Quantitative methods in syntax & semantics research

Ted Gibson
9.59 / 24.905

Acceptability judgments

The standard method in syntax / semantics research since at least Chomsky (1957):
The researcher's own intuitions about the acceptability of different sentences.

OK: the cat
* cat the

Examples from Mahowald et al. (2016), Linguistic Inquiry 2001-2010:
37.2.Sigurdsson: OK: They would have elected me.
*There would have been me elected.

32.1.Martin: OK: Pam likes soccer, and Rebecca does too.
* I consider Pam to like soccer, and I believe Rebecca to as well.

35.3.Hazout
OK: There seem to have appeared some new candidates in the course of the campaign.
*There seems to have appeared some new candidates in the course of the campaign.

32.3.Culicover
OK: There was a promise to Susan from John to take care of himself.
*There was a promise to Susan from John to take care of herself.

Acceptability judgments

But what about:
OK: John was sleeping.
*? Mary told Bill that Fred said that Arianna believed that John was sleeping.

OK: The girl ate the pizza.
*? The pizza ate the girl.

OK: The girl ate the pizza.
?* The girl ate the dugong.

Pre-theoretically we think that there is a ***lexicon, syntax (word order/ composition rules)***, and the meaning associated with each. Consider making a comparison of sentence a vs sentence b. If we want to argue that this effect shows an effect of syntax (word order/composition rules), we need to control for other factors. That is, we need to make sure that the meaning is controlled across the two, and the words are the mostly the same, and have the same frequency (familiarity).

Non-quantitative syntax / semantics: The single-subject/single-item method

The standard method in syntax / semantics research c. 2010 and before: The researcher's own intuitions about the acceptability of different sentences.

This worked ok when the field was developing:
e.g., *the big cat* vs. **cat big the*

But as the field progressed, the materials became more complex, and judgments are more subtle
e.g., *What do you wonder who saw?* vs. *I wonder what who saw.*

Furthermore, a researcher often doesn't natively speak the language that is being documented: How to evaluate those judgments?

Non-quantitative syntax / semantics: The single-subject/single-item method

Weaknesses of the single-subject/single-item method
(e.g., Schutze, 1996; Cowart, 1997; Wasow & Arnold, 2005; Ferreira, 2005; Featherston, 2007; Myers, 2009; Gibson & Fedorenko, 2010, 2012):

- **Cognitive biases** on the part of the researcher and participants
- The non-quantitative method presupposes that there is some categorical difference between "grammatical" and "ungrammatical". What if the difference is continuous, from completely unacceptable to very acceptable? *Using the non-quantitative method, we can't find probabilistic effects or relative effect sizes, or interactions among factors*
- ***Perhaps the biggest problem***: without quantitative methods, if researchers make ***any*** judgment errors, other researchers can never know which comparisons are ok, and which are not.

- (Note: problems with the experimental design — confound with lexicon / context etc — are problems for all methods)

Non-quantitative syntax / semantics: The single-subject/single-item method

Advantages of quantitative methods (controlled experiments or corpus analyses): (from class responses)

1. The current acceptable error rate in linguistics studies is too high
2. Informal linguistic experiments (judgments) make it difficult for researchers who either are from other fields or do not speak the target language in the materials
3. Cognitive biases: experiments are conducted using the experimenters' judgement to determine what is deemed correct or preferred or more grammatical
4. Only formal experiment can give detailed information on the size of effects in an objective manner
5. It is not always obvious that a contrast is obvious
6. All the reasons for adopting quantitative methods in linguistics research presented in this paper are convincing: it's hard to choose the strongest one…

Non-quantitative syntax / semantics: The single-subject/single-item method

Advantages of quantitative methods (controlled experiments or corpus analyses):

- allow the use of inferential statistics to evaluate the relative likelihoods of alternative hypotheses
- experimental participants are naive with respect to the hypotheses
- experimenter has control over the presentation of the experimental materials (e.g., can control for context effects by randomizing the order within and across participants)
- Language is *not* binary / thresholded. There is a continuum of difficulty. Presupposing a binary judgment is a weakness
- **Biggest advantage (?)**: other researchers have quantitative information about the quality of data: quantitative details enable an understanding of which comparisons support a theory, and which do not / might not

Response I to a plea for quantitative methods in syntax/semantics (G&F 2010)

Using quantitative methods would slow down research a great deal:

"It would cripple linguistic investigation if it were required that all judgments of ambiguity and grammaticality be subject to statistically rigorous experiments on naive subjects." (Culicover & Jackendoff, 2010, p. 234).

