Working memory and syntactic islands revisited

Edward Gibson, MIT Brain & Cognitive Sciences and Linguistics & Philosophy, egibson@mit.edu

Greg Scontras, Harvard Linguistics, scontras@fas.harvard.edu

What makes a long-distance extraction bad? (Ross, 1967)

* Who_i did [you criticize votes for [the impeachment of e_i]]?

* Who_i did [Emma doubt [the report [that we had captured e_i]]].

In Ivan gestures: What makes it hard to extract a wh-filler from way down there to way up here?

What is the nature of "syntactic islands"?

Purely syntactic accounts, going back to Ross & Chomsky:

- E.g., the **Subjacency** Condition (Ross, 1967; Chomsky, 1973; cf. Chomsky, 1962; Ross, 1967; Chomsky, 1986; Chomsky, 1995):
  - No rule may move a phrase from position Y to position X (or conversely) in:

    . . . X . . . [$_A$ . . . [$_B$ . . . Y . . .] . . .] . . . X . . .

    Where A and B are cyclic nodes.

  (1) * Who$_i$ did [you criticize votes for [the impeachment of e_i]]?

- **BUT**: Non-syntactic factors like plausibility and working memory demands affect extraction difficulty (e.g., Deane, 1991; Kluender, 1992; Sag et al., 2007; Hofmeister & Sag, 2010):

  (2) Who$_i$ did [you obtain votes for [the impeachment of e_i]]? (Deane, 1991)

- **Can syntactic island effects be explained by resource complexity or other non-syntactic factors?**

Open question in the syntax literature: The **domain-specificity / generality** of extraction constraints

- Proposal: **There is no principle in the grammar** that blocks extractions from long-distance sites (Deane, 1991; Kluender, 1992; Sag et al., 2007; Hofmeister & Sag, 2010).

  - Possibility 1: the unacceptability of extraction effects is *completely* explained by non-syntactic factors (plausibility; working memory effects; etc.)

  - Possibility 2: the unacceptability of extraction effects is *partially* explained by non-syntactic factors, but there is still a role for a purely syntactic factor: a low probability syntactic structure is more costly than a more probable one (Hale, 2001; Levy, 2008) (**but is not impossible / ungrammatical**)

Open question in the syntax literature: The **domain-specificity / generality** of extraction constraints

Experimental evidence: Complex NP-islands: Hofmeister & Sag (2010, Language)

One extreme:

I saw who Emma doubted the report that we had captured in the nationwide FBI manhunt.

Another extreme:

I saw which convict Emma doubted a report that we had captured in the nationwide FBI manhunt.

The type of embedded NP (definite / bare plural / indefinite) and the type of filler (who / which-NP) had large effects on the acceptability ratings and reading times in the predicted direction.

Open question in the syntax literature: The **domain-specificity / generality** of extraction constraints

More experimental evidence: Sprouse, Wagers & Phillips (2012, Language, **SWP**):

- Correlational study, using individual difference measures from two kinds of tasks:
  1. Acceptability ratings on extractions from islands
  2. A working memory measure (two tasks: n-back; serial recall)

- Result: No correlation whatsoever.

- *"We believe that the results of the experiments presented in this article provide strong support for grammatical theories of island effects because we can find no evidence of a relationship between processing resource capacity and island effects."*

Open question in the syntax literature: The **domain-specificity / generality** of extraction constraints

More experimental evidence: Sprouse, Wagers & Phillips (2012, Language, **SWP**):

- Correlational study, using individual difference measures from two kinds of tasks:
  1. Acceptability ratings on extractions from islands
  2. A working memory measure (two tasks: n-back; serial recall)

- Result: No correlation whatsoever.

BUT: Absence of evidence is not evidence of absence

Null correlation: how to interpret?

Would we be surprised if we found no correlation between Island ratings and participant height? Probably not.

How is this comparison similar / different?

