Oxford Bibliographies Online: Linguistics

Acceptability Judgments

Jon Sprouse
Assistant Professor
University of California, Irvine
Department of Cognitive Sciences
http://www.socsci.uci.edu/~jsprouse

INTRODUCTION

Acceptability judgments form a substantial portion of the empirical foundation of nearly every area of linguistics (e.g., phonology, morphology, syntax, and semantics), and nearly every type of linguistic theory. As the name implies, acceptability judgments are consciously reported perceptions of acceptability that arise when native speakers attempt to comprehend a (spoken or written) utterance, whether it be a syllable, (non)word, or sentence, and are asked to answer a question such as “How natural/acceptable/grammatical is this utterance?”. Although there is quite a bit of variation in both the form of the instructions given and the response scales employed during acceptability judgment collection, all acceptability judgment tasks share the following assumptions: (i) acceptability is a property of utterances, (ii) the grammatical status of the utterance strongly (but not necessarily uniquely) influences the acceptability of the utterance, (iii) native speakers can consciously perceive the acceptability of utterances, and (iv) native speakers can consciously report their perceptions of acceptability. For space reasons, this bibliography will focus exclusively on the use of acceptability judgments in the domain of syntax. Furthermore, because a complete bibliography of every use of acceptability judgments in the syntax literature would be impractical, this bibliography will focus on higher-level questions regarding the use of acceptability judgments for hypothesis testing. It should also be noted that even this restricted subset of the syntactic acceptability judgment literature does not easily lend itself to a linear, topic-by-topic bibliography, as most of the major works tend to cover several of the relevant

topics. For style reasons (such as the limit of one paragraph and 8 references per subsection), this has led to some subjectivity in the allocation of references in some sections. In general the goal has been to collect the most relevant references for each topic, with a preference for references that have appeared since the publication of previous comprehensive reviews.

GENERAL OVERVIEWS

Chomsky 1965 is generally cited as the first comprehensive argument in support of an acceptability-centric approach to syntactic data collection. Newmeyer 1983 provides a comprehensive introduction to the field of generative linguistics, with lucid discussions of its goals, scope, and the types of evidence that bear on generative theories. Schütze 1996 provides a comprehensive review of the acceptability judgment literature prior to the mid-1990s, and as such discusses many of the issues that will be covered in this bibliography in some detail. Acceptability judgments in syntax can generally be divided into two types based on the method of collection: traditionally collected judgments, which generally involve relatively informal collection procedures, and formally collected judgments, which generally involve the formal collection procedures familiar from experimental psychology. Through its discussion of the (participant and task) factors that may affect acceptability judgments Schütze 1996 provides a comprehensive introduction to the properties of both traditional judgment collection and formal judgment collection. Marantz 2005 presents one of the most accessible discussions of traditional judgment collection methods as both a behavioral experiment and evidence for syntactic theories (as part of a comprehensive discussion of the role of linguistic theories in cognitive neuroscience). Cowart 1997 is a textbook that provides a comprehensive introduction to formal judgment collection methods. Myers 2009b is a review article that covers many of the questions that have been raised concerning the reliability and sensitivity of both traditional and formal collection methods.

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

An influential book that discusses both the goals and methodology of generative approaches to syntax. Chapter 1 discusses the motivation behind the use of acceptability judgments as a primary source of syntactic data.

Cowart, W. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage.

This is the first (and currently only) textbook devoted solely to acceptability judgment collection. Topics include a discussion of the utility of formal judgment experiments, an introduction to experimental design and statistical analysis, and a tutorial in the use of Microsoft Excel for constructing experiments.

Marantz, A. 2005. Generative linguistics within the cognitive neuroscience of language. *The linguistic Review* 22.429-445.

This is a broad discussion of the role of linguistic theories in the field of cognitive neuroscience. Among other topics, it discusses the properties of traditional data collection methods, how they are a type of informal behavioral experiment, and how the non-standard collection and reporting customs in syntax may be obscure the relevance of syntactic theories to the cognitive neuroscience of language.

Myers, J. 2009b. Syntactic judgment experiments. *Language and Linguistics Compass* 3.406-423.

This article reviews many of the issues surrounding the collection of acceptability judgments, such as the reliability and sensitivity of both formal and traditional methods, and reviews many of the arguments in favor of formal experiments.

Newmeyer, F. 1983. *Grammatical Theory: Its limits and its possibilities.* Chicago: University of Chicago Press.

This book presents a comprehensive introduction to the goals and scope of generative linguistics, including the use of acceptability judgments as evidence for linguistic theories.

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

A comprehensive review of the acceptability judgment literature prior to 1996, with topics ranging from the evidential role of acceptability in theories of grammar, to concerns about the reliability of traditionally collected acceptability judgments.

TRADITIONAL JUDGMENT COLLECTION METHODS

As with any data type, the reliability of acceptability judgments is of central concern to the linguists that use them for theory construction. The informality with which acceptability judgments have traditionally been collected, and the lack of detailed descriptions of the collection methods in published articles, has led some to question the reliability of the traditionally collected data underlying syntactic theories. This section provides a list of references regarding those concerns and the current evidence bearing on them.

Concerns about the Reliability of Traditionally Collected Judgments

Hill 1961 is often cited as the first published article to raise concerns about the reliability of traditionally collected judgments. Schütze 1996 (see General Overviews) reviews many similar pieces from the earliest days of generative syntax, and discusses more recent examples of data-driven debates in the literature. Ferreira 2005, as part of a discussion of the role of linguistics in cognitive sciences, raises the concern that the informality with which acceptability judgments are collected (and reported) may lead researchers in allied areas of cognitive science to disregard syntactic theories. Featherston 2007 argues that relatively minor additions to the traditional method, specifically collecting judgments using multiple items and multiple participants, could substantially increase the reliability of acceptability judgment data, and provides several empirical examples. Gibson and Fedorenko 2010a and 2010b reiterate the concerns of Ferreira 2005 and Featherston 2007, and additionally criticize the use of professional linguists as participants in traditional judgment collection methods, suggesting that the theoretical biases of linguists could influence their judgments. Phillips 2009 provides a potential counterpoint to these articles. While agreeing in principle that more formal experimentation would be beneficial to the field, Phillips 2009 argues that the concerns may have been overstated, as many of the major phenomena do replicate under formal experimentation, and many of the major debates in syntactic theory are not over the status of data, but rather over the assumptions of specific theories. Sprouse and Almeida 2012b review the results of several empirical studies of the reliability of traditionally collected judgments in order to directly address the empirical questions raised by critics of traditional methods.

Featherston, S. 2007. Data in generative grammar: The stick and the carrot.

Theoretical Linguistics 33.269-318.
This article presents two types of arguments in favor of the adoption of more formal methods of judgment collection: the avoidance of potentially unreliable data, and the collection of potentially finer-grained data. Featherston presents several examples from his own research to empirically motivate the proposal.

Ferreira, F. 2005. Psycholinguistics, formal grammars, and cognitive science. *The linguistic Review* 22.365-380.
As the title suggests, this is a broad discussion of the role of linguistic theories in the field of psycholinguistics, as well as cognitive science generally. It suggests a number of ways that linguists could make formal grammars more useful to psycholinguists working on processing theories, including the adoption of formal experimental methods for judgment collection.

