Acceptability Judgments

James Myers

Summary

Acceptability judgments are reports of a speaker's or signer's subjective sense of the well-formedness, nativeness, or naturalness of (novel) linguistic forms. Their value comes in providing data about the nature of the human capacity to generalize beyond linguistic forms previously encountered in language comprehension. For this reason, acceptability judgments are often also called grammaticality judgments (particularly in syntax), although unlike the theory-dependent notion of grammaticality, acceptability is accessible to consciousness. While acceptability judgments have been used to test grammatical claims since ancient times, they became particularly prominent with the birth of generative syntax. Today they are also widely used in other linguistic schools (e.g., cognitive linguistics) and other linguistic domains (pragmatics, semantics, morphology, and phonology), and have been applied in a typologically diverse range of languages. As psychological responses to linguistic stimuli, acceptability judgments are experimental data. Their value thus depends on the validity of the experimental procedures, which, in their traditional version (where theoreticians elicit judgments from themselves or a few colleagues), have been criticized as overly informal and biased. Traditional responses to such criticisms have been supplemented in recent years by laboratory experiments that use formal psycholinguistic methods to collect and quantify judgments from non-linguists under controlled conditions. Such formal experiments have played an increasingly influential role in theoretical linguistics, being used to justify subtle judgment claims or new grammatical

models that incorporate gradience or lexical influences. They have also been used to probe the cognitive processes giving rise to the sense of acceptability itself, the central finding being that acceptability reflects processing ease. Exploring what this finding means will require not only further empirical work on the acceptability judgment process, but also theoretical work on the nature of grammar.

Keywords: acceptability, grammaticality, syntax, semantics, phonology, morphology, psycholinguistics, syntactic islands, gradience, frequency

1. What Acceptability Is and Is Not

Acceptability judgments are metalinguistic reports of a speaker's or signer's subjective sense of the acceptability of sentences, words, or discourse fragments. They have become particularly prominent in syntactic research, but they are used in all branches of theoretical linguistics, including pragmatics, semantics, morphology, and phonology. Since they reflect not just grammatical knowledge but also its development and cognitive underpinnings, acceptability judgments are also an important source of data in language acquisition, psycholinguistics, and clinical linguistics.

This section unpacks the notion of acceptability judgments. Section 2 describes the traditional methodology involved in collecting them and traditional criticisms of this methodology. Section 3 reviews formal psycholinguistic experiments for collecting acceptability judgments and some of the findings they have made possible. Section 4 discusses experimental explorations of the cognitive processes underlying acceptability judgments.

Acceptability is a cover term for others used in the literature, such as well-formedness, nativeness, and naturalness. Syntacticians often also use the term grammaticality judgments (though this is a misnomer, as will be explained shortly). In morphology and lexical phonology they are also called wordlikeness (or word-likeness) judgments and are often described as tests of productivity. In semantics they are sometimes called meaningfulness judgments or judgments of "availability", as when a reading (a particular semantic interpretation of a sentence) is judged as available (possible).

Acceptability is distinct from grammaticality, despite their widespread use as synonyms in the syntactic literature. Acceptability judgments, being overt expressions (whether statements or button presses in an experiment), are empirical data (see Dennett, 2003, for how judgments allow subjective conscious experience to be studied objectively). By contrast, grammaticality refers to the legality of a form with respect to to a linguistic hypothesis (or invented rules, in the case of artificial grammar experiments; Reber, 1989, Culbertson, 2012). In an attempt to avoid the confusion between acceptability and grammaticality, Chomsky (1965) introduced the term grammaticalness for the latter concept, but it never caught on, unlike his related term competence (tacit mental grammar).

Since it is impossible to directly observe a theoretical construct like grammar, acceptability judgments are also not the same as introspection (literally, looking inside oneself), despite confusion on this point among linguists themselves (as criticized by Wasow & Arnold, 2005). It is also somewhat misleading to call them intuitions (gut feelings), because acceptability judgments still involve conscious judgment-making, even if they are made on the basis of unconscious processes (Ludlow, 2011). Crucially, however, the judge need not be able to articulate how the judgments are arrived at (Dienes & Scott, 2005).

While acceptability judgments are thus not the "voice of competence" (as Devitt, 2006, p. 4, claims that linguists tend to conceive them to be; see Gross & Culbertson, 2011, for a critical response), they are nevertheless relevant to studying grammar because they are made on linguistic forms that may never have been encountered before, so that the judge must generalize beyond rote memory (see section 4 for discussion on the nature of grammar). Acceptability judgments on attested forms can still provide information about grammar, but the confounding of acceptability with familiarity makes their analysis more difficult.

The novelty criterion is not strictly enforced in syntax, semantics, and pragmatics, where linguists sometimes elicit judgments for corpus-attested sentences and discourse fragments (sometimes inadvertently, when the forms are short and simple), but the rich combinatorial powers of syntax and discourse help ensure that invented examples tend to be novel. In lexical research (morphology and lexical phonology) acceptability judgments are usually only elicited on unattested words (also called nonwords or nonce words) to avoid the powerful influences of lexical frequency and lexical semantics.

Acceptability judgments complement corpus data (recordings of natural language production). On the one hand, both provide evidence about grammar (see, e.g., Backus & Mos, 2011; Hoffman, 2006; Schütze, 2011). On the other hand, they provide different kinds of evidence about grammar. Corpora reflect language production, while acceptability judgments primarily reflect language comprehension (though judges may also take into account whether they would ever produce the test item). Corpora also include accidental forms (e.g., speech errors) that their producers would likely judge as unacceptable. Words and sentences in corpora are situated in ever-shifting discourse contexts, but acceptability judgments for them are usually elicited in isolation.

Perhaps most importantly, as elicited responses to linguistic stimuli, acceptability judgment data are experimental, whereas corpus data are observational. Experiments are the gold standard in most sciences because they allow one to distinguish causation from correlation (Woodward, 2016; though see Cleland, 2002, for arguments that observational data can also test causal hypotheses, and Stefanowitsch, 2006, for statistical techniques for inferring ungrammaticality from corpus absence). For grammatical research, a particular advantage of an experimental approach is that a corpus, no matter how large, is finite, whereas novel forms can always be invented to elicit novel responses (e.g., Ohala, 1986). This is important because distinguishing between competing grammatical hypotheses may depend on form types that are quite rare, though the force of this point depends on whether one sees the scope of theoretical linguistics as covering potential language use or only actual language use (Sampson, 2007; Gries, 2012).

The complementary roles of acceptability judgments and corpus data are also demonstrated by the fact that the former are more common in syntax, semantics, and pragmatics, and the latter more common in phonology and morphology. This contrast holds even of generative linguistics (perhaps surprisingly; see section 2.1): Chomsky (1957, 1965) cite judgments for numerous invented sentences, but Chomsky and Halle (1965) only mention phonological acceptability judgments in passing, and Chomsky and Halle (1968) and Aronoff (1976) both rely almost exclusively on dictionary data for their phonological and morphological analyses.

One reason for this contrast is that a relatively acceptable nonword, unlike a relatively acceptable novel sentence, may still be avoided by speakers merely for being nonlexical, thereby lowering its acceptability (Haspelmath, 2002, p. 99). A closely related reason is combinatorial richness. Whereas it is quite likely that even large corpora will be missing sentences just as acceptable as those attested, limitations on word length and morpheme and phoneme inventories

in natural language use allow even small corpora or dictionaries to provide rich evidence about word form patterns (Myers, 2012b).

Besides acceptability judgments, there are many other experimental tasks (what the participants [subjects] are asked to do) and paradigms (the overall experimental procedures) that theoretical linguists use. These include other non-chronometric (non-speeded) tasks (Derwing & de Aldemeida, 2009) that, like acceptability judgments, reflect only the final outcome of psychological processes; for example, in the wug test (Berko, 1958), so named because participants are asked to use morphological operations to produce novel words like the plural of the English nonword *wug*. There are also chronometric (speeded) tasks that reflect the time course of language processes (which may even be probed on the fly via methods like eye-tracking and some kinds of brain imaging); if grammar is relevant to these processes, studying them should also provide grammatical evidence (see section 4). As Derwing and de Almeida (2009) point out, knowledge claims, including claims about mental grammar, should be tested in multiple ways to distinguish true knowledge from mere test-taking ability.

Acceptability judgments are metalinguistic because they treat linguistic form as objects of discussion, not as means of communication or thought, unlike other experimental tasks like silent reading, picture description, or truth value judgments (see sections 3.3 and 4). Although metalinguistic tasks are less natural than normal language use, they are far from being unnatural. It is a useful skill to be able to identify when an utterance is not acceptable in some sense (Ludlow, 2011), or when a speaker or signer tends to produce such utterances (e.g., non-natives or children); “non-native speaker” is likely to be a universal folk linguistic concept (e.g., *baragada* is the Hausa term for non-native Hausa speech; Hunter, 1982), and more generally, so the treatment of language as an object, as in quotation (Saka, 1998). Bilinguals are also capable

of judging the acceptability of code-switched structures even if they do not code-switch themselves (Toribio, 2001). Even children four years and younger are capable of judging whether meaningful and true sentences (ostensibly produced by a doll learning to talk) are "silly" (Ambridge, 2012; McKercher & Jaswal, 2012).

Metalinguistic tasks have earned an important place in experimental psycholinguistics, where they have been staple tools for decades. In particular, the lexical decision task, which asks participants to decide if an item is or is not a lexical word, has proven extremely useful in confirming and complementing other tasks in the development of lexical access models. The metalinguistic nature of this task is made explicit by Hung, Tzeng, and Ho (1999), who criticize it as being inappropriate for Chinese, where the lack of orthographic word boundaries may make lexical status less accessible to awareness. Nevertheless, there is no evidence that the results of Chinese lexical decision experiments are any less reliable than those run on other languages (Myers, 2017).

Psycholinguists have also long used acceptability judgments, though they are rarely identified as such. Sentence parsing experiments usually require that all of the invented stimuli seem intuitively natural to some minimal degree, so that sentences predicted to be harder to process are not harder merely because they are ungrammatical. Lexical decision tasks are designed with nonword foils that are roughly equivalent in wordlikeness with the real words (often described as pseudowords or "pronounceable" nonwords), so that the decisions must be based solely on lexical status. Meaningfulness tasks are standard in tests of verbal working memory (e.g., Kane, Hambrick, Tuholski, Wilhelm, Payne, & Engle, 2004, p. 196). Native speaker judgments have also long been the implicit criterion for the accuracy of non-native productions, and in recent decades, increasing attention has been paid to the acceptability

judgments of the non-native learners themselves as evidence about their developing grammatical knowledge (Birdsong, 1989; Sorace, 1996; Davies & Kaplan, 1998; Sperlich, 2015). Acceptability judgments are also a basic tool for assessing the learning of artificial grammars, not just in recent research in theoretical linguistics (e.g., Culbertson, 2012), but in the cognitive psychology literature as well (e.g., Reber, 1989). Occasionally they are used in clinical linguistics as well, to evaluate of the nature of aphasic deficits (e.g., Gibson, Sandberg, Fedorenko, Bergen, & Kiran, 2016).

2. Traditional Acceptability Judgments

Acceptability judgments have been used to justify grammatical analyses since the beginnings of linguistics as an academic discipline. While the ostensible purpose of ancient Indian grammarians in the Pāṇinian tradition (ca. 500 BCE) was to describe the linguistic system of the Vedas, they went beyond this corpus to cite examples from their own or other dialects of Sanskrit to illustrate posited grammatical rules (Cardona, 1994). Apollonius Dyscolus (ca. 200 CE), a Greek grammarian of Latin, was among the first to report negative judgments of invented sentences to illustrate violations of grammatical principles (Householder, 1973), and this tradition continued under Renaissance grammarians like Sanctius (Breva-Claramonte, 1983).

In modern times, the American structuralist Leonard Bloomfield routinely wrote that such-and-such a form “cannot be used” (e.g., Bloomfield, 1933, p. 185). This is despite his behaviorist views on language, where the only valid data were held to be concrete instances of language use in context (forming stimuli-response pairs), not judgments about when forms cannot be uttered. Implicit acceptability judgments also remained crucial in linguistic fieldwork

during this period; Geary (1943, p. 150), for example, writes that certain Algonquin forms are "ungrammatical", not merely unattested.

The theoretically motivated prominence of attested language usage in American structuralism made the contrasting emphasis on acceptability judgments in generative syntax, starting with Chomsky (1957), seem radical and new. The anti-corpus rhetoric of the early generativists may have been more aggressive than was strictly necessary (Harris, 1993), and today generative syntacticians occasionally cite corpus data as well (e.g., Gordon & Hendrick, 2005; Newmeyer, 2010). Nevertheless, given the limitations of corpus data for studying syntax (see section 1), the shift to acceptability judgments made it possible to amass a much larger and more varied set of syntactic generalizations than had ever been reported. This holds for other schools like cognitive linguistics as well, where acceptability judgments are also widely used (see Gibbs, 2007, and Noonan, 1999, for discussion, and Lakoff, 1991, for an illustration).