WM measures and acceptability ratings:
Is rating a sentence the same as processing a sentence?

SWP implicitly assume that processing difficulty is reflected in acceptability ratings *which vary across individuals*, for a particular item

Alternative: an analogy between WM ability and **physical ability**.

People of different physical abilities would judge a marathon to be more challenging than a walk in the park.

Similarly, an individual with low WM might be able to judge which sentences sound easier vs. harder to process.

Rating a sentence may not be the same as processing a sentence.

New design

- **Main potential problem with SWP's design:** ***Rating a sentence is NOT the same as processing a sentence.***

New design: 3 tasks:

Task 1: Sentence completions of WM-complex materials: a general intelligence task, which correlates with IQ tasks (which also correlate with WM tasks)

Task 2: Sentence acceptability of WM-complex materials

Task 3: Sentence acceptability of islands.

Examine the relationship between task 1 and task 2:

- If they correlate (as SWP implicitly assume), then we can evaluate task 1's correlation with task 3 (SWP's question)
- If not, then SWP's evaluation tells us nothing about the relationship between WM and island acceptability.

Assessing linguistic WM / intelligence
Gibson & Fedorenko (2012)

General background:

Nested structures are difficult to understand (Yngve, 1960; Chomsky & Miller, 1963).

Singly-nested relative clause structure:

The manuscript that the student found was 200 pages long.

Doubly-nested relative clause structure:

The manuscript that the student who the dog bit found was 200 pages long.

Right-branching relative clause structure:

The dog bit the student who found the manuscript which was 200 pages long.

Gibson & Fedorenko (2012)

We focused on the materials that elicited completion performance in the dynamic range.
60 participants on Mechanical Turk
Design: 4 critical conditions (6 items each) + 2 control conditions
Task: Complete each preamble to make a complete sentence.

ORC-anim/SRC	The reporter who the professor who…
ORC-inan/SRC	The manuscript which the student who…
SC/ORC	The fact that the professor who the diplomat…
SC-verb/ORC	The rumor stating that the suspected mobster who the media…
Control 1: ORC	The veterinarian who the…
Control 2: SRC	The fencer who…

Gibson & Fedorenko (2012)

For analyses of the critical conditions we only included participants (n=55) who grammatically completed at least 5/6 of each of the control conditions (simple SRCs, simple ORCs).

As in Experiment 1, most incorrect completions involved omitting the middle VP:

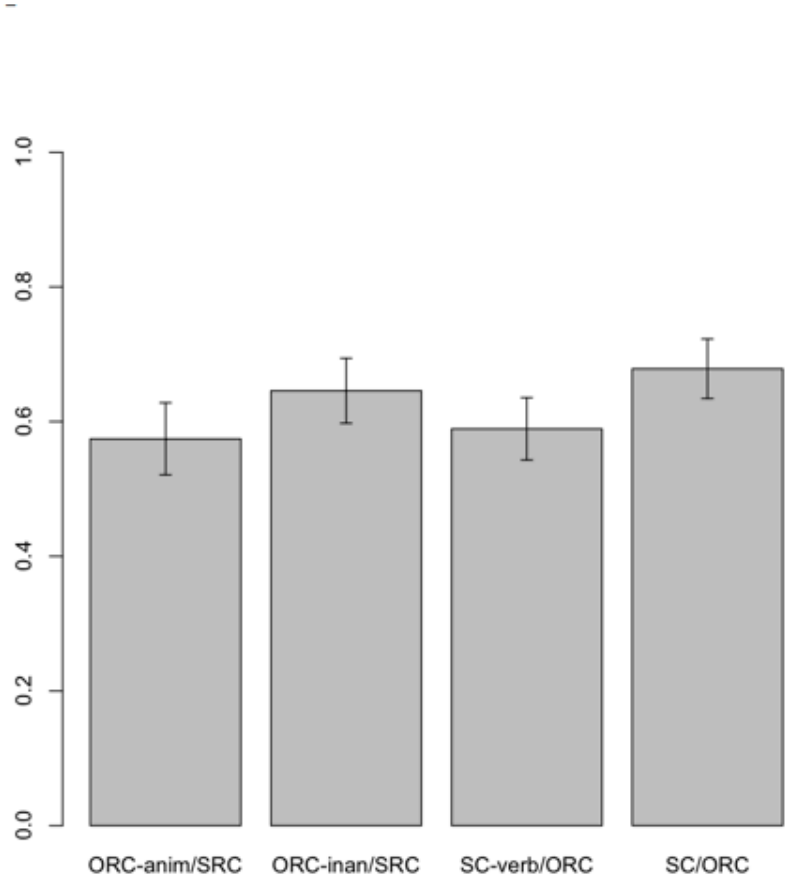
The fact that the professor who the diplomat ...

