(for Chomsky legacy volume, eds. Nick Allott, Terje Lohndal, and Georges Rey)

Linguistic Judgments As Evidence

Steven Gross
Johns Hopkins University

1. Introduction

Linguistics research is filled with observations such as the following: 'There are three green books on the table' is an acceptable sentence, but 'There are green three books on the table' is not. Such judgments—as well as judgments about co-reference, ambiguity, pronounceability, and more—form a significant part of the evidence base for linguistics. This is in large measure due to Chomsky, whose work has exemplified the fruitfulness of such evidence and whose *Aspects of the Theory of Syntax* (Chomsky 1965, chapter 1) is a *locus classicus* for theorizing about their status. The prominence of judgment data in contemporary linguistics is crucially tied to Chomsky's mentalist reconception of the field.

Judgment data were not completely absent prior to Chomsky's work. For example, field linguists did not always prescind from asking informants whether such-and-such was something they would say, and Chomsky's teacher Zelig Harris emphasized the importance for phonology of speakers' judgments concerning sound differences (Harris 1951). But the positivist, behaviorist, and structuralist positions that dominated American linguistics in the first half of the 20th century tended to view the use of judgment data with suspicion and focused rather on produced sentences.

The methodological strictures in part arose in reaction to problems encountered in earlier introspectionist psychology (discussed in section 3). But the focus on produced utterances reflected as well a particular conception of what languages are and thus what linguistics is about. Though linguists of this period differed in many ways, they shared a tendency to view languages as consisting in the totality of utterances speakers of that language can produce (an E-language in Chomsky's (1986) terminology); and much work focused on describing, analyzing, and taxonomizing languages so conceived—for example, the many Native American languages so apparently different from the Indo-European languages which were then more familiar to linguists. While such a conception does not of itself preclude the use of speakers' judgments (cf. the remarks on Devitt's (2006) views in section 2), it is more naturally combined with an emphasis on corpus data, especially given the methodological scruples already mentioned.

Conversely, judgment data find a natural home in Chomsky's mentalist reconception of linguistics—a reconception, according to Chomsky (1966), that is in fact a recovery and development of earlier ideas about language. On this approach, linguistics aims, not just to describe linguistic products, but to provide a cognitive explanation of various of their distinctive features. One of Chomsky's core hypotheses is that there is an innately constrained computational procedure realized in the mind-brain—so-called I-language—that is implicated specifically in linguistic phenomena and whose character explains some of their distinctive features. As with aspects of our cognition more generally, we cannot directly observe I-language but must infer it from the effects to which it contributes. The methodological claim relevant to this chapter is that judgment data prove particularly useful in this endeavor.

This is so for several reasons. First, judgment tasks allow linguists to collect data concerning items that rarely show up in a corpus of actually produced speech—for example, perfectly fine sentences that are structurally a bit complex. Such sentences have proven useful for investigating particular phenomena and hypotheses. Judgment tasks, like any experiment, thus allow linguists to draw upon data beyond what happens to occur naturally without intervention. Second, relatedly, judgment tasks can provide negative evidence regarding items that do not appear in the corpus specifically because they violate I-language constraints. Mere absence from a corpus can occur for any number of reasons. A judgment of unacceptability provides stronger evidence of ungrammaticality—insofar as reasonable alternative explanations can be ruled out (pragmatic oddity, processing difficulties, memory constraints, lexical awkwardness, etc.). This points to a third reason judgment data have proven useful. One can systematically vary the items presented in judgment tasks in ways that control for competing explanations. An inference back to the judgment's cause may then plausibly isolate features of I-language from among the various interacting aspects of cognition responsible for the judgment.

There is no claim that in principle *only* judgment data can serve this purpose. Indeed, in early work (Chomsky 1955/1975, pp. 101-3), Chomsky looked for behavioral measures that might serve instead. But as the efficiency and fruitfulness of judgment data became evident, they became the central source of evidence in theoretical syntax—a status they have retained, even as the evidential base of linguistics has broadened. Nevertheless, despite their success in allowing interesting phenomena to be identified, important questions to be raised, and substantial hypotheses to be tested, the use of judgment data in linguistics has attracted controversy from the start and continues to do so (e.g., Branigan and Pickering 2017)—in a way that the analogous use of judgment data in vision science does not.

After first looking more closely at what judgment data are and how they are supposed to provide evidence for linguistic hypotheses, we shall examine these objections, including recent attempts to address some of them experimentally. The research canvassed is part of a movement in recent decades to gather and analyze judgment data more formally than had been typical. We close by surveying some of the new paths this work has opened up.

