made sandwiches, and I believe that the boys did
too.
145 The bully threatened students triumphantly.
146 Sarah expected to receive a good grade.
147 Who did the executives appoint a friend of to the board?

The heat turned the meat rotted.
It seems a woman to be in the yard.
There is likely a man to appear.
It is unimaginable Mary to arrive on time.
To whom did you contribute what?
Some frogs and a fish is in the pond.
That much the more you give, the happier you will be.
Brain asked what who sent.
He met about whom he wrote several stories.
Him would have been fired.
There like to be storms at this time of year.
I seem eating sushi.
I tend taking vacations.
My parents appeared an unexpected appearance last night.

How much money there is in your account?
Into which room did walk three men?
Onto which floor fell the wine glass?
Mr. Ted said we should write five pages each night, but the
actual doing of so turned out to be impossible.
With the generals about a new approach the president
talked.
It will take from five seven days for your cat to feel better.
A few days, the newlyweds travelled, and then settled down
at a motel.
Who did Larry cook the meal?
There seems a box to be on your doorstep.
Dale was suspended from the principal.
The freezer defrosted to remove the ice build-up.
The chamber flooded intentionally.
John broke a cup, and Mary did so with a saucer.
Linda graduated high school her brother.
Leonard talks what Sally likes.
Eric began a romantic relationship with an employee supe-
rior to who later found himself.
What did who bring?
He envied me after the promotion my success.
Maggie is unlikely to be complimented without anyone em-
barrassing.
Who did that Mary was going out with bother you?
Who does Angela wonder which story about is online?
The teacher denied his extra credit to George.
In that room is given anyone who stays long enough horrible
headaches.
Before the award ceremony was picked the prom queen some
flowers.
The child that the clown gave the creeps at the party to is
still upset.
All the children are certain to have all been picked up.
Shannon stole for her family to enjoy.
The question was answered nervous.
I informed Stan that I must wash the car together.
The gardener painted the fence together.
There have all arisen in these negotiations the issues the
lawyers warned us about.
Caroline likes cupcakes because Lisa cookies.
Kelsey speaks Japanese more fluently than Jason English.
The girls made sandwiches, and I believe that the boys hot
dogs.
Threaten students, the bully did triumphantly.
Expect to receive, Sarah did a good grade.
Who did the executives appoint a friend of chairman of the
board?

148	When yesterday Marnie started to present that report, I	When that report Marnie started to present yesterday, I
	thought we were in for a lot of surprises.	thought we were in for a lot of surprises.
149	Bill knows that such books John only reads at home.	Bill asked if such books John only reads at home.
150	If George comes, the party will be a disaster.	If George probably comes, the party will be a disaster.