Gibson, E. and E. Fedorenko. 2010a. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14.233-234.
This letter and the related journal article (Gibson and Fedorenko 2010b) present several criticisms at traditional judgment collection methods, ultimately suggesting that traditional judgment collection methods may lead to unreliable data.

Gibson, E. and E. Fedorenko. 2010b. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*. 10.1080/01690965.2010.515080
This journal article and the related letter (Gibson and Fedorenko 2010a) present several criticisms at traditional judgment collection methods, ultimately suggesting that traditional judgment collection methods may lead to unreliable data.

Hill, A. A. 1961. Grammaticality. *Word* 17.1-10.
This article argues that acceptability judgments are affected by many factors, such as intonation and (lack of) context, that may interfere with their use as primary evidence for grammatical theories.

Phillips, C. 2009. Should we impeach armchair linguists? In *Proceedings from Japanese/Korean Linguistics* 17. S. Iwasaki, H Hoji, P. Clancy, and S.-O. Sohn, eds. Stanford, CA: CSLI Publications.
This article addresses many criticisms of traditional judgment collection. Particular arguments include the observation that most judgments replicate easily in formal experiments, that different data points carry different evidential weight in theory construction, and that debates in the field are rarely about the status of the data, but rather about the best theory for explaining the (generally agreed upon) data.

Sprouse, J. & D. Almeida. 2012b. The empirical status of data in syntax: A reply to Gibson and Fedorenko. *Language and Cognitive Processes*. 10.1080/01690965.2012.703782
This letter reviews several large empirical investigations of the reliability of traditionally collected judgments, and attempts to apply the results to the criticisms levied by Gibson and Fedorenko 2012b.

Evidence of the (Un)reliability of Traditionally Collected Judgments

The primary evidence for the unreliability of traditionally collected judgments comes from mismatches between the results of traditionally collected judgments and the results of formally collected judgments, or mismatches between two sets of traditionally collected judgments. Schütze 1996 (see General Overviews) presents several examples of controversial judgments published in the literature. Since then there have been at least four high profile examples of such mismatches. Wasow and Arnold 2005 presents formally collected judgments regarding the placement of objects in verb-particle constructions in English. Although the mean pattern of judgments matches those reported in the literature using traditional methods, they argue that the variability observed in the responses would not be accessible without formal experiments. Clifton et al. 2006 formally tests Kayne's (1983) claim that the addition of a third wh-word increases the acceptability of Superiority violations (the displacement of an object wh-word into a position that precedes a subject wh-word). Clifton et al. could not replicate Kayne's traditionally collected judgments, instead finding that the addition of a third wh-word has no impact on the mean judgments at all (which they admit may itself be unexpected given that the addition of extra wh-words usually results in a decrease in acceptability). Alexopoulou and Keller 2007 formally investigates the ability of resumptive pronouns in English to increase the acceptability of violations of syntactic island constraints involving wh-displacement, but finds no effect, contrary to the traditionally collected judgments of Ross (1967) and many others. This result is further corroborated by Heestand et al. 2011. Langendoen et al. 1973 formally investigated the claim by Fillmore (1965) that the first object of a double-object construction cannot be questioned in English. They found that this claim is true for about 80% of participants, but the remaining 20% could form such questions. Wasow and Arnold 2005 and Gibson and Fedorenko 2010a,b (see Concerns about the Reliability of Traditionally Collected Judgments) argue that this is further evidence for the unreliability of traditional methods. Gibson and Fedorenko 2010b also presents a novel experiment testing a claim from Chomsky 1986 that two Superiority violations may have distinct levels of acceptability. Gibson and Fedorenko failed to find significant differences using formal methods, suggesting that the traditionally collected judgments reported in Chomsky 1986 may be unreliable. Because single instances of mismatches between judgment methods do not provide any information about how common such mismatches are compared to matches between judgment methods, Sprouse and Almeida 2012a formally tests all 469 traditionally collected data points in a recent syntax textbook (Adger 2003) and finds that 98% of the data points replicated using formal experiments (using conservative statistical thresholds), suggesting that mismatches between the two methods may be relatively rare compared to matches.

Alexopoulou, T. & F. Keller. 2007. Locality, cyclicity, and resumption: At the interface between the grammar and the human sentence processor. *Language* 83.110-160.

This article reports a series of formal judgment experiments investigating the effect of resumptive pronouns on wh-dependencies in English, German, and Greek. The results suggest that resumptive pronouns do not ameliorate island effects in these languages, contrary to previous reports in the literature.

Clifton, Jr., C., G. Fanselow, & L. Frazier. 2006. Amnestying superiority violations: Processing multiple questions. *Linguistic Inquiry* 27.51-68.

This article reports a series of formal judgment experiments designed to investigate Kayne's (1983) claim that a third wh-word increases the acceptability of superiority violations. The results suggest that a third wh-word does not lead to higher acceptability. However, Clifton et al. do note that the addition of the third wh-word does not lead to lower acceptability either (as is generally the case when adding additional wh-words to a sentence).

Gibson, E. and E. Fedorenko. 2010b. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*. 10.1080/01690965.2010.515080

This article reports the results of a novel experiment comparing two Superiority violations from Chomsky 1986 that were reported to have different levels of acceptability. However, Gibson and Fedorenko fail to find significant differences between the two violations using formal methods.

Heestand, D., M. Xiang, & M. Polinsky. 2011. Resumption still does not rescue islands. *Linguistic Inquiry* 42.138-152.

This article reports a series of three experiments designed to further investigate the rescuing effect of resumptive pronouns in English islands by testing relative clauses in speeded grammaticality tasks. The results corroborate those of Alexopoulou and Keller 2007 in that no rescuing effect of resumptive pronouns was detected.

Langendoen, T. D., N. Kalish, & J. Dore. 1973. Dative questions: A study of the relation of acceptability to grammaticality of an English sentence type. *Cognition* 2.451-478.

This article investigates the claim that the first object of double object constructions cannot be questioned. The results suggest two groups of speakers, the 80% who cannot form such questions, and the 20% who can.

Sprouse, J. & D. Almeida. 2012a. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*. doi: 10.1017/S0022226712000011

This article attempts to estimate the reliability of traditional judgment collection methods by testing all 469 data points contained in a recent syntax textbook (Adger 2003). The results suggest that 98% of the data points in the textbook replicate using formal methods and relatively conservative statistical thresholds.

Wasow, T. & J. Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115.1481-1496.

This article presents a list of criticisms of traditional judgment collection, as well a formal experiment designed to test judgment patterns reported in *The logical structure of linguistic theory* (Chomsky 1955).