In their traditional use in the theoretical linguistic literature, some aspects of acceptability judgment "best practices" (Phillips & Wagers, 2007, p. 740) are the same as in standard non-chronometric psycholinguistic experiments. Linguistic forms are presented as stimuli to a participant who gives acceptability judgments in response. The stimuli include at least one control condition so that there is a basis of comparison. Conditions are designed to contrast only in the theoretical variable(s) of interest. For example, Chomsky and Halle (1965, p. 101) contrast the perceived English wordlikeness of the nonwords *blick* and *bnick*, which are identical except for the consonant clusters. Similarly, syntactic acceptability judgments are usually elicited for sentences that differ in syntactic structure but contain the same words, as far as possible, and syntacticians routinely adjust for semantic and pragmatic influences, so a sentence that seems unacceptable in a neutral context may be placed in a richer discourse environment (Newmeyer,

1983, pp. 55-57). Test items usually fall into a factorial design, usually involving just one factor (forming minimal pairs), though occasionally two binary factors are crossed. Crossing factors makes it possible to test for interactions, such as between syntax and the lexicon (e.g., two different verb types in two different structures) or between two aspects of the syntactic structure. A prototypical example of the latter is the testing of syntactic island constraints (where antecedents cannot be linked to gaps within certain types of syntactic constituents) by crossing two binary factors (Cowart, 1997; Sprouse, 2015): constituent type (island vs. non-island) and gapping (presence vs. absence of gap within the syntactic constituent). Figure 1 (a) illustrates this traditional factorial logic using examples from Ross (1967, p. 70), with Ross's original example numbers and the * indicating his reported judgment of unacceptability (_ represents a gap and [] mark an island, here a complex noun phrase).

[INSERT FIGURE 1 ABOUT HERE]

Other aspects of traditional acceptability judgments are quite different from psycholinguistic best practice. Very few items are tested, with perhaps only one or two reported per empirical claim. The response scale is not explicitly stated (Bard, Robertson, & Sorace, 1996); it usually seems to be binary (reject/accept, respectively marked with/without an asterisk * diacritic, or sometimes # for semantic judgments), though a gradient scale is sometimes implied (?, ??, and so on indicating gradually decreasing acceptability). Most notoriously, traditional acceptability judgments are elicited by theorists from themselves, making the sample not just small but also biased.

Moreover, as with any methodology, best practices are not always followed, with violations

of proper factorial design and inconsistencies in distinguishing empirical acceptability judgments from theoretical grammaticality claims (Myers, 2009b, 2012b; Wasow & Arnold, 2005). For example, Di Sciullo and Williams (1987, p. 33) cite the contrasting acceptability in 1 versus 2 in support of their claim that compounding blocks the assignment of thematic roles (where *bread* is the patient of *bake*), but Spencer (1991, p. 333) notes that *man* does not compound with verbs in general, which already leads to the judgment contrast in 3 versus 4. Myers (2012b) concludes that the Di Sciullo and Williams claim should have been tested in a factorial design defined by two factors: [+/-compound] and [+/-patient].

(1) a baker of bread

(2) * a bake-man of bread

(3) a baker

(4) * a bake-man

Even when applied properly, the traditional methodology of acceptability judgments has received a variety of criticisms. Those dealt with in section 1 are either misunderstandings (acceptability is not grammaticality and judgment making is not introspection) or a matter of taste (experiments provide a different kind of evidence from corpora). Two other major criticisms are harder to dismiss: traditional acceptability judgments are biased and noisy. They are biased because the judges, being linguists, have a stake in the outcome, or at least have considerable experience in making acceptability judgments, so their judgments may not generalize to those of non-linguists (Hill, 1961; Spencer, 1973; Labov, 1996; Dąbrowska, 1997, 2010). They are noisy because judgments vary across speakers and test times (e.g., Householder, 1965, p. 15), and this

variation cannot be modeled accurately by testing just a few items and speakers (Schütze, 1996; Cowart, 1997; Featherson, 2007; Gibson & Fedorenko, 2010, 2013).

These criticisms have received a variety of traditional responses: that theoretically crucial judgments are so clear that even crude methods suffice to collect them (Chomsky, 1965); that what seem to be disagreements over data are usually disagreements over analysis (Newmeyer, 1983) or at most due to cross-speaker variation or processing influences (Fanselow, 2007); that published judgment claims are checked not just by the theorist alone but also by potentially rival colleagues, conference talk audiences, and reviewers (Phillips, 2010); or even that expert linguists may be more qualified to make theoretically relevant judgments than non-linguists, just as judging wines requires a refined palate (Valian, 1982).

Traditional acceptability judgments have also been justified as being a more efficient way to collect reliable evidence about grammar than other kinds of experimental methods (Chomsky,1965; Ludlow, 2011; Cowart, 1997; Phillips & Lasnik, 2003). However, as critics have noted (e.g., Gibson & Fedorenko, 2013), the validity of this efficiency argument depends on the assumption of data reliability, which by its very nature must be tested against other data sources (see section 3 and 4).

3. Formal Acceptability Judgment Experiments

In recent decades, many theoretical linguists have come to recognize that the most straightforward and convincing response to criticisms of traditional acceptability judgments is to collect them from larger samples of ordinary people and items via standard psycholinguistic protocols (section 3.1). These kinds of experiments have not only addressed the noise and bias

criticisms (section 3.2) but have also led to new empirical discoveries and theoretical concepts (section 3.3).

3.1 Psycholinguistic Methodology

A psycholinguistic experiment is defined by its design, procedures, and statistical analysis (Kirk, 2012). The design is the logical structure of the conditions being compared (item types, in the case of acceptability judgment experiments). As noted in section 2, the traditional methodology generally already incorporates some features of good design, at least with regard to the fixed variables (i.e., the variables manipulated by the experimenter for testing hypotheses). The limitations of traditional acceptability judgments relate more to the random variables: the participants (judges) and items. These variables are treated as random because research hypotheses usually do not make predictions for particular people or forms, but rather generalize across them (though see the end of this section). This means that enough of each must be tested to make such generalizations robust, and if the goal is to generalize to ordinary people, the participants must also be ordinary people, not expert linguists.

The design must also keep other potential variables under control. In the typical psycholinguistic experiment, trials (stimulus-response cycles) are presented in a different random order for each participant, since order is known to influence responses (see section 3.3), and a fixed order would lead to confounds with the fixed variables. Similarly, items are usually distributed across participants in such a way that each participant is presented with all of the experimental conditions but never directly related items (called a Latin square design). For example, in testing the relative acceptability of /bl/ and /bn/ onsets for English speakers, no

participant would receive both *blick* and *bnick*, but instead one group might receive *blick* and *bnart* and another *blart* and *bnick*. As with randomizing presentation order, this kind of design prevents the response to one item from being influenced by responses to others (Cowart, 1997). Other design decisions may be justifiable, however; the forced choice paradigm, where participants are presented with minimal pairs and must choose the more acceptable item, has the advantage of being almost as statistically powerful as gradient judgment scales in small samples (Sprouse & Almeida, 2017). Finally, psycholinguists often include a large number of otherwise irrelevant filler items (Cowart, 1997, advocates giving each participant at least twice as many fillers as experimental items). Among other things, fillers serve to obscure the goals of the experiment from participants and provide a range of well-formedness within which the crucial items can be judged, a range that may be intentionally manipulated by the experimenter, as via the inclusion of real-word fillers in a wordlikeness judgment task (Goldrick, 2011). If indisputably good and bad fillers are included, they may also help the experimenter confirm if participants are following the instructions (Coward, 1997).

Regarding the experimental procedure, one consideration is what the participants are asked to do (Schütze, 2005). Asking non-linguists to judge "grammaticality" seems to be a particularly bad idea, as Newmeyer (1983) notes in his critique of Hill (1961); this is unsurprising given that even linguists use this term inconsistently. However, as long as participants understand that their judgments are meant to be intuitive and not prescriptive, the precise details of the instructions do not seem to have a large influence on the results (Aronoff & Schvaneveldt, 1978; Cowart 1997), though this assumption has yet to be investigated systematically (Schütze & Sprouse, 2013).

A related procedural question concerns the response scale, with several options used in the literature (Schutze & Sprouse, 2013, pp. 31-36): binary accept/reject, binary forced choice,

discrete multi-point scales (often called Likert scales, after American psychologist Rensis Likert), and open-ended continuous ratios relative to a baseline judgment (the magnitude estimation task, originally developed for psychophysics by Stevens, 1956; Bard et al., 1996; see Featherston, 2008, for variants and Sprouse, 2011a, for empirical challenges). When these response types have been explicitly compared on the same test items and the same speaker populations, they have proven to give very similar results (phonology: Greenberg & Jenkins, 1964; syntax: Bader & Haüssler, 2010, Weskott & Fanselow, 2011). In particular, binary scales can capture gradience just as well as numerical scales with as few as fifteen data points (estimated from Figure 6 in Sprouse & Almeida, 2017, p. 25). In Figure 1, (b) and (c) show mean magnitude estimation judgments and binary judgment acceptance rates, respectively, for an idealized experiment testing the four Ross (1967) sentences (see the appendices in Sprouse & Almeida, 2012, for actual experimentally elicited judgments for sentences like these). In both cases, the interaction is shown graphically by the fact that the lines linking conditions are not parallel: even though gaps and islands individually lower acceptability, the acceptability for gaps in islands is worse than would be expected if these two effects were simply summed (see Sprouse, 2015, and Sternberg, 1998, for more on what can be inferred from interactions).

Statistical analysis is a key part of psycholinguistic methodology because even when a generalization seems robust, establishing the degree of robustness requires quantification (Gibson, Piantadosi, & Fedorenko, 2013). Following Clark (1973), participants and items are both treated as random variables, traditionally through separate by-participant and by-item analyses of variance (ANOVA) that test if variation across the fixed variables is greater than expected by chance. In the past decade or so, psycholinguists have also begun to employ mixed-effects regression, which incorporates both fixed and random variables in the same

statistical model (Baayen, Davidson, & Bates, 2008). As a form of regression (a generalization of correlation), mixed-effects modeling is not restricted to data with factorial categorical independent (input) variables and a gradient (and ideally normally distributed) dependent (output) variable, as is ANOVA. The flexibility of mixed-effects modeling with regard to dependent variable types allows for trial-level analyses of binary responses (accept/reject, or forced choice), while still taking all fixed and random variables into account, via mixed-effects logistic regression (Jaeger, 2008; Myers, 2009b).

Its flexibility with independent variables has had an even greater impact, particularly in lexical research, where variables are often gradient and correlated (e.g., word frequency and word length); picking and choosing items merely to force a classic factorial design may result in a lexically nonrepresentative sample (Baayen, 2010; Forster, 2000). Lexical researchers have thus advocated regression-based designs (Balota, Yap, Hutchison, & Cortese, 2012), using sufficiently large random samples to allow the statistics to find the key patterns. In acceptability judgment experiments, this approach is more directly relevant for research on morphology and lexical phonology (e.g., Bailey & Hahn, 2001), but the power of regression models to take gradient variables into account has also been used in syntactic acceptability judgment experiments to model influences like constituent length (Bresnan & Ford, 2010), participant working memory capacity (Sprouse, Wagers, & Phillips, 2012a), and trial order (Myers, 2012a).

Responses are often rescaled in various ways before statistical analysis (e.g., Cowart, 1997; Bailey & Hahn, 2001), but this sometimes seems to be done more by convention. Schütze and Sprouse (2013) point out that the logarithm transform (commonly used to make reaction time distributions more normal by reducing the long right tail) actually increases the skew in judgment distributions. However, their recommendation (following Cowart, 1997, and others) to

apply separate *z*-score transforms to each participant (subtracting the mean and dividing by the standard deviation to put all participants on the same scale), is not the only way to deal with cross-participant variation in response scales. Mixed-effects modeling already accounts for the default response values of participants as random intercepts (where the individual participant regression lines cross the *y*-axis) and also accounts for by-participant variation in effect size and direction as random slopes (of the individual participant regression lines). Even traditional by-participant and by-item ANOVAs implicitly encode random intercepts and slopes (Barr, Levy, Scheepers, & Tily, 2013). Nevertheless, an overall rescaling to *z*-scores may make it easier for the mixed-effects algorithm to converge on the best-fitting model (Bates, Mächler, Bolker, & Walker, 2015). Moreover, if *z* scores are used not just for the responses but for the predictor variables as well, the model will yield standardized regression coefficients that can be interpreted as effect sizes for comparison, even across studies (Aiken & West, 1991; Menard, 2004).

