A standard scale of well-formedness: Why syntax needs boiling and freezing points

Sam Featherston

SFB441: Linguistische Datenstrukturen
Eberhard-Karls-Universität Tübingen

Linguistic Evidence 2008, Tübingen, 2nd February 2008

Talk overview

Three questions for judgements in empirical syntax

1. How can we gather judgements?
   - magnitude estimation as a standard method?
     Insights from psychophysics
   - what we do: thermometer judgements

2. Can we investigate language structure with judgements?
   Insights from psychophysics

3. Do we need any more scale than this?
   - the uses of a standard scale

Question 1

How do we gather judgements?

How do we gather judgements?

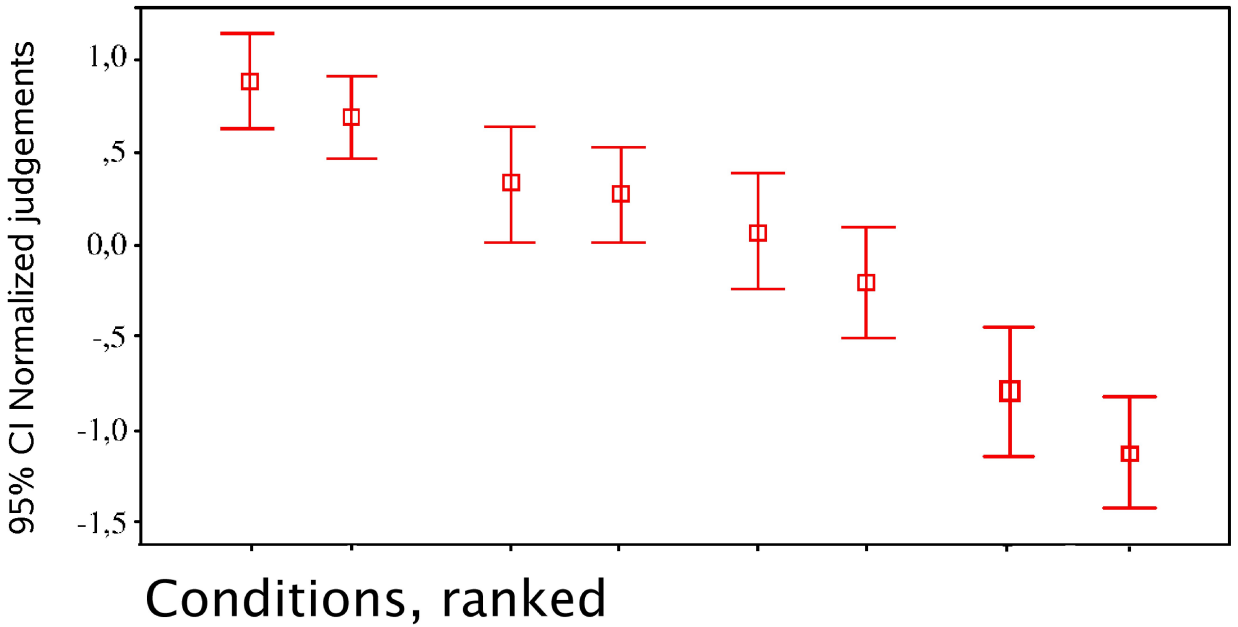
Magnitude estimation à la Bard et al (1996): "standard method"?

- Upsides:
  - provides good results
  - significant advance
  - enabled new work to be done

- Downsides:
  a) no magnitudes in results
  b) log conversions unmotivated
  c) integer preference near zero
  d) reference item too variable as normalization basis

Magnitude estimation: problem (a)

- a) Pattern of results: linear, interval scale (contra Sprouse 07)
  Apparently subjects cannot give magnitude judgements.
  How bad is this? Not too much effect on results, since subjects ignore instructions, but intellectually unsatisfying.



Magnitude estimation: problems (b), (c), & (d)

- b) Log conversions unnecessary
  (Featherston 2005, Sprouse 2007)
  How bad is this? Can falsify data pattern, cf Keller (2003).
  The fewer transformations the better.

- c) Floor effects near zero: single reference item.
  How bad is this? Only moderate distortion.

- d) Single reference item too variable as normalization basis
  How bad is this? Significant weakening of power.
  Subject means more stable basis for normalization.

Magnitude estimation: A validated method?

- But problems (b), (c), and (d) can be solved.
- Should we abandon a well-researched standard method?
- Validation of magnitude estimation:
  - psychophysics: stimulus measurable.
  - linguistics: stimulus is not independently measurable
  - so linguistics is dependent on psychophysical validation

Bard et al (1996)'s chief source

- Bard et al (1996) refer to S. S. Stevens (eg 1975)
  - head of the Havard psycho-acoustic laboratory
  - devised scale terms: nominal, ordinal, interval, ratio

- Stevens' psychophysics: the measurement of Sensation
  - Method: magnitude estimation
  - Finding: Power Law of Sensation and Stimulus
  - Method is validated by the consistency of the findings

- Sounds very convincing ...

Any counter evidence?

- Savage (1970) *The measurement of sensation*
  - on Stevens: 'his methods of psychophysical measurement [...] are spurious.'
  - psychophysics: "conceptually confused"
  - on measurement: number assignment not enough, we must be able to use a unit of measurement.

- Birbaum (1980): "psychological primitive" is stimulus difference, ratios derived from them.

- Shepard (1981) We must take the response function into account. So Stevens' power law conclusion is 'invalid'

Any more counter-evidence?

- Poulton (1989): Whole book *Bias in Quantifying Judgements*

  'Once most of Stevens' power functions are rejected because they are produced by a logarithmic response bias, there is no need to dwell on their other inadequacies.'