Linguist Participants versus Non-linguist Participants

One common concern about the reliability of traditional judgment collection methods concerns the reliance upon linguists as participants (as compared to the common use of non-linguists as participants in formal judgment collection studies). Ferreira 2005, Wasow and Arnold 2005, and Gibson and Fedorenko 2010a, 2010b (see Concerns about the reliability of traditionally collected judgments) all observe that professional linguist participants are more likely than non-linguist participants to recognize the theoretical relevance of experimental manipulations, and therefore their judgments could potentially be influenced by theoretical biases; however, they do not present any empirical group comparisons to asses this possibility. The earliest comparisons of judgments between linguists and non-linguists are Greenbaum 1973 and Spencer 1973. Greenbaum 1973 replicates, using non-linguist participants, a study Elliot, Legum, and Thompson (1969) that primarily relied upon linguist participants. The results do suggest some differences between the judgments of the two groups. Spencer 1973 re-tested 150 sentence types reported in 7 linguistics articles from the late 1960s using two non-linguist groups: undergraduate students

with no linguistics training and graduate students with exposure to some linguistics. The results suggest some differences in judgments between both groups of non-linguists and the traditionally collected judgments reported in the target articles (see Schütze 1996 in General Overviews for detailed discussions of these papers and several others from the 1970s and 1980s). More recently, Dąbrowska 2010 formally collected judgments on a series of long-distance wh-questions from two groups: 38 professional linguists and 38 undergraduate students. The results suggest that professional linguists tend to provide more categorical judgments (using the endpoints of the scale, with very few judgments in between), while non-linguists tend to use the entire range of possible ratings. Dąbrowska 2010 also reports a potential difference between types of professional linguists: self-identified generative linguists tend to rate complex NP constraint violations higher than self-identified functional linguists. Culbertson and Gross 2009 compared judgments for four groups: PhD linguists, undergraduate students with some exposure to syntax, undergraduate students with no exposure to syntax but exposure to cognitive science, and undergraduate students with no exposure to cognitive science at all. They found that the crucial factor in judgment reliability appears to be experience with cognitive science in general: participants within the three groups with cognitive science experience were more correlated with the other members of their respective groups, and more correlated across the three groups, than member of the inexperienced group. (This is part of a broader discussion of the nature of judgments in linguistics with Michael Devitt. For the full exchange see Devitt 2006, 2010, and Gross and Culbertson 2011). Finally, Sprouse and Almeida 2012a compares the traditionally collected judgments reported in a recent syntax textbook (Adger 2003) to the formally collected judgments of non-linguist participants. They find at most a 2% divergence between the two sets of judgments, suggesting relatively little differences between groups for the phenomena in the textbook.

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

This article discusses a number of issues surrounding the use of linguists as a source of data in linguistic theory. The included experimental study compares the judgments of three groups: linguists, non-linguists with previous experience as participants in psychology experiments, and non-linguists with no previous experiments in psychology experiments.

Dąbrowska, E. 2010. Naïve v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27.1-23.

This article presents a systematic comparison of the judgments of linguists and non-linguists for a range of wh-constructions of interest to theories of both syntax and sentence processing, including Complex NP island effects.

Devitt, M. 2006. Intuitions in Linguistics. *British Journal for the Philosophy of Science* 57.481-513.

This is primarily a discussion of Devitt's views of the evidential role of acceptability judgments in linguistics. One of Devitt's proposals is that all speakers perform a type of grammatical analysis when giving acceptability judgments, therefore speakers with more sophisticated grammatical analyses (such as linguists) will provide more reliable judgments (note that while other linguists have made similar claims about the reliability of linguists' judgments, Devitt's reason is a novel one). There is no experimental data, but this serves as the starting point for Culbertson and Gross 2009, which is an attempt to test one of the empirical predictions of Devitt's proposal.

Devitt, M. 2010. Linguistic intuitions revisited. *British Journal for the Philosophy of Science* 61.833-865.
This is Devitt's response to Culbertson and Gross 2009, as well as Fitzgerald 2010 (also in volume 61 of BJPS).

Greenbaum, S. 1973. Informant elicitation of data on syntactic variation. *Lingua* 31.201-212.
This article reports a replication of a previous study (Elliot, Legum, and Thompson 1969) that was designed to investigate a potential implicational hierarchy within the grammars of individuals. In contrast with the original study, this replication relied on non-linguist participants, and found potential differences in the judgments of the two groups.

Gross, S. & J. Culbertson. 2011. Revisited linguistic intuitions. *British Journal for the Philosophy of Science* 62.639-656.
This is a response to Devitt 2010, clarifying the claims from Culbertson and Gross 2009.

Spencer, N. J. 1973. Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research* 2.83-98.
This article compares 150 traditionally collected judgments contained in 7 prominent linguistics articles of the late 1960s to the formally collected judgments of non-linguist participants. Although all of the analyses are descriptive (rather than inferential), the results suggest potential differences between the published judgments and the non-linguists' judgments.

Sprouse, J. & D. Almeida. 2012a. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*. doi: 10.1017/S0022226712000011
This article compares all of the traditionally collected judgments reported in a recent syntax textbook (Adger 2003) to formally collected judgments from non-linguist participants. The results suggest that at least 98% of the syntactic phenomena in Adger 2003 replicate in formal experiments with non-linguist participants.

FORMAL JUDGMENT COLLECTION METHODS

Although traditional judgment collection methods are (currently) the dominant method in the syntax literature, many syntacticians agree that the rising popularity of formal judgment collection methods is a positive development for the field. This section provides a list of references regarding the actual methodology behind formal judgment collection: experimental design, data analysis, software tools, and the potential limitations of formal experiments that experimenters should consider when designing or interpreting a study.

Designing and Deploying Formal Judgment Experiments

Cowart 1997 is the only textbook devoted solely to acceptability judgment experiments. However, it should be noted that many of the methodological issues that are relevant to acceptability judgments are also relevant to sentence processing experiments, so psycholinguistic methodology textbooks may also be useful. Cowart 2012 presents an updated tutorial in the use of Microsoft Excel for constructing judgment experiments. Myers 2009a provides a comprehensive tutorial in the design, deployment, and analysis of small scale judgment experiments using Myers' MiniJudge software (see also Software for Deploying and Analyzing Judgment Experiments). Nagata 1992 and Cowart 1994 explore the effect that the overall composition of a judgment experiment can have on the judgments of individual items. Sprouse

2011a compares data collected from a large sample of participants in a laboratory to a large sample of participants collected using Amazon's Mechanical Turk online marketplace on a number of dimensions that may be relevant to syntactic studies, discusses several questions syntacticians might have about using the marketplace, and provides links to templates and analysis scripts that may be useful. Gibson et al. 2011 provides an additional tutorial and set of materials for the use of Amazon Mechanical Turk.

Cowart, W. 1994. Anchoring and grammar effects in judgments of sentence acceptability. *Perceptual and Motor Skills* 79.1171-1182.

This follow-up to Nagata 1992 presents evidence from a formal judgment experiment that suggests that while anchoring effects do alter the absolute ratings of sentences, the relative acceptability between sentence types remains unchanged. This suggests that relative acceptability patterns are relatively robust compared to absolute ratings.

Cowart, W. 1997*. Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage.

This is the first (and currently only) textbook devoted solely to acceptability judgment collection. Topics include a discussion of the utility of formal judgment experiments, an introduction to experimental design and statistical analysis, and a tutorial in the use of office software such as Excel for constructing experiments.