Quantified experimentation also makes it possible to treat participant groups as a fixed rather than random variable (Gervain, 2003). Studies on syntactic judgment idiolects (i.e., idiosyncratic variation in judgment patterns) include Langendoen, Kalish-Landon, and Dore (1973) and Gerken and Bever (1986) in English and Fanselow, Kliegl, and Schlesewsky (2006) in German (though these studies claim that the cross-speaker variation involves differences in processing rather than grammar per se), as well as Han, Lidz, and Musolino (2007) in Korean (see sections 3.3 and 4).

Despite the greater reliability made possible by quantification, large samples and sophisticated statistics will not prevent a poorly designed experiment from giving ambiguous or even misleading results (Culicover & Jackendoff, 2010; see also section 2). Moreover, qualitative methods are still worthy of respect (e.g., Heigham & Croker, 2009), and the very

informality of traditional acceptability judgments has its advantages; Henry (2005) argues that conversational back-and-forth is essential when eliciting judgments in non-standard dialects. Similarly, though Petronio and Lillo-Martin (1997) consulted eighteen American Sign Language (ASL) signers for their syntactic judgments, sufficient to run powerful statistics, these judgments were influenced by such subtle variables (e.g., as how open the eyelids were when making the nonmanual marker for *wh*-questions) that rushing into a quantified experiment would seem to be premature.

Even when quantitative precision is possible and desired, Gibson and Fedorenko (2013) admit that formal judgment experimentation can impose a greater burden on the researcher. Attempts have thus been made to reduce this burden, through tutorials (syntax: Cowart, 1997, 2012; Featherston, 2009; Gibson, Piantadosi, & Fedorenko, 2011; Sprouse, 2011b; phonology: Derwing & de Almeida, 2009; Frisch & Stearns, 2006; Hammond, 2012; Kawahara, 2011; Ohala, 1986), statistical analyses demonstrating that very small sample sizes suffice in many cases (Mahowald, Graff, Hartman, & Gibson, 2016), and software tools that automate some of the drudgery of experimental design, data collection, and statistical analysis (Erlewine & Kotek, 2016; Myers, 2009b). Linzen and Oseki (2015) and Myers (2016) also advocate the development of web-based platforms for sharing and cross-checking judgment data.

3.2 New Responses to Traditional Criticisms

Acceptability judgment experiments run in accordance with psycholinguistic best practices have directly confronted the criticisms that informal judgments are noisy and biased. Regarding noise, judges have been found to be self-consistent across different test times (phonology:

Greenberg & Jenkins, 1964; syntax: Cowart, 1997; Verhagen & Mos, 2016), though, unsurprisingly, this varies with the strength of the judgments themselves (Adli, 2007), and inter- and intra-speaker variation may itself be of theoretical interest (Verhagen & Mos, 2016). Regarding bias, some experiments comparing the syntactic judgments of linguists and non-linguists have found significant mismatches (Hill, 1961; Spencer, 1973; Dąbrowska, 1997, 2010; Gordon & Hendrick, 1997; Linzen & Oseki, 2015), while others have found strong agreement (Cowart, 1997; Sprouse & Almeida, 2012, Sprouse, Schütze, & Almeida, 2013). These conflicting results may relate to the greater sensitivity of the latter set of experiments, each of which included around ten times more non-linguist participants than the former set. The earliest experiments (Hill, 1961; Spencer, 1973) were also criticized by Newmeyer (1983) for their unclear instructions.

To the extent that the linguist effect is real, it may be due to register or dialect differences, insufficient vetting of published judgments (which is a particular problem for delicate judgments in languages other than English, as Linzen & Oseki, 2015, point out in their study on Hebrew and Japanese), or linguists' prior exposure to published judgment diacritics (shown experimentally to influence judgments by Luka, 1998). Moreover, linguist-like acceptability judgments of complex written sentences seem to increase with increasing education (Dąbrowska, 1997), though it seems that cognitive science experience has a greater influence than linguistics training per se (Culbertson & Gross, 2009). A language's social status also has an effect: Neidle, Kegl, MacLaughlin, Bahan, and Lee (2001) observe that ASL signers may rate English-like ASL structures higher if they have internalized the belief that the culturally dominant spoken language is superior, or lower if they want to highlight a separate cultural identity, whether or not they use these structures in their own natural signing. The expert effect is also reported in the

experimental philosophy literature, where Western philosophers' intuitions (e.g., about linguistic reference) have often been found to differ from those of non-philosophers, especially in non-Western cultures (e.g., Machery, Mallon, Nichols, & Stich, 2004; Haug, 2014).

Rather than undermining the value of acceptability judgments entirely, however, such findings merely emphasize the importance of checking delicate judgments on a wider range of items and participants, and designing experimental materials to take into account the influence of known extraneous variables (e.g., by crossing syntactic structure with semantic plausibility). As Ohala (1986, p. 10) emphasizes, if a result seems distorted by the experimental method itself, it is the job of the experimenter to design a new experiment that controls for the distorting influence as much as possible.

Since informal acceptability judgments played very little role in the traditional methodology of lexical research (phonology and morphology; see section 2), early formal experiments were intended to test dictionary-based grammaticality claims. Such experiments, mostly using wug tests rather than acceptability judgments, generally found that lexical patterns tend to be quite unproductive (see review in McCawley, 1986). Unlike the case with syntax, however, generative linguists never denied that the productivity of lexical patterns is constrained by the memorized nature of real words (Kiparsky, 1975, 1982). Moreover, when experiments have used the reception-oriented acceptability judgment task rather than the production-oriented wug task, even lexical phonological patterns have been shown to have significant effects in languages like English (Hayes & Wilson, 2008), Turkish (Zimmer, 1969), Arabic (Frisch & Zawaydeh, 2001), Japanese (Kawahara & Sano, 2014), Russian (Gouskova & Becker, 2013), and Sign Language of the Netherlands (Arendsen, van Doorn, & de Ridder, 2010). Acceptability judgments have also confirmed the generalizability of morphological regularities in languages like English (Aronoff

& Schvaneveldt, 1978) and Mandarin (Myers, 2007). Nevertheless, wug tasks still seem to dominate experimental approaches to theoretical phonology and morphology, and in some cases may have advantages; Kawahara (2015) reports that a forced-choice wug production task was more sensitive to a Japanese morphophonological pattern than gradient acceptability judgments.

3.3 New Discoveries and New Concepts

The reduction in noise and bias afforded by formal acceptability judgment experiments has made subtle syntactic patterns easier to detect (Gibson & Fedorenko, 2013; Myers, 2009a; Schütze & Sprouse, 2013). For example, Featherston (2005) found that German speakers are sensitive to a constraint parallel to the *that*-trace effect of English, which informal judgments had failed to detect (Haider, 1983; though see Fanselow, 2007, for a critique). Sprouse, Caponigro, Greco, and Cecchetto (2016) found that variation in the acceptability of island violations in English and Italian was more complex than had previously been reported on the basis of informal judgments (Rizzi, 1982). Sprouse, Fukuda, Ono, and Kluender (2011) discovered a reverse island effect in English (but not in Japanese), where it is more acceptable for a *wh*-phrase to remain within an island in multiple *wh*-questions (see Dillon, Staub, Levy, & Clifton, 2017, for another recent experimental study on subtle judgments involving English *wh*-phrases). Sufficiently large and careful experiments have also cast doubt on claims originally based on informal judgments. For example, a large number of studies in numerous languages (e.g., Alexopoulou & Keller, 2007; Francis, Lam, Zheng, Hitz, & Matthews, 2015) have found that filling gaps in syntactic islands with resumptive pronouns does not always make them more acceptable.

Experiments have been especially helpful in clarifying delicate judgments in semantics and

pragmatics, though many use the truth value judgment task rather than acceptability per se. Studies on English have examined topics like indefinite scope (Ionin, 2010), negative polarity items (Clifton & Frazier, 2010), and quantification (Lidz, Pietroski, Halberda, & Hunter, 2011, using truth value judgments); topics studied in other languages include scalar implicature in Greek (Papafragou & Musolino, 2003, using truth value judgments) and French (Chemla & Spector, 2011, using truth value judgments), ellipsis interpretation in Dutch (Koornneef, Avrutin, Wijnen, & Reuland, 2011), and a pragmatic constraint on a syntactic construction in Mandarin (Lin, 2004).

In a particularly striking study (albeit one that again used truth value judgments), Han et al. (2007) found that a surface ambiguity in Korean in the hierarchical position of the verb has led to two distinct semantic idiolects: around two thirds of the adults and children they tested systematically interpreted 5 with the quantified object (underlined) scoping over negation (bold), while the remaining one third interpreted it with negation scoping over the quantified object. Since Han, Musolino, and Lidz (2016) found that children may have a different semantic idiolect from their parents, it seems that these idiolects are not learnable from normal language experience, and thus are unlikely to be detectable without formal experimentation.

(5) Khwukhi Monste-ka <u>motun khwukhi-lul</u> **an** mek-ess-ta.

Cookie Monster-NOM <u>every cookie-ACC</u> **NEG** eat-PST-DECL

every > negation: ‘Cookie Monster ate none of the cookies.’

negation > every: ‘Cookie Monster didn’t eat every cookie.’

Acceptability judgment experiments have led to the discovery of new patterns in

morphology and phonology as well. For example, while primarily a wug production study, Albright and Hayes (2003) also elicited acceptability judgments for regular and irregular past tense forms for nonce English verbs, unexpectedly finding that even regularly inflected forms (argued to be generated on the fly by Prasada & Pinker, 1993) were judged better if they were phonologically similar to real regularly inflected forms (though cf. Ullman, 1999, discussed in section 4). Frisch and Zawaydeh (2001) not only confirmed known constraints in Arabic on the co-occurrence of consonants of the same place of articulation, but also found that acceptability was gradiently sensitive to the degree of consonant similarity. Similarly, in experiments on known consonant co-occurrence constraints in Japanese and their systematic exceptions, Kawahara and Sano (2014) found hitherto unsuspected effects of syllable identity. Formal judgment experiments also allowed Keller and Alexopoulou (2001) to explore the interaction between phonology and syntax in the realization of information structure (e.g., focus vs. ground) in Greek.

With new empirical methods have come new theoretical proposals as well. Particularly prominent are grammatical models of gradience. While acceptability judgments have long been recognized as gradient, and even Chomsky (1965, pp. 10-11) broached the possibility that grammar itself might be gradient, quantification has brought gradience into the theoretical mainstream. In syntax, Bard et al. (1996), Sorace and Keller (2005), Featherston (2007), Bresnan (2007), and Pullum (2013a, 2013b) have interpreted gradient acceptability as showing that grammar itself makes use of gradient representations, and similar claims have been made in phonology by Greenberg and Jenkins (1964), Ohala (1986), Coleman and Pierrehumbert (1997), Albright (2009), Hayes and Wilson (2009), and Goldrick (2011). Meanwhile, skeptics have suggested that acceptability gradience may be sufficiently explained by interactions of

categorical grammar with gradient temporal processes (e.g., Schlesewsky, Bornkessel, & McElree, 2006; see also Neeleman, 2013, critiquing Pullum, 2013a, and the response by Pullum, 2013b). It has even been argued that acceptability judgments are not in fact as gradient as had been assumed, with judgments tending to cluster at the top and bottom of the scales in syntax (Sprouse, 2007), in phonology (Coetzee, 2009; Gorman, 2013), and even in artificial grammar learning (Tunney & Shanks, 2003).

Formal acceptability judgment experiments have also given more prominence to factors that have traditionally been viewed as extra-grammatical. The most important of these is lexical frequency. In a study on the English dative shift, Bresnan (2007) and Bresnan and Ford (2010) report that the relative acceptability of English verbs taking one of the two syntactic realizations of the indirect argument (e.g., *give Mary the book* vs. *give the book to Mary*) depend on their relative corpus frequencies, motivating their claim that grammatical knowledge is represented in part in terms of such frequencies. Correlations between acceptability judgments and corpus type frequencies are also observed in morphology (e.g., Hay, 2002, and Teddiman, 2006, who used novel combinations of real English stems and suffixes, and Bermel & Knittl, 2012, who used real Czech words with variable case suffixes) and in phonology, where two distinct types of frequencies have been shown to have independent effects on judgments in languages like English (Bailey & Hahn, 2001; Shademan, 2007) and Cantonese (Kirby & Yu, 2007): neighborhood density (the number of lexical items very similar to a test item) and phonotactic probability (the frequency of the test item's substrings across lexical items). Figure 2 compares the relative influences of these two lexical phonological variables in these two languages.

[INSERT FIGURE 2 ABOUT HERE]

Another extra-grammatical influence that has attracted attention is trial order. Acceptability judgments for sentences tend to become less sharp as more and more items are judged; this phenomenon, dubbed syntactic satiation (by analogy with semantic satiation; Jakobovits & Lambert, 1964), has been observed in formal syntactic judgment experiments on English (Snyder, 2000; Hiramatsu 2000; Luka & Barsalou 2005; Braze 2002; Francom 2009) and other languages (Goodall, 2011; Hiramatsu, 2000; Nagata, 1988, 1989; Myers, 2012a); satiation has yet to be studied formally in phonology or morphology. However, satiation is inconsistent across the studies that report it, and sometimes it fails to appear (Sprouse, 2009); its explanation also remains elusive, in particular whether it is only a general processing issue or if it can be used as a grammatically relevant diagnostic. This unclarity is unsurprising; trial order effects remain understudied in psycholinguistics more generally, usually being treated as a mere nuisance (e.g., Keuleers, Diependaele, & Brysbaert, 2010) since it reflects confounds with at least three very different variables: practice, fatigue, and cross-trial priming.