2. What they are

Judgment data are meta-linguistic judgments—judgments about specific linguistic items, construed broadly to include language-*like* items (e.g., ungrammatical strings). Linguists advert to a wide variety of meta-linguistic judgments. We have already mentioned judgments about acceptability, co-reference, ambiguity, and pronounceability, but there are plenty more: judgments about truth-value, entailment, frequency of use, etc. Moreover, within these types, the judgments may be expressed in a variety of ways. As explained below, subjects may be asked to rate a sentence's acceptability on a Likert scale, for example, or to provide a magnitude estimation, or to express preferability by performing a forced choice among multiple sentences.

But not *all* meta-linguistic judgments are used as evidence: that 'I wonder what who saw' (but not 'I wonder who saw what') violates the Superiority condition is a meta-linguistic judgment, but one invoked in linguistic theorizing to explain data, not as a datum itself. Other cases can be less obvious—notably, judgments of (un)grammaticality. It was once common for linguists to advert to such judgments as evidence, but now they do so far less frequently. This in part reflects the use of 'grammar' as a technical term for I-language, so that a 'judgment about grammaticality' in that sense refers to a proposed explanation of evidence (cf. Chomsky 1965, p.

11, on acceptability vs. grammaticality). When linguists serve as their own subjects, there is a particular risk of not clearly distinguishing *explanans* and *explanandum*.

What then distinguishes meta-linguistic judgments of the kind used as evidence? In fact, it is not clear that there is an informative general answer that captures all and only the relevant cases. But canvassing some candidates helps nonetheless bring out some central, typical features.

One answer is that such judgments are in some sense intuitive; and indeed the judgments used as evidence are sometimes referred to as linguistic intuitions. Some linguists eschew this label, perhaps owing to unwanted suggestions not encouraged by the more neutral term 'judgment data'—for example, that subjects have a special intuitive faculty that provides them access to the causes of their judgment (in particular, their I-language), or that hunches about the causes of judgment may serve as useful data. Another reason could be that it is none too easy to spell out the intended sense of 'intuitive' in a way that captures the judgment data that linguists in fact use. For example, linguistic intuitions are sometimes characterized as relatively immediate, or unreflective, or not based on conscious inference. But it can take time and effort to get a reading of a sentence, and reflection and perhaps conscious inference to arrive at a truth-judgment; experimenters may find it informative to compare rushed and considered judgments; and, even if a linguist developed the skill to spot and label, for example, Superiority violations quickly and without reflection, it is not obvious that would increase the evidentiary value of the resulting judgments. Still, the characteristic features of "intuitive" judgments hold of a large range of judgment data, and their doing so helps explain their usefulness. For a relatively immediate, unreflective judgment not based on reasons is more likely to reflect the character of I-language—the object of inquiry—rather than the subject's beliefs *about* language or other intrusions from higher cognition.

Judgment data are also often characterized as introspective (sometimes critically, as we will see). But again it is not clear this applies to all the judgments linguists use as evidence. The question turns in part on what is meant by 'introspection'. On one common characterization, introspective judgments are non-inferential, non-exteroceptively-based judgments about one's own mental states (Schwitzgebel 2019). Some judgment data may be introspective in this sense, as when subjects report whether a sentence sounds good. But a judgment that a sentence is acceptable, or an utterance true, is about the sentence or the utterance, not about one's mental states concerning them. (Compare: if I report that it is raining, the report *expresses* my belief, but both the report and my belief are about the rain, not about my belief itself, nor about my perception of the rain.) Talk of introspection may be cashed out in other ways—for example, rather than requiring a distinctive kind of internally-directed content, one might require just a distinctive kind of access. But it is not obvious how to do so without either collapsing the distinction between perceptual and introspective judgments or omitting some of the judgments linguists use as evidence. An example of the former would be a liberal conception of introspection that included any judgment immediately based on a conscious experience. An example of the latter would be a conception that emphasizes the need to compare similarities and differences across mental states (Chirimuuta 2014), which would render forced-choice preferences introspective, but not acceptability judgments.

Both of these answers—in terms of intuition or introspection—involve in part an appeal to the etiology of the meta-linguistic judgments that serve as evidence in linguistics. Might some other etiological approach do the job? A problem is that the different kinds of judgment—concerning, recall, acceptability, pronounceability, truth-value, even frequency or likelihood of use (Labov 1996)—draw on different cognitive capacities. From the perspective of etiology,

judgment data thus appear to form something of a motley. To be sure, in all cases, it is assumed that the etiology involves aspects of our linguistic competence. This assumption is built into the logic of linguists' explanations, when they infer properties of this competence from such causal effects as speakers' judgments. But I-language also figures in the etiology of many judgments *not* used as evidence in linguistics—for example, whenever one comes to believe what one is told. Even if we limit ourselves to just one central kind of judgment—acceptability judgments—it is unclear that we are currently in a position to offer an informative account of their etiology, especially of that part of the causal story that follows upon (attempts at) parsing and comprehension. One natural suggestion is that the presence or absence of error signals of the sort hypothesized by theories of language monitoring may play a role as well in the generation of at least some meta-linguistic judgments (Matthews, unpublished; Sprouse 2018; Gross 2020). But this idea stands in need of development and empirical investigation. Moreover, it is probably not the full story: it is plausible that (in)comprehension also plays a role, as may feelings of effort or disfluency and other factors.