Any more counter-evidence?

- Poulton (1989): Whole book *Bias in Quantifying Judgements*

  'Once most of Stevens' power functions are rejected because they are produced by a logarithmic response bias, there is no need to dwell on their other inadequacies.'

  'Ratio judgements are biased and invalid.'

Any more counter-evidence?

- Poulton (1989): Whole book *Bias in Quantifying Judgements*

  'Once most of Stevens' power functions are rejected because they are produced by a logarithmic response bias, there is no need to dwell on their other inadequacies.'

  '... inadequacies in the design or conduct of the investigations.'

  'Ratio judgements are biased and invalid.'

  'Chapter 10 describes how investigators can use these and other techniques to obtain the results that they predict.'

Stevens' instructions: biased?

- Instructions (Stevens 1956)

' ..if the standard is called 10 what would you call the variable? [...] if the variable sounds 7 times as loud as the standard, say 70. If it sounds one fifth as loud, say 2; if a twentieth as loud, say 0.5, etc.'

'Try to make the ratios between the numbers you assign to the different tones correspond to the ratios of the loudnesses between the tones.'

Stevens comments on the methodology (Stevens 1956)

'... let me say that the success of the foregoing experiment was achieved only after much trial and error in the course of which we learnt at least some of the things *not* to do.'

...

3. 'Call the standard by a number, like 10, that is easily multiplied and divided.'

4. Use just one standard: 'If E assigns numbers to more than one stimulus, he introduces constraints of the sort that force O to make judgements on an interval rather than on a ratio scale.'

...

Stevens comments on the methodology (Stevens 1956)

'... let me say that the success of the foregoing experiment was achieved only after much trial and error in the course of which we learnt at least some of the things *not* to do.'

...

3. 'Call the standard by a number, like 10, that is easily multiplied and divided.'

4. Use just one standard: 'If E assigns numbers to more than one stimulus, he introduces constraints of the sort that force O to make judgements on an interval rather than on a ratio scale.'

...

Laming (1997): 'Reading between these lines of Stevens' advice, it is evident that even he found it easy to fail to get good power law data.' '[...] that result seems to be the very opposite of robust.'

Stevens and his troublesome 'observers'

'Another problem we encounter is due to the fact that some Os seem to make their estimates on an interval-scale, or even an ordinal scale, instead of on the ratio-scale we are trying to get them to use.' (Stevens 1956)

Stevens and his troublesome 'observers'

'Another problem we encounter is due to the fact that some Os seem to make their estimates on an interval-scale, or even an ordinal scale, instead of on the ratio-scale we are trying to get them to use.' (Stevens 1956)

How do you expect us to make progress if you produce judgments like that!
(from Poulton 1989)

Resistance to change....

'Stevens is such a strong and eloquent advocate of ratio judgements, that no investigator firmly articulates the reason for the discrepancy. [...] The invalid ratio judgements will be very difficult, if not impossible, to get rid of. ' (Poulton 1989)

'During the past 35 years, dozens of investigators from laboratories in various parts of the world have confirmed the power law [...]. If the experiment is conducted with care, magnitude estimation will inevitably be found to increase as a power function of stimulus intensity. Because of the consistency of this experimental outcome, the psychophysical power function has, for most psychophysicists, attained the status of an empirical law.' (Gescheider 1997)

So is MagEst a validated standard method?

- No. Highly controversial.
  - method per se? strong doubts
  - validated by consistency of results? strong doubts
- But the MagEst scale does have certain advantages ....
- ... accepted in the psychophysics literature
  eg *Functional measurement* Anderson (1962... 1992)
- For a specifically linguistic method:
  We can pick and choose the scale features à la carte.

Choosing the features of the scale we want

- Criteria:
  - scale must should impose minimum constraints on subjects
  - but still be easy to use for naive informants

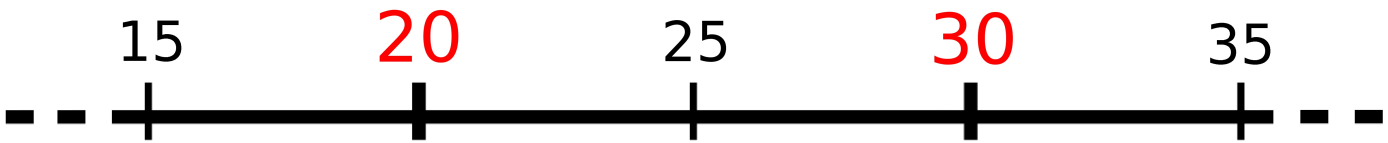
- Parameters:
  a) instructions: ratios or differences?
  b) anchors: where? how many? labels or reference items?
  c) end points: closed or open?
  d) scale type: continuous or category?
  e) scale numbers: location? range?

Building our own scale (a), (b), (c)

(a) Instructions: ratios or differences? Differences.
- perception is linear.
Birnbaum (1980), Poulton (1989), Anderson (1992), Laming (1997) ...