Finally, formal acceptability judgment experiments have supplemented artificial grammar learning experiments (e.g., Culbertson, 2012; Moreton & Pater, 2012a, 2012b) in the study of universal markedness (i.e., the degree of intrinsic linguistic naturalness). This use of acceptability judgments is relatively rare in syntax, where arguments from typology and the poverty of the stimulus (Chomsky, 1986) are more common, but one example is Saah and Goodluck (1995). They reported that Akan speakers accepted sentences with gaps in complex noun phrases, violating an otherwise extremely robust typological constraint (see Figure 1 for English), but the judges still gave them lower ratings than sentences that did not violate this constraint, suggesting the influence of universal markedness (though perhaps motivated by

processing constraints rather than grammar per se; Hawkins, 1999).

In phonology, however, acceptability judgment experiments have become a common paradigm in the study of markedness. Studies on languages like English (Hayes & White, 2013; Pinker & Birdsong, 1979), Tagalog (Zuraw, 2007), and Mandarin (Myers, 2015) have found that nonlexical forms obeying typologically rarer phonological patterns are judged as less acceptable than those obeying typologically more common patterns (Pertz & Bever, 1975, even found that English-speaking adults and children could make accurate guesses about the typological frequency of non-English consonant clusters). However, taking language-internal influences fully into account is not a trivial task. This problem is exemplified by Daland, Hayes, White, Garellek, Davis, and Norrmann (2011), who observed that English speakers find nonwords containing typologically less marked onset clusters like /bn/ more acceptable than those with typologically more marked onset clusters like /lb/, even though neither cluster is attested in English. Yet they then go on to show that this pattern can be learned directly from English lexical statistics via a computational model with no explicit encoding of markedness. The validity of their model is rejected by Berent, Wilson, Marcus, and Bemis (2012), who show that it makes a number of false predictions that can only be remedied in a model with innate architectural constraints, a point that Hayes and White (2013) seem to accept.

4. Processing Acceptability

Since even informal acceptability judgments (and corpus data) reflect mental knowledge and psychological processes, formal experiments are not necessary to establish “psychological reality” (Dresher, 1995). Nevertheless, while most of the debates over acceptability judgments

have revolved around their reliability (i.e., replicability), it is equally important, as Newmeyer (1983) observed, to know if they also have validity, that is, if they detect what we intend them to detect: mental grammar. After all, far from providing direct evidence about grammar, acceptability judgments, like all behaviors, are the final output of psychological processes that unfold over time in a physical brain. This is supposed to be a commonplace view in grammatical research, though the confusions reviewed in section 1 indicate that it bears repeating (see, e.g., Goldrick, 2011, for a reminder aimed at his fellow phonologists). The classic argument for this view came from Yngve (1960) and Miller and Chomsky (1963), who pointed out that sentences with multiply center-embedded structures (typical examples would include [*The mouse* [*the cat* [*the dog hates*] *chased*] *ran*]) are unacceptable despite being arguably grammatical, in the sense of conforming to syntactic generalizations that hold in acceptable sentences (like [*The mouse* [*the cat chased*] *ran*] and [*The cat* [*the dog hates*] *chased the mouse*]). The question of the validity of acceptability judgments as evidence for grammar thus immediately raises a larger one: what is the difference, if any, between grammar and language processing?

To some extent this is a matter of definition. Even the restriction on multiply center-embedded structures may be handled within grammar if grammar, as claimed by Bresnan (2007), refers to properties like frequency or length (to explain the increase in acceptability when different words are used in an otherwise identical structure, as in [*The two mice* [*the cat* [*I hate*] *chased*] *ran away*]). However, grammar is widely assumed to be defined by representations and operations that are domain-specific (i.e., specialized for human language) and structure-dependent (i.e., not superficially analogical) (Chomsky, 1971; Crain & Nakayama, 1987). If the ability to memorize arbitrary features in an arbitrary vocabulary involves domain-general memory processes relatively unhindered by structural constraints, the

unacceptability of multiply centered structures would have to be ascribed to domain-general processes, not grammar.

In practice, domain-specificity tends to be defined negatively, by ascribing to grammar linguistic patterns that cannot be completely explained by known domain-general processes. Häussler, Grant, Fanselow, and Frazier (2015) apply this logic in their argument that an English-like judgment contrast in German is due to mere processing constraints, so that the languages truly do differ in grammar. The logic is not straightforward to apply in general, however. Hofmeister and Sag (2010) (following Kluender & Kutas, 1993, and Hawkins, 1999) argued that apparent island constraints are actually due to independently established working memory limitations that make it harder to link gaps and antecedents the farther apart they are, showing experimentally that acceptability is increased by making nonstructural modifications to sentences without modifying the island structures themselves. Sprouse et al. (2012a) challenged this claim in an experiment that tested whether the acceptability of island violations would rise in participants with larger working memory capacities, as they should if island effects are reducible to memory limitations, but the correlations they found were very small, often nonsignificant, and inconsistent in direction (for more on this debate, see Hofmeister, Casasanto, & Sag, 2012a, 2012b; Sprouse, Wagers, & Phillips, 2012b; Yoshida, Kazanina, Pablos, & Sturt, 2014).

Such debates are inevitable because grammatical knowledge must somehow be implemented via processes, including domain-general processes, in real-time language use. There are two general ways in which this might work: either the grammar is a store of domain-specific structure-dependent rules, constraints, and principles that is consulted by the processor in real time (what Lewis & Phillips, 2015, call the two-systems hypothesis), or the grammar is not a separate module at all, but rather a domain-specific, structure-dependent aspect

of processing itself (the one-system hypothesis). While the two-systems hypothesis may seem more intuitive, the one-system hypothesis has both conceptual and empirical advantages. Conceptually it is a special case of the highly influential approach to cognition promoted by Marr (1982), whereby unitary systems can nevertheless have functionally distinct levels of description. Thus language processing can be described both at the level of real-time events and at the level of grammar, which Neeleman and van de Koot (2010) characterize as the overarching generalizations that allow it to serve as a code for communication and thought. They also demonstrate that the more the grammar is implicit in the processing itself, rather than having to be consulted from a separate store (e.g., to winnow out all ungrammatical sentences from a set randomly generated by domain-general processes), the more computationally efficient the system becomes.

Lewis and Phillips (2015) argue that a one-system view of grammar and processing has empirical support as well. They review many studies where real-time syntactic parsing is sensitive to the same structures and generalizations that affect acceptability judgments (Goldrick, 2011, observes a similar correlation between wordlikeness judgments and real-time phonological processing). For example, Phillips (2006) found a correspondence between the acceptability of islands with parasitic gaps licensed by a coreferential island-external gap (Engdahl, 1983) and the processing of such structures. He first established that a parasitic gap structure like 6, where the first gap is inside an island (an infinitival complement of a subject noun phrase) but is licensed by the following island-external gap, is just as acceptable as a non-island gap like that in 7 (_ marks the gaps, [] the island, underlining the antecedent, and bolding the pre-gap verbs). He then showed in a chronometric (self-paced reading) study that when the antecedent was semantically incompatible with the verb preceding a potential legal parasitic gap location, as in 8,

reading slowed at this verb. Thus even though the continuation of the sentence made it clear that there was no parasitic gap at that point, the parser seemed to "know" that there could have been one (see Huang & Kaiser, 2008, for a replication of these results in Chinese). It is difficult to see how such a highly structure-dependent effect could be reduced to domain-general processing (though see Hofmeister, Casasanto, & Sag, 2013, for critical discussion).

(6) The outspoken environmentalist worked to investigate what the local campaign to preserve the important habitats had **harmed** _.

(7) The outspoken environmentalist worked to investigate what [the local campaign to **preserve** _] had **harmed** _.

(8) The school superintendent learned which high school students [the proposal to **expand** drastically and innovatively upon the current curriculum] would **motivate** _ during the following semester.

While Lewis and Phillips (2015) admit that syntactic acceptability and sentence processing also show mismatches (as in multiply center-embedded structures), they claim that all such cases can be explained by known domain-general factors like limits on memory access (e.g., to store and retrieve the elements in a center-embedded structure) and control mechanisms (e.g., to find and link the correct elements in a center-embedded structure), rather than requiring domain-specific processes completely distinct from grammar itself. However, if the conclusions of studies like Langendoen et al. (1973), Gerken and Bever (1986), and Fanselow et al. (2006)

are correct (see section 3.1), it may still be possible for people to adopt different language processing habits within a single speech community, blurring the line between domain-general and domain-specific.

The line is also blurred by the effect of lexical frequency on syntactic acceptability. This effect may show that grammar is intrinsically lexicalized, as argued by Bresnan (2007) and Ford and Bresnan (2010) (see section 3.3), but alternatively it may be explained by domain-general processing. Unsurprisingly, however, syntactic acceptability cannot be reduced to frequency effects alone; mismatches have often been reported (Crocker & Keller, 2006; Divjak, 2008; Kempen & Harbusch, 2005), as have mismatches between frequency and parsing (Gibson & Schütze, 1999). Processing seems to be made easier, and acceptability thereby increased, not only when structures have been practiced, but also when structures conform to general principles (grammar-based or otherwise), regardless of practice. Further evidence that frequency effects are domain-general effects comes from aphasic patients, who, despite serious deficits in language production and/or comprehension, are often able to make the same syntactic acceptability judgments as their unimpaired peers; importantly, however, this is especially so for higher-frequency structures (Gibson *et al.*, 2016). The frequency effect here suggests that aphasic patients rely more on familiarity than do unimpaired speakers, who are also able to make use of the same processes used for language production and comprehension.

Even wordlikeness judgments are not reducible to frequency. Of course frequency influences (non)word processing, hence acceptability; the English nonword acceptability judgments of Bailey and Hahn (2001) are increased by the same lexical variable, neighborhood density, that slows down the English nonword rejection latencies in the lexical decision task of Yap, Sibley, Balota, Ratcliff, & Rueckl (2015), and Teddiman (2006) reports similar cross-task

correlations in responses to English nonwords composed of real English stems and affixes. Yet not only does wordlikeness seem to be influenced by universal markedness (see section 3.3), but its extra-lexical nature is already suggested by within-language effects.

For example, Berent and Shimron (1997) and Frisch and Zawaydeh (2001) found that acceptability reflected formal phonotactic constraints in Hebrew and Arabic, respectively, even when neighborhood density and phonotactic probability were controlled; Kager and Pater (2012) drew similar conclusions regarding a Dutch phonotactic pattern. Shademan (2007) argues that the statistical independence of phonotactic and neighborhood effects (i.e., neither can be reduced to the other in regression analyses) itself demonstrates that acceptability reflects grammatical (phonotactic) knowledge, and not merely analogy with lexical neighbors. Similar conclusions hold for morphological judgments. For example, Bermel and Knittl (2012) found that inflected Czech words were often more acceptable than their low corpus frequencies would predict, and Ullman (1999) found that ratings of real English regular past tense verb forms depended on the ratings for the stems, not on their own frequencies or lexical neighbors (though cf. Albright & Hayes, 2003, discussed in section 3.3). Similar to the aphasic syntax results, Teddiman (2006) found that the selectional restrictions of English suffixes had greater effects on judgments for morphologically complex nonwords in sparser morphological neighborhoods (i.e., smaller morphological family sizes; Schreuder & Baayen, 1997), as if the judges were forced to go beyond mere familiarity.

If grammar is what (domain-specific, structure-dependent) processing does, acceptability judgments provide reliable information about grammar via a very simple mechanism: acceptability increases when the processing is easier (with domain-general processing effects controlled or statistically extracted). Topolinski and Strack (2009) present an interesting general

model of how processing ease might then affect acceptability judgments. In a series of experiments, they asked participants to make different kinds of judgments (about the semantic coherence of a set of words, the identity of a blurred image, or whether letter strings conformed to an artificial grammar). They then crossed processing ease (e.g., modifying visual clarity to affect perception, or priming to improve memory recall) with induced emotional state (by using words with positive or negative affect, or flashing images of smiling or frowning faces). This allowed them to show that processing ease and affect have independent additive effects on judgments. Putting everything together, then, grammar is an aspect of language processing, processing ease may affect one's emotional state, and acceptability judgments may reflect this emotional state.

Further Reading

Articles and Book Chapters

Chomsky, N. (1965). Methodological preliminaries. Chapter 1 of *Aspects of the theory of syntax* (pp. 3-62). Cambridge, MA: MIT Press.

