Comparing three methods for sentence judgment experiments

Shin Fukuda
Dan Michel
Henry Beecher
Grant Goodall

3 response methods in experimental syntax

- Yes/No forced choice

	Does the sentence sound good?	
1. What does Helen want Gary to drink?	yes	no

- *n*-point numerical scale (Likert scale)

	very bad				very good
1. What does Helen want Gary to drink?	1	2	3	4	5

- Magnitude estimation

	How good does the sentence sound?	
Reference sentence: **What do you wonder whether Mary bought?**	**100**	
1. What does Helen want Gary to drink?	**200**	

Standard view of response methods

	Advantages	Disadvantages
Yes/No	Simple task for subjects	Coarse-grained
Numerical scale	-Familiar task for subjects -Somewhat fine-grained	-May not be true interval data. -Subjects may make more distinctions than scale allows.
Magnitude Estimation	-Subjects can make as many distinctions as needed -More powerful statistics	Unfamiliar, strange task for subjects

ME provides insights that other methods do not (Bard et al. 1996, Cowart 1997, Featherston 2005 etc.)

Recent critiques of ME

- Sprouse (2008):
  - The role of reference in ME at best questionable.
  - Participants are not able to make use of the alleged advantages of ME.

- Weskott & Fanselow (2008):
  - ME data not more informative.
  - ME data may have more spurious variance.

Structure of Weskott & Fanselow (2008)

- One set of subjects judges same stimuli with both Yes/No and 7-point scale.
- Another set judges same stimuli with both Yes/No and ME.
- Judgment tasks 2 weeks apart; order balanced
- Stimuli: ACC- and DAT-scrambling in German.

Structure of Weskott & Fanselow (2008)

- All three methods captured expected contrasts equally well (comparable effect sizes)
- With ½ of subjects, loss of effect size greater with ME than other methods

Comment on Weskott & Fanselow (2008)

- Potential drawback:
  - Same subjects do 2 tasks; experience with one could influence judgments on other.
  - Type of stimuli very limited (only scrambling).

Our study

- Three groups of subjects:
  - Yes/No (N=36)
  - 5-point scale (N=36)
  - ME (N=36)

- All UCSD students, randomly assigned to response method.

Our study

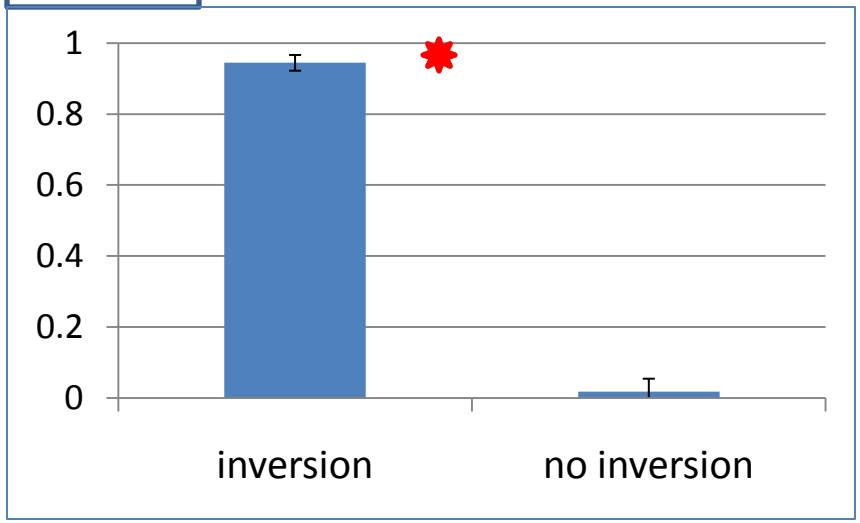
- Three subexperiments:
  - Presence/absence of inversion
  - *that*-trace effect
  - Extraction out of subject and object DPs
- Range of expected contrasts, from dramatic (inversion) to subtle (extraction out of subj/obj DPs).
- Are the three methods equally effective in capturing acceptability contrasts with different degrees of subtlety?