A.4 Conflict sentences

Proposed as acceptable	Proposed less acceptable	TARGET PAIRS	TARGET PAIRS	LIKERT	LIKERT
		INITIAL/WITHIN	BETWEEN	INITIAL/WITHIN	BETWEEN
Celia convinced Brad that he will have eaten dinner by the time she gets home.	Celia convinced Brad that he would have eaten dinner by the time she gets home.	~,x	~	~,~	~
How likely to sell the house is Diane?	How likely to be a stock market crash is there?	x,~	x	~,~	~
Richard may have been hiding, but Blake may have been doing so too.	Richard may have been hiding, but Blake may have done so too.	x,~	x	~,~	~
We believed Ben with all our hearts to be innocent.	We believed with all our hearts Ben to be innocent.	x,x	x	✓,✓	~
Melanie prefers for everyone to get a raise that you do.	Melanie prefers for everyone you do to get a raise.	x,x	~	~,~	~
There are leaves burnt.	There are leaves green.	x,x	~	✓,✓	✓
What do you complain that the neighbor turns on at night?	What do you notice if the neighbor turns on at night?	x,x	~	~,✓	~
The instructor put more solutions than the TA did on the board.	The instructor put more solutions on the board than the TA did on the handout.	x,x	x	x,~	~
It seemed at that time that Robert had confessed.	It seemed at that time Robert had confessed.	x,x	~	✓,✓	~
What they children believe is that they will get some candy.	What they children believe is they will get some candy.	x,x	x	~,~	~
They denied and we suspected that Sean would buy the car.	They denied and we suspected Sean would buy the car.	~,x	x	✓,~	~
That Addison bit the boy, Jena didn't believe.	Addison bit the boy, Jena didn't believe.	x,~	x	~,~	~
Eric wondered how Lisa learned to dance a certain dance, but it's not clear what dance.	Eric wondered how Lisa learned to dance, but it's not clear what dance.	x,x	x	~,~	~
They anticipate that everybody will contact Fred that you do.	They anticipate that everybody you do will contact Fred.	x,x	x	~,~	~
A dog bigger than my corgi started snarling.	A bigger dog than my corgi started snarling.	~,x	✓	~,~	~
There is likely to appear a man.	There is likely a man to appear.	x,x	~	~,~	~
What did you contribute to whom?	To whom did you contribute what?	x,~	x	~,~	~
Who did the executives appoint a friend of to the board?	Who did the executives appoint a friend of chairman of the board?	~,~	x	~,~	~
Bill knows that such books John only reads at home.	Bill asked if such books John only reads at home.	~,~	x	✓,✓	✓

This table lists only those pairs of sentences for which there is disagreement in the decisions made by different measures. Expert judgments from *Linguistic Inquiry* are taken as the reference standard, ✓ indicates that the formal measure endorses the expert-preferred item, × indicates the formal measure endorses the less preferred item, and ~ indicates the formal measure failed to reject the null hypothesis of no difference. Two relevant formal measures are given, TARGET PAIRS for its distinctively high power and LIKERT to represent the consensus of the other formal measures. Since this data set does not in general contain multiple items targeting the same phenomena, little can be said about any possible relation between particular sentence structures and the behavior of any one measure. What is important from a test-retest reliability viewpoint is that there are different ways in which disagreements can arise. Some appear to be simple sampling error, which seems to be the most likely explanation for the single instance of disagreement with expert judgment in the *Celia convinced Brad...* sentence pair (pair #1). Others appear to be instances where the high power of TARGET PAIRS allows it to pick up a counterintuitive effect too small to be detected with the other measures, such as *The instructor...* (pair #8). Finally, there do appear to be rare instances where presenting controlled contrasts directs people's attention to features of a construction that are not salient when the sentences involved are considered in isolation or contrasted with unrelated partners, such as *There are leaves...* (pair #6) or ... *Robert had confessed* (pair #9).

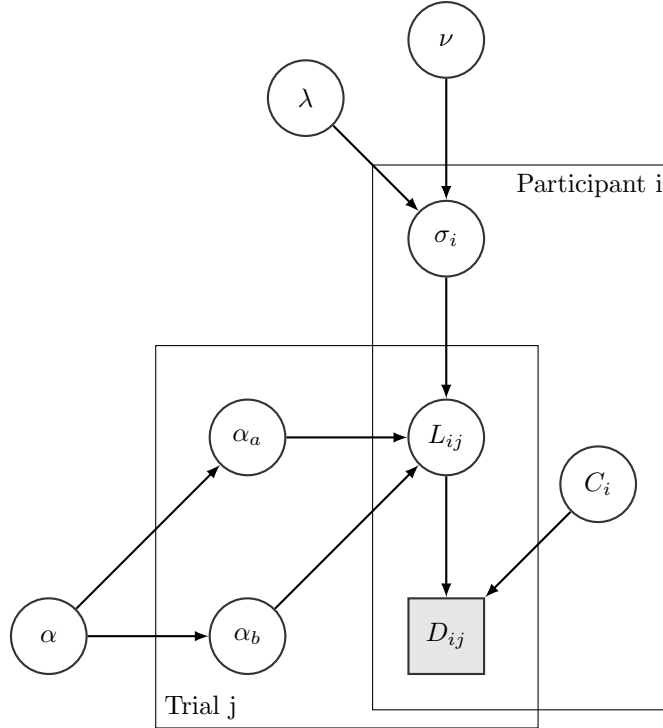


Figure 8.1: Model diagram showing the implementation of the THURSTONE model. Variable definitions are given in Table 8.3.