... had lunch with was true.

The manuscript which the writer who ...

... is well-known wrote was exellent.
... hangs out at Denny's sent in, was rejected.
... was sad wrote, won the prize.

... was sick was overdue.
... was poor was rejected.
... enjoyed opium was really very confusing.
... collaborated with the original book's author, was finished ahead of schedule.

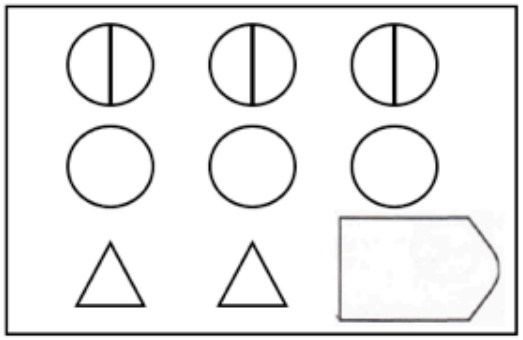


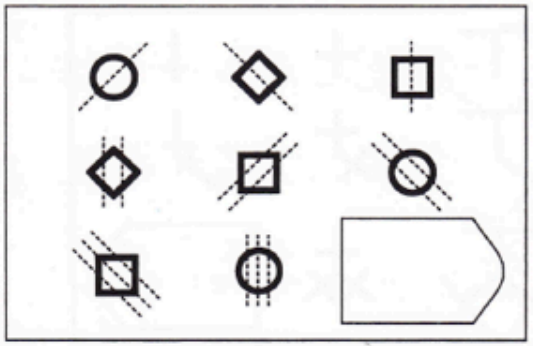
Gibson & Fedorenko (2012)

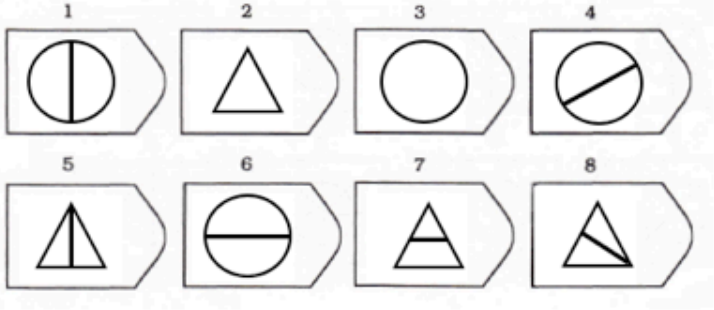
60 participants on Mechanical Turk

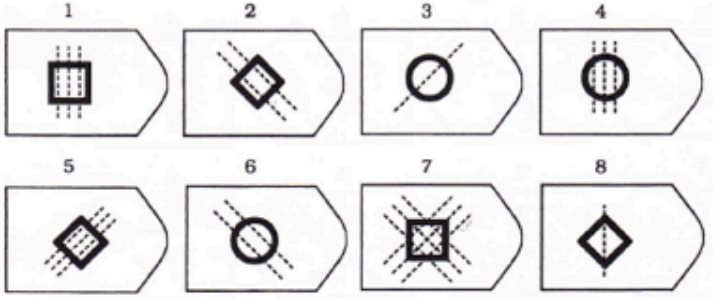
The non-language task: Ravens Advanced Progressive Matrices (RAPM; Raven et al., 1988) (a very general fluid intelligence task; has been shown to correlate with many WM and cognitive control tasks, like Stroop & digit span and many others)

Task: Choose a picture (out of 8 possibilities) that best completes the set.
40 trials.