It is useful here to pause briefly over two accounts of how meta-linguistic judgments fulfill their evidential function that provide *alternatives* to the dominant mentalist conception championed by Chomsky. These views have played a significant role in philosophical debates and illustrate how differing conceptions of judgment data and their evidential status are tied to differing conceptions of language and linguistics. But they also illustrate, more specifically, differing conceptions of judgment data's etiology.

According to Katz (1981), languages are abstract objects—akin to mathematical structures—with which we do not causally interact. The meta-linguistic judgments deployed in linguistics are thus not arrived at via perception or introspection, which would require causal interaction with instances of their object. Rather, we possess a faculty of intuition—the competence-grammar—that provides *a priori* access to these abstract, non-mental structures, including to their necessary properties. This access consists in our capacity to construct *representations* of sentences and their properties from tacit knowledge (innate and learned) of the grammatical principles of the language. The content of linguistic intuitions—for example, that some sentence S is grammatical—is thus directly supplied to judgment by this faculty. Note the contrast with Chomsky's view, as I have presented it, which in no way assumes that I-language itself issues meta-linguistic contents concerning grammaticality—or acceptability, for that matter.[i]

Devitt (2006), on the other hand, endorses a non-Platonist E-language conception of the object of linguistics. On his view, linguistics concerns the possible tokens of a language—not the psychology of speakers—with the tokens' linguistic properties determined by conventions among speakers. These conventions do impose constraints on the psychology of a speaker if she is to count as a competent member of the linguistic community: she must process language in a way that yields outputs whose properties *respect* these conventions. But Devitt denies that satisfying this constraint requires or involves an I-language or modularized language faculty. Thus, *contra* Katz, the contents of judgment data, according to Devitt, are not the direct product of a language module that serves as a faculty of linguistic intuition. Nor do they provide the basis for an inference to the best explanation concerning the properties of I-language. They reflect, rather, the "central systems" knowledge that language-users have of a language after many years of immersion in it—just as regular interaction can lead to fairly reliable knowledge of surrounding flora and fauna.

Critical discussion of these alternative conceptions of language is beyond our scope (but see, for example, Iten, Stainton, and Wearing 2006 on Katz, and Maynes and Gross 2013 and Rey 2020 on Devitt). Our point was to contrast them with the mentalist conception and thus indicate the range of views concerning the etiology of judgment data. It is not in general a condition on evidential status in the sciences that theorists possess a *complete* understanding of the etiology of data (Bogen and Woodward 1988). But further insight into the etiological details—in addition to its intrinsic interest and consequences for other debates in the philosophy of linguistics—could help address some of the worries about the use of judgment data in linguistics to which we now turn.

3. Objections to judgment data

We remarked that the use of judgment data has never been without critics. The objections have taken various forms. Earlier objections tended to deem judgments problematic as evidence *per se*, not just in linguistics but more generally. Later objections contend that the use of judgment data is problematic more specifically in the linguistic domain, in some cases on account of how they are in fact typically gathered or because of an over-reliance on them over other sources of evidence. No matter their scope, we limit ourselves to objections that target features of putative evidence that would be problematic no matter the type of scientific inquiry. These include failures of validity, reliability, sensitivity, or freedom from bias.[ii] We thus set aside worries that judgment data are, say, unscientific or subjective—*unless* cashed in these other terms. This accords with Chomsky's long-standing protests against a "methodological dualism" that would hold the sciences of the mind to different standards regarding evidence, often based on extra-scientific *a priori* philosophical concerns (e.g., Chomsky 2000, p. 142).

As noted in section 1, a main source of concern for linguists prior to Chomsky stemmed from problems encountered by introspectionist psychology. Introspection proved inadequate to ground a consensus on various of the main questions introspectionist psychology pursued—such as identifying the atoms of sensory experience and determining whether there is imageless thought (see Hatfield 2005). These failures led to a more general mistrust of introspection as a valid or reliable source of evidence—at least for settling illuminating questions (cf. Feest 2014). But was this generalization an instance of salutary methodological caution or an over-generalization? Introspection may well have been unfit for the purposes to which it was put. The atoms of sensory experience—if there are such—may be too fine for introspection to uncover; and the introspective techniques deployed in the imageless thought controversy tended to target mental processes (about which it is widely agreed that we lack reliable introspective access) rather than their products. It hardly follows that introspective data in general are methodologically suspect—let alone that judgment data are. (Recall from section 2 that judgment data form a broader class than introspective judgments on some common uses of the term 'introspection'.) Note, in particular, that the use of judgment data in linguistics is rather different from their use in the introspectionist examples just mentioned: the judgments in linguistics are not themselves *about* mental processes, nor are they used to illuminate sensory experience; rather, they are products of a mental process aspects of which the theorist attempts to *infer* from those products.

