# Experiments in syntax and philosophy: the method of choice?

Samuel Schindler and Karen Kiil Brøcker
Centre for Science Studies
Aarhus University

*This chapter will be published in the collection* Linguistic Intuitions, Evidence, and Expertise *with Oxford University Press. Please cite the final published version.*

**Abstract**

Within Chomskyan syntax, linguistic intuitions have traditionally been gathered informally from small samples of linguists. Since the mid-1990s, however, several linguists have called for more `scientific' methods, including the use of larger sample sizes of ordinary speakers and the use of statistics. In the first part of this chapter, we discuss whether such an "experimental approach" to obtaining syntactical intuitions is really methodologically superior to the informal approach, as sometimes claimed. We think the answer is: not always and not in all respects. In the second part, we turn our attention to another academic field in which intuitions arguably play an evidential role, namely philosophy. Also here, critics have demanded that intuitions be harvested more systematically and have even appealed to experimental syntax in order to support their cause. However, given our assessment, experimental methods in syntax can be a model for the promotion of experimental methods in philosophy only under certain conditions.

Key words: intuition, experimental syntax, experimental philosophy, acceptability judgement, grammaticality judgement

## 1 Introduction

In the Chomskyan tradition of the study of syntax, it is common practice to "informally" consult one's own or one's colleagues' syntactic intuitions when building and testing theories of grammar. More recently, however, critical questions have been raised about this approach by proponents of experimental syntax (XSyn). In their pioneering works,

Schütze (1996) and Cowart (1997) have argued that linguists should adopt a more "scientific" approach and put theories of grammar on a broader foundation by collecting linguistic intuitions from a large number of ordinary speakers and by applying well-established statistical methods.

Interestingly, this call for a more systematic approach in the practice of syntacticians is mirrored in recent discussions in metaphilosophy, that is, the study of philosophical methods. Like linguists, philosophers have traditionally used their intuitions (in thought experiments) to assess their theories. But starting in the early 2000s, some philosophers have criticised this informal method and campaigned for a more systematic investigation of intuitions from non-philosophers. This critical approach is generally known as experimental philosophy, or simply XPhi. Two pioneers and proponents of XPhi have recently sought to motivate XPhi by appealing to XSyn and its alleged benefits in linguistics (Machery and Stich 2013).

In this chapter we argue that claims about the superiority of experimental methods in syntax are not always justified. Experimental methods in linguistics can therefore not unconditionally serve as a model for the promotion of experimental methods in philosophy.

This is how we proceed. In Section 2, we review claims about the methodological superiority of XSyn in comparison to traditional, informal methods of using linguistic intuitions as evidence in syntactic research. In Section 3, we assess whether these claims are justified. In Section 4, we discuss Machery and Stich's appeal to XSyn in their championing of XPhi. In Section 5, we assess Machery and Stich's claims in the light of our discussion in Section 3. In Section 6, we conclude in favour of methodological pluralism.

## 2 Experimental Syntax

The way syntactic intuitions have traditionally been collected in linguistics can, in short, be characterised like this: linguists construct a sentence that contains some syntactic phenomenon that they are interested in. They then ask a native speaker, most commonly themselves, whether the sentence appears acceptable or not. Often, they might ask for the opinion of one or more colleagues as well and refine their analysis based on the colleagues' responses. The sentence might be considered in the context of other similar sentences, which differ from the focus sentence mainly regarding the phenomenon of interest (minimal or near-minimal pairs). Schütze and Sprouse (2013) mention five ways in which what they call traditional judgement experiments are different from standard practice in the neighbouring field of experimental psychology:

I. “relatively few speakers (fewer than ten);
II. linguists themselves as the participants;
III. relatively impoverished response options (such as just ‘acceptable’, ‘unacceptable’, and perhaps ‘marginal’);
IV. relatively few tokens of the structure of interest, and
V. relatively unsystematic data analysis.” (Schütze and Sprouse 2013, 30)

Prima facie, this approach looks utterly unscientific. Accordingly, a number of commentators have criticised this informal practice and called for a more systematic approach. This approach is generally called “experimental syntax” after Cowart’s (1997) book. Cowart focuses on how you can design and run syntactic experiments to avoid the problems of the traditional method. According to Cowart, the basic experimental set-up for collecting syntactic intuitions should consist of a questionnaire and use multiple informants and sentences that come in paradigm-like sets with multiple sentences of each type, with varying order of presentation between informants. Finally, the results should be subjected to relevant statistical tests (Cowart, 1997, 12-13). Similar recommendations for experimental work in syntax are found in Schütze (1996), Wasow and Arnold (2005), Featherston (2007), and Gibson et al. (2013). Another recommendation which reoccurs in the XSyn literature is that subjects should not know what the hypothesis being tested is, and so that linguists should not use their own intuitions as evidence (see, e.g., Wasow and Arnold 2005).

Proponents of XSyn argue that a change from the traditional method to an experimental approach as described above will bring about several improvements. Firstly, proponents of XSyn argue that adopting experimental methods will lead to more reliable data by weeding out error variance (random fluctuations across participants). Reliability in this sense means consistency across time and circumstances. By asking many participants to judge many sentences, random fluctuations cancel each other out, resulting in a higher reliability (consistency) of the results. Secondly, proponents of XSyn argue that experimental methods will lead to a higher degree of validity (ensuring we are actually investigating the phenomenon of our interest) by avoiding non-random, irrelevant effects e.g. experimenter bias, other unconscious biases, unwanted lexical effects etc. (see, e.g., Wasow and Arnold 2005; Myers 2009; Gibson and Fedorenko 2010). One way in which experimental studies can avoid such irrelevant effects lies in the design of experimental studies. For instance, by letting subjects judge multiple lexicalisations of a target structure, one can try to control for parsing issues. For example, in addition to presenting subjects with the sentence "the lawyer visited on Tuesday was a mess", which might be judged

unacceptable by subjects, one can present (the same or other) subjects with the structurally similar sentence "the factory visited on Tuesday was a mess", in which the semantics arguably blocks the wrong parse (factories cannot go on visits).[1] Thirdly, proponents of XSyn argue that experimental studies allow for obtaining more nuanced data. Featherston (2007, 275) argues that traditional methods do not cast sufficient light on cases where multiple grammatical phenomena interact or on cases where effects are relatively small. Another benefit of using the more sensitive experimental methods, according to some proponents of XSyn, is that they allow for the detection of gradience in acceptability and grammaticality (see, for instance, Featherston 2007).

Lastly, proponents of XSyn argue that a change from the traditional method to experimental methods will yield more scientific methodology and data practices in general. Some commonly mentioned aspects of this more scientific practice are objectivity, rigour, and transparency (Ferreira 2005; Cowart 1997; Myers 2009; Gibson, Piantadosi, and Fedorenko 2013). We can thus summarize the motivations of XSyn, as compared to more traditional methods:

1. Better reliability: less error variance and noise in the data;
2. Better validity: less theoretical bias and fewer irrelevant factors;
3. Higher sensitivity and richer data: better detection of more aspects of the phenomena;
4. Overall better, more scientific, methodology.

In the following section we shall discuss whether these motivations are well-founded.