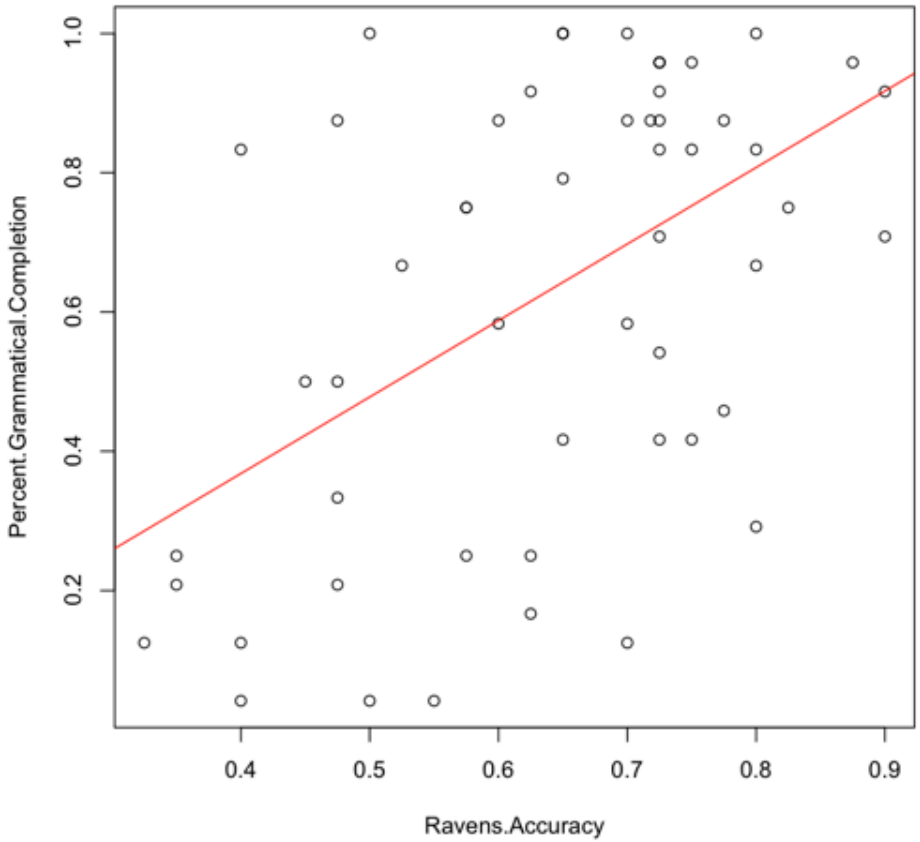




Gibson & Fedorenko (2012)
Correlation between completions and Ravens

Correlation between proportion of grammatical completions and Ravens accuracy:
n=55; r=.505; p<.001

Significant effect of RAPM score predicting the completion score (β=5.16; t=3.31; p=.0017), but no reliable effect of work time (β=-0.17; t=-0.68; p=.50) nor average filler length (β=0.30; t=1.31; p=.20).



New design

60 participants on M Turk (run twice, with similar results in each run)

(1) a language / fluid intelligence measure that has been shown to correlate strongly with fluid intelligence: complex sentence completion (Gibson & Fedorenko, 2012); (WM correlates highly with fluid intelligence)

(2) a rating study consisting of

(a) nested vs. non-nested sentences, a contrast thought to be related to WM demands; and

(b) adjunct and NP islands, similar to SWP's, but with 8 trials per condition per participant, and with plausibility-matched control conditions.