The attempt to generalize beyond introspectionist psychology is further blunted by a comparison with the use of judgment data in contemporary vision science. As with much judgment data in linguistics, the judgments used in vision science may be based on conscious experience and may even ascribe mind-dependent properties, while yet being about external

objects and their features, as opposed to subjects' experience. (Though, to be sure, judgments about how things look or appear also play a role in vision science—for example, in work on color constancy.) Because such judgments are based on experience, one may reasonably infer something about the subject's experience from them. But the research question driving a vision scientist need not concern conscious experience. As with the linguist, the vision scientist's concern may be earlier aspects of the judgment's etiology—for instance, aspects of early visual processing. Importantly, in vision science there are no general methodological qualms about using such judgments as evidence. The tremendous success of vision science provides a powerful reply to any general suspicion of judgment data (and, to the extent it relies on judgments that *are* deemed introspective, to any general suspicion of introspection).

This leaves open the possibility that judgment data are invalid or unreliable in particular domains or in particular deployments, as they proved to be in some introspectionist psychology. Language differs from vision in various ways, for example in the degree and kind of individual variation and in the mechanisms by which processing provides material for judgments. Generalizing from success in one domain to success in another requires caution just as generalizing from failure in one domain to failure in another does. The obvious response is that the tremendous success of linguistics based on judgment data likewise makes a strong case for their appropriateness in this domain.

Presenting this case in detail—that is, reviewing significant tracts of linguistics—is of course beyond our scope, but also perhaps not to the point, at least so far as more recent critics are concerned. For they are familiar with this work and yet their worries persist. Instead, we can consider how such a case could be strengthened in response to two more specific objections sometimes raised about the use of judgment data in linguistics. The first is that the validity and reliability of this data is threatened by the "informal" way the judgments are typically gathered in practice. Much judgment data consists in linguists reporting their own responses. It is objected that data collected in this manner are subject to confirmation bias, not properly controlled, and not amenable to statistical analysis owing to the small number of subjects (often just one!). Second, critics sometimes bemoan the disconnect between work in, for example, theoretical syntax and work in psycholinguistics and other area of cognitive science (Ferreira 2005). The case for judgment data in linguistics would be strengthened to the degree that other converging sources of evidence helped demonstrate that judgment data are in effect sufficiently "calibrated" to the phenomena they are used to understand. Let's consider each of these challenges in turn.

One strategy in response to worries concerning how judgment data are in practice collected is to note that the informal process of gathering judgment data actually involves more than this objection allows (Phillips 2009). Data that make it to print typically receive corroboration by seminar attendees, conference audiences, journal referees, and others—some of whom may be biased towards other hypotheses. As for controls, linguists often look for "minimal pairs": strings that differ so far as possible only in one way relevant to the hypothesis in question ('John is eager to please' vs. 'John is easy to please'). Even when only a single string is presented in a published text, many variants may also have been considered—and expert consumers of the publication will consider many variants as well in reflecting on what they read. Indeed, "informal" methods allow for the consideration of many more variants than do more "formal" methods involving many subjects. Finally, strong linguistic judgment data share a feature with the judgment reports used in vision science: readers are often in a position to readily judge for themselves. For example, in a highly cited paper on visual adaptation to numerosity, Burr and Ross (2008) reported the judgments of four subjects, two of whom were the authors.

While there has been debate concerning the interpretation of their results, no one has challenged the data: for anyone can view the figures and experience the results for themselves. Similarly, in linguistics, judgment data are unlikely to be taken up further in the literature if readers find the judgments dubious.

There is, however, another way one can respond to worries about how judgment data are in-practice collected—namely, by changing how one collects data. Indeed, the last few decades have witnessed a significant growth in "formally" collected judgment data—judgment data gathered from large numbers of naïve subjects using methods common across the social sciences. Work in experimental syntax was catalyzed in part by the publication of Schütze (1996) and Cowart (1997) which laid out, and showed how to avoid, the more subtle confounds to which judgment data can be subject—for example, order effects, where previously considered stimuli affect subjects' response to subsequent stimuli. (For an example of how order effects can be relevant to judgments that serve as evidence for Binding Theory, see Gordon and Hendrick 1997.) Formal experiments are not themselves immune to methodological pitfalls; no process of gathering data is (cf. Culicover and Jackendoff 2010). Confounds may not be controlled for; undetected bias can infect results; subjects can misunderstand instructions; etc. But anticipating and avoiding such problems to the extent possible is at the core of experimentalists' training; and their standards for reporting results are intended to facilitate the further uncovering of design issues and alternative interpretations.