Presence/absence of inversion

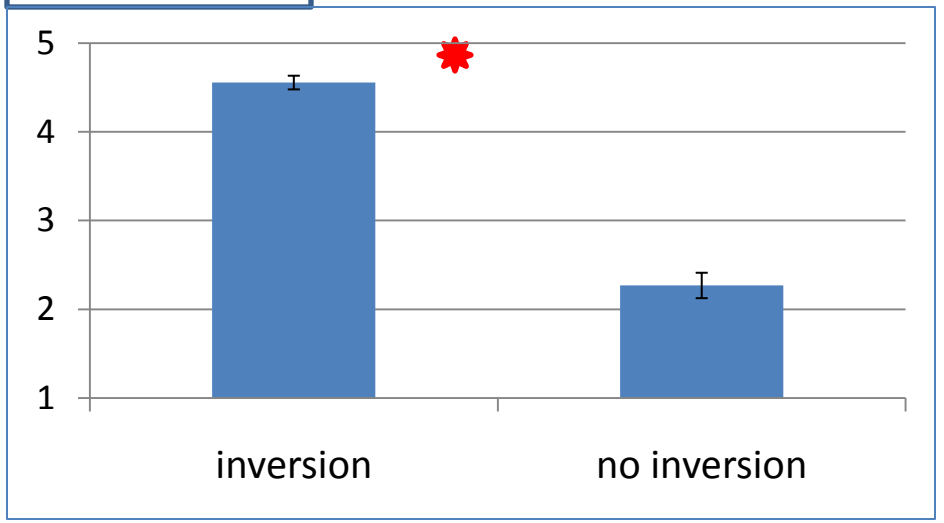
1. *What will **you** watch on Thursday?*
2. *What will **he** watch on Thursday?*
3. *What will **the man** watch on Thursday?*
4. *What **you** will watch on Thursday?*
5. *What **he** will watch on Thursday?*
6. *What **the man** will watch on Thursday?*