Participants also answered 2 comprehension questions about each sentence.

Result 1

Complex sentence completion:

Replicating previous results, we found stable completion rates within individuals but varying across individuals. (cf. Gibson & Fedorenko, 2012; Gibson & Thomas, 1999; Frazier, 1978; Christiansen & Chater, 2001; Vasishth et al., 2008)

Lots of incomplete completions: ~50%

Sample completions:

The reporter who the professor who …

3 VPs:

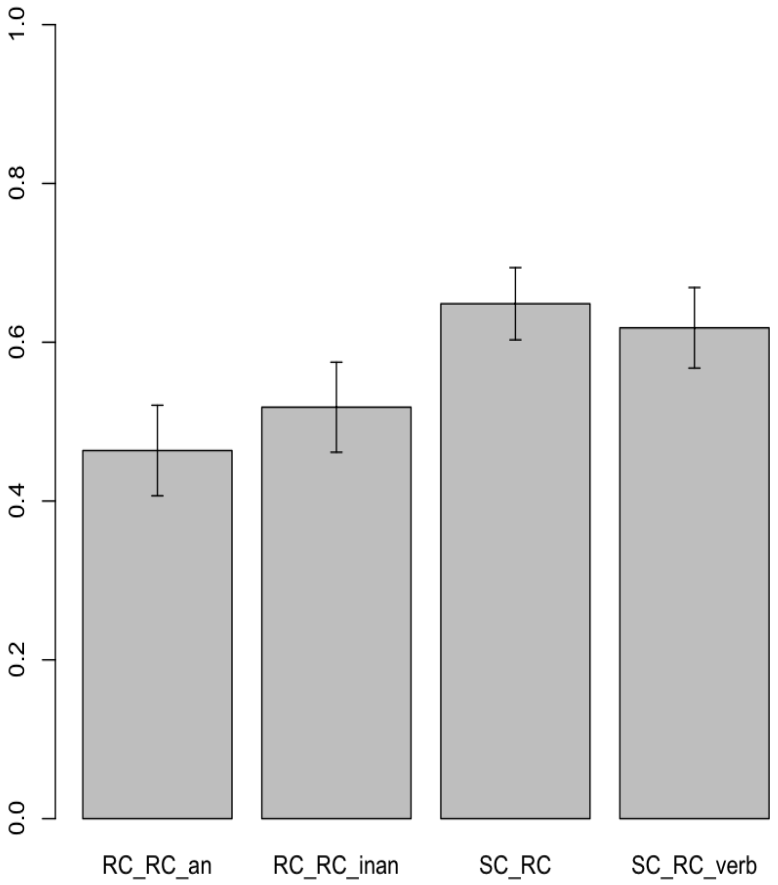
…taught English called was late to their appointment.

…taught physics, was friends with asked me some questions.

2 VPs:

…was dating him, thought that they were out of line.

…was intelligent, didn't like each other.



Result 2

Complex sentence completion:

This measure correlated reliably with the ability to answer comprehension questions in the rating study (r=.510; p<.001).

The reporter who the professor who …

High WM / IQ participant typically completes with 3 VPs:

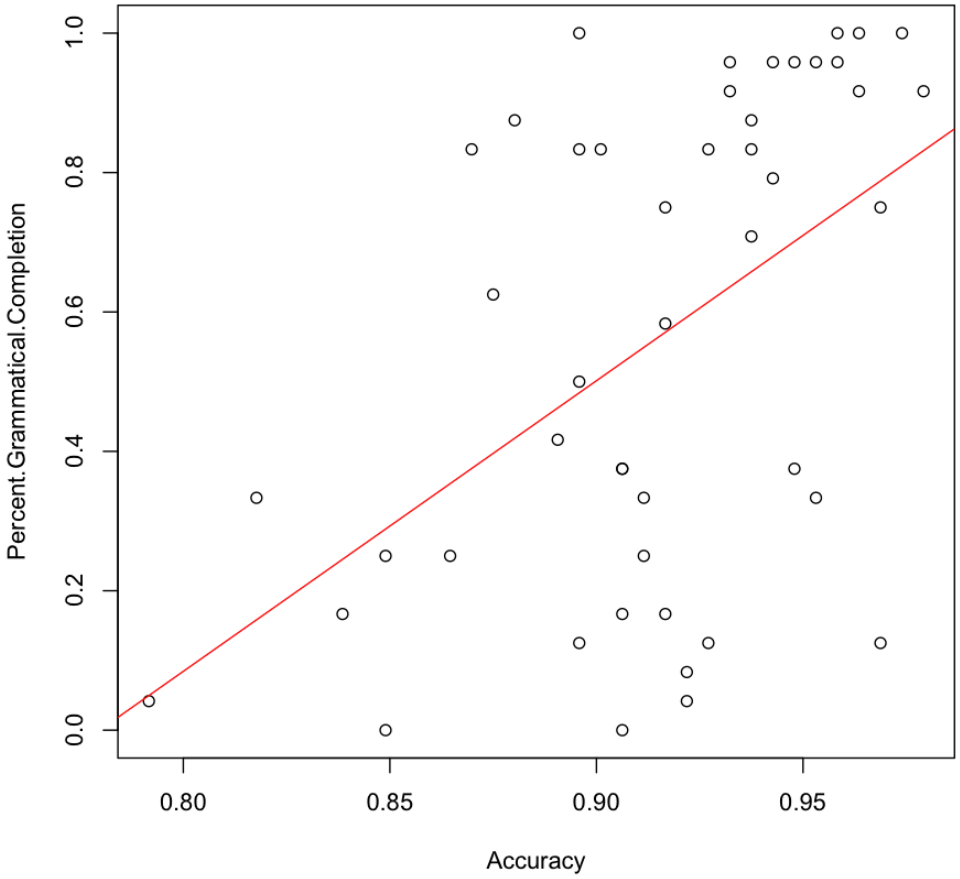
…taught English called was late to their appointment.

…taught physics, was friends with asked me some questions.

Low WM / IQ participant typically completes with 2 VPs:

…was intelligent, didn't like each other.

…was dating him, thought that they were out of line.



Result 3

Reliable differences in ratings were observed for each contrast:

- nested vs. non-nested:

  2.12 vs. 3.17 out of 5; $p<.001$

  The building which the architect who the contractor met designed was commissioned by the city.

  The contractor met the architect who designed the building which was commissioned by the city.

- adjunct-island vs. pronoun-dependency:

  1.72 vs. 2.22; $p<.001$

  The neighbor wonders what you sneeze if the dog owner leaves open at night.

  The neighbor wonders what the thing is, such that you sneeze if the dog owner leaves it open at night.

- NP-island vs. pronoun-dependency:

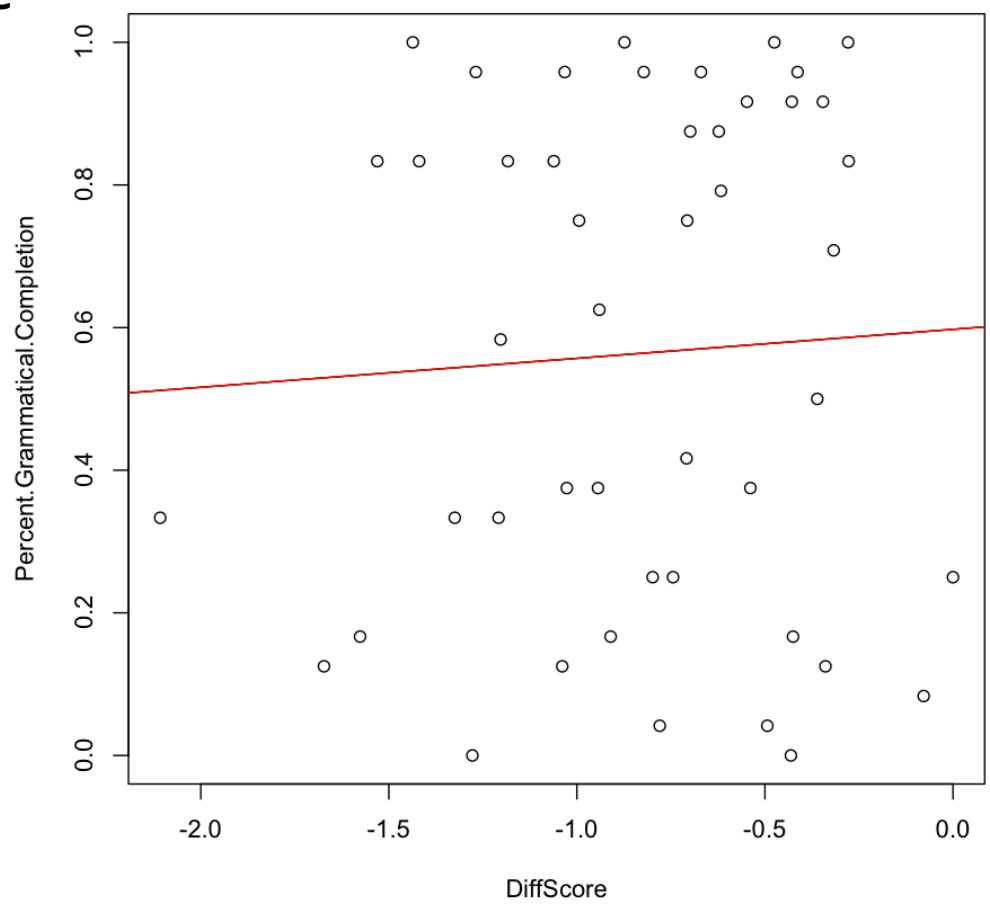
  1.89 vs. 2.27; $p<.001$

  Joe wondered what the chef heard the statement that Jeff baked.

  Joe wondered what the thing was, such that the chef heard the statement that Jeff baked it.

Result 4

Critically, we found no correlation between our IQ / WM measure and the nested vs. non-nested difference score in the ratings (r =.052), suggesting that SWP's assumption about acceptability ratings reflecting WM demands is not warranted.