Not only does the use of formal methods provide a response to worries concerning informally gathered data, it also enables the empirical investigation of data-gathering methods themselves. For example, we noted earlier that subjects may in various ways attribute various properties to strings of words. A subject may express the acceptability of a string by rating it on a Likert scale or by providing a magnitude estimate. In the former case, she may select a natural number from 1 to 5, with 1 indicating complete unacceptability and 5 indicating full acceptability. In the latter case, after a numerical starting point is established with respect to an initial string, further strings are placed on an open-ended numerical scale in accordance with their judged distance in acceptability from the initial string. (More specifically, subjects are instructed to rate the further strings as a multiple or fraction of the initial standard—see Bard, Robinson, and Sorace 1996.) Among other advantages, magnitude estimation allows subjects to introduce as many distinctions among strings as they like. Further investigation, however, has suggested that, unlike with other features such as luminance in vision tasks, subjects are unable to judge ratios of acceptability—perhaps owing to the difficulty of establishing a zero point for full unacceptability (Featherston 2008). Perhaps they complete such tasks by tacitly converting them into a more standard rating task, similar to rating along a Likert scale (Sprouse 2011). A methodological inquiry like this can only be done using a more formal approach.

Of particular interest to us is that formal methods can be used to investigate *in*formal methods and thus provide yet another response to worries concerning them—now by potentially vindicating them, as opposed to supplying an alternative to them. For example, Sprouse and Almeida (2012) took all 469 strings that were provided an acceptability rating in Adger's (2003) syntax textbook and tested them on 440 naïve subjects. Using conservative criteria, they found that 98% of the informal judgments replicated. Similarly, Sprouse, Schütze, and Almeida (2013) used a random sampling of 300 informal acceptability judgments drawn from papers published the previous decade in the journal *Linguistic Inquiry* to test naive subjects on three different judgment tasks. Again, using conservative criteria, they found that 95% of the informal judgments replicated. More generally, comparisons of informally and formally gathered

judgment data so far appear to vindicate the former, at least relative to the latter (cf. Culbertson and Gross 2009; Mahowald, Graff, Hartman, and Gibson 2016). It should be noted, though, that the work to date has concentrated on acceptability judgments in English (though cf. Linzen and Oseki 2018, albeit still on acceptability judgments).

In addition to investigating consistency between informally and formally gathered judgments, other relevant features have been explored as well—in particular, sensitivity and bias. Sprouse and Almeida (2017) compared the sensitivity (the probability of detecting an effect) of four different judgment tasks. This speaks to the number of subjects needed to achieve statistical power: recall the worry that informally gathered judgments in practice involve too few subjects. The most sensitive task—forced-choice judgments—reached 80% power on strong effects with just 11 participants. As for the possibility of unconscious confirmation bias, Sprouse (2020) notes that one way such bias might reveal itself would be via sign reversals in a comparison between naïve and expert judgments—that is, cases with a change in direction between the effect reported by the two groups. But in fact the data reported in Sprouse, Schütze, and Almeida (2013) revealed very few such cases—about 1-2%—and thus little evidence for confirmation bias (cf. also Dabrowska 2010).

We have suggested that there is little intrinsic reason to reject judgment data generally and that specific reasons for worry about *linguistic* judgment data based on how they are in practice collected can be avoided by changing those practices and in any event are perhaps no cause for great concern. Let's return now to the issue of calibration and more generally of mesh with other methods and theoretical domains. It is a common complaint that there is an over-reliance among linguists on judgment data—at least in theoretical syntax. This is an objection, specific to the linguistic domain, about how such judgments are *used*, not an objection to such data *per se* or to how they are in practice gathered. The other side of the coin for such over-reliance would be a dearth of converging evidence from other sources and, in particular, a lack of calibration of judgment data by other methods. As a consequence, we would be permitted less confidence that our evidence provides an accurate measure of the domain we use it to investigate.

Note that the worry pertains to what evidence linguists in *practice* most rely on. It is not claimed that linguists endorse a *principled* restriction of the data to meta-linguistic judgments. Recall Chomsky's objection to methodological dualism. As we saw, it provides a response to certain methodological strictures *against* judgment data—those based on extra-scientific philosophical concerns. But it also undermines any principled restriction *to* judgment data.