Table 8.3: Parameter descriptions

Variable	Meaning	Distributed as
D_{ij}	Observed decision	1 if $L_{ij} > C$, -1 if $L_{ij} < -C$, 0 otherwise
L_{ij}	Decision variable for trial j	$N(\alpha_a - \alpha_b, \sigma_i)$
C	Participant decision criterion	$U(1, 50)$
σ	Participant variability	$Weibull(\lambda, \nu)$
λ	Variability hyperparameter	$ N(2, 5) $
ν	Variability hyperparameter	$ N(0.02, 0.04) $
α	Item acceptability	$N(0, 20)$

B Model implementation

Choice task responses were modeled using a hierarchical variation of the traditional Thurstonian model, fit using JAGS (Plummer, 2003) as outlined in Figure 8.1. Participants responded to a series of sentence pairs. For each pair, three response options were available: one for each of the sentences indicating that sentence is more acceptable, plus one to indicate no difference. Given a collection of responses, the model infers an acceptability strength for each item under the assumptions that the language community is homogeneous, and that acceptability is transitive.

The acceptability strength scores are inferred using the following decision process mapping acceptability strength to choice responses: When presented with test items a and b on trial j , participant i samples an acceptability strength for each item. By definition, these samples come from normal distributions with means centered at the group’s consensual acceptability score for each item, α_a and α_b . Each participant has their own degree of variability around the shared mean, σ_i . The participant then takes the difference between the perceived acceptability scores on this trial, L_{ij} , and compares it to their criterion level C_i . The criterion determines how each participant maps their perception of the acceptability difference onto the response options. If the absolute difference is greater than the criterion value, the participant indicates the preference indicated by the sign of the difference, otherwise they indicate no preference. Although different in some respects, this mapping process follows similar motivating logic to a scheme previously described by Bader and Häussler (2010) for mapping between graded sentence acceptability responses and binary ones.

The model requires priors on sentence acceptability, participant variability, and participant criterion values. Some aspects of these were informed by our understanding of the acceptability judgment task, such as the general shape of participant variability distribution. In cases where

we were forced to choose between a number of plausible alternatives, we tested the sensitivity of outcomes to changes in the specification of the model to examine the impact of these decisions.

The sentence acceptability scale has arbitrary units, making it relatively difficult to specify a completely terrible prior, since the primary requirement is simply that all acceptabilities appear on the same scale. However there are issues with the way the boundaries of acceptability are set. A uniform prior over the real number line is not possible, and a truncated distribution (such as a uniform distribution) can introduce distortions for sentences with extreme acceptabilities, exaggerating acceptability estimates at the ends of the scale when the prior is relatively wide and compressing them when the prior is relatively narrow (where 'wide' and 'narrow' are defined relative to the degree of variation in the acceptability of the sentences tested, which is in general unknown). Also, when the prior is very wide relative to the true variability in sentence acceptability, the model is unidentifiable, with multiple solutions corresponding to identical relative arrangements of acceptability scores falling in different regions of the scale. We considered various solutions to this problem, including forcing identifiability by fixing a reference distance between a reference pair of sentences, but eventually settled on a normal prior over acceptability, which has a number of useful properties.

Most importantly, with a normal prior over acceptability the width of the prior has no meaningful impact on the final estimates. Changing the width of the prior simply changes all acceptability estimates by a multiplicative scaling factor, as shown in Figure 8.2, which shows the perfect correlation between acceptability scores from models fit to the same data using different width priors on acceptability. Although the scores are not numerically identical (models with wider priors give numerically wider acceptability ranges) there is no meaningful difference in the information presented.