Cowart, W. 2012. Doing Experimental Syntax: Bridging the gap between syntactic questions and well-designed questionnaires. In *In Search of Grammar: Experimental and Corpus-based Studies*. J. Myers (ed).67-96.

This chapter provides an updated tutorial in the use of Excel for the construction of experimental stimuli. Topics include the creation of matching sentence sets, distribution of sentence sets into lists using a Latin Square procedure, and randomization of items within lists.

Gibson, E., S. Piantadosi, and K. Fedorenko. 2011. Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass* 5.509-524.

This article is both a criticism of traditional judgment collection (along the lines of Gibson and Fedorenko *in press*) and a tutorial in the use of Amazon Mechanical Turk for the collection of judgments.

Myers, J. 2009a. The design and analysis of small-scale syntactic judgment experiments. *Lingua* 119.425-444.

Myers proposes a type of formal experiment designed to provide valid statistical inference with minimum effort. He calls these experiments *small-scale*, as compared to the experiments proposed by others, they consist of relatively few tokens of each condition, relatively few participants, and a relatively intuitive task (e.g., a categorical yes-no task). This article presents evidence of the reliability of these experiments, as well as a tutorial in the use of tools that Myers developed for the construction, deployment, and analysis of small-scale experiments.

Nagata, H. 1992. Anchoring effects in judging grammaticality of sentences. *Perceptual and Motor Skills* 75.159-164.

This article reports a series of three formal experiments designed to investigate the effect of anchoring on acceptability (i.e., the impact of the rating of surrounding sentences on a target sentence). The results suggest that very unacceptable anchors can increase the absolute ratings of target sentences, whereas very acceptable anchors can decrease the absolute ratings of target sentences.

Sprouse, J.. 2011a. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43.155-167.

This article presents a direct comparison of judgments collected in the laboratory and on Amazon Mechanical Turk along several quantitative dimensions likely to be of interest to syntacticians (statistical power, data quality, etc). Several issues related to the use of Mechanical Turk are discussed, and links are provided to templates and R scripts designed to facilitate future use of Mechanical Turk.

The Statistical Analysis of Judgment Data

As a behavioral response, acceptability judgments are generally analyzed according to the best practices of inferential statistics for social and behavioral sciences. However, there are also several statistics textbooks that have been written explicitly for linguistics. Baayen 2007 provides a general introduction to statistics and the R statistical computing language with a particular focus on lexical decision experiments. Johnson 2008 provides an introduction to statistics and the R language by focusing on characteristic data sets from several different subfields of linguistics such as phonetics, sociolinguistics, psycholinguistics, and syntax. Gries 2010 also provides an introduction to statistics and the R language, with a particular focus on the nature of hypothesis testing in linguistics. In addition to basic statistical tests, syntacticians should also be aware of the so-called “language-as-fixed-effects-fallacy”, which in terms of acceptability judgments is a debate about whether the sentences in an experiment should be treated as a random effect (meaning that the sentences were chosen randomly from a larger population of sentences) or as a fixed effect (meaning that the sentences represent all of the possible sentences that could be tested). The difficulty is that sentences are most likely somewhere in between the two, but there is no way to tell exactly where, and no way to capture the in-between status mathematically. Clarke 1973 argues in favor of treating sentences as random effects, because this is the more conservative option. Wike and Church 1976 argues in favor of treating sentences as fixed effects, because this is the more statistically powerful option. There are several other responses in the same volume as Wike and Church 1976 that further explore the risks and rewards of the two approaches. Raaijmakers et al. 1999 and Raaijmakers 2003 re-evaluates this debate, and argues that there are ways to control for item variability in the experimental design that would allow experimenters to take advantage of the increased statistical power of the fixed-effects approach. Baayen et al. 2008 discusses linear mixed effects models, which allow both fixed and random effects to be specified simultaneously. While this doesn’t solve the debate about fixed versus random effects, for those who believe in specifying random effects, linear mixed effects models provide a relatively straightforward method for performing the analysis. Bates and Maechler’s (2012) LME4 is a freely available package for R that allows for the easy construction of linear mixed effects models. Baayen’s (2007) LANGUAGER is a freely available package for R that contains a useful function for calculating *p*-values from linear mixed effects models.

Baayen, R. H. 2007. *Analyzing Linguistic Data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

This textbook provides a comprehensive introduction to both statistics and the R statistical computing language that is tailored to a linguistics audience (with a particular focus on lexical decision experiments). This book also introduces the *languageR* package, which provides a useful function for the calculation of *p*-values from linear mixed-effects models. The *languageR* package is freely available from the online repository of R packages (CRAN).

Baayen, R. H., D. J. Davidson & D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59.390–412.

This article provides an introduction and tutorial in the use of linear mixed-effects models for experimental designs that are highly relevant to linguistics (i.e., repeated measures experiments with both subject and item random effects).

Bates, D. M. & M. Maechler. 2012. Lme4: Linear mixed-effects models using S4 classes. http://cran.r-project.org/web/packages/lme4/.

This is a freely available package for R that allows one to easily construct linear mixed-effects models using R's standard formula notation.

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.

This paper raises concerns about the treatment of items as fixed effects in the statistical practices of linguists and psycholinguists. Clark suggests a method for treating both subjects and items as random effects by calculating both by-subject and by-item *F* statistics, and using these to calculate the *min F′* statistic.

Gries, S. Th. 2010. *Statistics for linguistics with R: A practical introduction*. Mouton de Gruyter.

This textbook provides a comprehensive introduction to both statistics and the R statistical computing language that is tailored to a linguistics audience, with a particular focus on the nature of hypothesis testing within linguistics.

Johnson, K. 2008. *Quantitative methods in linguistics*. Wiley-Blackwell.

This textbook provides a comprehensive introduction to both statistics and the R statistical computing language that is tailored to a linguistics audience, with chapters that analyze characteristic data sets from several subfields of linguistics, such as phonetics, sociolinguistics, psycholinguistics, and syntax.

Raaijmakers, J. G., J. M. C. Schrijnemakers, & F. Gremmen. 1999. How to deal with the "Language-as-Fixed-Effect Fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language* 41.416-426.

This article discusses the impact of Clark's (1973) suggestions on the actual statistical practice of linguists and psycholinguists. Raaijmakers et al. criticize the practice of reporting both and only F_1 and F_2 and the associated assumption that a significant F_2 value indicates generalizability across subjects, as neither suggestion is actually contained in Clark 1973 (where F_1 and F_2 are intermediate steps toward the calculation of min F'). They also discuss various experimental designs that explicitly control for item variability, making Clark's suggestion moot (i.e., F_1 is the correct statistic for these designs). Despite these criticisms of the actual use of Clark's suggestions, ultimately Raaijmakers et al. endorse his criticisms.

Raaijmakers, J. G. 2003. A further look at the "Language-as-Fixed Fallacy". *Canadian Journal of Experimental Psychology* 57.141-151.

This article further revisits some of the issues regarding fixed and random effects that were originally discussed in Raaijmakers et al. 1999 (such as the interpretation of F_1 and F_2 as tests of the generality to subjects and items respectively).