(b) Anchors: Two reference items at 25% & 75% of normal range
- closer reference points better (Laming 1997)
- difficult to find extremely bad reference examples
- descriptive metalinguistic labels problematic (Schütze 1996)

(c) End points closed or open? Open
- difficult to find unreachable end points in linguistics
- avoids end point distortion (Stevens 1956, Poulton 1989)

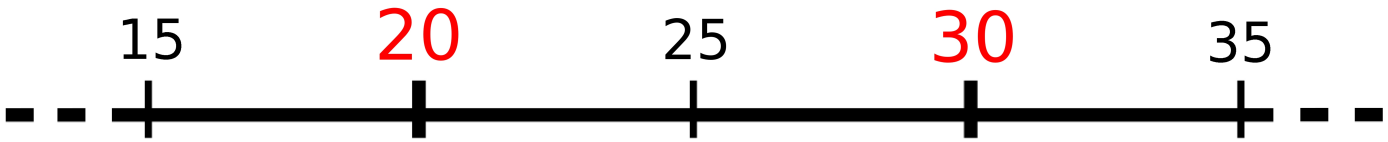


Building our own scale (d), (e)

(d) Scale type: continuous or category? Continuous
- contains more information (Anderson 1992)
- interval scale for inferential stats

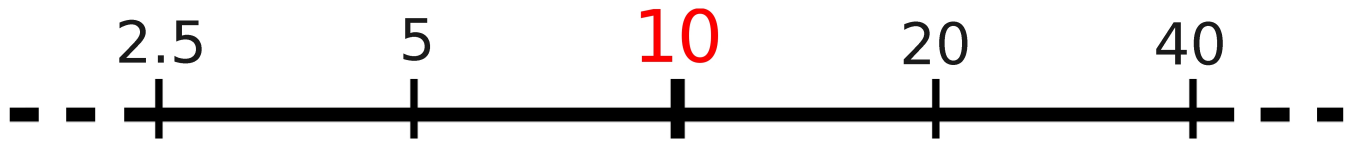
(e) Scale numbers: 25% reference item is at 20, 75% one at 30
- gives sufficient space (Anderson 1992: 20-point scale)
- allows informants to use integers
- avoids zero point distortion

■ Our scale is a bit like a thermometer scale ...
... so we call the method *Thermometer Judgements*

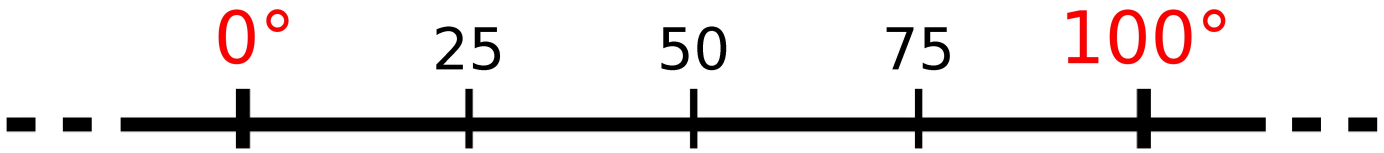


Four scales in comparison

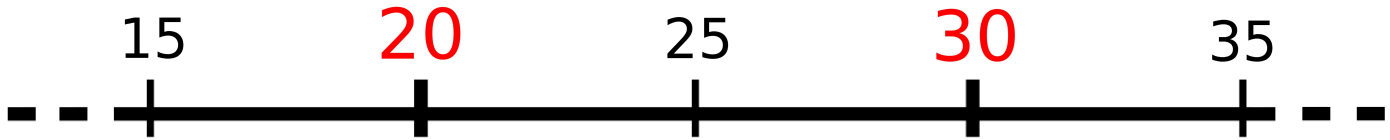
MagEst:



Celsius:



Thermometer:



7-point:

Methods conclusion

- Thermometer judgements
  - have the advantages of MagEst ...
    ... but not the disadvantages
  - related to 5-point or 7-point category scale ...
    ... but without the disadvantages

- If your experiment design is fairly simple ...
  ... you can use a category scale with no loss of information

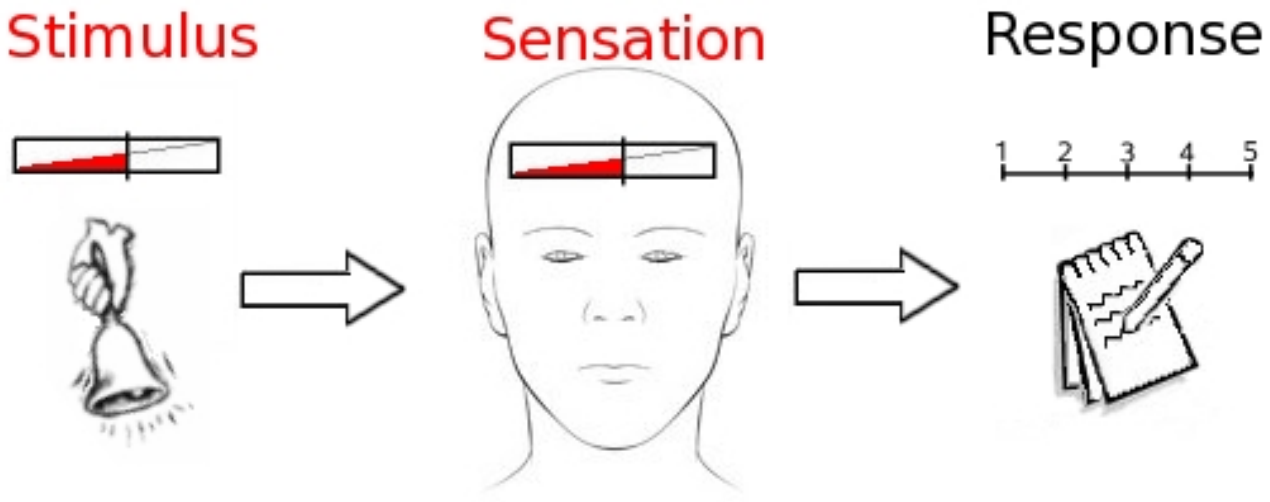
- If you want to learn about the architecture of the grammar...
  ... don't you want the least constrained data possible?

Question 2

Can we investigate language structure with judgements?