Kawahara, S. (2016). Psycholinguistic methodology in phonological research. Oxford Bibliographies Online.

Lewis, S., & Phillips, C. (2015). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research*, *44*(1), 27-46.

Myers, J. (2009). Syntactic judgment experiments. *Language and Linguistics Compass 3*(1), 406-423.

Schütze, C. T. (2011). Linguistic evidence and grammatical theory. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(2), 206-221.

Schütze, C. T., & Sprouse, J. (2013). Judgment data. In R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 27-50). Cambridge, UK: Cambridge University Press.

Sprouse, J. (2013). Acceptability judgments. Oxford Bibliographies Online.

Books

Birdsong, D. (1989). *Metalinguistic performance and interlanguage competence*. Berlin: Springer.

Cohn, A. C., & Fougeron, C. (2012). *The Oxford handbook of laboratory phonology*. Oxford: Oxford University Press.

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

Devitt, M. (2006). *Ignorance of language*. Oxford: Oxford University Press.

Ludlow, P. (2011). *The philosophy of generative linguistics*. Oxford: Oxford University Press.

Noveck, I. A., & Sperber, D. (Eds.). (2004). *Experimental pragmatics*. Basingstoke: Palgrave Macmillan.

Runner, J. T. (Ed.) (2011). *Syntax and semantics, vol. 37: Experiments at the interfaces*. Bingley, UK: Emerald Group Publishing.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Links to Digital Material

English wordlikeness judgment datasets used by Gorman (2013)

Mandarin Wordlikeness Project: Database used in Myers (2015)

MiniJudge: Factorial judgment experiment software used in Myers (2009b)

Turktools: Interface to Amazon Mechanical Turk for web-based judgment experiments (direct payments possible only for participants in the US or India) introduced in Erlewine and Kotek

(2016)

Worldlikeness: Web platform for running and sharing (wordlikeness) judgment experiments across the world, as sketched in Myers (2016)

References

Adli, A. (2005). Gradedness and consistency in grammaticality judgments. In S. Kepser, & M. Reis (Eds.), *Linguistic evidence: Empirical, theoretical and computational perspectives* (pp. 7-25). The Hague: Mouton de Gruyter.

Aiken, L S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. London: Sage.

Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology, 26*(1), 9-41.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition, 90*(2), 119-161.

Alexopoulou, T., & Keller, F. (2007). Locality, cyclicity and resumption: At the interface between the grammar and the human sentence processor. *Language 83*(1), 110-60.

Ambridge, B. (2012). Assessing grammatical knowledge. In E. Hoff (Ed.), *Research methods in child language: A practical guide* (pp. 113-132). New York, NY: Wiley-Blackwell.

Arendsen, J., van Doorn, A. J., & de Ridder, H. (2010). Acceptability of sign manipulations. *Sign Language & Linguistics*, *13*(2), 101-155.

Aronoff, M. (1976). *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Aronoff, M., & Schvaneveldt, R. (1978). Testing morphological productivity. *Annals of the New York Academy of Sciences, 318*(1), 106-114.

Baayen, R. H. (2010). A real experiment is a factorial experiment. *The Mental Lexicon*, *5*(1), 149-157.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412.

Backus, A., & Mos, M. (2011). Islands of (im) productivity in corpus data and acceptability judgments. In D. Schönefeld (Ed.) *Converging evidence: Methodological and theoretical issues for linguistic research* (pp. 165-195). Amsterdam: John Benjamins.

Bader, M., & Häussler, J. (2010). Toward a model of grammaticality judgments. *Journal of Linguistics 46*(2), 273-330.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language, 44*(4), 568-591.

Balota, D. A., Yap, M. J., Hutchison, K.A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.). *Visual word recognition, Vol. 1* (pp. 90-115). London: Psychology Press Psychology Press.

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language 72* (1), 32-68.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68* (3), 255-278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1). doi: 10.18637/jss.v067.i01.

Berent, I., & Shimron, J. (1997). The representation of Hebrew words: Evidence from the obligatory contour principle. *Cognition, 64*(1), 39-72.

Berent, I., Wilson, C., Marcus, G. F., & Bemis, D. K. (2012). On the role of variables in phonology: Remarks on Hayes and Wilson 2008. *Linguistic Inquiry, 43*(1), 97-119.

Berko, J. (1958). The child's learning of English morphology. *Word, 14*(2-3), 150-177.

Bermel, N., & Knittl, L. (2012). Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech. *Corpus Linguistics and Linguistic Theory, 8*(2), 241-275.

Birdsong, D. (1989). *Metalinguistic performance and interlanguage competence*. Berlin: Springer.

Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart, and Winston.

Braze, F. D. (2002). *Grammaticality, acceptability and sentence processing: A psycholinguistic study*. Storrs: University of Connecticut dissertation.

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherson & W. Sternefeld (Eds.), *Roots: Linguistics in search of its evidential base* (pp. 75-96). Berlin: Mouton de Gruyter.

Bresnan, J., & Ford, M. (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language, 86*(1), 168-213.

Breva-Claramonte, M. (1983). *Sanctius' theory of language: A contribution to the history of Renaissance linguistics*. Amsterdam: John Benjamins.

Cardona, G. (1994). Indian linguistics. In G. Lepschy (ed.), *History of linguistics, Volume I: The Eastern traditions of linguistics* (pp. 25-60). London: Longman.

Chemla, E., & Spector, B. (2011). Experimental evidence for embedded scalar implicatures. *Journal of Semantics, 28,* 359-400.

Chomsky, N., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics, 1*(2), 97-138.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row.

Chomsky, N. (1957). *Syntactic structures.* Mouton, The Hague.

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

Chomsky, N. (1971). *Problems of knowledge and freedom*. London: Fontana.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. New York: Praeger.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior 12,* 335-359.

Cleland, C. E. (2002). Methodological and epistemic differences between historical science and experimental science. *Philosophy of Science, 69*(3), 474-496.

Clifton Jr, C., & Frazier, L. (2010). When are downward-entailing contexts identified? The case of the domain widener *ever*. *Linguistic Inquiry, 41*(4), 681-689.

Coetzee, A. W. (2009). Grammar is both categorical and gradient. In S. Parker (Ed.). *Phonological argumentation: Essays on evidence and motivation* (pp. 9-42). London: Equinox.

Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In *Computational phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology,* pp. 49-56. Somerset, UK: Association for Computational Linguistics.

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

Cowart, W. (2012). Doing experimental syntax: Bridging the gap between syntactic questions and well-designed questionnaires. In J. Myers (Ed.) *In search of grammar: Experimental and corpus-based studies* (pp. 67-96). Language and Linguistics Monograph Series 48. Taipei, Taiwan: Language and Linguistics.

Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language, 63*(3), 522-543.

Crocker, M. W., & Keller, F. (2006). Probabilistic grammars as models of gradience in language processing. In G. Fanselow (Ed.) *Gradience in grammar: Generative perspectives* (pp. 227-245). Oxford: Oxford University Press.

Culbertson, J. (2012). Typological universals as reflections of biased learning: Evidence from artificial language learning. *Language and Linguistics Compass, 6*(5), 310-329.

Culbertson, J., & Gross, S. (2009). Are linguists better subjects? *British Journal for the Philosophy of Science, 60,* 721-736.

Culicover, P. W., & Jackendoff, R. (2010). Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences, 14*(6), 234-235.

Dąbrowska, E. (1997). The LAD goes to school: A cautionary tale for nativists. *Linguistics, 35*, 735-766.

Dąbrowska, E. (2010). Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review, 27*(1), 1-23.

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Norrmann, I. (2011). Explaining sonority projection effects. *Phonology, 28*(2), 197-234.

Davies, W. D., & Kaplan, T. I. (1998). Native speaker vs. L2 learner grammaticality judgements. *Applied Linguistics, 19*(2), 183-203.

Dennett, D. (2003). Who's on first? Heterophenomenology explained. *Journal of Consciousness Studies, 10*(9-10), 19-30.

Derwing, B. L., & de Almeida, R. G. (2009). Non-chronometric experiments in linguistics. In D. Eddington (Ed.), *Experimental and quantitative linguistics* (pp. 234-282). Munich: Lincom.

Devitt, M. (2006). *Ignorance of language*. Oxford: Oxford University Press.

Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological Research, 69*, 338-351.

Dillon, B., Staub, A., Levy, J., & Clifton Jr, C. (2017). Which noun phrases is the verb supposed to agree with? Object agreement in American English. *Language, 93*(1), 65-96.

Di Sciullo, A. M., & Williams, E. (1987). *On the definition of word*. Cambridge, MA: MIT Press.

Divjak, D. (2008). On (in)frequency and (un)acceptability. In B. Lewandowska-Tomaszczyk (Eds.) *Corpus linguistics, computer tools, and applications: State of the art* (pp. 213-233). Frankfurt am Main: Peter Lang.

Dresher, E. (1995). There's no reality like psychological reality. *Glot International, 1* (1), 7.

Engdahl, E. (1983). Parasitic gaps. *Linguistics and Philosophy, 6*(1), 5-34.

Erlewine, M. Y., & Kotek, H. (2016). A streamlined approach to online linguistic surveys. *Natural Language & Linguistic Theory, 34*(2), 481-495.

Fanselow, G. (2007). Carrots - perfect as vegetables, but please not as a main dish. *Theoretical Linguistics, 33*(3), 353-367.

Fanselow, G., Kliegl, R., & Schlesewsky, M. (2006). Syntactic variation in German *wh*-questions: Empirical investigations of weak crossover violations and long *wh*-movement. In P. Pica, J. Rooryck, & J. van Craenenbroeck (Eds.) *Linguistic variation yearbook 2005* (pp. 37-63). Amsterdam: John Benjamins.

Featherston, S. (2005). That-trace in German. *Lingua, 115* (9), 1277-1302.

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. In C. M. Riehl & A. Rothe (Eds.) *Was ist linguistische Evidenz?* (pp. 69-89). Aachen: Shaker Verlag.

Featherston, S. (2009). Relax, lean back, and be a linguist. *Zeitschrift für Sprachwissenshaft, 28*, 127-132.

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition, 28*(7), 1109-1115.

Francis, E. J., Lam, C., Zheng, C. C., Hitz, J., & Matthews, S. (2015). Resumptive pronouns, structural complexity, and the elusive distinction between grammar and performance: evidence from Cantonese. *Lingua, 162*, 56-81.

Francom, J. (2009). *Experimental syntax: Exploring the effect of repeated exposure to anomalous syntactic structure: Evidence from rating and reading tasks*. PhD. dissertation, University of Arizona.

Frisch, S. A., & Stearns, A. M. (2006). Linguistic and metalinguistic tasks in phonology: Methods and findings. In G. Fanselow, C. Féry, R. Vogel, & M. Schleswsky (Eds.), *Gradience in grammar: Generative perspectives* (pp. 70-84). Oxford: Oxford University Press.

Frisch, S. A., & Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language*, *77* (1), 91-106.

Geary, J. A. (1943). The Proto-Algonquian form for ‘I-thee’. *Language*, *19* (2), 147-151.

Gerken, L., & Bever, T. G. (1986). Linguistic intuitions are the result of interactions between perceptual processes and linguistic universals. *Cognitive Science, 10*(4), 457-476.

Gervain, J. (2003). Syntactic microvariation and methodology: Problems and perspectives. *Acta Linguistica Hungarica, 50*(3-4), 405-434.

Gibbs, Jr., R. W. (2007). Why cognitive linguists should care more about empirical methods. In M. Gonzalez-Marquez, I. Mittleberg, S. Coulson, & M. J. Spivey (Eds.), *Methods in cognitive linguistics* (pp. 2-18). John Benjamins.

Gibson, E., & Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences, 14*(6), 233-234.

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

Gibson, E., & Schütze, C. T. (1999). Disambiguation preferences in noun phrase conjunction do not mirror corpus frequency. *Journal of Memory and Language, 40*(2), 263-279.

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.

Gibson, E., Sandberg, C., Fedorenko, E., Bergen, L., & Kiran, S. (2016). A rational inference approach to aphasic language comprehension. *Aphasiology, 30*(11), 1341-1360.

Goldrick, M. (2011). Using psychological realism to advance phonological theory. In J. Goldsmith, J. Riggle, & A. C. L. Yu (Eds.) *The handbook of phonological theory, second edition* (pp. 631-660). Chichester, UK: Wiley-Blackwell.

Goodall, G. (2011). Syntactic satiation and the inversion effect in English and Spanish *wh-* questions. *Syntax, 14*(1), 29-47.

Gordon, P. C., & Hendrick, R. (1997). Intuitive knowledge of linguistic co-reference. *Cognition, 62*(3), 325-370.

Gordon, P. C., & Hendrick, R. (2005). Relativization, ergativity, and corpus frequency. *Linguistic Inquiry, 36*(3), 456-463.

Gorman, K. (2013). Generative phonotactics. Doctoral dissertation, University of Pennsylvania.