Answer to Response 1

No: Crowd-sourcing (e.g., *Amazon.com's Mechanical Turk*) makes it easy to conduct experiments these days

- cheap, reliable, fast labor
- Free software (e.g., *Turkolizer*, Gibson, Piantadosi & Fedorenko, 2011)
- An experiment can be completed in a couple of hours

E.g., Sprouse, Schutze & Almeida (2013) tested 148 pairs of examples from *Linguistic Inquiry (*2001-2010), using 3 different methods, in a few months; and S&A tested every example from Adger's (2003) textbook

Mahowald, Graff, Hartman & Gibson (2016, Language) tested 101 further examples from *Linguistic Inquiry* (2001-2010), 2 methods, with 12 exemplars of each, in a few weeks

Response 2 to a plea for quantitative methods in syntax/semantics (G&F 2010)

Phillips (2008)

> *"In order for there to be a crisis, however, it would need to be the case that intuitive judgments have led to generalizations that are widely accepted yet bogus... Carefully controlled judgment studies would solve these problems."*

Phillips' claim: *There are **few enough incorrect judgments** in the literature such that adopting quantitative standards wouldn't solve an existing problem.*

(cf. The **biggest problem** of non-quantitative methods: If researchers make ***any*** judgment errors, other researchers can never know which comparisons are ok, and which are not.)

Answer to Phillips (2008): Judgment errors really do occur

Examples of the kind that Phillips claims do not exist (Gibson & Fedorenko, 2010, LCP; Gibson & Fedorenko, 2012, LCP):

(a) *What do you wonder who saw?*

(b) **I wonder what who saw.*

Chomsky (1986, p. 48):
(a) is more acceptable than sentences that violate the Superiority condition like (b), due to a process of "vacuous movement"

BUT: (a) is actually judged as less acceptable than (b) (Gibson & Fedorenko, 2012).

Answer to Phillips (2008): Judgment errors really do occur

Furthermore, Sprouse, Schutze & Almeida (2013) find that approximately 5% of 146 *Linguistic Inquiry* contrasts from 2001-2010 were not ratified, with 1-2% reliable in the opposite direction

Thus judgment errors really do occur.

Following up Sprouse et al.: Mahowald, Graff, Hartman & Gibson (2016)

Mahowald et al. (2016): 100 randomly sampled comparisons from SSA's *Linguistic Inquiry* set, ones that SSA did not run. (12 items / comparison; 60 participants / expt; 2 methods: forced choice; ratings)

Results: 11% of judgments do not show a sig. result in at least one of the two methods; 5% do not show a sig. expected result in both methods. In the forced choice experiment, 2 judgments are sig. in the opposite direction.

Examples:

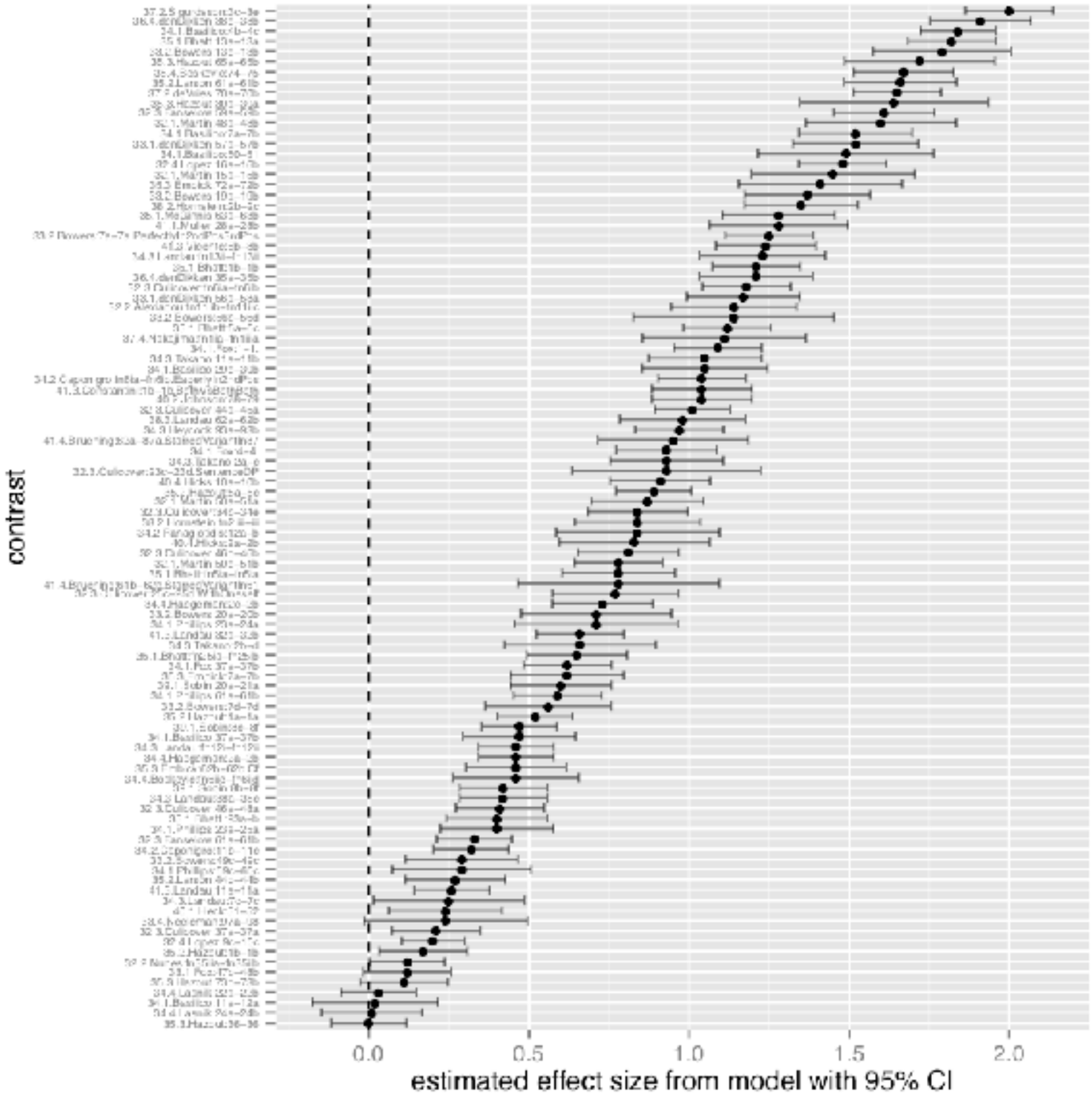
Fox (2002):
**I read something yesterday John recommended.* vs. *I read something yesterday John did.*
Hazout (2004):
*There seem / *seems to have appeared [some new candidates] in the course of the presidential campaign.*
Lasnik (2003):
?The detective asserted two students to have been at the demonstration during each other's hearings. vs.
*?*The detective asserted that two students were at the demonstration during each other's hearings.*
Nunes (2001):
We proved Smith to the authorities to be the thief. vs. **We proved to the authorities Smith to be the thief.*

Following up Sprouse et al.: Mahowald, Graff, Hartman & Gibson (2016)

Method: acceptability judgement 1-7, z-scored within individuals. obtain mean z-score for each item in each contrast, and averaged these to give an overall z-score for the 'acceptable' sentence and for the 'unacceptable' sentence in each contrast.

The effect size is the difference between these two z-scores.

(Effect size: Cohen's d is a measure of effect size that is equal to the difference in means between the two conditions, in z-scores.)



z-score: (value-mean)/ sd

Source: Mahowald, Kyle, Peter Graff, Jeremy Hartman, and Edward Gibson. "SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments." Language 92, no. 3 (2016): 619-635.

Response 3 to a plea for quantitative methods in syntax/semantics (G&F 2010)

Sprouse, Schutze & Almeida (2013): Judgment errors are too rare to matter.

A 5% error rate is the acceptable standard in cognitive psychology experiments. Therefore, this should also be acceptable in linguistics judgments.

Answer to Sprouse & Almeida (2012): 5% errors is too many errors

(1) 5% is actually **no longer standardly acceptable** in psychology experiments: many failures to replicate (e.g., Nosek et al., the Open Science Collaboration, 2015)

Abandon quantitative methods? No!

Quantitative objective replication is critical. *The error rate can then be made arbitrarily small, with more data (e.g., $p < .00001$ or smaller, for real effects).*

Answer to Sprouse & Almeida (2012): 5% errors is too many errors

(2) A p < 0.05 *false-positive* ***threshold*** for null hypothesis significance testing (NHST) in behavioral experiments is not comparable to a 5% *false-positive* ***rate*** in published acceptability judgments.

The NHST paradigm assumes that one has performed statistical significance testing for each effect; the p < 0.05 threshold is an easy way to classify the results, but it does not substitute for the quantitative information.

Furthermore a 5% error rate in linguistic acceptability judgments suggests that 5% of all judgments would diverge from the results of a formal experiment. But there is no sampling being done; the method provides no quantitative information about any individual effect.