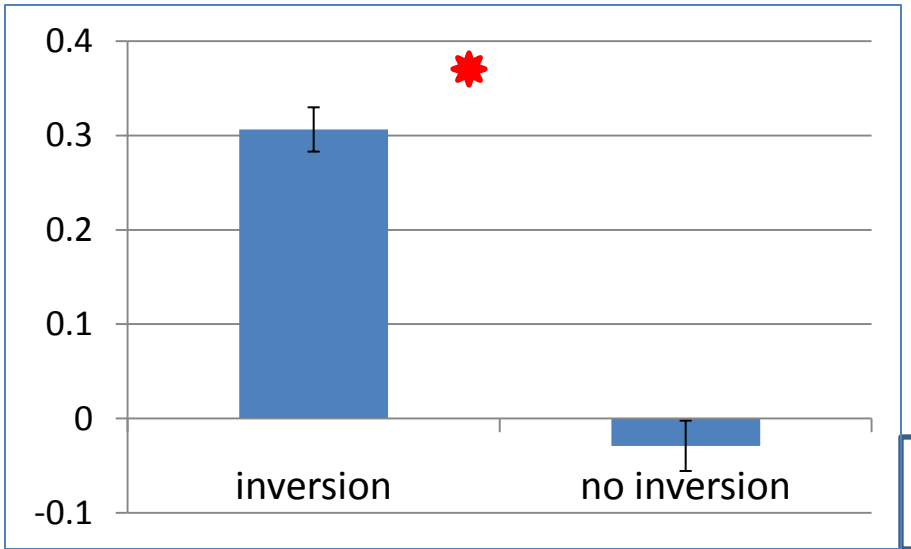
Results: Inversion

Y/N



5-point





ME

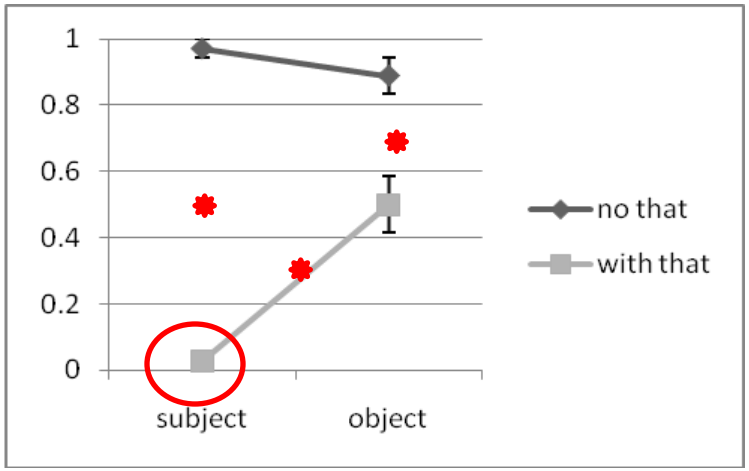
$p < .05$

that-trace effect

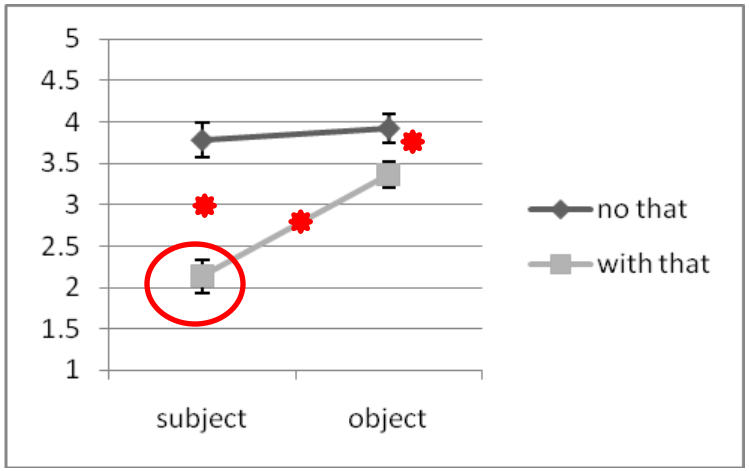
1. ***Who*** *do you feel* ***that*** *__insulted Pat at the theater?*
2. ***Who*** *do you feel* ***that*** *Pat insulted __at the theater?*
3. *Who do you feel __insulted Pat at the theater?*
4. *Who do you feel Pat insulted __at the theater?*