Wike, E. L. & J. D. Church. 1976. Comments on Clark's "The Language-as-Fixed-Effect Fallacy". *Journal of Verbal Learning and Verbal Behavior* 15.249-255.

This is a direct response to Clark's (1973) paper. Wike and Church argue that Clark's proposal is mistaken on several levels, including (i) the definition of random effects, (ii) the properties of quasi-*F*-ratios, and (iii) the best method for assessing the generalizability of experimental results.

Software for Deploying and Analyzing Formal Judgment Experiments

There are several pieces of freely available software that can assist in the deployment and analysis of formal acceptability judgment experiments. WebExp2 is a Java-based program designed to facilitate the collection of behavioral responses over the internet, including acceptability judgments and reaction times that was developed by Neil Mayo, Martin Corley, and Frank Keller (www.webexp.info). MiniJudge is a Java-based or JavaScript-based program designed to facilitate the design, deployment, and analysis of small scale judgment experiments (www.ccunix.ccu.edu.tw/~lngproc/MiniJudge.htm). It was developed by James Myers to fill the gap between traditional judgment collection and the full-scale formal judgment collection of WebExp2. Amazon Mechanical Turk is online marketplace where syntacticians can post judgment experiments for users to complete for payment (www.mturk.com). Experiments can be posted directly on Mechanical Turk as a webform (using basic HTML), or links can be provided to server-based programs like WebExp2 and MiniJudge. R is a software environment (programming language) for statistical computing (www.r-project.org/). It is free and open-source, and there is a large user community that has developed an enormous number of free packages to perform any number of useful functions, from cutting-edge statistical analyses, to print-quality graphics, and even text manipulation (for designing experiments or performing corpus analyses).

WebExp2: www.webexp.info

Free Java-based software for the collection of behavioral responses, such as acceptability judgments and reaction times, over the internet

MiniJudge: www.ccunix.ccu.edu.tw/~lngproc/MiniJudge.htm

Free Java-based or JavaScript based software for the design, deployment, and analysis of small scale experiments.

Amazon Mechanical Turk: www.mturk.com

An online marketplace that can be used for the rapid collection of acceptability judgments.

R: www.r-project.org

This is a free and open-source statistical computing language that is widely used in many sciences for statistical analysis, graphing, and text manipulation.

Potential Limitations of Formal Judgment Collection Methods

Nearly every syntactician believes that the increased use of formal experimental methods for the collection of acceptability judgments is a positive development. However, it is also important to remember that formal experiment methods are a tool like any other, and therefore have both benefits and limitations. For example, as part of a response to Featherston 2007 (see General Overviews), Fanselow 2007 observes that formal experiments can only report values; experiments cannot tell a researcher how to interpret variation between speakers, how to interpret variation between constructions within a single speaker, or whether an acceptability effect is driven by grammatical constraints or properties of sentence processing; only logical arguments, often predicated upon potentially controversial assumptions, can do that. Fanselow and Féry 2008 provides a concrete example of this problem by offering a reanalysis of the alleged Superiority effect observed in Featherston 2005a and 2005b. Grewendorf 2007 and Haider 2007, both also responses to Featherston 2007, argue that discrepancies between traditionally collected and formally collected judgments should not be automatically resolved in favor of formally collected judgments, as the results of formal experiments are susceptible to certain types of confounds, such as careless materials creation, unmotivated participants, individual differences between participants, and other performance factors. Den Dikken et al. 2007 observes that the common practice of averaging judgments from multiple participants in formal experiments could potentially obscure variation across individuals, which is a potentially rich source of information for linguistic theories. Culicover and Jackendoff 2010 makes similar arguments about the need for care in experimentation, and the potential to lose data when averaging across multiple participants, as part of a response to Gibson and Fedorenko 2010a. Finally, Featherston 2009 provides advice about when formal experiments may be useful, when they may not be useful, which factors appear to affect acceptability judgments (and therefore should be controlled), and which factors appear not to affect acceptability judgments (and therefore may be less important to control if doing so would present an undue hardship).

Culicover, P. W. & R. Jackendoff. 2010. Quantitative methods alone are not enough: Response to Gibson & Fedorenko. *Trends in Cognitive Sciences* 14.234–235.

This reply to Gibson and Fedorenko (2010a) argues that while formal experimental methods would be a welcome addition to the field of syntax, formal experiments by no means ensure reliable (and relevant) data collection. They argue that experimental results are only as meaningful as the controls incorporated into the materials, and the interpretations licensed by the results.

den Dikken, M., J. Bernstein, C. Tortora & R. Zanuttini. 2007. Data and grammar: means and individuals. *Theoretical Linguistics* 33.335–352.

This reply to Featherston 2007 touches upon several topics, with a particular focus on the standard practice in formal experiments to average results across a group of speakers. The authors argue that uniformly adopting such a procedure could obscure interesting patterns of grammatical variation across individuals – a rich source of information for linguistic theories.

Fanselow, G. 2007. Carrots – perfect as vegetables, but please not as a main dish. *Theoretical Linguistics* 33.353-367.

This is a comprehensive response to the two primary claims in Featherston 2007: that formal experiments will help resolve complicated issues in grammatical theory and that grammars are likely gradient rather than categorical. Fanselow argues that relatively few theoretical debates hinge on controversial data, and when they do, formal experiments have done little to resolve

them. He suggests that variation between speakers, variation between constructions within a single speaker, and extra-grammatical factors such as processing difficulty may be the cause of these controversial cases, as well as the gradience observed in acceptability judgments.

Fanselow, G. & C. Féry. 2008. Missing Superiority effects: Long movement in German (and other languages). In *Elements of Slavic and Germanic grammars: A comparative view.* J. Witkos & G. Fanselow (eds). Frankfurt: Lang, 67-87.

This article presents a series of acceptability judgment experiments that suggest that the decrease in acceptability that Featherston (2005a, 2005b) describes as a Superiority effect is instead a general (and small) decrease associated with long-distance, crossing dependencies in German. The authors consequently argue that there is no Superiority effect in German.

Featherston, S. 2009. Relax, lean back, and be a linguist. *Zeitschrift für Sprachwissenshaft* 28.127-132.

This chapter provides advice on the use and construction of formal judgment experiments, with a particular focus on the utility of formal experiments within the syntactic enterprise, and the factors that do and do not affect acceptability judgments.

Grewendorf, G. 2007. Empirical evidence and theoretical reasoning in generative grammar. *Theoretical Linguistics* 33.369-381.

This is another comprehensive response to the claims in Featherston 2007. Grewendorf argues that discrepancies between the results of formal experiments and traditional methods should not be automatically resolved in favor of formal experiments, as many aspects of the experiments, from materials construction to individual variation (and averaging across individuals), could lead to the discrepancies. Grewendorf also discusses the role of experimental evidence in linguistic reasoning.

Haider, H. 2007. As a matter of facts – comments on Featherston's sticks and carrots. *Theoretical Linguistics* 33.381–395.