# 3 Is XSyn methodologically superior?

The formal methods of data collection preached by the XSyn movement have not fallen on deaf ears, and even for those sympathetic to traditional methods, they have become important tools of investigation (Sprouse 2015). However, informal methods of collecting linguistic intuitions are still very much predominant in syntactic research.[2] This could of course be due to purely pragmatic reasons. That is, linguists might stick to informal methods because it is simply the most convenient, least time-consuming, least expensive way to obtain syntactic intuitions. In a suboptimal world, linguists would stick to traditional methods, because they are convenient, *despite experimental methods being*

[1] Thanks to an anonymous referee for bringing up this issue and for providing the examples.

[2] A search in the comprehensive database *Linguistics and Language Behavior Abstracts* reveals that in the past 10 years, there have been only 27 peer-reviewed journal articles that contain the keyword "experimental syntax" (in the English language). Note that this is likely to underestimate the use of experimental methods, as of course not all papers will necessarily use the term.

*superior*. Do we live in such a world? In order to assess this question, we will discuss each of the four aforementioned claims made by XSyn proponents.

### 3.1 Better reliability of data gathered by XSyn?

In a seminal contribution to the debate about whether the reliability of formal methods of collecting linguistic intuitions are superior to informal ones, Sprouse and Almeida (2012) tested all (469) acceptability judgements (about English sentences) found in a popular syntax textbook with a large sample of ordinary speakers. They conservatively estimated that at least 98% of the judgements of linguists and ordinary speakers converge (interpreting all failures to replicate as true negatives). Likewise, Sprouse et al. (2013) tested 148 randomly sampled judgements from a leading linguistics journal and estimated a convergence rate of 95%, with a margin of error of 5.3-5.8%. In a discussion note, Sprouse and Almeida (forthcoming) remark about both of these results:

> These high (conservative) convergence rates suggest that the sample sizes used by linguists (whatever they are) have historically introduced little error to the empirical record, either because (i) the samples are larger than what critics claim, (ii) the effect sizes are so large that small samples still yield good statistical power, or (iii) [acceptability judgement] results are highly replicated before and after publication (e.g., Phillips, 2009), or any combination of (i-iii).

Thus, it seems that the acceptability judgements obtained by informal methods from mostly linguists themselves reliably reflect the acceptability judgements obtained by more formal methods from large amounts of ordinary speakers.

Even though we consider Sprouse and Almeida's studies and conclusions convincing, one may want to know not only *that* the traditional method is reliable and valid, but also why it is. Unfortunately, there hasn't been much discussion of this question. Nevertheless, some of the widely-held views amongst generative linguists regarding the etiology of syntactic judgments could provide an explanation. More specifically, the view that native speakers are competent in their language entails that every speaker should in principle be well suited to make reliable acceptability judgements. However, it should be noted that in this justification of the reliability of acceptability judgments, there is a leap from reliable language production/comprehension to the ability to make reliable judgements *about* language. This leap is made routinely by linguists without much argument. Philosophers, however, have hotly debated this issue (see chapters X, Y, Z of this collection).

There is another widely-entertained belief amongst generative linguists which seems to be in tension with this standard justification: that of a so-called I(nternal)-language, viz. the idea that every speaker has their own, idiosyncratic grammar rules that cause (ever so slight) differences in language production. If every speaker has their own I-

language (and their own grammar) and their judgements are informative of their I-language but not anybody else's, then it is hard to see how linguists could ever be able to build a grammar of English, say. It's also not clear how speakers' judgements could be evidence for such a grammar of English, rather than just evidence *only* of their own I-language.[3]

## 3.2 Better validity?

One advantage that is often emphasised by the proponents of XSyn is that the intuitions of ordinary subjects are not as much subject to theoretical biases as the intuitions of linguistic experts (Schütze 1996, Cowart 1997, Featherston 2007, Gibson and Fedorenko 2010, 2013). There is however another set of concerns which pulls in the opposite direction with regards to the decision whether to use expert linguists or laypeople, which is sometimes underestimated by proponents of XSyn, namely pragmatic factors. For example, ordinary subjects, who for the first time in their lives are asked to provide an acceptability judgment, might be confused about the purpose of the task. They might, for example, misunderstand the task as asking for whether the sentence in question accords to their language community's etiquette. Similar concerns hold for other instructions such as "does this sound natural to you", which might be understood to ask about the relative frequency of structures like the one presented. More sophisticated strategies have been developed for making sure that ordinary speakers understand the purpose of the task presented to them (see e.g. Gibson and Fedorenko 2013, and Sprouse 2015). However, with all the cautionary steps one might take, there is a case to be made for linguists simply knowing best what acceptability judgement tasks are about.

Another set of concerns relates to performance factors such as parsing and memory constraints. Many ordinary subjects would for example judge the (by now famous) sentence "The horse raced past the barn fell" unacceptable [for easier reference, we will refer to this sentence simply as HORSE in what follows]. Linguists, on the other hand, would recognise that this judgement arises from processing limitations in the parsing of the sentence (a so-called garden path effect) and that the sentence is nevertheless perfectly grammatical. It is this judgement that the sentence is grammatical which would then be used as evidence for or against theories of grammar (rather than the plain acceptability judgement per se). Similarly, most subjects would find centre-embedded phrases like "A

[3] In chapter XX of this volume, Cowart writes that even linguists who accept the I-language view seem to report their results as being about some particular (E-)language (say, Finnish). This shows, he thinks, that they use the E-languages as convenient labels for those aspects of the grammar that the diverse I-languages within the relevant linguistic community have in common.

man that a woman that a child that a bird that I heard saw knows loves" unacceptable, even though it is perfectly grammatical (it derives from: "A man that a woman loves"). Again, a grammar consistent with this sentence would be seen as supported, not contradicted, by the linguistic evidence. As mentioned in Section 2, proponents of XSyn have claimed that performance factors "cancel out" in appropriately designed experiments that use large numbers of ordinary subjects. In order to control for performance effects in sentences like HORSE or centre-embedded sentences as the one considered earlier, one will have to vary their lexicalisation. For example, a sentence like "the paint daubed on the wall stank" (Collins 2008) is presumably parsable without problem also for people unfamiliar with the garden-path effect. This will not work for all targets, though; multiply centre-embedded sentences are cases in point. Also, the controls can only be as good as what the experts manage to come up with: the best performance of the folk is thus constrained by the ingenuity of the designers of the experiment. But the design may not be good enough to extract the grammaticality judgments of interest in a given case. And of course, even with very good controls things can still go wrong. In contrast, professional linguists, qua their training and experience, know what kinds of extraneous effects to look out for and have honed their skills with extremely many examples. Well-designed experiments may make it more likely to successfully control for confounders, but even when everything goes well, lay subjects may arguably at best do just as well as the experts.

This is not to say that there is no epistemological advantage to using laysubjects instead of linguistic experts. On the contrary, experiments with subjects without any theoretical stakes in debates about grammar and without relevant grammatical presuppositions seem to be much better suited for shielding against theoretical bias. We do not think that performance factors are a priori and intrinsically more problematic than theoretical bias is. But we do think that *both* theoretical bias and performance factors are important errors and that the risks of both must be carefully weighed against each other, as there are both costs and benefits in using linguistic experts vs. ordinary speakers.