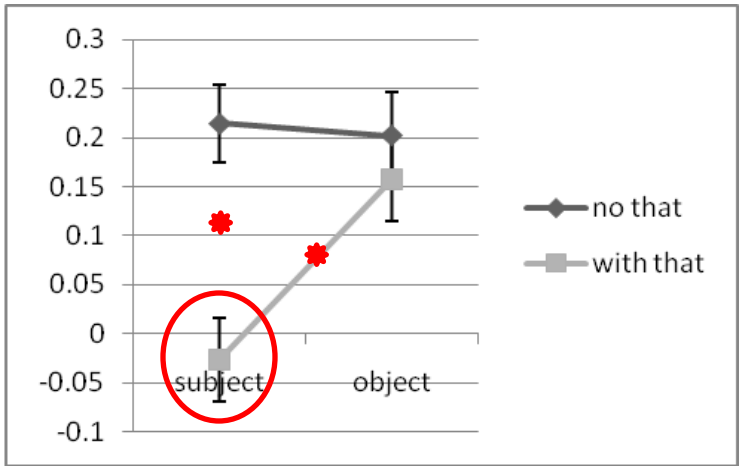
Results: *that*-trace

Y/N



5-point





ME

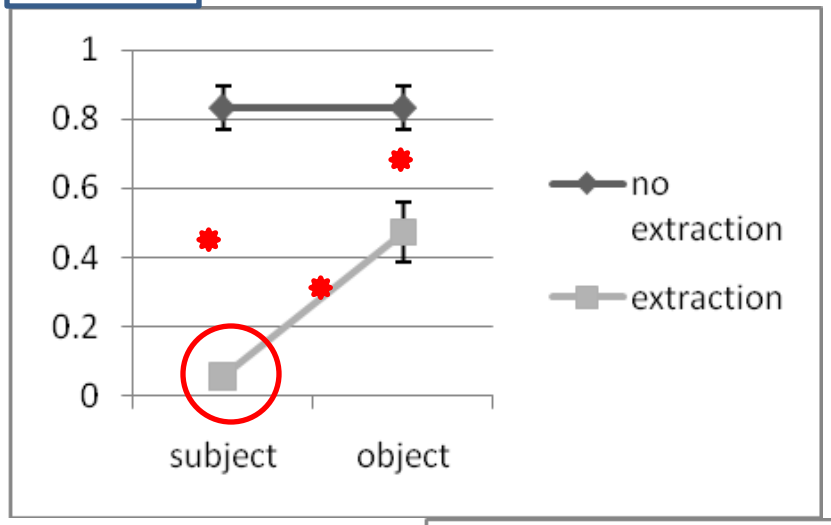
✸ $p < .05$

Extraction out of subject and object DPs

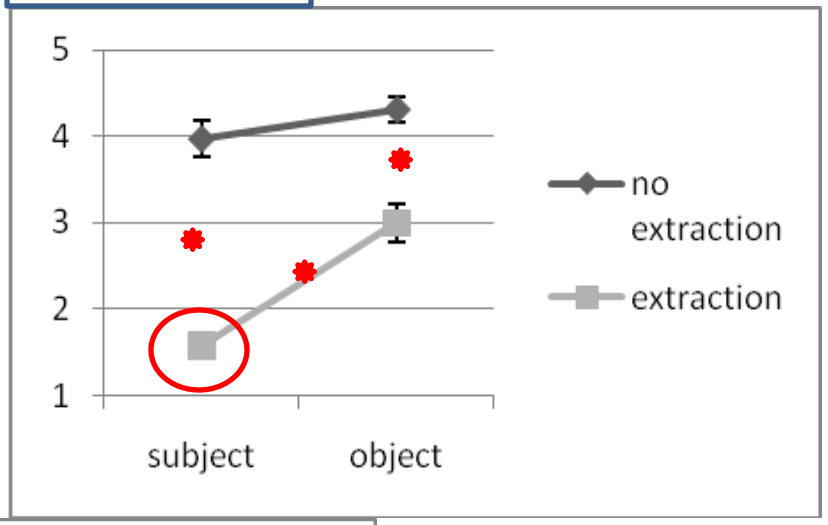
1. ***What*** *do you think* ***[pictures of ___]*** *will be on the website?*
2. ***What*** *do you think the website will post* ***[pictures of ___]****?*
3. *Do you think* ***pictures of the new car*** *will be on the website?*
4. *Do you think the website will post* ***pictures of the new car****?*

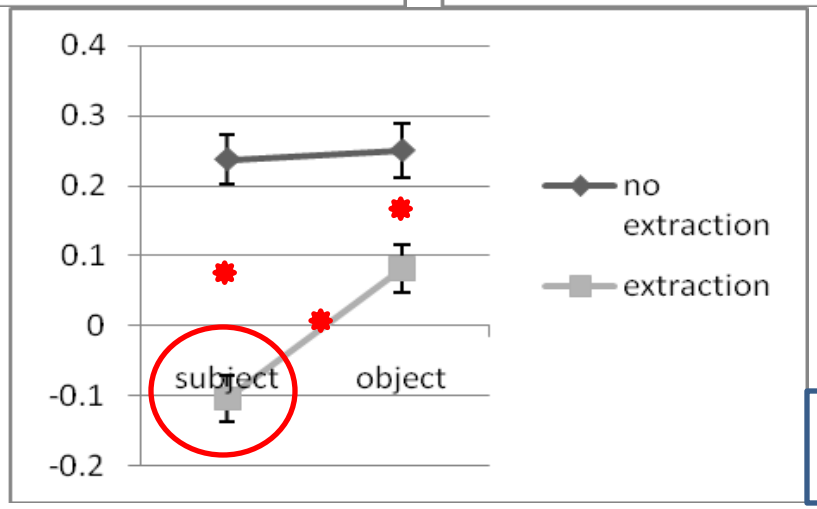
Results: Extraction from subject/object DPs

Y/N



5-point





ME

✸ $p < .05$

Interim conclusion

- Yes/No and 5-point scale able to capture contrasts just as well as ME.

Informativity (effect size)

- Significance is virtually the same across all three methods.
- But are the results of the three methods equally informative?
- We can measure this with η^2 (Cohen 1973, Cowart 1997).
- Shows how much of variation between two means is accounted for by factor of interest (as opposed to random variation).