Gouskova, M., & Becker, M. (2013). Nonce words show that Russian yer alternations are governed by the grammar. *Natural Language & Linguistic Theory, 31*(3), 735-765.

Greenberg, J. H., & Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English. *Word, 20* (2), 157-177.

Gries, S. Th. (2012). Corpus linguistics, theoretical linguistics, and cognitive/psycholinguistics: towards more and more fruitful exchanges. In J. Mukherjee & M. Huber (Eds.), *Corpus linguistics and variation in English: Theory and description* (pp. 41-63). Amsterdam: Rodopi.

Gross, S., & Culbertson, J. (2011). Revisited linguistic intuitions. *The British Journal for the Philosophy of Science, 62,* 639-656.

Haider, H. (1983). Connectedness effects in German. *Groninger Arbeiten zur Germanischen Linguistik, 23,* 82-119.

Hammond, M. (2012). Empirical methods in phonological research. In J. Myers (Ed.) *In search of grammar: Experimental and corpus-based studies* (pp. 29-66). Language and Linguistics Monograph Series 48. Taipei, Taiwan: Language and Linguistics.

Han, C.-H., Lidz, J. & Musolino, J. (2007). V-raising and grammar competition in Korean: evidence from negation and quantifier scope. *Linguistic Inquiry 38*(1), 1-47.

Han, C. H., Musolino, J., & Lidz, J. (2016). Endogenous sources of variation in language acquisition. *Proceedings of the National Academy of Sciences, 113*(4), 942-947.

Harris, R. A. (1993). *The linguistics wars*. Oxford: Oxford University Press.

Haspelmath, M. (2002). *Understanding morphology*. London: Arnold.

Haug, M. C. (2014). (Ed.) *Philosophical methodology: The armchair or the laboratory?* New York: Routledge.

Häussler, J., Grant, M., Fanselow, G., & Frazier, L. (2015). Superiority in English and German: Cross- language grammatical differences? *Syntax, 18*(3), 235-265.

Hawkins, J. A. (1999). Processing complexity and filler-gap dependencies across grammars. *Language*, *75*(2), 244-285.

Hay, J. (2002). From speech perception to morphology: Affix ordering revisited. *Language, 78*(3), 527-555.

Hayes, B., & White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic Inquiry, 44*(1), 45-75.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry, 39*(3), 379-440.

Heigham, J., & Croker, R. A. (2009). *Qualitative research in applied linguistics: A practical introduction*. Basingstoke, UK: Palgrave Macmillan.

Henry, A. (2005). Non-standard dialects and linguistic data. *Lingua, 115*(11), 1599-1617.

Hill, A. A. (1961). Grammaticality. *Word, 17*, 1-10.

Hiramatsu, K. (2000). *Accessing linguistic competence: Evidence from children's and adults' acceptability judgements*. Storrs, CT: University of Connecticut dissertation.

Hoffmann, T. (2006). Corpora and introspection as corroborating evidence: The case of preposition placement in English relative clauses. *Corpus Linguistics and Linguistic Theory, 2*(2), 165-95.

Hofmeister, P., Casasanto, L. S., & Sag, I. A. (2012a). How do individual cognitive differences relate to acceptability judgments?: A reply to Sprouse, Wagers, and Phillips. *Language, 88*(2), 390-400.

Hofmeister, P., Casasanto, L. S., & Sag, I. A. (2012b). Misapplying working-memory tests: A reductio ad absurdum. *Language, 88*(2), 408-409.

Hofmeister, P., Casasanto, L. S., & Sag, I. A. (2013). Islands in the grammar? Standards of evidence. In J. Sprouse & N. Hornstein (Eds.) *Experimental syntax and the islands debate* (pp.

42-63). Cambridge: Cambridge University Press.

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language 86*(2), 366-415.

Householder, F. W. (1965). On some recent claims in phonological theory. *Journal of Linguistics, 1*(1), 13-34.

Householder, F. W. (1973). On arguments from asterisks. *Foundations of Language*, *l0*(3), 365-376.

Huang, Y.-C., & Kaiser, E. (2008). Investigating filler-gap dependencies in Chinese topicalization. In M. K. M. Chan & H. Kang (Eds.) *Proceedings of the 20th North American Conference on Chinese Linguistics*. Columbus, OH: Ohio State University.

Hung, D. L., Tzeng, O. J. L. & Ho, C.-Y. (1999). Word superiority effect in the visual processing of Chinese. In O. J. L. Tzeng (Ed.), *Journal of Chinese Linguistics Monograph Series No. 13: The biological bases of language* (pp. 61-95). Taipei: Academia Sinica.

Hunter, L. (1982). Silence is also language: Hausa attitudes about speech and language. *Anthropological Linguistics, 24* (4), 389-409.

Ionin, T. (2010). The scope of indefinites: an experimental investigation. *Natural Language*

Semantics, 18(3), 295-350.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59* (4), 434-446.

Jakobovits, L. A., & Lambert, W. E. (1964). Stimulus-characteristics as determinants of semantic changes with repeated presentation. *The American Journal of Psychology, 77*(1), 84-92.

Kager, R., & Pater, J. (2012). Phonotactics as phonology: Knowledge of a complex restriction in Dutch. *Phonology, 29*(1), 81-111.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*(2), 189-217.

Kawahara, S. (2011). Experimental approaches in theoretical phonology. In M. van Oostendorp, C. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell companion to phonology* (pp. 2283-2303). Oxford: Wiley-Blackwell.

Kawahara, S. (2015). Comparing a forced-choice wug test and a naturalness rating test: An exploration using rendaku. *Language Sciences, 48*, 42-47.

Kawahara, S., & Sano, S. I. (2014). Identity avoidance and Lyman's Law. *Lingua, 150*, 71-77.

Keller, F., & Alexopoulou, T. (2001). Phonology competes with syntax: experimental evidence for the interaction of word order and accent placement in the realization of information structure. *Cognition 79(*3), 301-372.

Kempen, G., & Harbusch, K. (2005). The relationship between grammaticality ratings and corpus frequencies: A case study into word order variability in the midfield of German clauses. In S. Kepser & M. Reis (Eds.), *Linguistic evidence: Empirical, theoretical and computational perspectives* (pp. 329-349). Berlin: Mouton de Gruyter.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords. *Frontiers in Psychology, 1*, 174.

Kiparsky, P. (1975). What are phonological theories about? In D. Cohen and J. R. Wirth (Eds.), *Testing linguistic hypotheses* (pp. 187-209). New York: John Wiley and Sons.

Kiparsky, P. (1982). Lexical phonology and morphology. Linguistics in the morning calm. In Linguistics Society of Korea (Ed.) *Linguistics in the morning calm: Selected papers from SICOL-1981* (pp. 3-91). Seoul: Hanshin Publishing Company.

Kirby, J. P., & Yu, A. C. L. (2007). Lexical and phonotactic effects on wordlikeness judgments in

Cantonese. *Proceedings of the International Congress of Phonetic Sciences, 16*, 1389-1392.

Kirk, R. E. (2012). *Experimental design: Procedures for the behavioral sciences, 4th edition*. Thousand Oaks, CA: Sage Publications.

Kluender, R., & Kutas, M. (1993). Subjacency as a processing phenomenon. *Language and Cognitive Processes, 8*(4), 573-633.

Koorneef, A. W., Avrutin, S., Wijnen, F., & Reuland, E. (2011). Tracking the preference for bound-variable dependencies in ambiguous ellipses and *only*-structures. In J. T. Runner (Ed.) *Syntax and semantics, vol. 37: Experiments at the interfaces* (pp. 67-100). Bingley, UK: Emerald Group Publishing.

Labov, W. (1996). When intuitions fail. *CLS 32: Papers from the Parasession on Theory and Data in Linguistics 32*, 77-105. Chicago Linguistics Society.

Lakoff, G. (1991). Cognitive versus generative linguistics: How commitments influence results. *Language & Communication, 11*(1-2), 53-62.

Langendoen, D. T., Kalish-Landon, N., & Dore, J. (1973). Dative questions: A study in the relation of acceptability to grammaticality of an English sentence type. *Cognition, 2*(4), 451-478.

Lewis, S., & Phillips, C. (2015). Aligning grammatical theories and language processing models.

Journal of Psycholinguistic Research, *44*(1), 27-46.

Lidz, J., Pietroski, P., Halberda, J., & Hunter, T. (2011). Interface transparency and the psychosemantics of *most*. *Natural Language Semantics, 19*(3), 227-256.

Lin, P. (2004). The semantic connotation of the Chinese *bei*-construction. *Chinese Journal of Psychology, 46* (2-3), 197-212.

Linzen, T., & Oseki, Y. (2015). The reliability of acceptability judgments across languages. New York: New York University ms. http://ling.auf.net/lingbuzz/002854

Ludlow, P. (2011). *The philosophy of generative linguistics*. Oxford: Oxford University Press.

Luka, B. J. (1998). Example sentences in syntax articles: The influence of asterisks and affinity toward author. *Proceedings of the 34th Annual Meeting of the Chicago Linguistic Society*. University of Chicago: Chicago Linguistic Society, 269–280.

Luka, B. J., & Barsalou, L. W. (2005). Structural facilitation: mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language 52*(3), 436-459.

Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004). Semantics, cross-cultural style. *Cognition*, *92*(3), B1-B12.

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.

Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman and Company.

McCawley, J. D. (1986). Today the world, tomorrow phonology. *Phonology, 3*(1), 27-43.

McKercher, D. A., & Jaswal, V. K. (2012). Using judgment tasks to study language knowledge. In E. Hoff (Ed.) *Research methods in child language: A practical guide* (pp. 149-161). Oxford: Blackwell.

Menard, S. (2004). Six approaches to calculating standardized logistic regression coefficients. *The American Statistician, 58* (3), 218-223.

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology, vol. II* (pp. 419-491). New York: John Wiley and Sons.

Moreton, E., & Pater, J. (2012a). Structure and substance in artificial- phonology learning, part I: Structure. *Language and Linguistics Compass, 6*(11), 686-701.

Moreton, E., & Pater, J. (2012b). Structure and substance in artificial-phonology learning, part II: Substance. *Language and Linguistics Compass, 6*(11), 702-718.

Myers, J. (2007). Generative morphology as psycholinguistics. In G. Jarema & G. Libben (Eds.), *The mental lexicon: Core perspectives* (pp. 105-128). Amsterdam: Elsevier.

Myers, J. (2009a). Syntactic judgment experiments. *Language and Linguistics Compass 3*(1), 406-423.

Myers, J. (2009b). The design and analysis of small-scale syntactic judgment experiments. *Lingua 119*(3), 425-444.

Myers, J. (2012a). Testing adjunct and conjunct island constraints in Chinese. *Language and Linguistics 13*(3), 437-470.

Myers, J. (2012b). Methods in search of grammar, grammar in search of methods. In J. Myers (Ed.) *In search of grammar: Empirical methods in linguistics* (pp. 1-27). Language and Linguistics Monograph Series 48. Taipei, Taiwan: Language and Linguistics..

Myers, J. (2015). Markedness and lexical typicality in Mandarin acceptability judgments. *Language & Linguistics, 16*(6), 791-818.

Myers, J. (2016). Meta-megastudies. *The Mental Lexicon, 11*(3), 329-349.

Myers, J. (2017). Morphological processing of compounds, behavioral studies. In R. Sybesma, W. Behr, Y. Gu, Z. Handel, C.-T. J. Huang, & J. Myers (Eds.), *Encyclopedia of Chinese language and linguistics*, vol. 3 (pp. 94-100). Leiden, Netherlands: Brill.

Nagata, H. (1988). The relativity of linguistic intuition: The effect of repetition on grammaticality judgments. *Journal of Psycholinguistic Research, 17*(1), 1-17.

Nagata, H. (1989). Effect of repetition on grammaticality judgments under objective and subjective self-awareness conditions. *Journal of Psycholinguistic Research, 18*(3), 255-269.

Neeleman, A. (2013). Comments on Pullum. *Mind & Language, 28*(4), 522-531.

Neeleman, A., & van de Koot, H. (2010). Theoretical validity and psychological reality of the grammatical code. In M. Everaert, T. Lentz, H. De Mulder, Ø . Nilsen, & A. Zondervan (Eds.), *The linguistics enterprise: From knowledge of language to knowledge in linguistics* (pp. 183-212). Amsterdam: John Benjamins Publishing Company.

Neidle, C., Kegl, J., MacLaughlin, D., Bahan, B., & Lee, R. G. (2001). *The syntax of American Sign Language: Functional categories and hierarchical structure*. Cambridge, MA: MIT Press.

Newmeyer, F. J. (1983). *Grammatical theory: its limits and its possibilities*. Chicago: University of Chicago Press.

Newmeyer, F. J. (2010). What conversational English tells us about the nature of grammar: A critique of Thompson's analysis of object complements. In K. Boye & E. Engberg-Pedersen (Eds.), *Language usage and language structure* (pp. 3-44). Berlin: Walter de Gruyter.