The normal distribution also results in identifiable acceptability solutions without needing to pre-select reference sentences or reference differences. For any particular relative arrangement of sentences consistent with the data, solutions centering this arrangement around the mean of the prior are preferred.

Although the width of the acceptability prior does not impact the final estimates, the steepness of the distribution tails does. An inevitable feature of comparison data is that there is relatively little constraint over the placement of extreme items which either always win or always lose comparisons. While it's clear these items belong at the ends of the scale, there is little or no information constraining the distance between them and the next most extreme items. Since the data contain little relevant information, these distances are strongly influenced by the shape of the prior distribution, in particular how quickly it decreases in the tails.

We tested this in simulation by using a t-distribution as the prior over acceptability and varying the degrees of freedom.

Figure 8.3 shows the characteristic pattern on one particular simulated data set where the true distribution of sentence acceptability scores was uniform. Acceptability priors with fatter tails (fewer degrees of freedom) resulted in acceptability estimates for extreme items that were further from the mean. This tendency for the extreme items to 'fly away' can obscure the true distribution of acceptability scores (in this example, uniform).

The current version of the model attempts to mitigate this problem in two ways. One is the use of the normal prior on acceptability, which is quite steep and performed well in simulation recovery tests. The other, by analogy with a similar trick commonly used in the signal detection literature (McNicol, 2005), was to add two 'sacrificial' items to the data before fitting the model, each of which participated in a single comparison with every item in the data set, and won or lost every one. These imaginary items were then removed before proceeding to any data analysis. The presence of these imaginary items during the fitting process does not strongly constrain the width of the inferred acceptability distribution, but is helpful in moderating the scores given to extreme sentences. In particular, they make it impossible for any actual sentence to achieve a perfect record of only one type of response, which would be consistent with any distance from the rest of the sentences so long as it is on the appropriate side of the scale. Informally, the sacrificial items soak up the tendency for the ends of the scale to 'fly away', without imposing strict constraints on the form of the acceptability estimates.

The prior on participant criterion values encodes the expected rate of endorsements for the option 'these two sentences are equally acceptable'. The main consideration in choosing this prior was the scale size, defined by the width of the prior on sentence acceptability. The uniform(1,50) prior we used allows the proportion of equal responses to vary from near zero to near 100%,

without any particular expectation that any particular proportion is more likely, and without any expectation that the proportion of 'equal' responses will be similar between people. The relationship between criterion values and responses is complicated by the unknown distributions of sentence acceptability and participant variability. Given the responses observed over the course of this project, a more realistic prior for future work using this decision rule might express the expectation that people use largely similar criterion values and endorse the 'equal' option on around 25% of responses.

It was clear in simulation that a uniform prior for participant variability performed relatively poorly, but there are a range of plausible alternatives to consider (see for example Gelman (2006)). The Weibull distribution we ended up using has a number of nice properties, among them simplicity of implementation. More importantly, it flexibly expresses with only two parameters a range of distributions we felt to be plausible for participant variability in the sentence acceptability context: it respects the constraint to be positive, does not strongly constrain the location of the central tendency, and can exhibit strong clustering, quite flat distributions, or distributions with a long right tail. As it turns out, the typical Weibull distributions generated from the priors usually underestimated the amount of variation needed to be consistent with the plausible values of the criterion and acceptability components of the model. A future version of this model would benefit from a prior over variability more similar to the posterior observed here after fitting to the INITIAL dataset.

The priors and posteriors for each component of the model are summarized in Figure 8.4. Together, these plots give a relatively complete picture of the model's summary of the data set. The main text focuses on the reliability properties of these results, and finds they are highly reliable.