Inversion

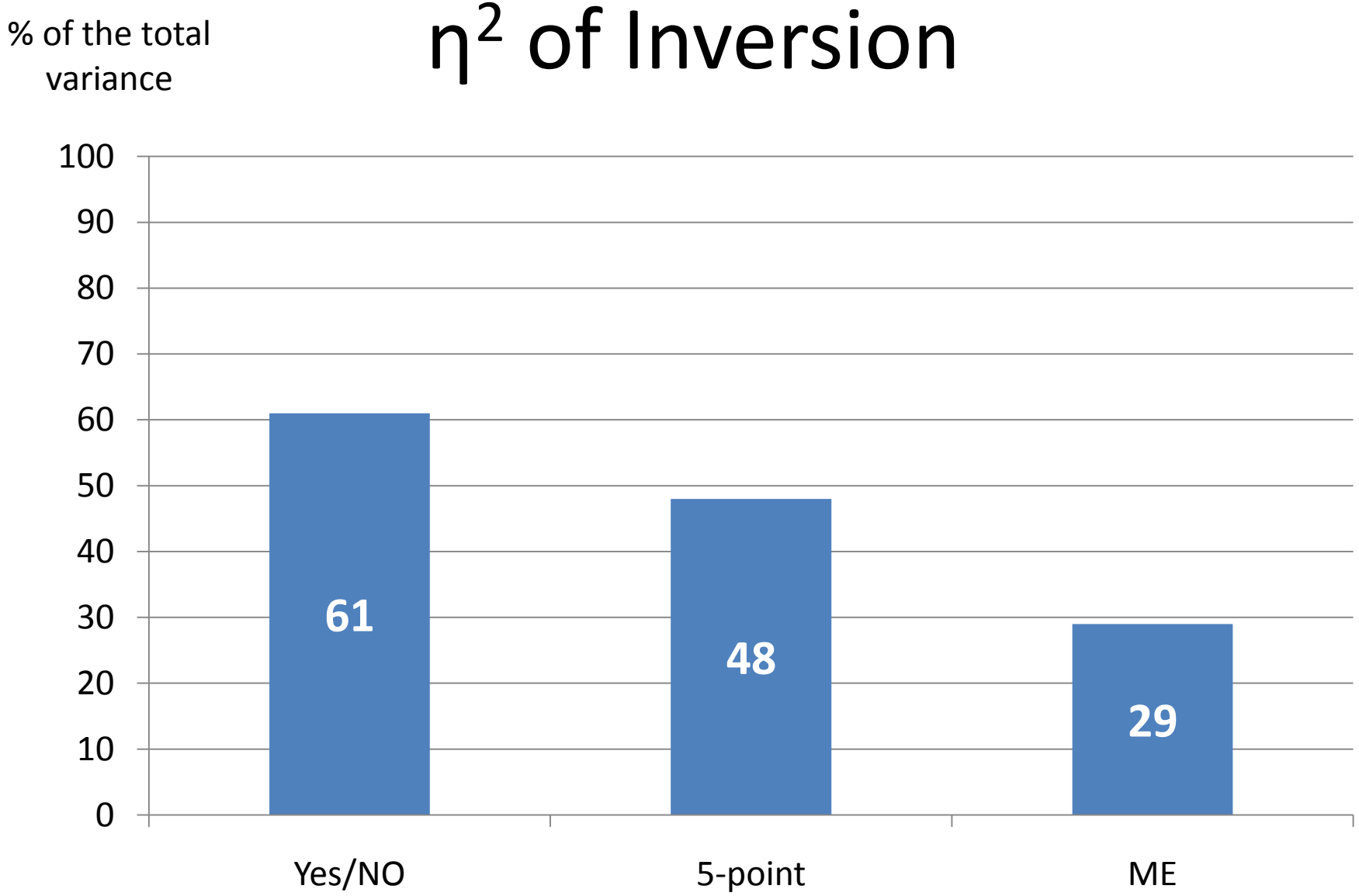
- All three methods found significant difference:

  *What <u>will **you**</u> watch on Thursday?*

  *What <u>**you** will</u> watch on Thursday?*

- But how much of this difference is due to presence/absence of inversion?