Noonan, M. (1999). Non-structuralist syntax. In M. Darnell, E.Moravcsik, F. Newmeyer, M. Noonan, & K. Wheatley (Eds.) *Functionalism and formalism in linguistics* (pp. 11-31). Amsterdam: John Benjamins.

Ohala, J. J. (1986). Consumer's guide to evidence in phonology. *Phonology Yearbook 3*, 3-26.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics–pragmatics interface. *Cognition, 86*(3), 253-282.

Pertz, D. L., & Bever, T. G. (1975). Sensitivity to phonological universals in children and adolescents. *Language, 51*(1), 149-162.

Petronio, K., & Lillo-Martin, D. (1997). WH-movement and the position of Spec-CP: Evidence from American Sign Language. *Language, 73*(1), 18-57.

Phillips, C. (2006). The real-time status of island phenomena. *Language, 82*(4), 795-823.

Phillips, C. (2010). Should we impeach armchair linguists? In S. Iwasaki, H Hoji, P. Clancy, &

S.-O. Sohn (Eds.) *Proceedings from Japanese/Korean Linguistics 17* (pp. 49-64). Stanford, CA: CSLI Publications.

Phillips, C., & Lasnik, H., (2003). Linguistics and empirical evidence: Reply to Edelman and Christiansen. *Trends in Cognitive Science, 7*(2), 61-62.

Phillips, C., & Wagers, M. (2007). Relating structure and time in linguistics and psycholinguistics. In M. G. Gaskell (Ed.), *Handbook of psycholinguistics* (pp. 739-756). Oxford: Oxford University Press.

Pinker, S., & Birdsong, D. (1979). Speakers' sensitivity to rules of frozen word order. *Journal of Verbal Learning and Verbal Behavior, 18*(4), 497-508.

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes, 8*(1), 1-56.

Pullum, G. K. (2013a). The central question in comparative syntactic metatheory. *Mind & Language, 28*(4), 492-521.

Pullum, G. K. (2013b). Consigning phenomena to performance: A response to Neeleman. *Mind & Language, 28*(4), 532-537.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology:*

General, 118(3), 219-235.

Rizzi, L. (1982). Violations of the *wh*-island constraint and the subjacency condition. In L. Rizzi (Ed.) *Issues in Italian syntax* (pp. 49-76). Dordrecht, NL: Foris.

Ross, J. R. (1967). *Constraints on variables in syntax*. Cambridge, MA: MIT Ph.D. thesis.

Saah, K. K., & Goodluck, H. (1995). Island effects in parsing and grammar: Evidence from Akan. *Linguistic Review, 12*, 381-409.

Saka, P. (1998). Quotation and the use-mention distinction. *Mind, 107*(425), 113-135.

Sampson, G. R. (2007). Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory 3*(1), 1-32.

Schlesewsky, M., Bornkessel, I., & McElree, B. (2006). Decomposing gradience: Qualitative and quantitative distinctions. In G. Fanselow, C. Féry, R. Vogel, & M. Schleswsky (Eds.), *Gradience in grammar: Generative perspectives* (pp. 124-142). Oxford: Oxford University Press.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language, 37*(1), 118-139.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and*

linguistic methodology. Chicago: University of Chicago Press.

Schütze, C. T. (2005). Thinking about what we are asking speakers to do. In S. Kepser & M. Reis (Eds.) *Linguistic evidence: Empirical, theoretical, and computational perspectives* (pp. 457-485). Berlin: Mouton de Gruyter.

Schütze, C. T. (2011). Linguistic evidence and grammatical theory. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(2), 206-221.

Schütze, C. T., & Sprouse, J. (2013). Judgment data. In R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 27-50). Cambridge, UK: Cambridge University Press.

Shademan, S. (2007). *Grammar and analogy in phonotactic well-formedness judgments*. Doctoral dissertation, UCLA.

Snyder, W., (2000). An experimental investigation of syntactic satiation effects. *Linguistic Inquiry, 31*(3), 575-582.

Sorace, A. (1996). The use of acceptability judgements in second language acquisition research. In W. C. Ritchie and T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 375-409). San Diego, CA: Academic Press.

Sorace, A., & Keller, F., (2005). Gradience in linguistic data. *Lingua 115*(11), 1497-1524.

Spencer, A. (1991). *Morphological theory: An introduction to word structure in generative grammar*. Oxford: Basil Blackwell.

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

Sperlich, D. (2015). Assessing anaphoric relations via the phased choice methodology. *International Review of Applied Linguistics in Language Teaching, 53*(4), 355-388.

Sprouse, J. (2007). Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics, 1*, 123-134.

Sprouse, J. (2009). Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry 40*(2), 329-341.

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. (2015). Three open questions in experimental syntax. *Linguistics Vanguard, 1*(1),

89-100.

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), 14.1-32.

Sprouse, J., Caponigro, I., Greco, C., & Cecchetto, C. (2016). Experimental syntax and the variation of island effects in English and Italian. *Natural Language & Linguistic Theory, 34*(1), 307-344.

Sprouse, J., Fukuda, S., Ono, H., & Kluender, R. (2011). Reverse island effects and the backward search for a licensor in multiple *wh*- questions. *Syntax, 14*(2), 179-203.

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.

Sprouse, J., Wagers, M., & Phillips, C. (2012a). A test of the relation between working-memory capacity and syntactic island effects.*Language, 88*(1), 82-123.

Sprouse, J., Wagers, M., & Phillips, C. (2012b). Working-memory capacity and island effects: A

reminder of the issues and the facts. *Language, 88*(2), 401-407.

Stefanowitsch, A. (2006). Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory, 2*(1), 61-77.

Sternberg, S. (1998). Discovering mental processing stages: The method of additive factors. In D. Scarborough & S. Sternberg (Eds.), *An invitation to cognitive science, vol. 4: Methods, models, and conceptual issues* (pp. 703-863). MIT Press.

Stevens, S. S. (1956). The direct estimation of sensory magnitudes - loudness. *American Journal of Psychology, 69*, 1-25.

Teddiman, L. (2006). On the processing of novel root+suffix combinations. In C. Gurski & M. Radisic (Ed.s), *Proceedings of the 2006 Annual Conference of the Canadian Linguistic Association*. http://westernlinguistics.ca/Publications/CLA2006/Teddiman.pdf

Topolinski, S., & Strack, F. (2009). The architecture of intuition: Fluency and affect determine intuitive judgments of semantic and visual coherence and judgments of grammaticality in artificial grammar learning. *Journal of Experimental Psychology: General, 138*(1), 39-63.

Toribio, A. J. (2001). Accessing bilingual code-switching competence. *International Journal of Bilingualism, 5*(4), 403-436.

Tunney, R. J., & Shanks, D. R. (2003). Subjective measures of awareness and implicit cognition. *Memory & Cognition 31*(7), 1060-1071.

Ullman, M. T. (1999). Acceptability ratings of regular and irregular past-tense forms: Evidence for a dual-system model of language from word frequency and phonological neighbourhood effects. *Language and Cognitive Processes, 14*(1), 47-67.

Valian, V. (1982). Psycholinguistic experiment and linguistic intuition. In T. W. Simon & R. J. Scholes (Eds.), *Language, mind, and brain* (pp. 179-188). Hillsdale, NJ: Lawrence Erlbaum.

Verhagen, V., & Mos, M. (2016). Stability of familiarity judgments: Individual variation and the invariant bigger picture. *Cognitive Linguistics*, *27*(3), 307-344.

Wasow, T., & Arnold, J. (2005). Intuitions in linguistic argumentation. *Lingua, 115*(11), 1481-1496.

Weskott, T., & Fanselow, G. (2011). On the informativity of different measures of linguistic acceptability. *Language, 87*(2), 249-273.

Woodward, J. (2016). Causation and manipulability. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition). https://plato.stanford.edu/archives/win2016/entries/causation-mani/.

Yap, M. J., Sibley, D. E., Balota, D. A., Ratcliff, R., & Rueckl, J. (2015). Responding to nonwords in the lexical decision task: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 597.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society, 104*(5), 444-466.

Yoshida, M., Kazanina, N., Pablos, L., & Sturt, P. (2014). On the origin of islands. *Language, Cognition and Neuroscience*, *29*(7), 761-770.

Zimmer, K. E. (1969). Psychological correlates of some Turkish morpheme structure constraints. *Language*, *45*, 309-321.

Zuraw, K. (2007). The role of phonetic knowledge in phonological patterning: corpus and survey evidence from Tagalog infixation. *Language, 83*(2), 277-316.

Figure captions.

Figure 1. The results of idealized factorial syntactic acceptability judgment experiments on syntactic islands. (a) Traditional acceptability judgment report (Ross, 1967, p. 70). (b) Idealized results for cross-participant mean judgments from a magnitude estimation experiment on the same sentences. (c) Idealized results for cross-participant mean binary accept/reject judgment rates on the same sentences.

Figure 2. Effects of neighborhood density and phonotactic probability on random samples of auditory nonlexical syllables in English (Bailey & Hahn 2001; original data shared by Todd Bailey) and Cantonese (Kirby & Yu, 2007; original data shared by James Kirby), computed in by-language linear regression models on by-item judgment means (both studies used Likert scales that were arcsine square root transformed, a widely used procedure that attempts to give a better fit for linear models, and rescaled to the range of zero to one). For the purpose of this cross-study comparison, both lexical variables and judgments were converted to by-language *z* scores. The slopes can thus be interpreted as effect sizes; line lengths indicate data ranges.

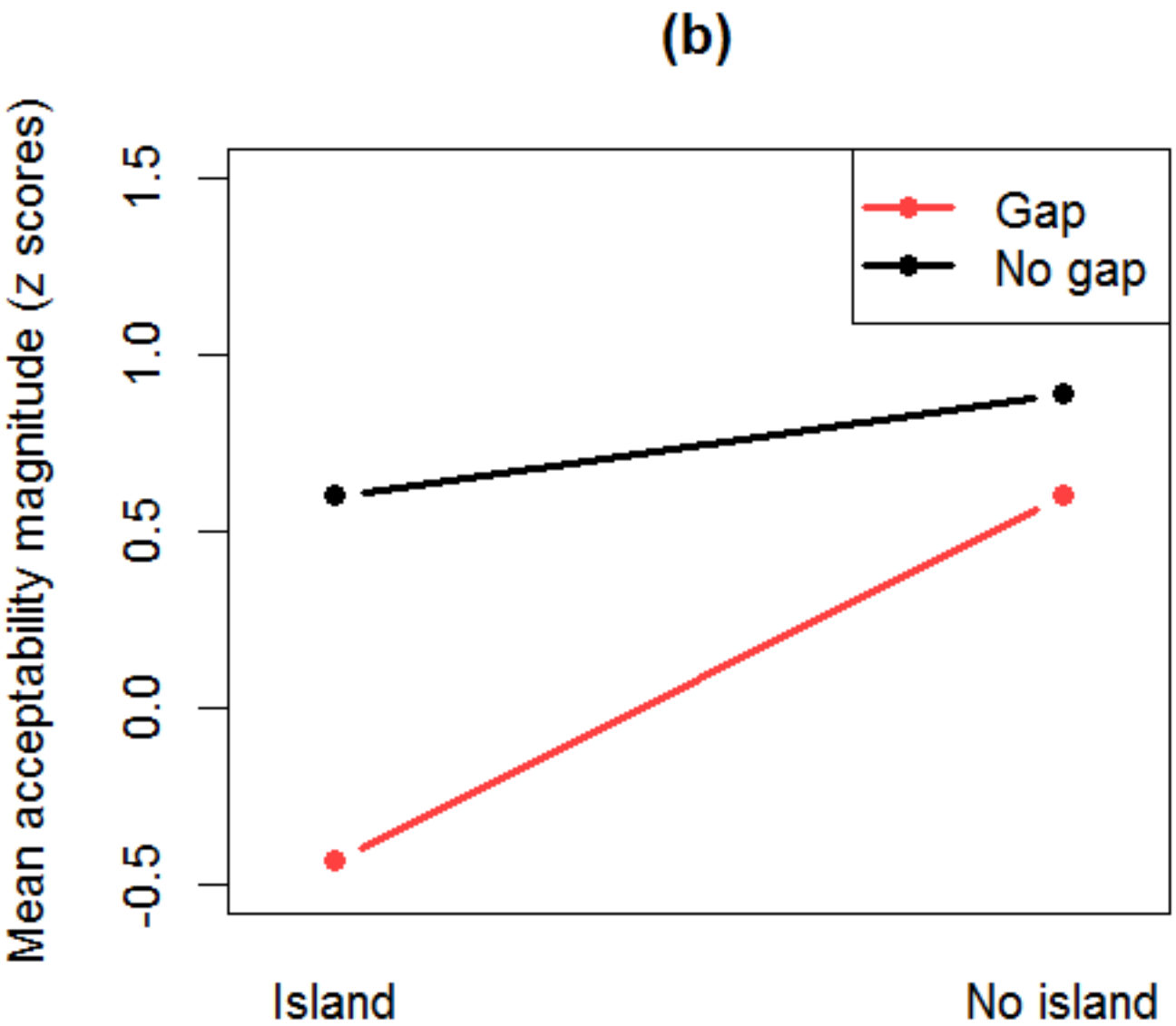
(a)