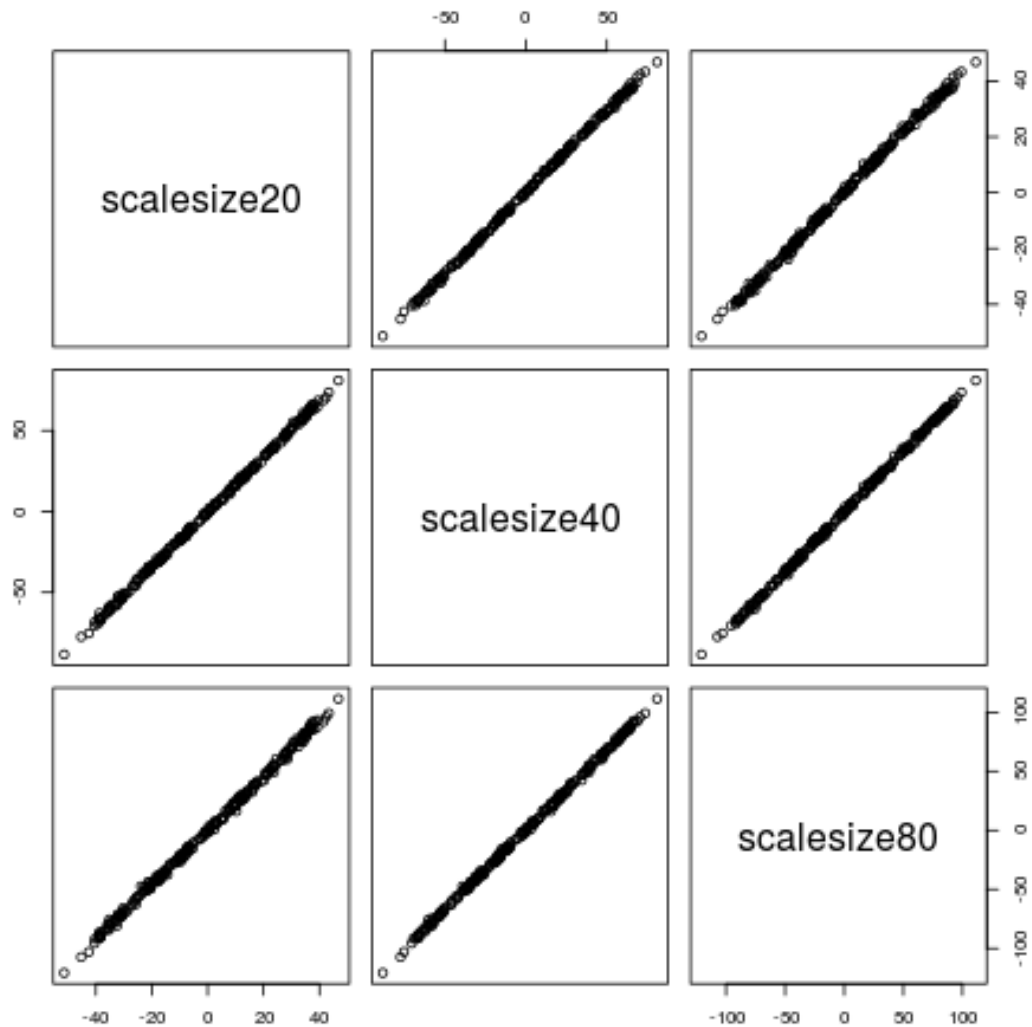


Figure 8.2: Mutual correlation between acceptability scores for the same data set from models with different width priors over sentence acceptability. All priors were normal and centered at zero, "scalesize" refers to their standard deviation. The resulting acceptability estimates were numerically different, with wider priors resulting in estimates that covered a wider range, but they were near-perfectly correlated, encoding the same information about acceptability differences between sentences. The particular data set used for this figure was a simulation one with acceptability uniformly distributed over a 100 point range.