Sprouse and Almeida have also commented on the risk of theoretical bias in acceptability judgements. They argue that if theoretical bias were a real concern, one would expect "sign reversals" between expert and naive subject populations (Sprouse and Almeida forthcoming). That is, if linguists' judgements were biased, there should be many instances in which linguists judge a sentence acceptable whereas ordinary speakers don't,

and vice versa. Again, they find no evidence for this idea in their textbook study (no sign reversals) and very few instances in journal study (1-3%).[4]

Sprouse and Almeida make another interesting point with regard to the risk of theoretical bias: although syntacticians have constructed many substantially, or even radically, different *kinds* of syntactic theories, this divergence is "rarely based on different data sets" (Sprouse and Almeida 2012, 631). Instead, they conclude: "whatever disagreements there are in linguistics literature, they appear to obtain mostly at the level of interpreting, not establishing, the data" (cf. Phillips 2009).

### 3.3 Richer data?

It is generally accepted that acceptability judgements exhibit gradience (more vs. less acceptable) rather than categoricity (acceptable vs. unacceptable). It is also accepted that XSyn is a good means for revealing gradience. Yet, it is controversial whether gradience in acceptability indicates real degrees of grammaticality, or whether acceptability judgements exhibit gradience *despite* a categorical grammar (see also Haider 2007, Fanselow 2007, Phillips 2009).[5] Given the path linguistics has taken, it would seem that most linguists believe the second disjunct: although Chomsky himself initially thought that his theories could and should capture degrees of grammaticality (Chomsky 1957, 1964), and although there have been several attempts to develop grammars which do allow for grammatical gradience (Bard, Robertson, and Sorace 1996; Sorace and Keller 2005; Keller 2000; Featherston 2005; Fanselow et al. 2006), gradience has not played a major role in "mainstream" theories of grammar .[6]

In order to detect gradience in acceptability judgments, non-categorical scales must be used. The seemingly most straightforward way to do so is by using Likert scales (with *n*-points). The use of such scales, however, is intrinsically problematic. First, there is a possibility that for any chosen *n*, subjects have gradience intuitions which are larger than

[4] A referee of this paper remarked (rightly we think) that avoiding sign reversals may be too low a standard. In other disciplines like psychology, the fact that effects do not replicate or are just weaker than the original results is already reason for concern. The equivalent in linguistics, we suppose, would be weaker acceptability judgements.

[5] It is interesting to note that even judgements about clearly categorical concepts, such as number oddity, yield *stable* gradience judgments (see Gleitman and Gleitman 1983).

[6] In a recent survey by one of us (Brøcker 2019), generative linguists were asked whether gradient results in acceptability judgment experiments could be due to a graded grammar or whether such results are more likely due to extra-grammatical factors. There was no significant difference in the frequency with which each option was chosen. This result seems somewhat at odds with the fact that categorical grammars are so prevalent in the literature. As the discussion in the rest of this section indicates, though, there may be good reasons for this prevalence.

*n*. Second, there is no guarantee that the distances between the n points is equi-distant: subjects might e.g. use the scale-difference between 1 and 2 in such a way that it implies a larger (psychological) distance than the scale distance between 3 and 4.

Bard et al. (1996) were the first to suggest the application of so-called *magnitude estimation* to acceptability judgements. In magnitude estimation, which were developed in psychophysics, subjects are presented with a stimulus (such as a light source of a certain brightness) to which they are asked to assign a standard (e.g. "100"). They are then asked to assess other stimuli of the same kind in comparison to the standard. For example, if a subject believes that a light source is twice as strong as the standard, it would be assigned the value 200. The advantage of magnitude estimation is apparent: in contrast to Likert scales, subjects can choose the grain of the scales for themselves and the distances between the units are stably defined in terms of the standard. The indefinite grain of magnitude estimation presumably is much better suited for reflecting subjects' own degrees of gradience (which may vary).

Although magnitude estimation seems a powerful tool for probing acceptability judgments, critics have noted that the gain in sensitivity with magnitude estimation as compared to other tasks is effectively insubstantial, as the results of several studies using magnitude estimation have been shown to be representable equally well with Likert scales (Bader and Häussler 2010; Weskott and Fanselow 2011, 2008). More problematically, Sprouse (2011) has demonstrated that subjects fail to make the ratio judgments required by the commutativity assumption for stimuli that underlies magnitude estimation ($p*(q*X) \approx q*(p*X)$, where X is the standard, and p and q multiples relating the standard to other stimuli). Sprouse speculates that "acceptability judgements may not have a true zero point representing the absence of all acceptability the way that physical stimuli such as loudness have a true zero point representing the absence of all sound" (285). Thus, given Sprouse's results, magnitude estimation seems inapplicable to acceptability judgments in linguistics. But since both magnitude estimation and Likert scales have their specific problems as methods of detecting gradience in acceptability judgements, experimentalists who argue for the relevance of gradience will have to either find ways to ameliorate these problems, provide arguments for why gradience may be used despite these shortcomings, or develop new, more appropriate methods for detecting gradience.

Another problem in the works of some proponents of XSyn is the relation between acceptability judgements and grammaticality. Featherston (2007), for example, distinguishes between three types of grammaticality judgments: judgments of "perceived well-formedness" (essentially, subjects' acceptability judgements using magnitude estimations), "traditional binary grammaticality judgement[s]" (which he considers to

reflect relative frequency in language use), and judgements according to some theoretical notion of grammaticality, which "is dependent on linguistic knowledge and related to particular assumptions about what a structure *should* be like" (294-5; emphasis added). Featherston (2007) restricts his analysis to the first type of judgment. It is however not at all clear that these types of judgements really can be taken to be informative of grammar rather than just acceptability. If they are just acceptability judgements, then results showing that subjects use graded judgements do not directly challenge traditional categorical grammars. In fact, it is not obvious that this question would be solvable empirically. Featherston suggests one should heuristically preclude only those acceptability judgements from grammar building "which can be accounted for by known performance or processing factors" (312; see also Keller 2000 and Schutze 1996).[7] Featherston's suggestion would implausibly enlarge the set of grammaticality judgments, as often the underlying psychological mechanisms for acceptability judgements (that do not reflect grammaticality) are ill-understood. Waiting for psychology to sort out these mechanisms would bring linguistics to a grinding halt. It thus seems indispensable for the practice of linguistics to use *theoretical* linguistic considerations in order to disambiguate acceptability from grammaticality.

For theoretical considerations to be used successfully to assess acceptability judgements, these considerations must have some normative force. Consider again HORSE: even though the sentence appears unacceptable, it is widely regarded to be grammatical. In order to make grammaticality judgments that "correct" acceptability judgements in this way, one needs *normative* grammatical theories: sentence X *ought to be* grammatical by the lights of well-confirmed grammatical theory T, despite X appearing unacceptable. But it's not clear that theories of grammatical gradience with such normative force exist. On the contrary, the theories of grammatical gradience that have hitherto been developed seem to lack it, as they take as input unfiltered acceptability judgements.[8]

---

[7] See Keller (2000, 29): "Given that no systematic performance explanation for gradience is available, we will work on the assumption that gradience is best analyzed in terms of linguistic competence". Schutze (1996, 6ff) also recommends to first get a better grasp of performance factors by building models of those *before* using acceptability judgments in the construction of grammars.