Linguists of course do draw upon a wide variety of evidence: corpus data, developmental and acquisition work, evidence from congenital and acquired deficits, etc. Some of this data—and success in developing theories exploiting it alongside judgment data—does indeed provide a kind of indirect "calibration" of judgment data. But when it comes to the more fine-grained specifics of an individual language's grammar, more direct calibration is difficult to find. A casebook example of a failure to find such calibration comes from the downfall of the Derivational Theory of Complexity, which unsuccessfully attempted to relate grammatical transformations to response times in production and comprehension (Fodor, Bever, and Garrett 1974). Arguably, however, the source of this failure was incorrect theory, not any problem with the judgments on which it was based (Marantz 2005). More recent reading-time studies provide a more successful example of mesh between judgment data and behavioral data, and more generally between the methods dominant in theoretical syntax and those more typical of psycholinguistics. For example, Stowe (1986) found that whether reading time slows at crucial

junctures in sentences with filler-gap dependencies depends on whether an active filler strategy would violate a syntactic island constraint (see Phillips 2006 for an overview of reading-time studies). With the increasing use of other experimental methods and the growing *rapprochement* between theoretical syntax and psycholinguistics, there is the possibility that further sources of specific converging evidence will be uncovered (see Sprouse and Schütze 2020 for a discussion of the relations more generally among judgment data and other data).

4. The future of judgment data

Informally gathered judgments have supplied linguists with large amounts of easily-obtained data directly relevant to their most pressing questions. A glance at leading linguistics journals suggest that they continue to do so—though see Marantz (2005, p. 438) for some skepticism on this score. Supposing that they do, and given the legitimation so far of informal methods by more formal inquiry, the question arises whether and when the use of formal methods is called for or might prove fruitful. The formal gathering and analysis of judgment data, after all, is not only subject to its own methodological pitfalls (mentioned above), but is also more time-consuming and costly—though less so than it was, owing to the advent of crowd-sourcing platforms such as Amazon Mechanical Turk. We conclude by briefly remarking on a few possibilities.

- One important use of formal judgment tasks has already been noted: to investigate differences among judgment tasks themselves. We mentioned comparisons between formal and informal methods and between Likert-scale ratings and magnitude estimations. Another example is the investigation of how varying task instructions can affect performance (Cowart 1997).
- Various authors propose using formal methods to investigate strings that are hard to judge or where informal judgments diverge (e.g., Linzen and Oseki 2018), whether to settle the status of the case or to better understand the source of the divergence. For example, formal methods can help reveal when variability at the group level masks interesting clusters among sub-groups (Kush, Lohndal, and Spouse 2018, 2019).
- It has long been recognized that acceptability is gradient; formal methods have made this especially clear. Some point to these results in advocating gradience in grammars (Sorace and Keller 2005—cf. Chomsky 1965, p. 11). But more work is required to determine whether gradience in acceptability reflects grammatical gradience or rather arises from other sources. Formal judgment tasks will no doubt play a role.
- Indeed, formal judgment tasks, in tandem with tasks of other sorts, have been used to explore how aspects of processing independent of I-language may affect judgments of acceptability. For example, it has long been assumed that working memory limitations can cause a grammatical sentence to be judged unacceptable, with multiply center-embedded sentences providing the parade case. But efforts to establish further working memory effects (surveyed in Sprouse 2018) have so far failed to yield unequivocal results. On the other hand, there has been more success investigating cases—so-called 'grammatical illusions'—where features of processing can cause ungrammatical sentences to be judged acceptable (Wagers, Lau, and Phillips 2009, Wellwood et al. 2018).
- The work on grammatical illusions points to another emerging role for formal judgment data. Judgment tasks have traditionally been deployed in theoretical syntax to understand I-language. But formal judgment tasks are proving useful for

understanding questions concerning linguistic processing (Phillips et al. forthcoming). More generally, formal judgment tasks hold out the promise of helping linguists better understand the relation between "competence" and "performance"—in particular, between I-language and the linguistic processing that I-language is variously said to be a "part of", or "embodied in", or an "abstraction from" (cf. Lewis and Phillips 2015).

Judgment data, no matter how collected, have a definite future in linguistics. That other methods now also play a large role is all to the good—and something in part made possible by the advances judgment data allowed. Hitting upon this large vein and displaying its riches is among Chomsky's enduring legacies.

References

Adger, D. 2003. *Core Syntax: A Minimalist Approach*. Oxford: Oxford University Press.

Bard, E., Robertson, D., and A. Sorace. 1996. "Magnitude Estimation of Linguistic Acceptability." *Language* 72: 32-68. DOI:10.2307/416793.

Bogen, J., and J. Woodward. 1988. "Saving the Phenomena." *The Philosophical Review*, 97: 303–52.