η^2 of Inversion



that-trace effect

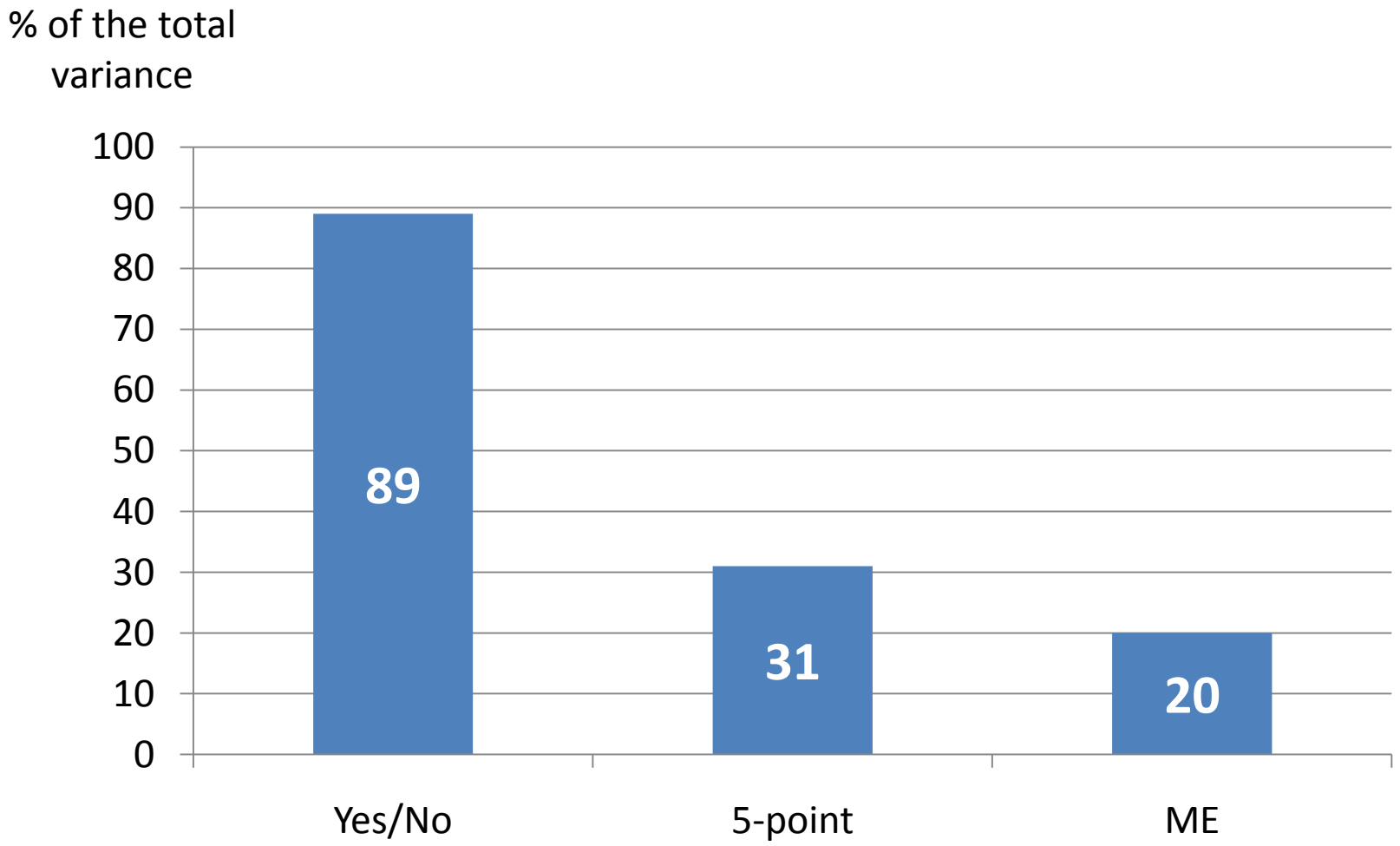
- All three methods found significant difference:

  Who do you feel that __ insulted Pat at the theater?

  Who do you feel __ insulted Pat at the theater?

- But how much of this difference is due to presence vs. absence of *that*?