[8] Most theories of grammatical gradience are based on versions of optimality theory, in which a 'competition' between candidate structures selects one candidate as optimal / grammatical, when it best satisfies multiple grammatical constraints. In its most advanced form, namely linear optimality theory, this approach comes with a learning algorithm which estimates weights for the grammatical constraints it takes as input. These weights are computed on the basis of training sets which contain candidate structures which are associated with a grammaticality score. The algorithm then determines an optimal set of constraint

Although the use of theoretical considerations about the underlying grammar for assessing the data can of course be problematic (see the previous section), it need not be. There are situations in which theoretical bias can be methodologically positive, namely when the theory generating the bias is a theory that has independent empirical support. For example, when it was found a few years ago that neutrinos travel faster than the speed of light, it was prudent of physicists to exercise scepticism towards this result, given that it contradicted one of the most well-confirmed theories in modern physics, namely Einstein's special theory of relativity (Schindler 2013).

Already Chomsky (1965) expressed doubt that there ever could be any "operational criteria" for determining grammaticality and believed that sometimes ambiguous cases should be disambiguated by accepted grammars (see also Schutze 1996, 22f.). Although XSyn proponents are not necessarily out for such operational criteria, they still would like to obtain more theory-neutral, empirically-driven, ways of discerning grammaticality from acceptability. While the search for operational criteria is not to be scoffed at, it is questionable whether this can be done *only* from the bottom-up, so to say, by first determining performance factors in order to determine grammaticality, as the XSyn proponents suggest. In fact, there are reasons to think that such an approach would set the bar too high. As we shall see in the next section, not even the model science of physics can be said to satisfy it.

### 3.4 More scientific?

Even though XSyn's call for systematic and controlled experiments seems prima facie much more scientific than the more informal ways of obtaining acceptability judgements, we've seen here that there are significant complications to such an approach. First and foremost, there are certain pragmatic reasons that may speak against using acceptability judgements of ordinary speakers as evidence for or against theories of grammar. This is partly because lay subject may need more instructions and more elaborate materials than linguists who already know the task at hand. Another reason is that, even when gathered in great numbers and systematically, acceptability judgements still have to be analysed by linguists before we can say whether the judgements are likely to be due to grammaticality

---

weights for a given training set (see Keller 2000 and Sorace and Keller 2005 for details). The weights ideally represent grammatical gradience. The problem, though, is that if the grammaticality score used in the determination of the weights comes from acceptability judgements (as they do in XSyn experiments), then gradience in the form of weights of grammatical constraints may reflect merely extra-grammatical processing constraints. Again, additional arguments are required for concluding that the gradience in judgements actually reflects degrees of grammaticality. It should also be noted that there is a risk of "overfitting" the weights of the constraints to the used training set so that the grammaticality models become poor predictors of 'unseen' / new data, as appreciated by Keller (2000, 272).

or to some extra-grammatical factors. These considerations must be weighed against any possible advantage in guarding against theoretical bias, on which XSyn has put much focus. Although XSyn proponents have emphasised the detection of gradience as a selling point of experimental methods, it is not at all clear how such gradience (which undoubtedly exists) is best and reliably to be measured, as we saw in Section 3.3. Lastly, it has been shown that informal methods produce judgments which seem highly representative of judgments obtained amongst ordinary speakers.

In sum, the case seems weak that XSyn would be more scientific and provide better data than informal methods. But are informal methods scientific in the first place? The fact that linguists' acceptability judgements are representative may suggest so. Still, one may ask again how this representativeness is achieved. One reason is surely a pedestrian one: the vetting that acceptability judgements undergo at conference and seminar presentations and by reviewers in journals (Phillips 2009).

There are other characteristics of using linguistic intuitions as evidence that deserve highlighting in a discussion of scientificity. Chomsky himself has often compared the building of grammars on the basis of acceptability and grammaticality judgments to the scientific method introduced by Galileo (Chomsky 1980; Chomsky and Saporta 1978). Botha (1982), in his review of Chomsky's talk about the 'Galilean style' of science, identifies three elements: the construction of abstract models of grammar, the mathematical (or formal) nature of models, and the belief that abstract models have a higher degree of reality than "the ordinary world of sensation". The last of these features is expressed through what Botha calls "epistemological tolerance", namely what Chomsky has described as "a willingness to set aside apparently refuting evidence" and "a readiness to tolerate unexplained phenomena or even as yet unexplained counterevidence" (Chomsky 1980, 9-10). Chomsky himself evokes Galileo's struggle to prove that Earth is a moving planet (on the basis of a number of astronomical observations) in a time when no plausible theory for terrestrial physics on a moving earth was yet available (Chomsky and Saporta 1978).

There are several examples in which Chomsky and other linguists have used epistemological tolerance (Botha 1982; Riemer 2009; Behme 2013), but one important application arguably are acceptability judgements: as we have seen also in our discussion, not all acceptability judgments give data about grammars. Although all acceptability judgements are potential evidence of grammars, an acceptability judgment might say more about certain performance factors than about the grammar itself. Acceptability judgements that conflict with (theoretically driven) grammaticality judgments are therefore to be treated with caution. Again, the line one has to tread here is a thin one, but

any theory-development requires "breathing space" from *apparent* refuting evidence, which later may turn out to be false (Lakatos 1970; Botha 1982; Feyerabend 1975). Sometimes one may even lack full explanations of why not all relevant phenomena can be accommodated by one's theory. Galileo and Newton, for example, pioneered our modern understanding of physics by discovering fundamental principles of motion *without* having a theory of so-called 'fudge factors' such as friction and air resistance. Instead, the difference between the predictions derived from those principles remained theoretically unaccounted for (cf. Koertge 1977). It would have been a tremendous loss to science had their theories been dismissed because they couldn't accommodate the phenomena in their entirety. Even the fundamental principles of more recent theories in physics often serve primarily explanatory purposes, and need to be amended with 'phenomenological' corrections in order to fully capture the phenomena (cf. Cartwright 1983). It seems ill-advised to set higher standards for linguistics in the discovery of grammatical principles and demand that linguists must first present fully-fledged theories of memory limitations and the like that account for acceptability judgments at odds with grammaticality judgements.

Despite these reservations, there are undoubtedly further seemingly more scientific benefits of XSyn. Some proponents of experimental methods argue that formal/quantitative methods, in addition to providing data, also provides us with information *about* the data. This meta-information ideally allows us to infer whether a specific result is trustworthy or not. For example, Myers (2009) points out that formally collected quantitative results are usually reported with a measure of statistical significance, which, "in turn, is related to the probability of future replications" (Myers 2009: 409). That is, the result of a formal judgement experiment comes with a measure of the experiments' reliability. This is not the case for informal judgment experiments. However, it should be emphasised that such experiments and reliability measures are first and foremost about acceptability judgments. That is, even with a perfectly reliable experiment and very stable and repeatable data, theoretical reasons can weigh in so strongly that the produced data have no bearing whatsoever on the grammatical theories in question. The produced data would then be data about parsing or memory limitations and the like.