If the average linguistics paper has thirty-three examples, divergences are uniformly distributed, and if the divergence rate is 5%, then every paper is likely to contain ~1-2 questionable judgments.

So perhaps the biggest problem with non-quantitative methods: if researchers make ***any*** judgment errors, other researchers have **no information** about which comparisons are ok, and which are not.

Judgments in other languages: Linzen & Oseki (2015)

Most readers speak / read English, so the judgment rate is likely to be better than for languages for which most readers don't speak.

Evaluation: Linzen & Oseki (2015): Hebrew & Japanese

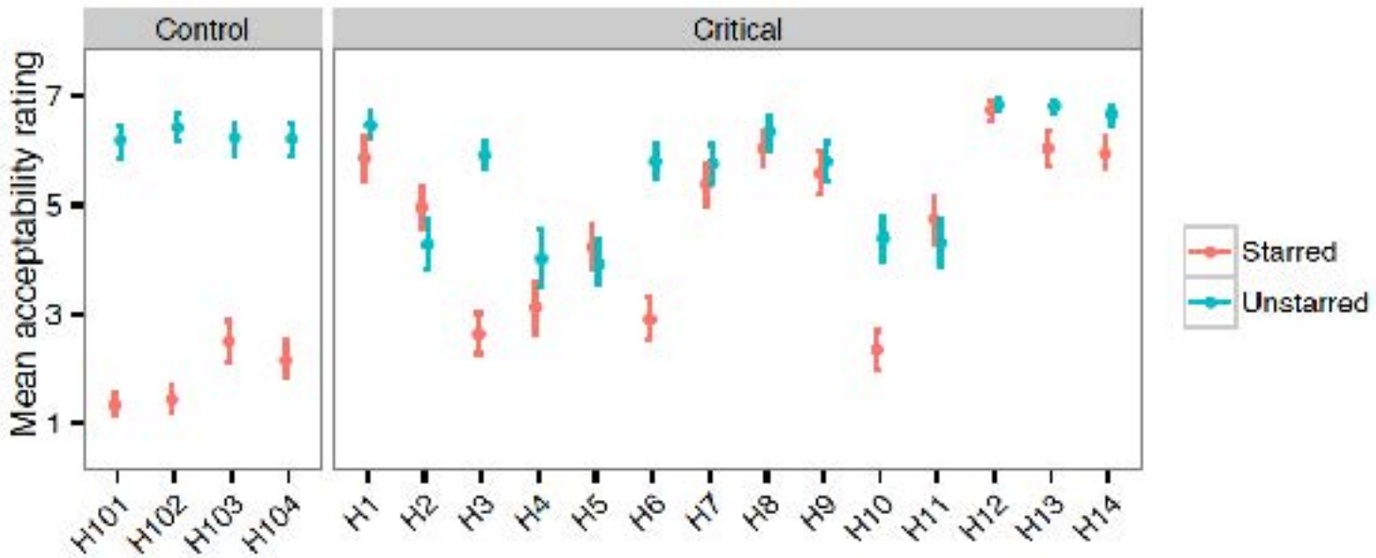
Selected 4 "obvious" control comparisons and 14 comparisons which they were less sure of in each language.

Results:
5 of 14 were reliably different in predicted direction in Hebrew;
7 of 14 were reliably different in predicted direction in Japanese

Overall, only 12 of 28 (~40%) were ratified

Judgments in other languages: Linzen & Oseki (2015)