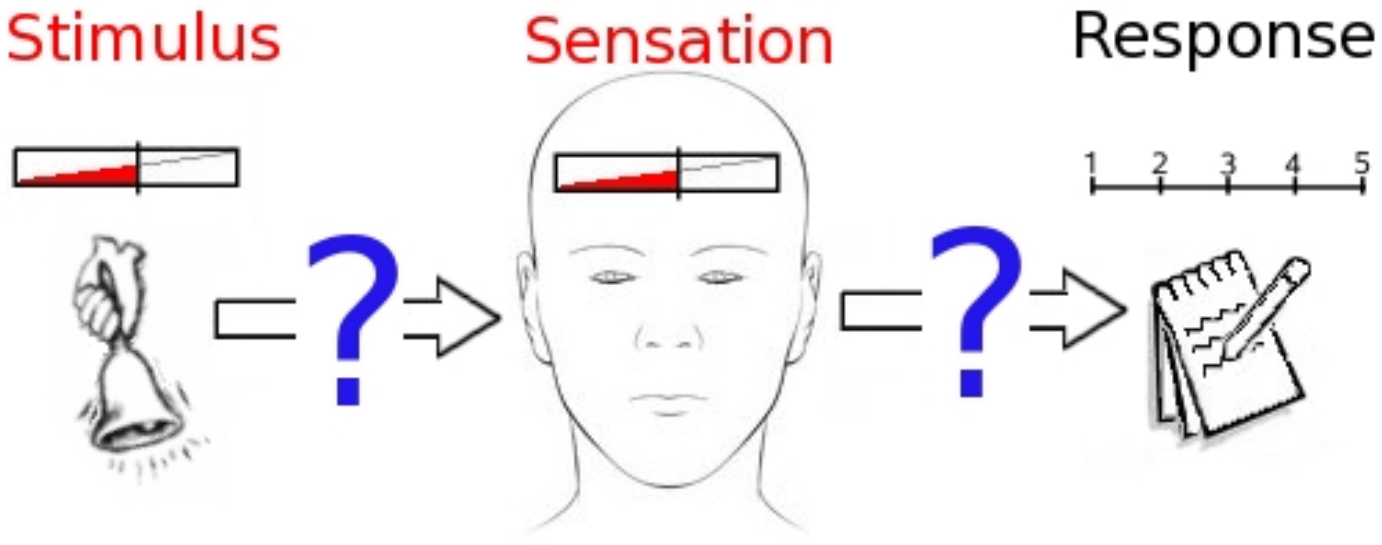
What do psychophysicists measure? Three views

- Traditional psychophysics (eg Stevens 1975): Relationship between stimulus and sensation



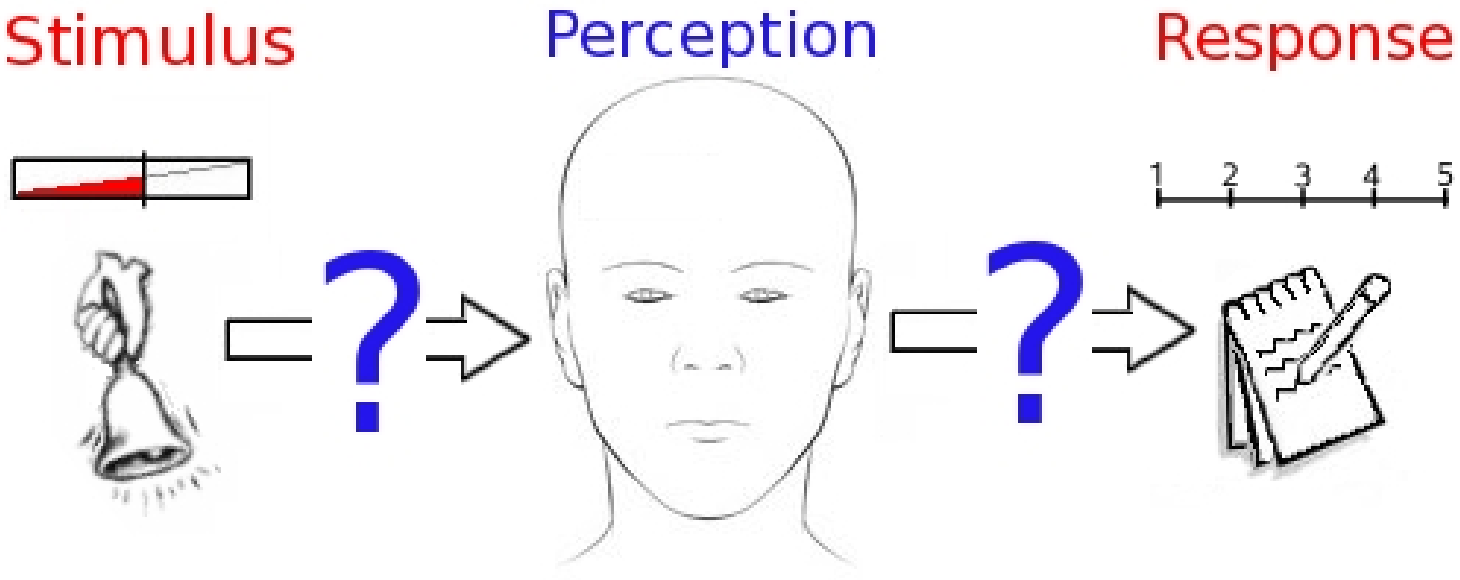
What do psychophysicists measure? Three views

- Methodological critiques of Stevens (eg Poulton 1989)
  Stimulus – sensation relationship plus perception factors



What do psychophysicists measure? Three views

- Psychophysics beyond sensation (Anderson 1992, Laming 1997, Kaernbach et al 2004):
  There is no internal sensation to be measured: just stages of perception and reporting of a physical stimulus.