Another measure that comes with experimental methods is effect size. Sprouse and Almeida (2013) show that most of the phenomena in their samples have very large effect sizes -- a result only to be obtained with experimental methods (Sprouse and Almeida 2012). One *could* take this result as *generally* legitimizing using only a handful of subjects when (informally) collecting intuitive judgements. Roughly speaking, if the effect size is

larger, fewer subjects are needed, and vice versa. Sprouse and Almeida seem to suggest as much when they point out that the phenomena of the textbook and journal articles they used are highly representative of the kinds of linguistic phenomena discussed by linguists in general. Yet, Gibson et al. (2013) deny this. Instead, they claim that "cutting edge" and "forefront" syntactic research often debates more exotic phenomena not covered by the examples investigated by Sprouse and Almeida. Against such a claim, Phillips (2009) has argued that in the vast majority of cases it is not the phenomena that are controversial, but rather the interpretations of the phenomena (as grammatical or not).

Still, Gibson et al. are not satisfied with the idea that the studies by Sprouse and Almeida have established for once and for all that the effect sizes studied by linguists are big. Instead, they demand that the effect sizes are *always* checked experimentally, and not just estimated from the armchair. They argue that even if a large majority of phenomena studied in syntax have large effect sizes, this does not help the researcher investigating a new phenomenon. Only once a quantitative study has been done for *that particular phenomenon* can one confidently make assertions about the effect and sample sizes for that phenomenon.

Another potential benefit of using formal methods is that with formal methods come more regimented procedures for collecting and analyzing data. For instance, it is not standard practice for those who use the informal method to report, e.g. how many colleagues they consulted for their judgements, how many agreed or disagreed with the judgement presented by the author, which lexicalisations they considered, if they discarded any particular lexicalisations, and why. Myers (2009) argues that this kind of transparency about the procedure that lies behind results produced with the informal method would allow readers to more easily judge the trustworthiness of those results (this is especially relevant in cases where the reader is not a native speaker of the language the judgments are about).

Others point out that a scientific test should first and foremost give valuable insights, not necessarily be maximally methodologically rigorous or objective according to some standard (Grewendorf 2007). However, as argued above, there *are* relevant insights to be gained from using more rigorous methods, at least potentially: insights *about* the data. Whether those insights are worth the added effort might vary with how controversial the judgements are, as well as the potential grammatical importance of the phenomenon in question.

# 4 Experimental Philosophy: common motivations

The practice of using one's intuitions as evidence for theorizing can also be found in philosophy. In the so-called method of cases, philosophers consider hypothetical scenarios, make judgements about these scenarios, and draw conclusions for their theories about the mind, language, knowledge, ethics, etc. These scenarios are often simply referred to as "cases" or thought experiments. Moreover, just like linguists, philosophers have started to systematically test the intuitions of non-philosophers ("the folk") (Weinberg, Nichols, and Stich 2001; Machery et al. 2004).

As an example of the method of cases, consider Goedel cases. In these cases, it is imagined that it wasn't Kurt Goedel, but rather a man called Schmidt who proved the incompleteness theorem of arithmetics. But Schmidt died under mysterious circumstances and Goedel somehow got hold of the proof and successfully claimed credit for it. The common intuition is that the name "Goedel" refers to the person who got hold of the proof and claimed credit for it, not the person who discovered the proof, despite the fact that the proof had come to be credited to Goedel, not Schmidt. Such cases were first brought up by Kripke in his famous *Naming and Necessity*, which is widely viewed as making a convincing case for so-called "causal" theory of reference. The method of cases is used throughout many areas of philosophy, such as the philosophy of mind , epistemology, moral philosophy, philosophy of language, metaphysics, and others. Famous thought experiments are Gettier cases, Mary's room, Searle's Chinese Room, fake barn cases, trolley cases, split brain cases, and more (Brown and Fehige 2017). Because judgments made in cases like these appear fairly immediate, they are often referred to as "intuitions", although several philosophers prefer not to describe them in terms of their subjective phenomenology (Williamson 2007, 2011; Machery 2017). We shall here adopt Machery's minimalist conception of intuitive judgements as simply "case judgements", i.e., judgements made in cases (Machery 2017).

There have been many attempts to justify the reliability of case judgements theoretically: as a priori necessary judgements (Bealer 1998), as possibility judgements (Malmgren 2011), as counterfactual judgments (Williamson 2007), or as quasi-perceptual mental states (Bengson 2015; Chudnoff 2013, 2011). Unlike in linguistics, however, there is no agreed upon or even widely shared view of *why* the use of case judgements as evidence in philosophical reasoning might be justified. Some philosophers have even denied that intuitions play any significant role in philosophical practice (Cappelen 2012).

Experimental philosophers, like experimental syntacticians, have criticised the reliance of philosophers on their own intuitions without taking into consideration the

intuitions of the folk. Several of the motivations of experimental philosophers mirror those of the experimental syntacticians. Machery and Stich (2012), for example, emphasise the risk of theoretical bias in philosophers' practice of using their own intuitions to assess theories of linguistic reference in the aforementioned Goedel cases; they are concerned that philosophers of the Kripkean persuasion (the majority of analytic philosophers) are biased toward the judgement consistent with the causal theory of reference.

In one of the first XPhi studies, Machery et al. (2004) presented evidence that Chinese undergraduate students (of unspecified field of study) have intuitions in Goedel cases that surprisingly align with descriptivist theories of reference (which in the Goedel case would be that "Goedel" refers to the person who actually discovered the proof). Machery and Stich (2012) see this study as demonstrating that philosophers cannot blindly rely on their own intuitions when constructing theories of reference and cite this study as a good example of the kind of method that philosophers should follow. They conclude by posing a dilemma to philosophers of language: either ordinary speakers' intuitions matter for theories of reference or they don't. In the former case, theories of reference must be "substantially modified to accommodate the variation in reference determination" (509). They cite Reimer's (2009) attempt to do so approvingly. In the latter case, they claim, philosophers' intuitions are not sufficient to justify the assumption that "proper names have a semantic reference", rather than just "speaker's reference", i.e., reference *intended* by the speaker in communication. In support of this claim, they cite Chomsky and other linguists and philosophers of language, who have expressed scepticism towards the existence of semantic reference (cf. Chomsky 2000). They also point out that the intuitions of philosophers, in contrast to the intuitions of experts in other fields, are not externally validated (as e.g., a doctor's intuitive judgement that the collarbone is broken can be externally validated by taking an x-ray picture).

Machery and Stich explicitly appeal to XSyn to support their conclusion and to motivate the XPhi approach more generally. Apart from the risk of theoretical bias, which is also a central motivation of experimental syntacticians, Machery and Stich mention the risk of the traditional method of cases ignoring the "diversity of intuitions" and compare this to the risk of ignoring "dialectical variation" in syntactic intuitions amongst non-linguists. In philosophy, one could for example be interested in potential differences in intuitions of subjects from different cultural backgrounds, differences of intuitions between men and women, etc. Such differences are arguably better investigated by experimental means, and not from the armchair.

However, Machery and Stich's analogy between such potential differences and dialectal variation in syntax does not seem entirely apt. First, in contrast to the traditional

practice in philosophy, syntacticians have investigated idiolects with traditional, informal methods. Second, contrary to how Machery and Stich make it sound, detecting dialectical variation has not played a major role in XSyn. In fact, one of the critiques launched against the experimental syntacticians by the traditionalists is that using large samples of speakers and averaging across the results puts one at risk of overlooking individual variation (Den Dikken et al. 2007; Fanselow 2007).