Branigan, H., and M. Pickering. 2017. "An Experimental Approach to Linguistic Representation." *Behavioral & Brain Sciences*, 40: e282. DOI:10.1017/S0140525X16002028.

Burr, D., and J. Ross. 2008. "A Visual Sense of Number." *Current Biology*, 18: 1-4. DOI: 10.1016/j.cub.2008.02.052.

Chiramuuta, M. 2014. "Psychophysical Methods and the Evasion of Introspection." *Philosophy of Science*, 81: 914-26. DOI: 10.1086/677890.

Chomsky, N. 1955/1975. *The Logical Structure of Linguistic Theory*. Chicago: University of Chicago Press.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. 1966. *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*. New York: Harper & Row.

Chomsky, N. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

Chomsky, N. 2000. *New Horizons in the Study of Language and Mind*. Cambridge University Press.

Cowart, W. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Thousand Oaks, CA: Sage Publications.

Culbertson, J., and S. Gross. 2009. "Are Linguists Better Subjects?" *British Journal for the Philosophy of Science*, 60: 721-36. DOI: 10.1093/bjps/axp032.

Culicover, P., and R. Jackendoff. 2010. "Quantitative Methods Alone Are Not Enough: Response to Gibson and Fedorenko." *Trends in Cognitive Sciences*, 14: 234-5. DOI: 10.1016/j.tics.2010.03.012.

Dabrowska, E. 2010. "Naive v. Expert Intuitions: An Empirical Study of Acceptability Judgments." *The Linguistic Review*, 27: 1-23. DOI:10.1515/tlir.2010.001.

Devitt, M. 2006. *Ignorance of Language*. Oxford: Oxford University Press.

Featherston, S. 2008. "Thermometer Judgments as Linguistic Evidence." In *Was ist linguistische evidenz?*, edited by C. M. Riehl and A. Rothe, 69-90. Aachen: Shaker Verlag.

Feest, U. 2014. “Phenomenal Experiences, First-Person Methods, and the Artificiality of Experimental Data.” *Philosophy of Science*, 81: 927-39. DOI: 10.1086/677689.
Ferriera, F. 2005. “Psycholinguistics, Formal Grammars, and Cognitive Science.” *The Linguistic Review*, 22: 365–80. DOI:10.1515/tlir.2005.22.2-4.365.
Fodor, J., Bever, T., and M. Garrett. 1974. *The Psychology of Language*. New York: McGraw Hill.
Gordon, P., and R. Hendrick. 1997. “Intuitive Knowledge of Linguistic Co-Reference.” *Cognition*, 62: 325–370. DOI:10.1016/S0010-0277(96)00788-3.
Gross, S. 2020. “Linguistic Intuitions: Error Signals and the Voice of Competence.” In *Linguistic Intuitions,* edited by S. Schindler, A. Drożdżowicz, and K. Brøcker. Oxford: Oxford University Press.
Harris, Z. 1951. *Methods in Structural Linguistics*. Chicago: University of Chicago Press.
Hatfield, G. 2005. “Introspective Evidence in Psychology.” In *Scientific Evidence: Philosophical Theories and Applications*, edited by P. Achinstein. Baltimore: Johns Hopkins University Press.
Iten, C., Stainton, R., and C. Wearing. 2006. “On Restricting the Evidence Base for Linguistics.” In *Handbook for the Philosophy of Science. Volume 12: Philosophy of Psychology*, edited by Paul Thagard. Amsterdam: Elsevier. DOI:10.1016/B978-044451540-7/50024-4.
Katz, J. 1981. *Language and Other Abstract Objects*. Totowa, NJ: Rowman & Littlefield.
Kush, D., Lohndal, T., and J. Sprouse. 2018. “Investigating Variation in Island Effects: A Case Study of Norwegian Wh-Extraction.” *Natural Language and Linguistic Theory*, 36: 743-779. DOI:10.1007%2Fs11049-017-9390-z.
Kush, D., Lohndal, T., and J. Sprouse. 2019. “On the Island Sensitivity of Topicalization in Norwegian: An Experimental Investigation.” *Language*, 95: 393-420. DOI:10.1353/lan.0.0237.
Labov, W. 1996. “When Intuitions Fail.” *Chicago Linguistics Society: Papers from the Parasession on Theory and Data in Linguistics*, 32: 76-106.
Lewis, S., and C. Phillips. 2015. “Aligning Grammatical Theories and Language Processing Models.” *Journal of Psycholinguistic Research*, 44: 27-46. DOI:10.1007/s10936-014-9329-z.
Linzen, T., and Y. Oseki. 2018. “The Reliability of Acceptability Judgments Across Languages.” *Glossa*, 3: 1-25. DOI: 10.5334/gjgl.528.
Mahowald, K., Graff, P., Hartman, J., and E. Gibson. 2016. “SNAP judgments: A Small N Acceptability Paradigm (SNAP) for Linguistic Acceptability Judgments.” *Language*, 92: 619-635. DOI: 10.1353/lan.2016.0052.
Marantz, A. 2005. “Generative Linguistics within the Cognitive Neuroscience of Language.” *The Linguistic Review*, 22: 429-45. DOI:10.1515/tlir.2005.22.2-4.429.
Matthews, R. unpublished. “Linguistic Intuition: An Exercise of Linguistic Competence.”
Maynes, J., and S. Gross. 2013. “Linguistic Intuitions.” *Philosophy Compass*, 8: 714-30. DOI:10.1111/phc3.12052.
Phillips, C. 2006. “The Real-Time Status of Island Phenomena.” *Language*, 82: 795-823. DOI:10.1353/lan.2006.0217.
Phillips, C. 2009. “Should We Impeach Armchair Linguistics?” In *Japanese/Korean Linguistics*, 17. Edited by S. Iwasaki, 49-65. Palo Alto: CSLI Publications.
Phillips, C., Gaston, P., Huang, N., and H. Muller. forthcoming. “Theories All the Way Down: Remarks on ‘Theoretical’ and ‘Experimental’ Linguistics.” In *Cambridge Handbook of Experimental Syntax*, edited by G. Goodall.