η^2 of presence vs absence of *that*



Extraction out of subject and object DPs

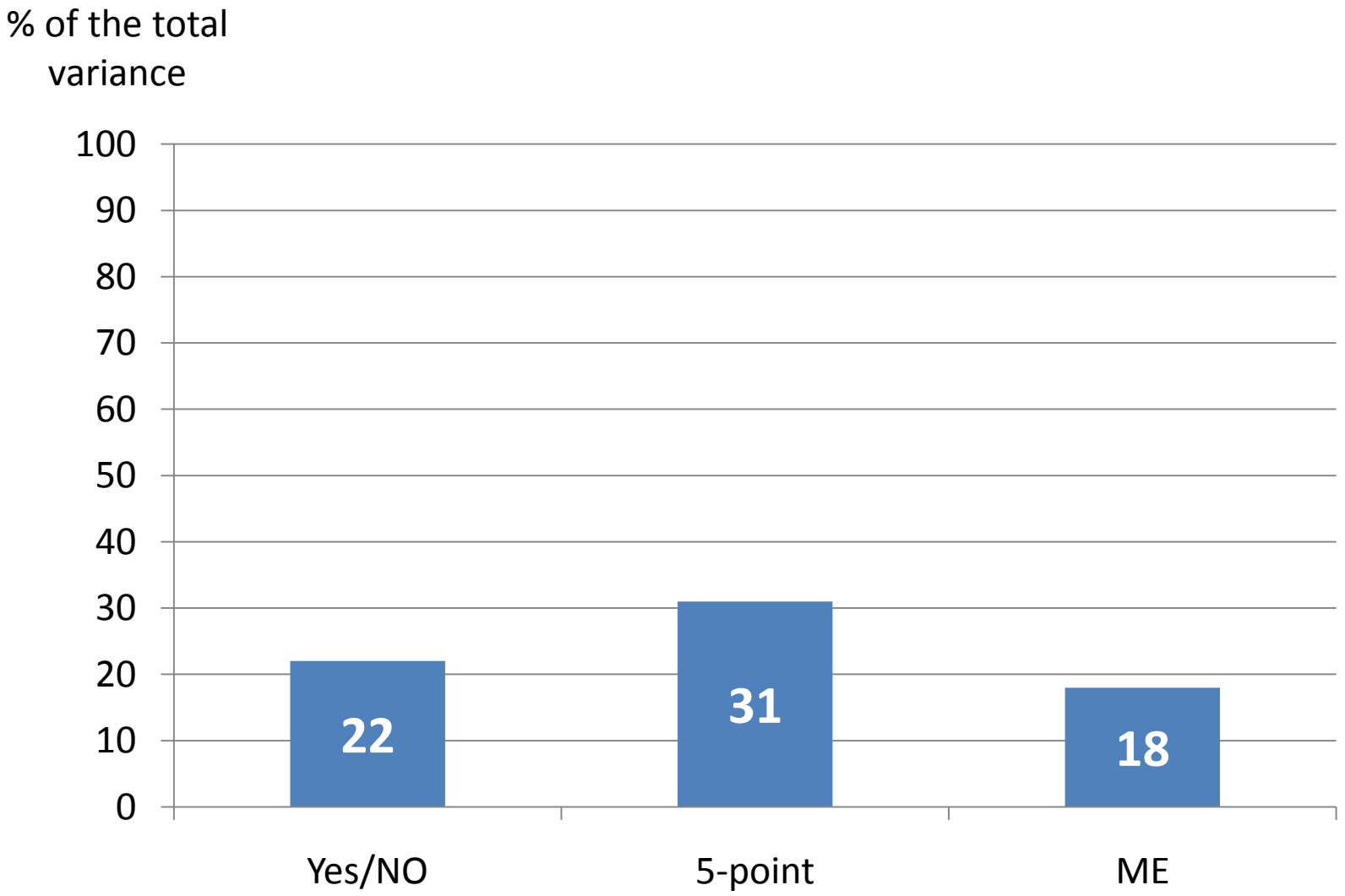
- All three methods found significant difference:

  What *do you think* ***[pictures of __]*** *will be on the website?*

  What *do you think the website will post* ***[pictures of __]****?*

- But again, how much of this difference is due to argument position?

η^2 of argument with extraction



What to conclude from this?

- ME does allow finer-grained distinctions, but…
  - This may lead to spurious variance in some cases (cf. Wescott & Fanselow 2008),
  - rather than contributing to our understanding of the contrast.

General implications of this study

- For the working syntactician:
  - Any of these response methods works.
  - More familiar techniques (both to researcher and to subjects) are just as sensitive as ME.
  - ME may introduce some spurious variance.

And perhaps more importantly…

- "Fear of ME" should not be a roadblock to more widespread adoption of experimental techniques.

- Y/N forced-choice & *n*-point numerical scale can capture even subtle contrasts with well-designed experiments.

(see also Myers 2009, Philips 2009)

Thank you!