Psychophysics beyond sensation:Implications for linguists

- Psychophysicists: Response = Stimulus x Perception factors
- For linguists, Stimulus = representation of language structure
- Psychophysics can pin down perception factors

  Response (known)= Stimulus (known) x Perception (unknown)
- But linguists can still operate with only one unknown ...
  ... if they pay attention to the psychophysicists' work

  Response (known)= Stimulus (unknown) x Perception (known)
- Careful judgement studies should illuminate psychologically relevant portions of language structure.

Interim conclusions

- Use of judgements to study language structure seems to be a valid psychophysical approach.
- We seem to know as much what language factors we are measuring with relative judgements as with other data types.
- Ultimately, we can only measure differences between conditions.
- But ... we need to take account of the psychophysics literature.

Question 3

Do we need any more scale than this?

Judgements are relative to a scale

- Psychophysics: Poulton (1989):
  Single most important factor in accuracy is a familiar scale
- Linguists: want absolute handles on relative judgements
  'How good are these relative to known examples?'
- What we need is a standard scale of well-formedness.

Cardinal well-formedness values

- Grounded scale for relative judgements
  - *five-group grammaticality*
  - partly arbitrary, like boiling and freezing points
  - finer grained than grammatical/ungrammatical contrast

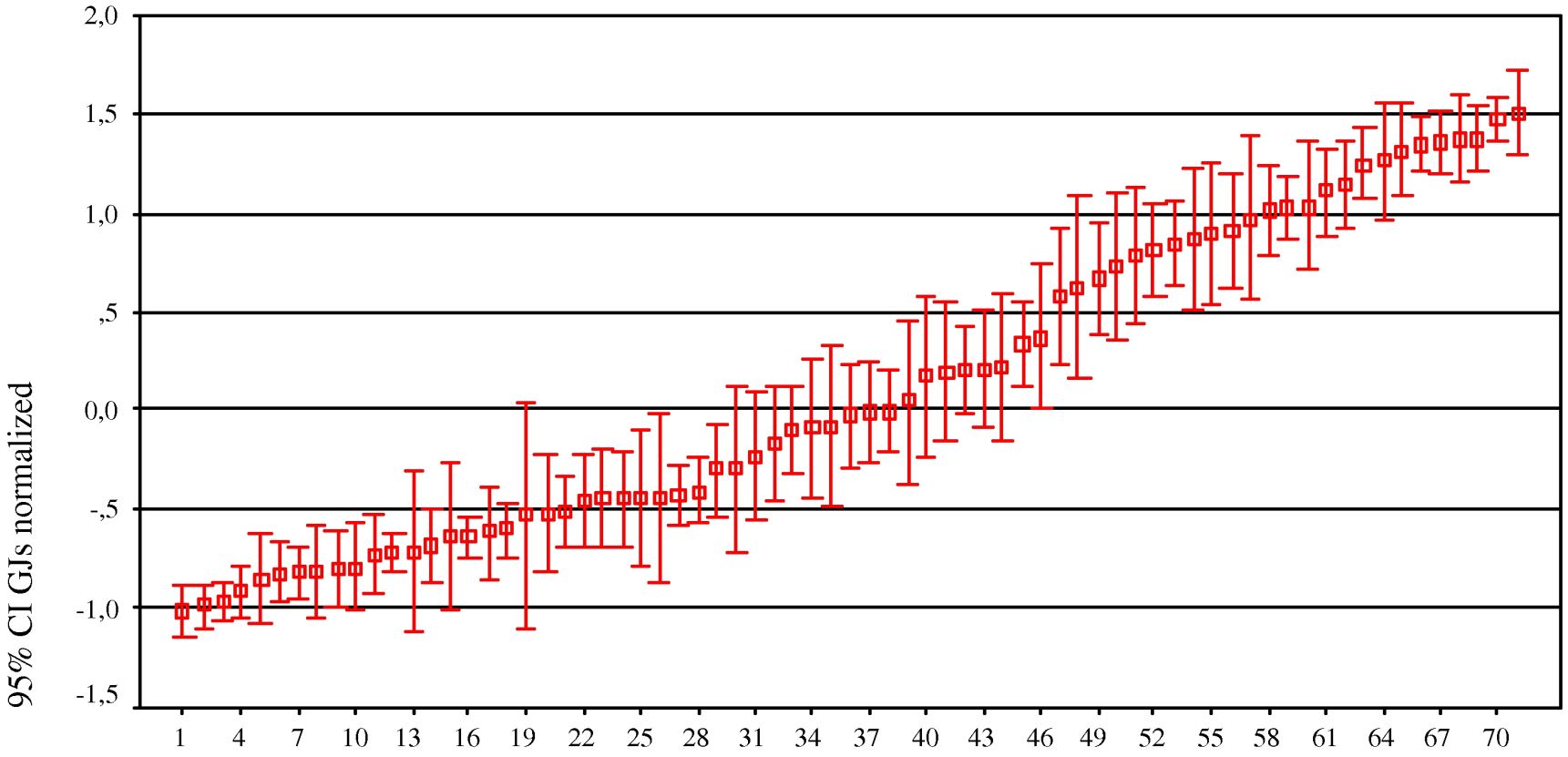
- Potential advantages:
  - provides local reference points (cf 10°C vs 20°C)
  - could sharpen individual's intuitions
  - make experiment results transferable
  - make intuitions more like objective phenomena
  - help communication of intuitions