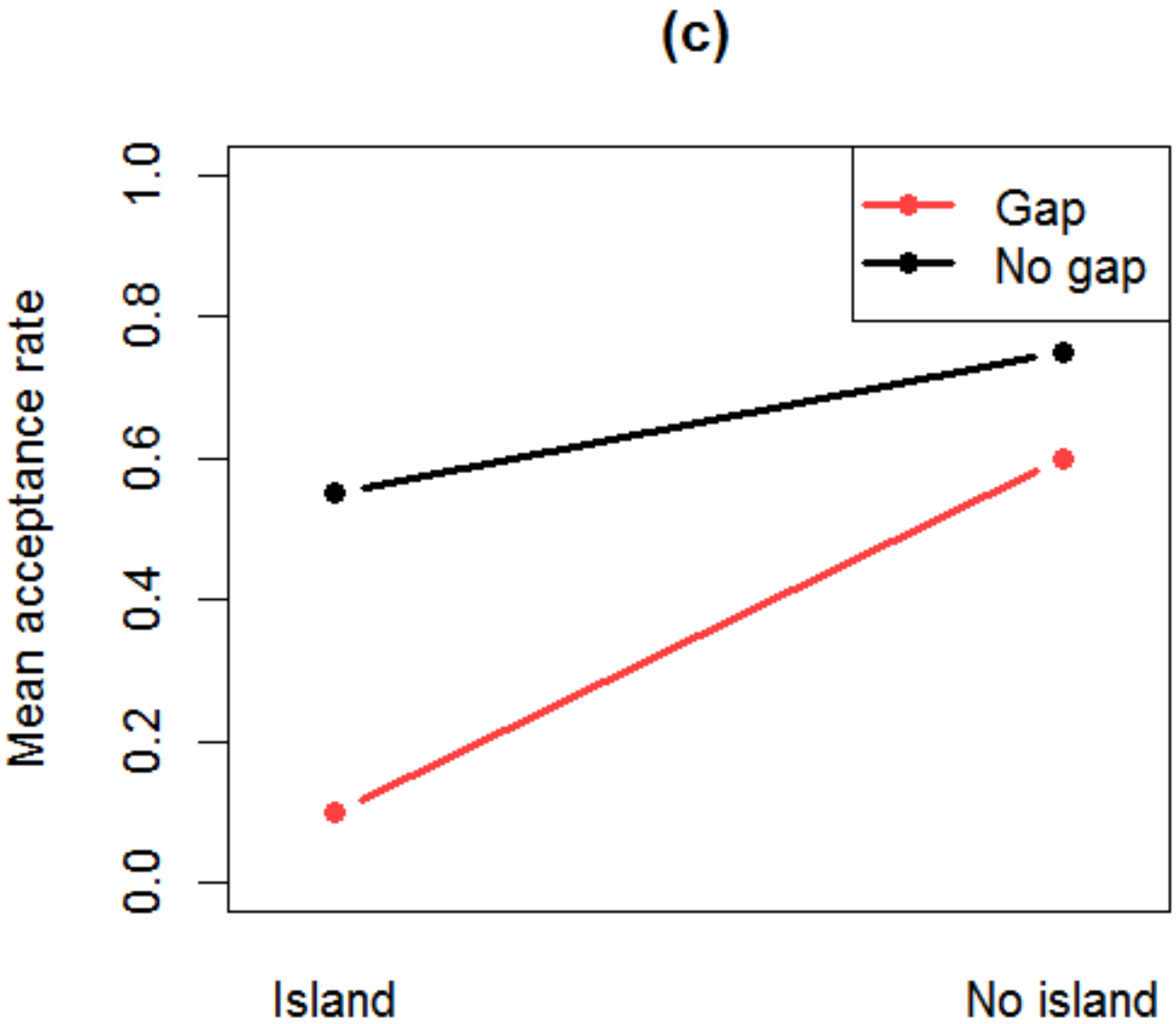
(4.17a) [-Gap] [+Island] I believed [the claim that Otto was wearing this hat]

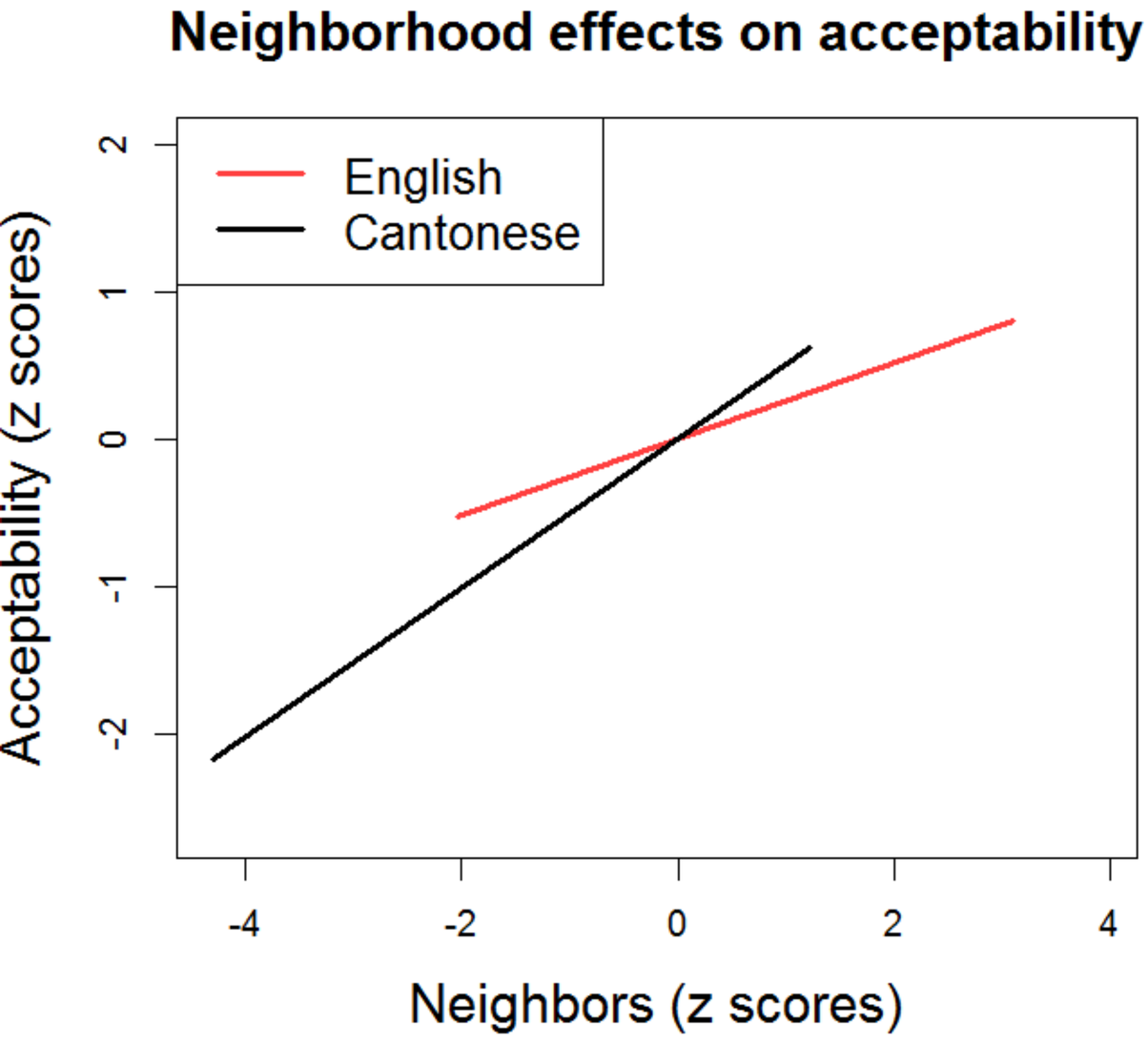
(4.17b) [-Gap] [-Island] I believed that Otto was wearing this hat

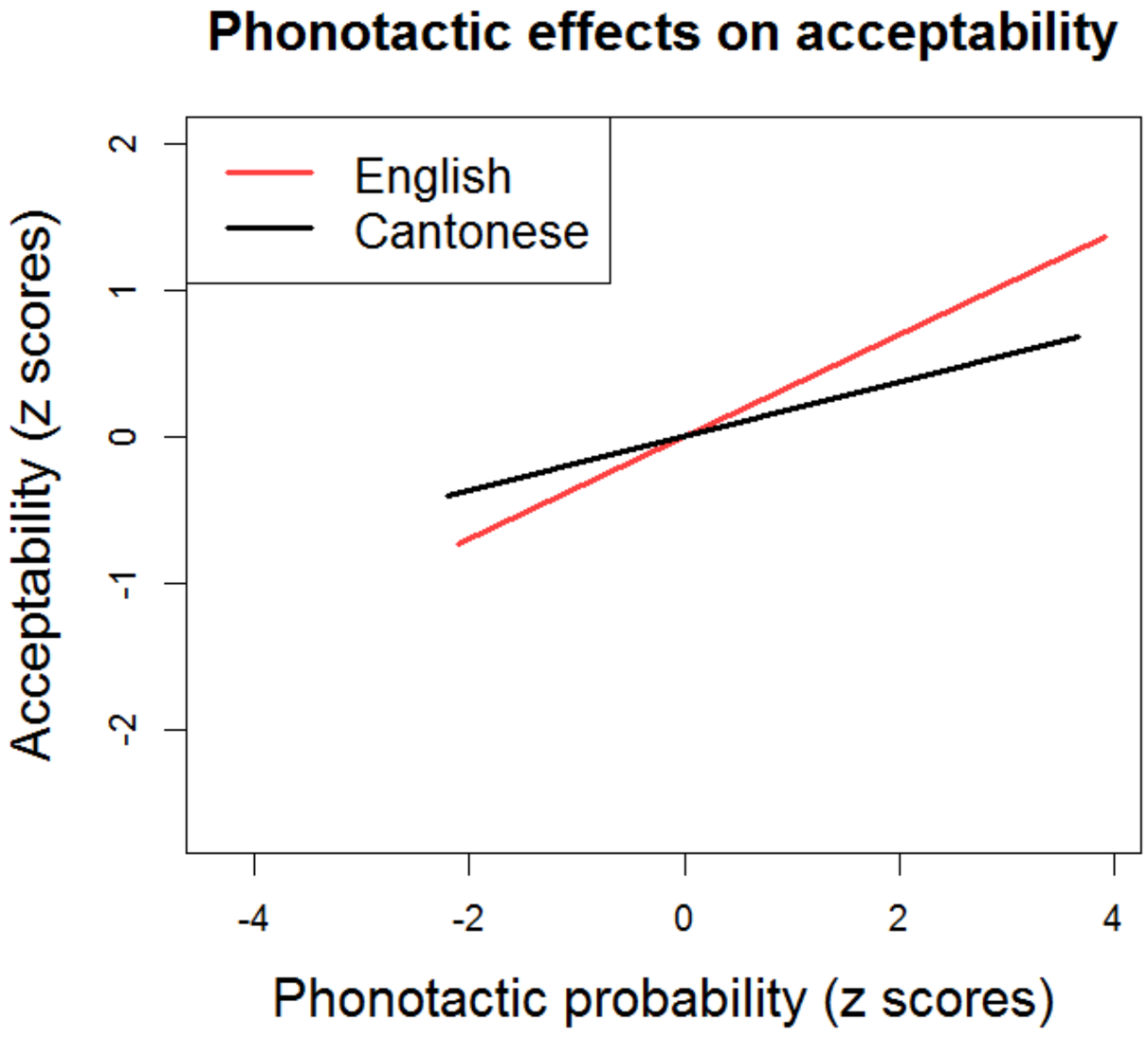
(4.18a) [+Gap] [+Island] * The hat which I believed [the claim that Otto was wearing _] is red

(4.18b) [+Gap] [-Island] The hat which I believed that Otto was wearing _ is red


(b)
Mean acceptability magnitude (z scores)
-0.5
0.0
0.5
1.0
1.5
Gap
No gap
Island
No island


(c)
Mean acceptance rate
0.0
0.2
0.4
0.6
0.8
1.0
Gap
No gap
Island
No island


Neighborhood effects on acceptability
English
Cantonese
Acceptability (z scores)
2
1
0
-1
-2
-4
-2
0
2
4
Neighbors (z scores)


Phonotactic effects on acceptability
English
Cantonese
Acceptability (z scores)
2
1
0
-1
-2
-4
-2
0
2
4
Phonotactic probability (z scores)