Although dialectical variation is not a concern of current XSyn, there is another motivation that drives proponents of XSyn, which seems exploitable for Machery and Stich's purposes, at least in principle. This concerns gradience in acceptability judgments. Detection of gradience in the judgement of ordinary speakers, just like the detection of diversity of philosophical intuitions, requires the systematic investigation of larger samples of judgements.

Hales (2006), Williamson (2007, 2011), Ludwig (2007), Devitt (2011, 2006), Horvath (2010), and others have argued, contrary to XPhi proponents, that the intuitions obtained from the folk do not directly bear on theories of reference, knowledge, and other objects of philosophical interest. The reason these critics cite is that the intuitions of non-philosophers are less reliable than the ones of philosophers. Philosophers, it is claimed, are simply better trained in processing the vignettes of thought experiments, in considering possible scenarios far removed from the actual world, etc. Alternatively, it has been argued that philosophers are better at applying concepts such as knowledge, belief, understanding, etc. to the scenarios described. That philosophers are better subjects for thought experiments is also known as the *expertise defense*. Although the equivalent of the expertise defense in linguistics (namely that professional linguists are better subjects for making acceptability judgements than ordinary speakers) has not been entertained by anybody in the debate about XSyn, it has been mentioned occasionally as a possible position (Gibson and Fedorenko 2013, Sprouse 2015) and has been defended by Devitt (see Chapter X of this collection).

Finally, proponents of XPhi, just like champions of XSyn, have claimed that the methods of XPhi are more scientific than the informal ones, and therefore also more reliable. Alexander and Weinberg (2014), for example, write that "philosophers need to continue to *improve* the methods used to study philosophical cognition, combining survey methods with more advanced statistical methods and analyses, and supplementing survey methods with a wider variety of methods from the social and cognitive sciences … too many questions pertinent to evaluating the trustworthiness of epistemic intuitions can only be addressed properly with some substantial reliance on *scientific* methods" (138ff.; added emphasis).

In sum, the four motivations driving XSyn can also be found in XPhi (more reliability, less theoretical bias, higher sensitivity and richer data, overall better, more scientific methodology). Already in Section 3 we saw that none of the motivations of XSyn would justify an all-round claim to superior methodology of XSyn. In what follows we will try to draw the appropriate lessons for XPhi by assessing the analogies that have been made between XPhi and XSyn.

# 5 Lessons for XPhi from XSyn

Machery and Stich (2012) have argued that "philosophers should emulate linguists, who are increasingly replacing the traditional informal reliance on their own and their colleagues' intuitions with systematic experimental study of ordinary speakers' intuitions" (495). In the face of the more accurate picture of the controversial value of experimental methods in syntax we presented here, XSyn cannot serve as a model for the promotion of XPhi, at least not unconditionally.

## 5.1 Better reliability of data gathered by XPhi?

In contrast to linguistics, where the debate has centred on the question of whether formal methods are more reliable than informal ones, the shape the debate has taken in philosophy is rather different. Although XPhi practitioners presume that their method is a more reliable and constitutes a more scientific way of collecting case judgments, they do not claim that the data obtained by these methods are more reliable than the case judgments by philosophers. On the contrary, they have argued on the basis of the results obtained by XPhi methods (with mostly subjects without philosophical training) that case judgements *per se* are unreliable (Alexander and Weinberg 2007, 2014; Machery 2017). More specifically, they argue that intuitive judgments (found in the folk) vary with factors which seem extraneous or irrelevant to the task at hand, such as demographic variables (e.g., culture, gender) or presentation effects (e.g., order of presentation) (ibid.). Several XPhi proponents have argued that case judgements should be used only with extreme caution -- if at all (Weinberg et al. 2001; Alexander and Weinberg 2007, 2014; Machery 2017).

As mentioned in Section 4, proponents of the traditional method have questioned XPhi's sceptical conclusions by criticising the use of subjects who they deem unqualified, namely subjects without philosophical training. Proponents of this expertise defense have demanded that experimental philosophers provide proof that not only the folk, but also professional philosophers are subject to such extraneous effects (Williamson 2011). XPhi proponents, in turn, have questioned the very idea that philosophers possess expertise for

making case judgements and that philosophers' judgements would not be subject to extraneous factors, even if they did possess an expertise in making these judgements (Weinberg et al. 2010, Machery 2017). Experimental philosophers have also questioned why it is philosophers who should be trusted rather than the folk in cases of disagreement (Machery 2017). Machery et al. (2004) therefore believe that the expertise defense "smacks of narcissism in the extreme".

In attempts to break this stalemate, experimental philosophers have started to do experiments with professional philosophers themselves. These experiments have presented evidence that even professional philosophers are subject to extraneous effects (Schwitzgebel and Cushman 2015, 2012; Tobia, Buckwalter, and Stich 2013; Schulz, Cokely, and Feltz 2011). But because these experiments have been carried out mostly in the realm of moral philosophy, and because intuitive case judgements may not form a single kind but instead form a motley bunch and be reliable to different degrees (Nado 2014), it is still a live option that philosophers may be better judges of cases in other areas of philosophy.

Lastly, in support of traditional judgements in epistemology, and in particular Gettier cases, it has been shown experimentally that the folk's intuitions actually do converge with the ones made by philosophers when the folk are guided in comprehending the different steps required for making a case judgement (Nagel, San Juan, and Mar 2013; Turri 2013).[9] This would seem to suggest that the reasons for previous results indicating non-standard responses in the folk may have to be sought in extraneous factors such as imperfect understanding of the vignette.

## 5.2 Better validity of XPhi?

Is the risk of theoretical bias a good reason for adopting experimental methods in philosophy? Like in XSyn, we believe, there is not only a risk of theoretical bias in the "traditional" method, but there is also a risk of performance errors in using laypeople. That is, just like in XSyn, there is a risk of naive subjects not understanding the task instructions properly and/or making mistakes resulting from the lack of frequent exposure to philosophical thought experiments. The problem of speaker's reference (discussed by Machery and Stich 2012) is just one example. That is, instead of understanding that the task is about the (fixed) reference of proper names, subjects might take the task to be asking about who *they* intend to refer to. But as we mentioned in the previous section, experimental philosophers have identified a whole range of extraneous variables with which intuitive judgements *of the folk* vary. Again, even though some studies show that

[9] Even outright armchair critics have experimentally confirmed some standard philosophical judgements in cases (Machery et al. 2017).

philosophers' intuitions also vary in this problematic way, there is still a lot of work to be done to show that philosophical case judgements are unreliable *in general,* for both the folk and the philosophers.

Again, we do not believe that the risk of theoretical bias is obviously more severe than the risk of performance errors. On the contrary, it seems, that theoretical bias is more easily controlled for within the standard practice of philosophers. Philosophers are known for holding often radically different views and for defending them ferociously. It thus seems unlikely that their theories would bias their judgements in such a way that these judgements would converge *because* of the theories philosophers hold. Indeed, we think that something analogous to XSyn is true: philosophers tend to agree on many judgements in thought experiments *despite* the fact that there is such a diversity of philosophical views. For example, there is a plethora of views regarding the possibility of strong AI in Searle's famous Chinese Room argument.[10] Yet, all parties of the debate seem to agree with the judgement that the man in the room does not understand Chinese, even though they might disagree that the entire system of the room, or variations of the room, exhibits understanding. Although there exists no systematic survey amongst philosophers that would show that there is a consensus in the judgements of important thought experiments[11], even critics of the "method of cases" concede that there are very robust case judgements, which are widely shared in the philosophical community (Machery 2017).[12]

In sum, just as in linguistics, performance errors must be a concern for those advocating the use of non-philosophers in their desire to forego the risk of theoretical bias when using experts. And theoretical bias seems no great *actual* risk, given that philosophers with radically different views often share the same intuitions.