In this response to Featherston 2007, Haider generally agrees with Featherston that traditional data collection methods in syntax could use more care; however, like the other responses, he disagrees with the claim that formal experiments will resolve controversial cases, as the results of formal experiments could still be contaminated by extra-grammatical factors or individual variation. Haider also disagrees with the claim that grammar is gradient.

SPECIFIC TOPICS IN ACCEPTABILITY JUDGMENTS

Beyond comparisons of the two judgment collection methods, the rising interest in the investigation of acceptability judgments has opened up several lines of research into the nature of acceptability and its role as evidence for grammatical theories. This section provides a list of references regarding these lines of research: the sensitivity of magnitude estimation, the role of processing factors in acceptability judgments, the role of satiation as evidence, and the relationship between gradient acceptability and grammatical theories.

Magnitude Estimation

Although the magnitude estimation task has been in use in psychophysics since at least the 1940s, and various social sciences since the 1970s, it was first introduced to the field of syntax by

Bard et al. in 1996 and Cowart 1997. The magnitude estimation task asks participants to judge the magnitude of a stimulus, in this case the acceptability of a sentences, as a multiple of the magnitude of a reference stimulus (also known as a standard) – e.g., ½ as acceptable, twice as acceptable, etc. Bard et al. 1996 demonstrates that magnitude estimation allows participants to distinguish more levels of acceptability than ordinal response tasks like the 7-point scale task, and that participants' ratings using the task were reliably across response modalities, suggesting an internal consistency in judgments. Cowart 1997 demonstrates that magnitude estimation could be used to uncover previously unreported differences in acceptability (as part of an investigation of the *that*-trace effect), and illustrates how fill-in-the-bubble response sheets could be used to semi-automate the analysis of magnitude estimation responses. Keller 2000 leverages the finer-grained ratings made possible by magnitude estimation to uncover previously unreported differences in acceptability for a number of phenomena, ultimately leading to the construction of a gradient version of an Optimality Theoretic grammar. In a similar vein, Featherston 2005a and 2005b demonstrate the potential sensitivity of magnitude estimation by uncovering a previously unreported Superiority effect in German that holds among sentences that are all reported to be categorically acceptable. Bader and Haüssler 2010 compares magnitude estimation and yes-no ratings of 16 sentence types in German in order to investigate the claim that magnitude estimation provides more sensitive ratings than categorical rating tasks. They find that both tasks yield similar results, both descriptively and inferentially, for the 16 sentence types, suggesting that the claims regarding the sensitivity of magnitude estimation may have been overstated. Weskott and Fanselow 2011 compares magnitude estimation to both the yes-no task and the 7-point scale task in a similar investigation of 7 additional sentence types in German, and also finds that the three tasks yielded similar results, again suggesting that the claims regarding the sensitivity of magnitude estimation may have been overstated. Sprouse 2011b investigates the underlying cognitive assumptions of magnitude estimation in an attempt to determine whether it is even logically possible for magnitude estimation to yield qualitatively different data than other tasks, and found that one of the fundamental cognitive assumptions (that participants can make ratio judgments of acceptability) is not met, suggesting that participants likely treat the magnitude estimation task like other rating tasks (e.g., the 7-point scale task). This provides a possible explanation for the lack of increased sensitivity. Featherston 2008 proposes a new task, the thermometer task, which is more in line with the cognitive abilities of participants, yet also takes advantage of a continuous response scale to potentially provide finer-grained responses than ordinal response scales (like the 7-point scale).

Bader, M. & J. Haüssler. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46, 273–330.

This paper presents a comparison of magnitude estimation and yes-no judgments both with time pressure (speeded judgments) and without (offline judgments). The results suggest a strong correlation between the two measures, as well as the potential to relate the two types of judgments using Signal Detection Theory.

Bard, E. G., D. Robertson & A. Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72.32-68.

This is the first published article to propose the use of magnitude estimation for acceptability judgment collection. Bard et al. provide a series of case studies to demonstrate the potential increased sensitivity of ME, as well as a cross-modal matching study to demonstrate its internal validity.

Cowart, W. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage.

This is the first published book to propose the use of magnitude estimation for acceptability judgment collection. As a textbook, it provides detailed instructions about how to design, deploy, and analyze formal acceptability judgment experiments.

Keller, F. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. PhD. dissertation, University of Edinburgh.

This dissertation is the first large-scale set of studies to use magnitude estimation in syntax. Keller presents a series of experiments demonstrating the sensitivity of magnitude estimation (for both standard acceptability judgments and co-reference judgments), as well as the continuous nature of acceptability. He also develops a type of gradient grammar based on an extension of Optimality Theory called Linear Optimality Theory to capture the continuous acceptability effects.

Featherston, S. 2005a. Magnitude estimation and what it can do for your syntax: some wh-constraints in German. *Lingua* 115.1525-1550.

This paper (and its sister below (2005b)) explores the use of magnitude estimation for acceptability judgment collection through a case study in Superiority effects in German. The results reveal a complex pattern of relative acceptability differences that are nearly identical to the Superiority effect observed in English, except that all of the sentences in German are traditionally reported to be categorically acceptable by native speakers.

Featherston, S. 2005b. Universals and grammaticality: wh-constraints in German and English. *Linguistics* 43.667-711.

This paper builds off of the results of Featherston 2005a to develop a theory of how constraints such as Superiority can manifest in one language as a complex pattern of relative acceptability that crosses a categorical boundary between acceptable and unacceptable sentences (e.g., English), and can manifest in another language as a complex pattern of relative acceptability among sentences that are all categorically acceptable (e.g., German).

Featherston, S. 2008. Thermometer judgments as linguistic evidence. In *Was ist linguistische Evidenz?* C. Riehl & A. Rothe (eds). Shaker Verlag.

This paper proposes a new task that combines the continuous scale of magnitude estimation with the interval steps inherent in the 7-point scale task. A typical thermometer experiment involves presenting participants with two sentences of differing acceptability, each assigned a different number (e.g., 20 and 30). Participants are then asked to rate target sentences relative to these numbers, such that a sentence that is more acceptable than the 30-sentence by the same amount that the 30-sentence is better than the 20-sentence would be rated 40.

Sprouse, J. 2011b. A test of the cognitive assumptions of Magnitude Estimation: Commutativity does not hold for acceptability judgments. *Language* 87.274–288.

The results of this study suggest that one of the foundational cognitive assumptions of magnitude estimation, that participants can make ratio judgments of acceptability, does not hold. Given the reliability of acceptability judgments reported during magnitude estimation tasks, this suggests that participants are instead treating the magnitude estimation task as another simple scaling task, such as the 7-point scale task.

Weskott, T. & G. Fanselow. 2011. On the Informativity of Different Measures of Linguistic Acceptability. *Language* 87, 249–273.

This paper presents a direct comparison of magnitude estimation, 7-point, and yes-no tasks, ultimately finding a strong correlation between the three tasks, with no evidence of increased information in magnitude estimation experiments. The results also corroborate the intuition that continuous response scales (e.g., magnitude estimation) can introduce more variance than scales with fewer response options (e.g., 7-point scales).