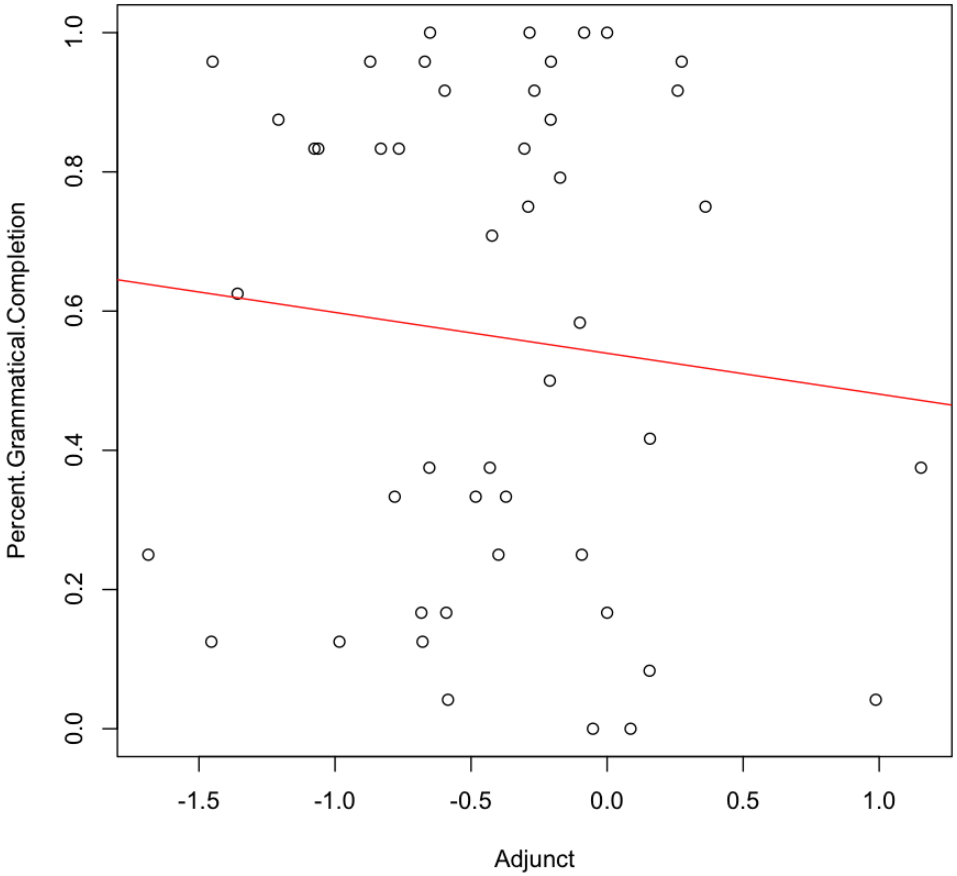
Result 5

Replicating SWP, we found no correlation between acceptability ratings for island conditions and our IQ / WM measure.

r = -.097 for adjunct-islands

r = -.089 for NP-islands

But this is now unsurprising, given that we also find no correlation with our WM measure for judgments based on nested vs. non-nested sentences



Summary & Conclusions

There is no correlation between our IQ / WM measure and the nested vs. non-nested difference score in the ratings (r =.052).

This shows that SWP's assumption about acceptability ratings reflecting WM demands is not warranted:

Rating a sentence is NOT the same as processing a sentence.

The results from Sprouse, Wagers & Phillips (2012) therefore **have no bearing** on the question of whether island-extraction complexity may be explained by WM or other factors (in whole or in part).

Where to go from here?

Even the controls for the island extractions are generally poorly rated:

- nested vs. non-nested:

  2.12 vs. 3.17 out of 5; $p<.001$

  The building which the architect who the contractor met designed was commissioned by the city.

  The contractor met the architect who designed the building which was commissioned by the city.

- adjunct-island vs. pronoun-dependency:

  1.72 vs. 2.22; $p<.001$

  The neighbor wonders what you sneeze if the dog owner leaves open at night.

  The neighbor wonders what the thing is, such that you sneeze if the dog owner leaves it open at night.

- NP-island vs. pronoun-dependency:

  1.89 vs. 2.27; $p<.001$

  Joe wondered what the chef heard the statement that Jeff baked.

  Joe wondered what the thing was, such that the chef heard the statement that Jeff baked it.

Island-extraction and implausible information structure?

Maybe many islands are not perfectly fine thoughts: perhaps their discourse structure is malformed, and this explains their low acceptability, even when formulated with constructions using pronominal coreference as opposed to wh-fronting.

BCI: Backgrounded Constructions are Islands
(Erteschik-Shir 1979, 1998; Goldberg, 1995; Ambridge & Goldberg, 2008; and others)

Explains the difference between:
Who did Bill think that Sandy saw?

? Who did Bill mumble that Sandy saw?

No extraction of backgrounded material. (Independent measurement of backgrounding in Ambridge & Goldberg, 2008)

Acknowledgements

- Greg Scontras (co-author)
- Richard Futrell & Adele Goldberg (collaborators on further island projects)
- Ev Fedorenko
- Colin Phillips (healthy discussion)

Acknowledgements

Ivan Sag: motivation for doing the studies