### 5.3 Richer data?

Although Machery and Stich (and earlier Machery et al. 2004) advertise the detection of cultural differences as good reason for adopting experimental methods -- as mentioned --

---

[10] In Searle's Chinese Room 'argument', it is imagined that a man without knowledge of the Chinese language whatsoever sits in a closed room and receives strings of Chinese symbols from one end and outputs other strings of Chinese symbols on the other end on the basis of symbol manipulations detailed in a big rule book of Chinese. Even though the outputs would appear perfectly well-formed (and meaningful) to a Chinese speaker located outside the room, the man in the room does not understand Chinese. Searle used this thought experiment to argue against Strong AI: although machines may be capable of perfectly simulating (linguistic) intelligence by syntactical manipulations, they lack the semantics required for true intelligence.

[11] However see our recent survey (Schindler and Saint-Germier preprint).

[12] The survey by Bourget and Chalmers (2014) provides data for philosophers' judgements on some thought experiments, including the Zombie argument and trolley cases. These seem to be less stable.

much of the XPhi proponents' work has followed a different agenda. This agenda is largely "negative", in the sense that XPhi proponents have used their experiments to argue for the unreliability of case judgments. Thus, many of the effects detected by XPhi seem to be of a kind that is generally deemed irrelevant for philosophical purposes by proponents of experimental and traditional methods alike (Alexander and Weinberg 2007). In that sense, these effects are akin to gradience in linguistic acceptability judgments which are not generally agreed to be of relevance to theories of grammar (although some proponents of XSyn disagree). But there is also work in a more positive vein, which is interested in the psychological mechanisms underlying case judgements (Knobe and Nichols 2008) and in better understanding the (confounding) factors that might contribute to our making judgements in a particular way (Mortensen and Nagel 2016; Alexander and Weinberg 2014). This use of experimental methods in philosophy, we think, may help improve the evidential base of philosophy and our understanding of intuitive judgements. The goal of this approach is not so much to replace the traditional armchair method but to understand how it can be put to use and when it is safe to apply it, and it shows how both formal and informal methods can be used alongside each other.

### 5.4 More scientific?

Just like in linguistics, it seems that methods should be chosen for the purposes they serve best. Whether the best method is the traditional "armchair" method or XPhi may have no unequivocal answer. At the very least, it depends on the purpose that is pursued. If the purpose is to obtain case judgements that are least subject to performance factors, then, just like in linguistics, armchair methods might be more suited (as the proponents of the expertise defense have argued). On the other hand, if one is interested in what the folk think (as Machery and colleagues have been arguing we should), then XPhi does seem the method of choice.

## 6 Conclusion

In this chapter, we argued that although XSyn and XPhi share many motivations (better reliability, better validity, richer data, more scientificity / objectivity), claims about the intrinsic superiority of these experimental approaches to collecting intuitions in these two fields are not justified. Although there is a case to be made for experimental methods being better suited for controlling theoretical bias, there are other errors (specifically performance errors) for whose control the use of experts seems more advantageous. As we have suggested in section 3, this is prima facie compatible with the idea that in some cases experimental methods may be helpful in uncovering and countering some forms of

performance error. Formal methods should thus not be used blindly and any method should be assessed for what it can achieve for the purpose at hand.

## References


Alexander, Joshua, and Jonathan M. Weinberg. 2007. "Analytic Epistemology and Experimental Philosophy." *Philosophy Compass* 2 (1): 56–80.

———. 2014. "The 'unreliability' of Epistemic Intuitions." In *Current Controversies in Experimental Philosophy*, edited by Edouard Machery and Elizabeth O'Neill, 128–45. New York: Routledge.

Bader, Markus, and Jana Häussler. 2010. "Toward a Model of Grammaticality Judgments." *Journal of Linguistics* 46 (2): 273–330.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. "Magnitude Estimation of Linguistic Acceptability." *Language* 72 (1): 32–68.

Bealer, George. 1998. "Intuition and the Autonomy of Philosophy." In *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, edited by Michael DePaul and William Ramsey, 201–40. Lanham: Rowman & Littlefield.

Behme, Christina. 2013. "Noam Chomsky , The Science of Language. Interviews with James McGilvray . Reviewed by." *Philosophy in Review* 33 (2): 100–103.

Bengson, John. 2015. "The Intellectual given." *Mind; a Quarterly Review of Psychology and Philosophy* 124 (495): 707–60.

Botha, Rudolf P. 1982. "On 'the Galilean Style' of Linguistic Inquiry." *Lingua. International Review of General Linguistics. Revue Internationale de Linguistique Generale* 58 (1): 1–50.

Bourget, David, and David J. Chalmers. 2014. "What Do Philosophers Believe?" *Philosophical Studies* 170 (3): 465–500.

Brøcker, Karen Kiil. 2019. "Justifying the Evidential Use of Intuitive Judgements in Linguistics." PhD.

Brown, J. R., and Y. F. Fehige. 2017. "Thought Experiments." Edited by Edward N. Zalta. *The Stanford Encyclopedia of Philosophy (Summer 2017 Edition)*. https://plato.stanford.edu/archives/sum2017/entries/thought-experiment/.

Cappelen, Herman. 2012. *Philosophy without Intuitions*. Oxford: Oxford University Press.

Carson, Sprouse Jon Schütze. 2013. "Judgment Data." In *Research Methods in Linguistics*, edited by R J Podesva, 27–50. Cambridge: Cambridge University Press.

Cartwright, Nancy. 1983. "How the Laws of Physics Lie." Oxford: Oxford University Press.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Massachusetts: MIT Press.

———. 1980. *Rules and Representations*. New York: New York University Press.

———. 2000. *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

Chomsky, Noam, and Sol Saporta. 1978. *An Interview with Noam Chomsky*. Vol. 4. Working Papers in Linguistics . Seattle: Department of Linguistics, University of Washington.

Chudnoff, Elijah. 2011. "The Nature of Intuitive Justification." *Philosophical Studies* 153 (2): 313–33.

———. 2013. "Intuitive Knowledge." *Philosophical Studies* 162 (2): 359–78.

Cowart, Wayne. 1997. *Experimental Syntax*. SAGE.

Den Dikken, Marcel, Judy B. Bernstein, Christina Tortora, and Raffaella Zanuttini. 2007. "Data and Grammar: Means and Individuals." *Theoretical Linguistics* 33 (3): 699.