Choosing exemplars of cardinal values

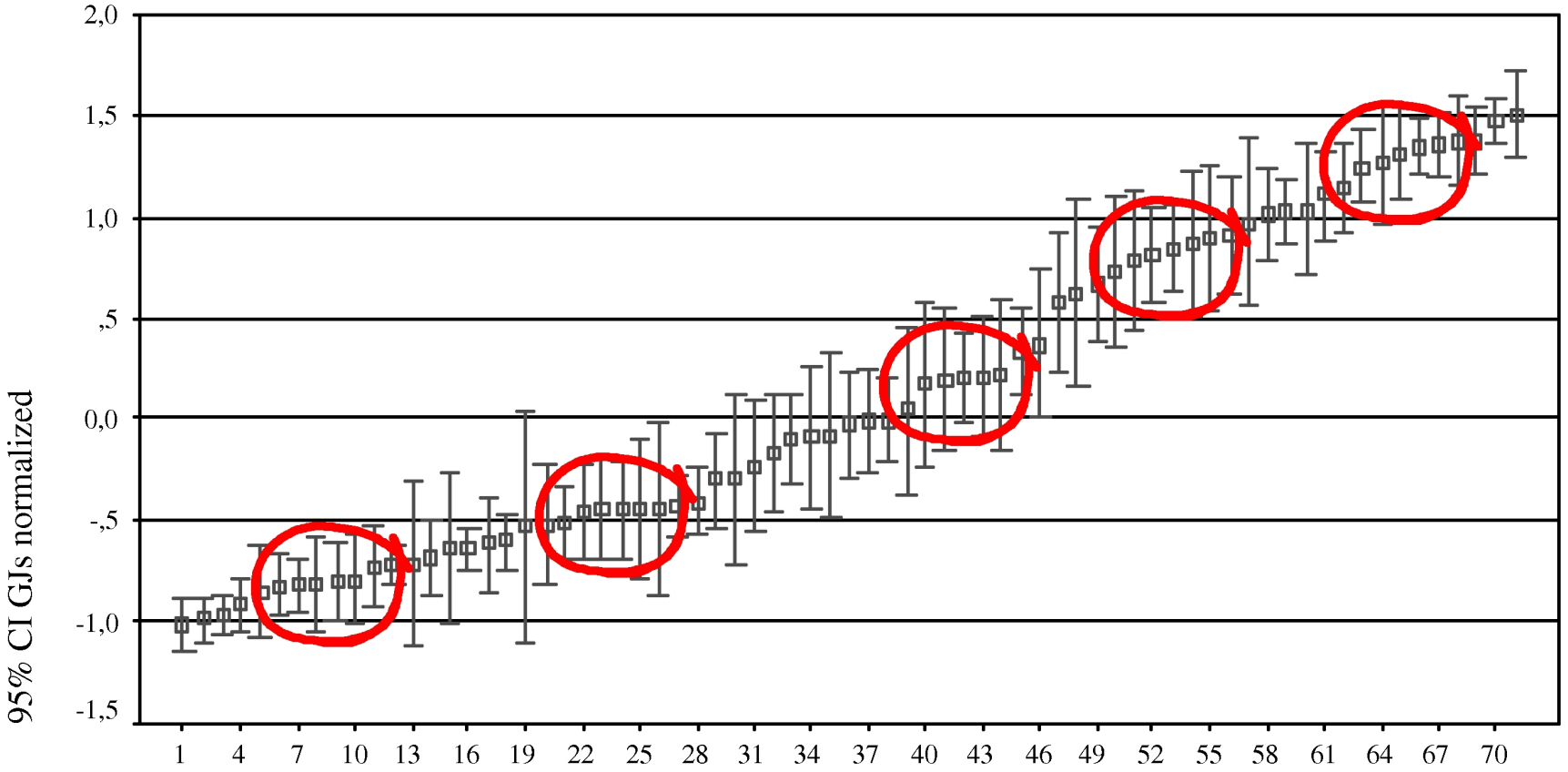
- Experiment 1: selecting candidates for groups
  - presented wide range of examples
  - varying violations, varying extenuating circumstances
  - German 71 items, English 60 items

- Experiment 2 + 3: can linguists can assign them to groups?
  Answer: yes, to within one group error

Choosing exemplars of cardinal values (1)



Choosing exemplars of cardinal values (2)



First results choosing English exemplars

Group A

This is the boy that Mary thinks will drop out of the course.
The girls like to come to town with me on Saturday morning.

Group B

You can't say that to me, who does most of the work here!
Who do you doubt will finish the marathon inside four hours?

Group C

I offered Jack to come to my party, but he said he was busy.
I insist that this sort of behaviour, we just cannot tolerate.

Group D

It is Joan that Mary thinks that is in charge of the campsite.
The three friends like to meet and play in the evening poker.

Group E

In northern Italy are very violent thunderstorms in summer.
The school children have finally finishing their drawings.

Cardinal well-formedness examples from German

Group A:

In der Mensa essen viele Studenten zu Mittag.
Nur sehr selten hört man den leisen, krächzenden Ruf eines Schwans

Group B:

Welche Zahnpasta hat der Zahnarzt welchem Patienten empfohlen?
Sie hofft, das Finanzamt hat den Betrüger überlistet.

Group C:

Was ich wissen will, ist wen wer in dieser Affäre betrügt.
Ich habe dem Kunden sich selbst im Spiegel gezeigt.