The Influence of Processing Factors on Acceptability

Because acceptability judgments can only be made after (at least partial) sentence processing has been completed, several non-syntactic factors can potentially influence acceptability judgments, from the plausibility of the meaning of the sentence to the ease or difficulty with which the sentence was processed. The role of processing difficulty has been of particular interest to linguists and psycholinguists, as the status of acceptability judgments as the output of (attempted) sentence processing entails that acceptability effects could be driven, in whole or in part, by properties of the sentence processing system rather than properties of the grammatical system. The seminal paper for this idea is Miller and Chomsky 1963, which proposes that the unacceptability of doubly center-embedded sentences is not due to a syntactic constraint on center embedding, but rather a constraint on the operation of the sentence processing system. Bever 1970 is presents a comprehensive discussion of two ways in which sentence processing mechanisms can influence acceptability judgments: direct influence on the judgments, or indirect influence by way of shaping the grammar. Gerken and Bever 1986 provides a classic example of the potential obscuring effect of sentence processing mechanisms based on judgments of pronoun coreference. Fanselow and Frisch 2006 demonstrates that local ambiguity can both increase and decrease the acceptability of globally acceptable sentences. Alexopoulou and Keller 2007 proposes that decreases in acceptability for long-distance wh-movement out of embedded clauses (in both non-island and weak island contexts) can best be understood as reflecting constraints on working memory resources. Sprouse 2008 suggests that some temporary processing difficulties (such as temporary syntactic ungrammaticality) affect acceptability judgments, but not others (such as temporary semantic implausibility). Hofmeister et al. 2011 proposes that Superiority violations may be the result of processing difficulty rather than grammatical constraints. Sprouse et al. 2011 suggests that the pattern of acceptability observed with syntactic island effects in wh-in-situ questions in English may be due to a parsing operation that searches backward through previously parsed material for a licensor for the in-situ wh-word. And Sprouse et al. 2012 investigates the role of individual working memory capacity on the acceptability of syntactic island effects in English wh-questions, ultimately finding no correlation between the two (but see the response by Hofmeister et al. 2012 for a different interpretation of the results).

Alexopoulou, T. & F. Keller. 2007. Locality, cyclicity, and resumption: At the interface between the grammar and the human sentence processor. *Language* 83.110-160.

As part of an investigation of wh-movement and resumption cross-linguistically, this article reports decreases in the acceptability of wh-movement out of multiple embedded clauses that may suggest an influence of constraints on working memory resources during sentence processing.

Bever, T. G. 1970. The influence of speech performance on linguistic structure. In *Advances in Psycholinguistics*. G. B. Flores d'Arcais & J. M. Levelt (eds). Amsterdam: North Holland Publishing Company. 65-88.

This classic chapter discusses various ways in which processing mechanisms may influence acceptability judgments independently of the grammatical status of the sentences in question. This chapter also discusses ways in which processing mechanisms may influence the shape of the grammar itself, thus indirectly affecting acceptability. Bever argues that both issues must be explored in order to arrive at a fuller understanding of the architecture of the language faculty.

Fanselow, G. & S. Frisch. 2006. Effects of processing difficulty on judgments of acceptability. In *Gradience in Grammar*. G. Fanselow, C. Féry, M. Schlesewsky & R. Vogel (eds). Oxford University Press, 291-316.

This chapter presents a series of acceptability judgment experiments that demonstrate the effect of processing difficulty on acceptability through local ambiguity. The results suggest that local ambiguity can both increase and decrease the global acceptability of the sentence depending on the grammatical status of the locally preferred reading, even when that reading is later abandoned.

Gerken, L. & T. G. Bever. 1986. Linguistic intuitions are the result of interactions between perceptual processes and linguistic universals. *Cognitive Science* 10.457-476.

This article reports a specific example of processing mechanisms influencing acceptability judgments. The authors argue that variability in acceptability judgments of pronoun co-reference can be explained by assuming that the grammatical constraint is present for all speakers, but that the parsing process that is employed for the sentence differs across speakers. This suggests an interesting interaction between the grammatical principles and processing mechanisms in determining sentence acceptability.

Hofmeister, P., T. F. Jaeger, I. Arnon, I.Sag, and N. Snider. 2011. The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments. *Language and Cognitive Processes*. 10.1080/01690965.2011.572401

This article uses a series of experiments to argue that Superiority effects in English may arise due to processing difficulty rather than a grammatical constraint.

Hofmeister, P., L. Staum Casasanto, & I. Sag. 2012. How do individual cognitive differences relate to acceptability judgments? A reply to Sprouse, Wagers, and Phillips. *Language* 88.390-400.

This letter is response to Sprouse et al. 2012. Hofmeister et al. argue that the methods and conclusions of the target article were problematic, suggesting that constraints on working memory may still be the cause of the unacceptability that characterizes syntactic island effects.

Miller, G. & N. Chomsky (1963). Finitary models of language users. In Luce, R.; Bush, R. and Galanter, E. (eds.) *Handbook of Mathematical Psychology, Vol 2.* New York: Wiley. 419-93.

This classic paper investigates the adequacy of Markov models for language acquisition. It is also the first analysis to suggest that the unacceptability of doubly center-embedded sentences is due to constraints on the sentence processing system rather than grammatical constraints.

Sprouse, J. 2008. The differential sensitivity of acceptability to processing effects. *Linguistic Inquiry* 39.686-694.

This article investigates the role of processing difficulty on the global acceptability of otherwise grammatical sentences. The results suggest that temporary syntactic ungrammaticality decreases acceptability, but temporary semantic implausibility and simple syntactic reanalysis do not.

Sprouse, J., S. Fukuda, H. Ono, and R. Kluender. 2011. Reverse island effects and the backward search for a licensor in multiple wh-questions. *Syntax* 14.179-203.

This article investigates the acceptability of multiple wh-questions that span syntactic islands in English, as well as single wh-in-situ questions in Japanese. The results of a series of formal acceptability judgment experiments suggest that there may be a backward search for a wh-in-situ licensor in English.

Sprouse, J., M. Wagers & C. Phillips. 2012. A test of the relation between working memory capacity and island effects. *Language* 88.82-123.

This article compares individual differences in working memory span to individual differences in acceptability judgments in an effort to evaluate to what extent the unacceptability that characterizes syntactic island effects can be reduced to limitations in working memory capacity.

Satiation

Repeated exposure to identical sentences or identical syntactic structures may affect the ratings that participants assign to those sentences. This phenomenon is often known as *satiation* in the syntax literature, although it is also called syntactic priming in the sentence processing literature (where the effects can be measured in reaction times as well). The primary goal of the satiation literature is to determine under what circumstances satiation occurs (either as an increase or decrease in acceptability), and to determine to what extent satiation might provide information about syntactic representations, the processing of syntactic structures, or the process of making acceptability judgments. Nagata 1988 and 1989 are two of the first systematic investigations of satiation in acceptability judgments, although the concept of satiation as a potential confound in experiments is mentioned as early as Greenbaum 1973 (see Linguist Participants vs. Non-linguist Participants). Snyder 2000 investigates satiation effects for a series of potential constraints on wh-dependencies (including syntactic island and *that*-trace effects), suggesting that satiation may provide evidence regarding the source of the constraint (i.e., grammatical constraints will not satiate, but constraints on sentence processing may). Hiramatsu 2000, as part of a broader investigation of children's acquisition of constraints on wh-dependencies, partially replicates Snyder's (2000) results, and expands the set of constraints investigated. Luka and Barsalou 2005 presents a series of experiments that suggest that several moderately unacceptable constructions may satiate. Sprouse 2009 presents a series of direct and related replications (using magnitude estimation) of Snyder 2000 and Hiramatsu 2000, but reports no reliable satiation effects for any of the syntactic island constraints. Francom 2009 presents a series of acceptability judgment and reading time experiments designed to systematically investigate the disparate results between Snyder 2000, Hiramatsu 2000, and Sprouse 2009. The results suggest that the interpretability of the sentences may contribute to their susceptibility to satiation.