Devitt, Michael. 2006. “Intuitions in Linguistics.” *The British Journal for the Philosophy of Science* 57 (3): 481–513.
———. 2011. “Experimental Semantics.” *Philosophy and Phenomenological Research* 82 (2): 418–35.
Fanselow, Gisbert. 2007. “Carrots--Perfect as Vegetables, but Please Not as a Main Dish.” *Theoretical Linguistics* 33 (3): 353–67.
Fanselow, Gisbert, Caroline Féry, Matthias Schlesewsky, and Ralf Vogel, eds. 2006. *Gradience in Grammar: Generative Perspectives*. Oxford: Oxford University Press.
Featherston, Sam. 2005. “The Decathlon Model of Empirical Syntax.” *Linguistic Evidence*, 187–208.
———. 2007. “Data in Generative Grammar: The Stick and the Carrot.” *Theoretical Linguistics* 33 (3): 1.
Ferreira, Fernanda. 2005. “Psycholinguistics, Formal Grammars, and Cognitive Science.” *The Linguistic Review* 22 (2-4). https://doi.org/10.1515/tlir.2005.22.2-4.365.
Feyerabend, Paul. 1975. *Against Method*. London: Verso.
Gibson, Edward, Steven T. Piantadosi, and Evelina Fedorenko. 2013. “Quantitative Methods in Syntax/semantics Research: A Response to Sprouse and Almeida (2013).” *Language and Cognitive Processes* 28 (3): 229–40.
Grewendorf, Günther. 2007. “Empirical Evidence and Theoretical Reasoning in Generative Grammar.” *Theoretical Linguistics* 33 (3): 383.
Hales, Steven D. 2006. *Relativism and the Foundations of Philosophy*. Cambridge, MA: MIT Press.
Horvath, Joachim. 2010. “How (not) to React to Experimental Philosophy.” *Philosophical Psychology* 23 (4): 447–80.
Keller, Frank. 2000. “Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality.” University of Edinburgh.
Knobe, Joshua, and Shaun Nichols. 2008. “An Experimental Philosophy Manifesto.” In *Experimental Philosophy*, edited by Knobe Joshua Nichols Shaun, 3–14. New York: Oxford University Press.
Koertge, Noretta. 1977. “Galileo and the Problem of Accidents.” *Journal of the History of Ideas* 38 (3): 389–408.
Lakatos, Imre. 1970. “Falsification and the Methodology of Scientific Research Programmes.” *Criticism and the Growth of Knowledge* 4: 91–196.
Ludwig, Kirk. 2007. “The Epistemology of Thought Experiments: First Person versus Third Person Approaches.” *Midwest Studies in Philosophy* 31 (1): 128–59.
Machery, Edouard. 2017. *Philosophy within Its Proper Bounds*. Oxford University Press.
Machery, Edouard, Ron Mallon, Shaun Nichols, and Stephen P. Stich. 2004. “Semantics, Cross-Cultural Style.” *Cognition* 92 (3): B1–12.
Machery, Edouard, and Stephen Stich. 2013. “The Role of Experiment.” In *Routledge Companion to Philosophy of Language*. Routledge.
Machery, Edouard, Stephen Stich, David Rose, Amita Chatterjee, Kaori Karasawa, Noel Struchiner, Smita Sirker, Naoki Usui, and Takaaki Hashimoto. 2017. “Gettier across Cultures.” *Noûs* 51 (3): 645–64.
Malmgren, Anna-Sara. 2011. “Rationalism and the Content of Intuitive Judgements.” *Mind; a Quarterly Review of Psychology and Philosophy* 120 (478): 263–327.
Mortensen, Kaija, and Jennifer Nagel. 2016. “Armchair-Friendly Experimental Philosophy.” *A Companion to Experimental Philosophy*, 53–70.
Myers, James. 2009. “Syntactic Judgment Experiments.” *Language and Linguistics Compass* 3 (1):

406–23.
Nado, Jennifer. 2014. "Why Intuition?" *Philosophy and Phenomenological Research* 89 (1): 15–41.
Nagel, Jennifer, Valerie San Juan, and Raymond A. Mar. 2013. "Lay Denial of Knowledge for Justified True Beliefs." *Cognition* 129 (3): 652–61.
Phillips, Colin. 2009. "Should We Impeach Armchair Linguists." *Japanese/Korean Linguistics* 17: 49–64.
Reimer, M. 2009. "Jonah Cases." In *Empty Names*, edited by A. Everett. Oxford: Oxford University Press.
Riemer, N. 2009. "Grammaticality as Evidence and as Prediction in a Galilean Linguistics." *Language Sciences* 31 (5): 612–33.
Schindler, Samuel. 2013. "Theory-Laden Experimentation." *Studies in History and Philosophy of Science. Part A.* 44 (1): 89–101.
Schindler, Samuel, and Pierre Saint-Germier. preprint. "Putting Philosophical Expertise to the Test."
Schulz, Eric, Edward T. Cokely, and Adam Feltz. 2011. "Persistent Bias in Expert Judgments about Free Will and Moral Responsibility: A Test of the Expertise Defense." *Consciousness and Cognition* 20 (4): 1722–31.
Schütze, Carson T. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.
Schwitzgebel, Eric, and Fiery Cushman. 2012. "Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-philosophers." *Mind & Language* 27 (2): 135–53.
———. 2015. "Philosophers' Biased Judgments Persist despite Training, Expertise and Reflection." *Cognition* 141: 127–37.
Sorace, Antonella, and Frank Keller. 2005. "Gradience in Linguistic Data." *Lingua. International Review of General Linguistics. Revue Internationale de Linguistique Generale* 115 (11): 1497–1524.
Sprouse, Jon. 2011. "A Test of the Cognitive Assumptions of Magnitude Estimation: Commutativity Does Not Hold for Acceptability Judgments." *Language*, 274–88.
———. 2015. "Three Open Questions in Experimental Syntax." *Linguistics Vanguard* 1 (1): 89–100.
Sprouse, Jon, and Diogo Almeida. Forthcoming. "Setting the Empirical Record Straight: Acceptability Judgments Appear to Be Reliable, Robust, and Replicable." *The Behavioral and Brain Sciences*.
———. 2012. "Assessing the Reliability of Textbook Data in Syntax: Adger's Core Syntax." *Journal of Linguistics* 48 (03): 609–52.
Sprouse, Jon, C. Schütze, and Diogo Almeida. 2013. "Assessing the Reliability of Journal Data in Syntax: Linguistic Inquiry 2001–2010." *Lingua. International Review of General Linguistics. Revue Internationale de Linguistique Generale*. http://linguistics.ucla.edu/people/cschutze/assessing.pdf.
Tobia, Kevin, Wesley Buckwalter, and Stephen Stich. 2013. "Moral Intuitions: Are Philosophers Experts?" *Philosophical Psychology* 26 (5): 629–38.
Turri, John. 2013. "A Conspicuous Art: Putting Gettier to the Test." *Philosophers' Imprint* 13 (10): 1–16.
Weinberg, Jonathan M., Shaun Nichols, and Stephen Stich. 2001. "Normativity and Epistemic Intuitions." *Philosophical Topics* 29 (1/2): 429–60.
Weskott, Thomas, and Gisbert Fanselow. 2008. "Variance and Informativity in Different Measures of Linguistic Acceptability." In *Proceedings of the 27th West Coast Conference on Formal*

*Linguistics (WCCFL)*, 431–39. Cascadilla Press Somerville, MA.
———. 2011. "On the Informativity of Different Measures of Linguistic Acceptability." *Language*, 249–73.
Williamson, Timothy. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.
———. 2011. "Philosophical Expertise and the Burden of Proof." *Metaphilosophy* 42 (3): 215–29.