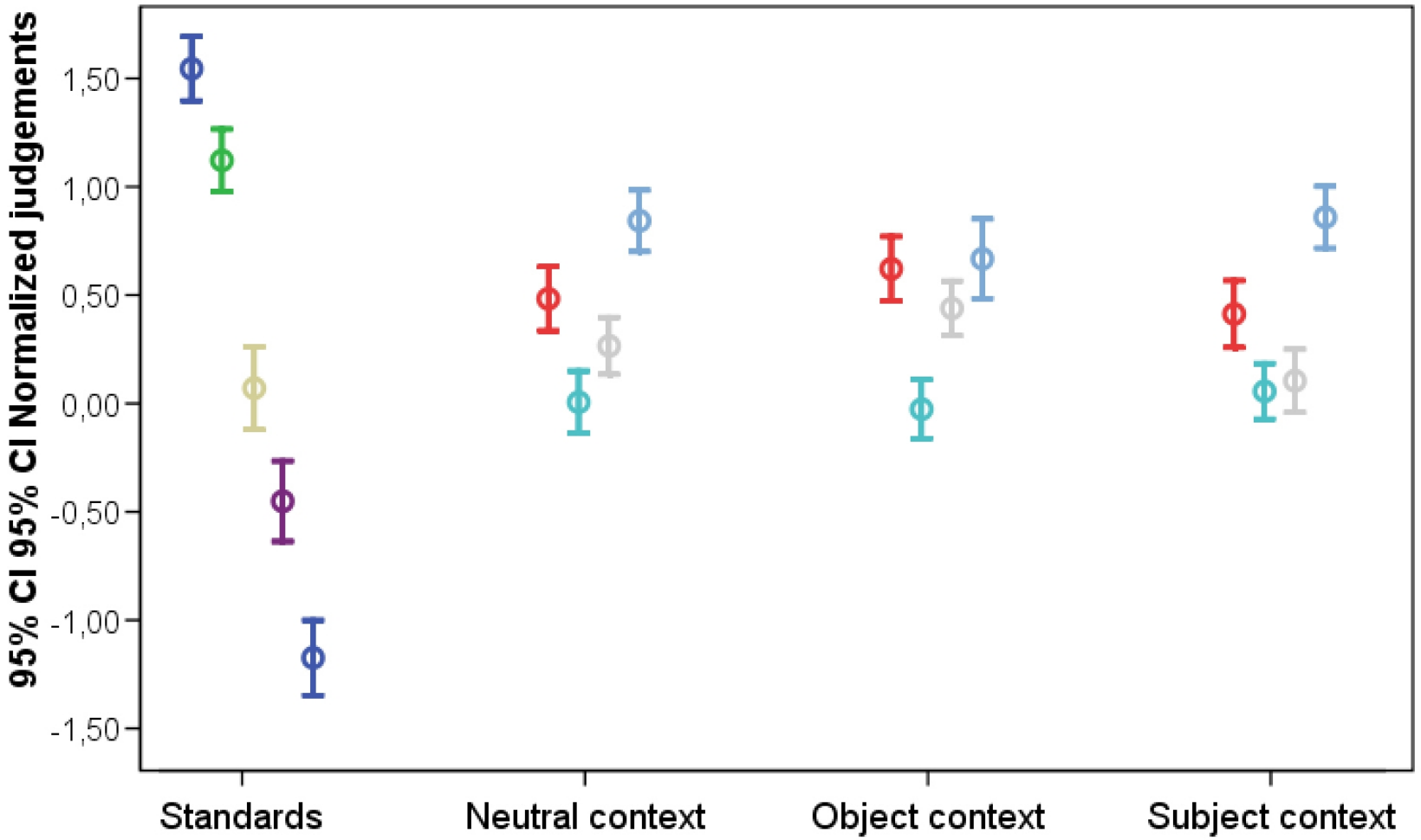
Group D:

Der Komponist hat dem neuen Tenor es zugemutet.
Welches Zimmer weißt du nicht wo sich befindet?

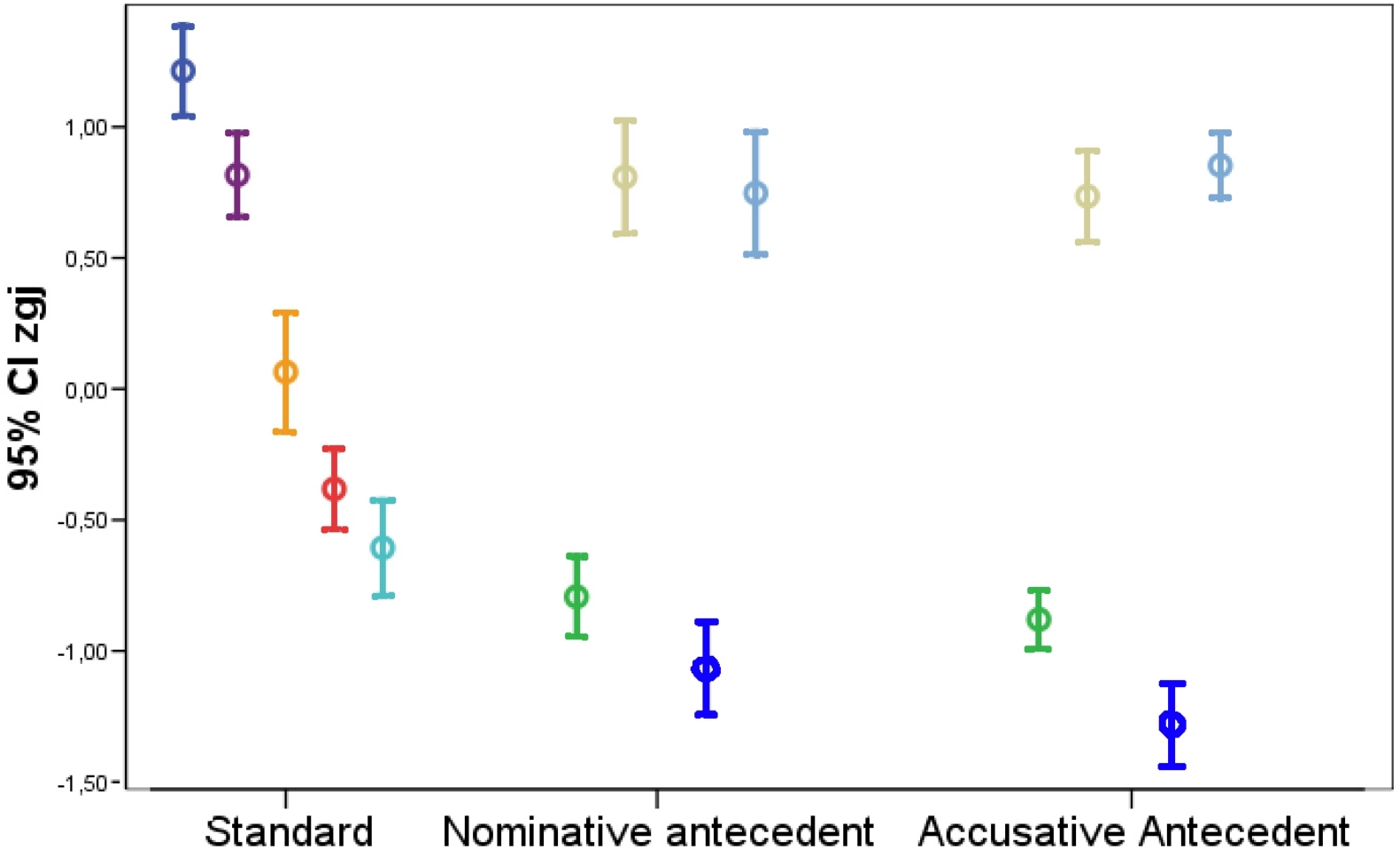
Group E:

Der Waffenhändler glaubt er, dass den Politiker bestochen hat.
Wen fragst du dich, ob Maria nicht kennenlemen sollte?

Standard items in studies (1): locating ranges



Standard items in studies (2): distinctions in the abyss …



Well-formedness scale: does it work?

- Useful in experiments:
  - provides reference points
  - makes relative judgements transferable
  - quality control

- More general uses:
  - not yet taught in primary schools

Conclusions

- MagEst not suitable as standard
  - especially in the specific variant of Bard et al (1996)
    (log conversion, normalization by single reference item)
  - little positive reason to use MagEst

- Otherwise take your pick:
  - thermometer judgements likely to give more detail,
  - 7-point scale is perhaps simpler
  - little positive reason to use binary scale

- Linguists should read the psychophysics literature

- Please use the standard scale of well-formedness

Thank you

References

Anderson N. (1992) Integration psychophysics and cognition. In: Algom D. (ed) *Psychophysical Approaches to Cognition, 13-113* Amsterdam: North Holland

Bard E., Robertson D. & Sorace A. (1996) Magnitude estimation of linguistic acceptability. *Language 72* (1), 32-68

Birnbaum D. (1980) Comparison of two theories of 'difference' and 'ratio' judgments. *Journal of Experimental Psychology: General 109*, 304-319

Featherston S. (2005) Magnitude estimation and what it can do for your syntax. Some wh-constraints in German. *Lingua 115 (11)*, 1525-1550

Gescheider G. (1997) *Psychophysics: The Fundamentals* (3rd edition). Mahwah, New Jersey: Lawrence Erlbaum

Kaernbach C., Schröger E. & Müller H. (eds) (2004) Psychophysics beyond Sensation: Laws and Invariants of Human Cognition. Mahwah, New Jersey: Lawrence Erlbaum

Keller F. (2003) A psychophysical law for linguistic judgments. Talk at AMLaP 2003, exists as paper too.

Laming, D. (1997) *The Measurement of Sensation*. Oxford: OUP

Savage C. W. (1970) *The Measurement of Sensation*. Berkeley: University of California Press

Poulton E. C. (1989) *Bias in Quantifying Judgments*. Hove & London: Lawrence Erlbaum.

Sprouse J. (2007) A program for experimental syntax: Finding the relationship between acceptability and grammatical knowledge. PhD dissertation, University of Maryland

Shepard R. N. (1981) Psychological relations and psychophysical scales: On the status of 'direct' psychophysical measurement. *Journal of Mathematical Psychology 24*, 21-57

Stevens S. S. (1956) The direct estimation of sensory magnitudes – loudness. *American Journal of Psychology 69*, 1-25

Stevens S. S. (1975) Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects. [posthumous, ed. Stevens G.]. New York: Wiley

SFB 441

A standard scale of well-formedness: Why syntax needs boiling and freezing points

Eberhard Karls Universität Tübingen

Thermometer judgements: practice 1

Reference line 1: this line is 20 long:

Reference line 2: this line is 30 long:

Your task is to judge the lengths of further lines relative to these two reference lines.

The length of this red line is 25.

And this blue line is worth 36.

Thermometer judgements: judging sentences

Reference example 1: This fairly unnatural sentence is worth 20.

The father fetches for the children it.

Reference example 2: This fairly natural sentence is worth 30.

The father fetches the sick children the food.

Judge all further examples relative to these.

This blue example might be worth 24.

The father fetches the children it.

Thermometer judgements: judging sentences

Reference example 1: This fairly unnatural sentence is worth 20.

The father fetches for the children it.

Reference example 2: This fairly natural sentence is worth 30.

The father fetches the sick children the food.

Judge all further examples relative to these.

This green example might be worth 33.

The father fetches the food for the children.

Thermometer judgements: judging sentences

Reference example 1: This fairly unnatural sentence is worth 20.

The father fetches for the children it.

Reference example 2: This fairly natural sentence is worth 30.

The father fetches the sick children the food.

Judge all further examples relative to these.

This red example might be worth 18.

The father fetches the food the children.