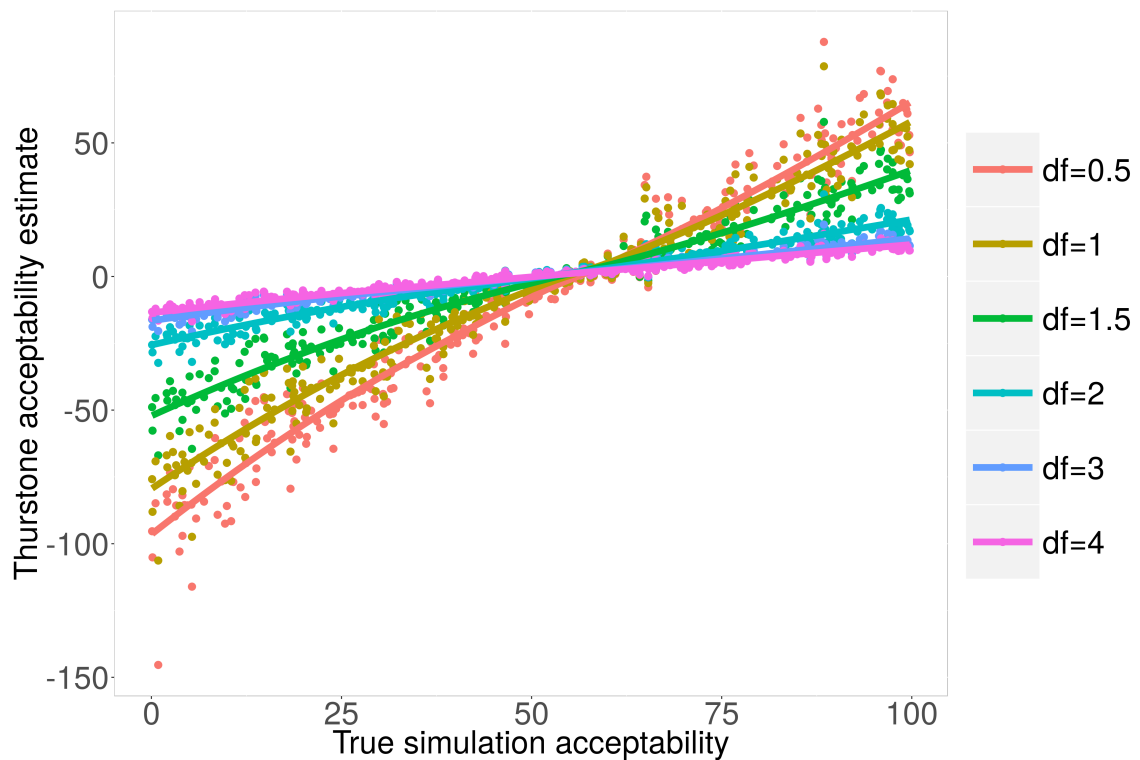


Figure 8.3: Testing the impact of different acceptability prior shapes. Although the width of the prior distribution over sentence acceptability was found not to influence the model’s acceptability estimates, the steepness of the distribution tails can. This is shown here by substituting a series of t-distributions with different degrees of freedom for the normal prior, and fitting each to the same simulated data set with uniformly distributed sentence acceptability. Lower degrees of freedom correspond to fatter tailed distributions and cause exaggeration in the acceptability distances at the extremes of the scale. Priors with tighter tails recover the simulation truth better in this case, motivating the normal prior.