Hebrew



Japanese

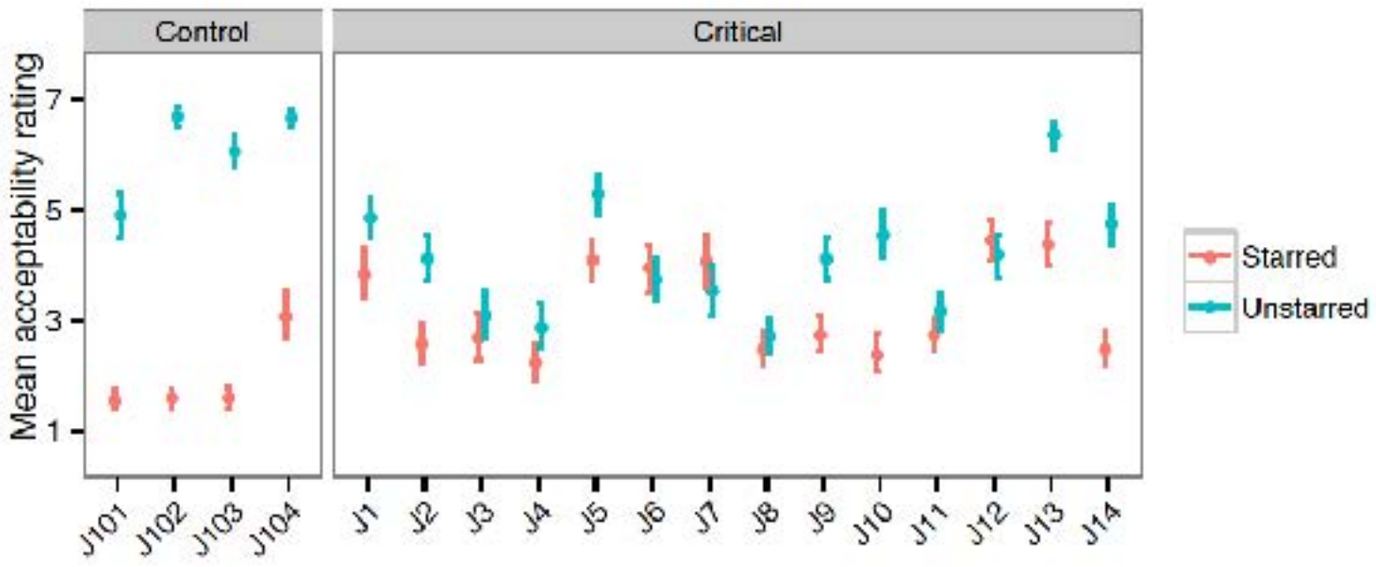


Figure 1: Results of the experiments: (a) Hebrew and (b) Japanese. Error bars represent bootstrapped 95% confidence intervals.

Courtesy of Tal Linzen and Yohei Oseki. Used with permission. Source: Linzen, Tal, and Yohei Oseki. "The reliability of acceptability judgments across languages." New York: New York University, ms (2015).

Summary:
Reasons to do quantitative research

• In non-quantitative work, because there are some judgment errors, other researchers can never know which comparisons are ok, and which are not. Even if we don't require arbitrarily low error rates, the details of a quantitative experiment provide *some* evidence about how strongly to believe the effect. **Objectivity**.

• Intuitions are not reliable for **interactions** among factors.

• Difficult / impossible to maintain **consistency** of judgments across many pairs of judgments.

• Can learn about **effect sizes**, which can often be used to determine if some factor is theoretically important.

Possible project

- There is evidence that the lexical decision task is affected by the context in important ways: instructions etc.
- For sentences, how does the context affect the judgments?
- If the distractor materials vary, does this affect the judgments in an interesting way?

Already known: no major differences among Likert scale judgments vs. magnitude estimation judgments

(or even simple forced choice: effectively a 2-point scale)

as long at there are lots of items

Non-quantitative syntax / semantics: The single-subject/single-item method

Disadvantages of quantitative methods (controlled experiments or corpus analyses): (from class responses)

1. Quantitative methods/experiments require either access to mechanical turk or funding. This can hinder researchers outside of the US, or researchers anywhere without financial backing.
2. What matters is the effect size, not just statistical significance
3. Unnecessary a lot of the time: the experimenters' intuition towards which tendencies will be preferred are correct.
4. More difficult to do an experiment: creating materials + money
5. Data from quantitative methods would not account for a researcher's failure to address exceptions to syntactical generalizations that stem from situational or wording-related factors

Further qs

1. The tested sentences in Mahowald et al were originally designed to show the same contrast as the original paired-data from Linguistic Inquiry.. How were the pairs of sentences constructed by students in a class?
2. Under many circumstances, linguists provide more than one pairs of sentences with different contrasts (independent evidence of distinct types) to support one step of their reasoning. In such cases, one pair of unreliable language data does not necessarily impair their reasoning. To consider this factor, it might be useful to categorize the linguistic data extracted from Linguistic Inquiry into several subcategories. Several pairs of language data which are contributed to the same argument might be marked as 'parallel data' which are separated from other language data which is the sole evidence for a certain argument.
3. Bayesian statistics? Confidence intervals for the SNAP judgments? Not for now…
4. With the SNAP judgements, I was a little bit unsure whether each decision of the 5 would have the same two tendencies and the test was if 5 people chose the same tendency over the other five times out of five.
5. potential differences between the magnitude-estimation and the rating study
6. What qualitative-research approaches are commonly employed in linguistics research?
7. How does this relate to other forms of linguistics research? Is the SNAP judgment paradigm only applicable to experiments investigating sentence acceptability?

9.59J/24.905J Lab in
Psycholinguistics Spring 2017