Rey, G. 2020. *Representation of Language: Philosophical Issues in a Chomskyan Linguistics*. Oxford: Oxford University Press.
Schütze, C. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.
Schwitzgebel, E. 2019. "Introspection." In *The Stanford Encyclopedia of Philosophy*, edited by E. Zalta. URL = <https://plato.stanford.edu/archives/win2019/entries/introspection/>.
Sorace, A., and F. Keller. 2005. "Gradience in Linguistic Data." *Lingua*, 115: 1497-524. DOI:10.1016/j.lingua.2004.07.002.
Sprouse, J. 2018. "Acceptability Judgments and Grammaticality, Prospects and Challenges." In *Syntactic Structures after 60 Years: The Impact of the Chomskyan Revolution in Linguistics*, edited by N. Hornstein, C. Yang, and P. Patel-Grosz. Berlin: Mouton de Gruyter.
Spouse, J. 2020. "A User's View of the Validity of Acceptability Judgments as Evidence for Syntactic Theories." In *Linguistic Intuitions,* edited by S. Schindler, A. Drożdżowicz, and K. Brøcker. Oxford: Oxford University Press.
Sprouse, J., and D. Almeida. 2012. "Assessing the Reliability of Textbook Data in Syntax: Adger's *Core Syntax*." *Journal of Linguistics*, 48: 609–652. DOI:10.1017/S0022226712000011.
Sprouse, J., Schütze, C., and D. Almeida. 2013. "A Comparison of Informal and Formal Acceptability Judgments Using a Random Sample from *Linguistic Inquiry* 2001-2010." *Lingua*, 134: 219-248. DOI:10.1016/j.lingua.2013.07.002.
Sprouse, J., and C. Schütze. 2020. "Grammar and the Use of Data." In *The Oxford Handbook of English Grammar*, edited by B. Aarts, J. Bowie, and G. Popova, 40–58. Oxford: Oxford University Press. DOI:10.1093/oxfordhb/9780198755104.013.28.
Stowe, L. 1986. "Parsing WH-Constructions: Evidence for On-Line Gap Location." *Language and Cognitive Processes*, 1: 227–245. DOI:10.1080/01690968608407062.
Wagers, M., E. Lau, and C. Phillips. 2009. "Agreement Attraction in Comprehension: Representations and Processes." *Journal of Memory and Language*, 6: 206-37. DOI:10.1016/j.jml.2009.04.002.
Wellwood, A., Pancheva, R., Hacquard, V., and C. Phillips. 2018. "The Anatomy of a Comparative Illusion." *Journal of Semantics*, 35: 543-83. DOI:10.1093/jos/ffy014.

[i] To be fair, one can find remarks by Chomsky that suggest otherwise, as when he writes that there are mechanisms in the mind that permit deduction-like computations from I-language to specific judgments (Chomsky 1986, p. 270). I am articulating the dominant, considered view, which has emerged more clearly in part owing to these debates.
[ii] These terms are used in their psychometric senses. Evidence is valid just in case it measures what it is intended to measure; and reliable just in case relevantly similar conditions elicit a similar response. ('Validity' and 'reliability' have different technical meanings in philosophy.) Sensitivity concerns the true positive rate. Reliability is necessary for validity, but not sufficient: reliably sinking in water was not a valid test of witchiness.