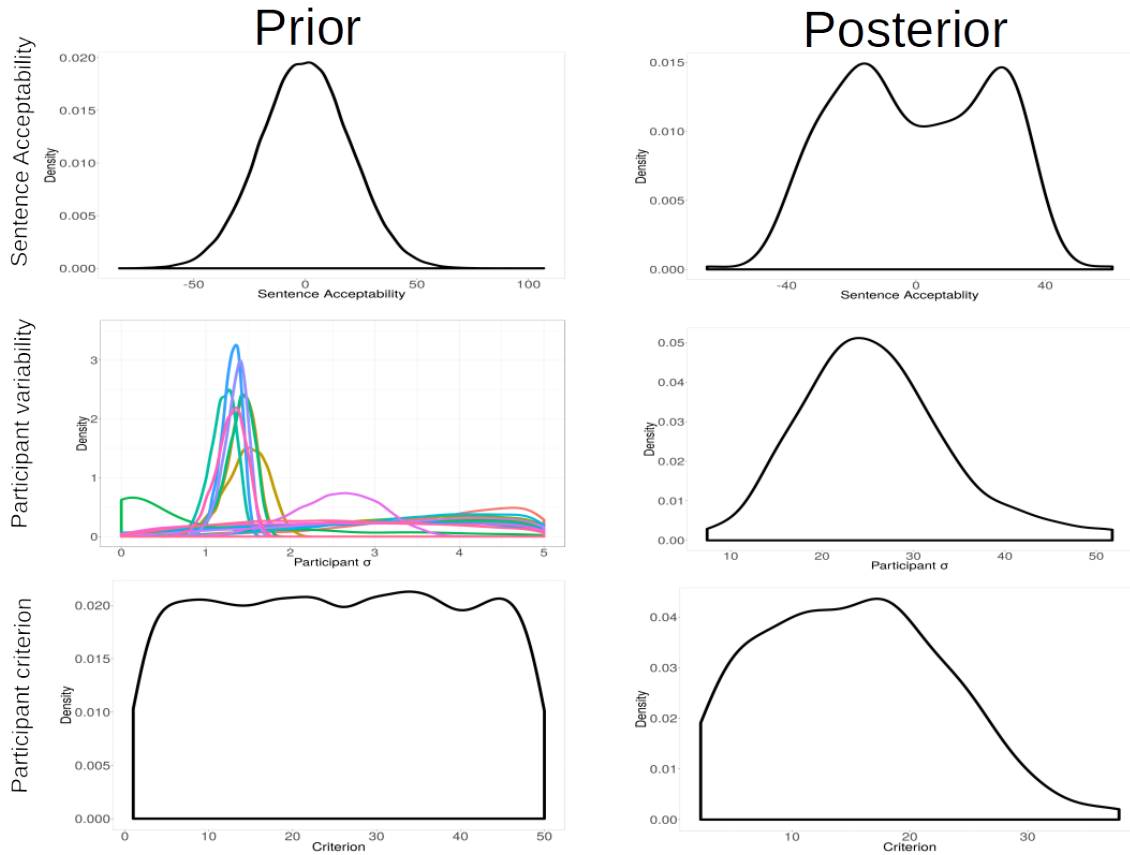


Figure 8.4: Priors and posteriors for key model parameters. The prior over sentence acceptability was $N(0,20)$. The model's acceptability estimates did not depend on the width of this prior, but were sensitive to the thickness of the distribution tails with relatively thin-tailed distributions such as the normal performing better on simulated data. The posterior is bimodal, reflecting distinct 'acceptable' and 'unacceptable' groups of sentences. This particular distribution of acceptabilities reflects the process by which these sentences were selected, as constructed contrasting pairs. Since this was a hierarchical model, the prior on participant variability was actually a family of distributions corresponding to different settings of the hyperparameters. Some randomly selected members are plotted here. The prior also puts some probability over distributions with extreme values for kurtosis and location of the mean, which have been filtered from this plot so that the more typical members of the family are clearly visible. Looking at the posterior, it appears that the means of typical distributions drawn from the prior are an order of magnitude too small for the acceptability and criterion values used in this model, but that the model was flexible enough and the data informative enough to infer a variability distribution on a scale compatible with them. The prior of participant criterion values was uniform between 1 and 50, expressing a lack of knowledge about how popular the 'equal' option would be. The data suggests the 'equal' option was relatively unpopular, resulting in a posterior distribution shifted towards the smaller end of this range.