Experimental Syntax Lab, UCSD

Fall 2009 Experimental Syntax class, UCSD

Sara Cantor, Carleton College

W&F vs. this study (η^2)

W&F

Exp1	YN1	YN2	7-point	ME
ACC-scrambling	.95	.95	.96	.98
Exp2	**YN1**	**YN2**	**7-point**	**ME**
DAT-scrambling	.91	.83	.92	.96

This study

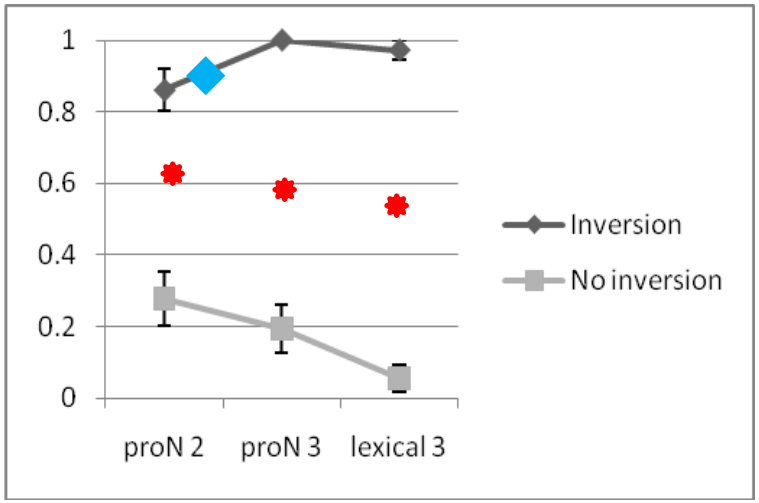
	YN	5-point	ME
Inversion	.61	.48	.29
that-trace	.89	.31	.20
Extraction out of Subj/Obj	.22	.31	.18

η^2 and partial η^2

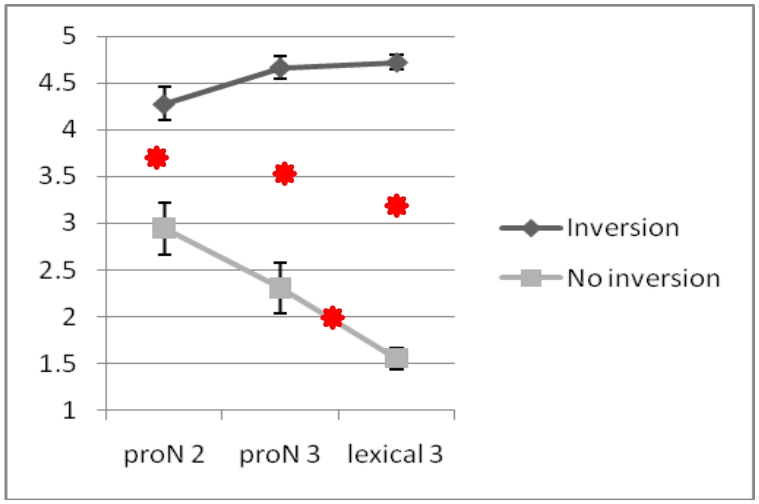
- $\eta^2 = SS_{factor} / SS_{total}$
- Partial $\eta^2 = SS_{factor} / SS_{factor} + SS_{error}$

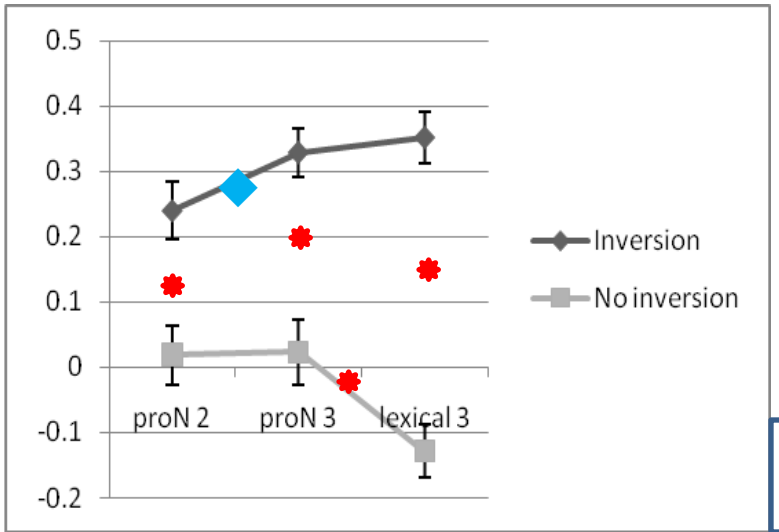
Inversion:
Break-down by subject type

Y/N



5-point





ME

$p < .01$

$p < .05$