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.

This dissertation presents three acceptability experiments and two reading time experiments designed to isolate the properties of sentences and/or experiments that give rise to repetition (satiation) effects. The results suggest that the interpretability of sentences may contribute to their susceptibility to satiation, and that failure to control interpretability may explain the divergent results of previous studies.

Hiramatsu, K. 2000. Accessing linguistic competence: Evidence from children's and adults' acceptability judgments. PhD. dissertation, University of Connecticut.

This dissertation contains a series of formal acceptability judgment experiments, including several designed to replicate and expand Snyder's (2000) evidence for repetition effects (satiation) in judgment experiments.

Luka, B. J. & L. W. Barsalou. 2005. Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language* 52.436-459.

This article presents five experiments that investigated the effect of repetition (satiation) on acceptability judgments. The results suggest that the ratings for several sentence types may increase with repeated exposure.

Nagata, H. 1988. The relativity of linguistic intuition: The effect of repetition on grammaticality judgments. *Journal of Psycholinguistic Research* 17.1-17.

This article presents three experiments that investigate the effect of repetition (satiation) on acceptability judgments. Although the results are mixed, there is some indication that the absolute value of judgments did increase with repeated exposure, although the relative pattern remained the same.

Nagata, H. 1989. Effect of repetition on grammaticality judgments under objective and subjective self-awareness conditions. *Journal of Psycholinguistic Research* 3.255-269.

This article presents three experiments that investigate how the effect of repetition (satiation) may interact with different mental states, in this case objective self-awareness (e.g., facing a mirror) and subjective self-awareness (not facing a mirror).

Snyder, W. 2000. An experimental investigation of syntactic satiation effects. *Linguistic Inquiry* 31.575-582.

This article presents a formal Yes-No experiment designed to investigate the effect of repetition (satiation) on acceptability judgments. The results suggest that the acceptability of *whether*-islands and complex NP islands may increase with repeated exposure.

Sprouse, J. 2009. Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry* 40.329-341.

This article presents nine experiments designed to investigate the effect of repetition (satiation) on acceptability judgments using both magnitude estimation and yes-no tasks for several different syntactic island violations. None of the results suggest that the ratings for these island violations change with repeated exposure.

Gradience in Acceptability and Grammaticality

Acceptability judgments are well known to be gradient, that is, they exist on a continuum. One potentially interesting question is where the gradience in acceptability comes from. One possibility is that grammatical theories are categorical in nature, and that the gradience in acceptability arises due to the gradience of other cognitive systems involved in making acceptability judgments. Another possibility is that grammars themselves are gradient in a way that directly gives rise to the gradience of acceptability judgments. Chomsky 1964 may be the earliest proposal to capture the gradience of acceptability with the grammar itself (in this case a grammatical system that gives rise to 7 categories of grammatical violations). Keller 2000 proposes a new form of Optimality Theory to capture the gradient effects of acceptability in a gradient grammar. Sorace and Keller 2005 reviews a series of results that indicate gradient acceptability, and argues that these effects are best captured with a gradient grammar along the lines of Keller's (2000) Linear Optimality Theory. Featherston 2005c proposes a different grammatical architecture to capture the gradience in acceptability, which combines a constraint ranking system with a probabilistic selection mechanism. Newmeyer 2007, in a response to Featherston 2007 (see Concerns about the Reliability of Traditionally Collected Judgments), argues that the ability to detect subtle, gradient effects in acceptability does not logically entail that such effects are due to the grammar. Sprouse 2007 attempts to find empirical evidence in formal judgments that might suggest categorical differences in grammaticality. Bresnan 2007 and Bresnan and Ford 2010 propose that grammatical knowledge is probabilistic, and that this probabilistic knowledge drives both the frequency of occurrence of constructions and the gradience of acceptability judgments.

Bresnan, J. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In *Roots: Linguistics in Search of Its Evidential Base.* S. Featherston & W. Sternefeld. Berlin: Mouton de Gruyter.75-96.

This article presents two acceptability judgment experiments that are designed to test whether native speakers judgments of the two forms of the dative alternation correlate with the probabilities of those two forms that are predicted by a probabilistic model of the dative alternation derived from corpus data. The results suggest that the there is a correlation.

Bresnan, J. & M. Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86.186-213.

This article compares the predictions of the probabilistic model of the dative alternation from Bresnan 2007 (derived from corpus data) to native speakers' acceptability judgments, reading times, and sentence completions. By further comparing two varieties of English, this study suggests that speakers' probabilistic knowledge is contingent upon input rather than categorical grammatical rules.

Chomsky, N. 1964. Degrees of Grammaticalness. In *The Structure of Language*. J. A. Fodor and J. J. Katz (eds). Prentice-Hall Inc. Englewood NJ. 384-389.

This is perhaps the first proposal to capture gradient acceptability with the grammar. Chomsky proposes three levels of grammatical analysis, with the possibility of a violation at each level, for a total of 7 possible categories of grammatical violation.

Featherston S. 2005c. The Decathlon Model of empirical syntax. In *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*. M. Reis & S. Kepser (eds). Berlin: Mouton de Gruyter. 187-208.

This chapter proposes a syntactic architecture known as the Decathlon model that combines a constraint ranking system with a probabilistic output selection component in an attempt to account for both gradient acceptability data and syntactic frequency data.

Keller, F. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. PhD. dissertation, University of Edinburgh.

This dissertation develops a type of gradient called Linear Optimality Theory to capture gradient acceptability effects.

Newmeyer, F. 2007. Commentary on Sam Featherston, 'Data in generative grammar: The stick and the carrot'. *Theoretical Linguistics* 33.395-399.

This is a response to Featherston 2007. Newmeyer discusses the logical dissociation between the claim that syntax would be well-served by adopting formal experimental methods for judgment collection, and the claim that gradient acceptability judgments are best explained by gradient grammatical architectures.

Sorace, A. & F. Keller. 2005. Gradience in linguistic data. *Lingua* 115.1497-1524.

This article surveys the evidence for gradience in acceptability judgments, and argues (following Keller 2000) that this gradience is best explained with Linear Optimality Theory.

Sprouse, J. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics* 1.118-129.

This article presents two pieces of evidence from formal judgment experiments that may suggest a categorical distinction between grammatical and ungrammatical sentences.