An Empirical Comparison of Measurement Scales for Judgements of Linguistic Acceptability

Brian Murphy[1] & Carl Vogel[2]

[1]Centre for Mind/Brain Sciences, University of Trento

[2]Department of Computer Science, Trinity College Dublin

brian.murphy@unitn.it
vogel@cs.tcd.ie

Magnitude estimation is an increasingly popular methodology for gathering judgements of linguistic acceptability in studies of syntax and semantics (see e.g. Keller and Asudeh, 2001, Featherston, 2005, Myers, 2007), and has a long history of use in other areas of linguistics (Brennan et al., 1975, Shapiro, 1997, Russell et al., 2005). Unlike categorial methods such as a Likert scale or simple pairwise comparison, magnitude estimation does not constrain participant responses to an arbitrary range, or to an arbitrary degree of granularity, and it appears to give privileged access to the internal cognitive scale that is operative during acceptability tasks (Bard et al., 1996). For this reason it is expected to provide larger effect magnitudes than a bounded scale would give (relative to data variance), and so provide more robust statistical inferences – that is, a smaller chance of drawing a false inference, a smaller chance of missing a genuine effect, and smaller margins of error. Magnitude estimation is also claimed to yield interval scale data, allowing the use of stronger parametric statistical tests (Stevens, 1946, 1975).

However, magnitude estimation is known to have issues of face-validity: though participants quickly adapt to the scale, many find it "bizarre" initially (Bard et al., 1996, p.41); and it demands a certain level of numeracy among subjects. Further, empirical comparisons of the relative utility of magnitude estimation in other non-physical domains find that it yields data that is equally or less informative than categorial methods (see e.g. Kaplan et al., 1979, Lawless, 1989, Southwood and Flege, 1999, Orth and Wegener, 2006). To the author's knowledge, no published work makes such a comparison for judgements of linguistic acceptability.

To evaluate the relative informativeness of these scales a large scale web-administered magnitude estimation experiment in English (ca. 50 participants, 150 stimuli, 1500 judgements) is replicated with a group of participants who are blindly assigned one of three measurement scales: magnitude estimation; a seven-point Likert scale; and pairwise comparison to an arbitrary reference sentence (total 140 participants). The phenomena of interest are semantic and pragmatic restrictions on

the productivity of the dative, benefactive and passive alternations in English. Dialogue extracts from popular cinema are used to provide familiar and authentic material, and explicit context is given to control participant interpretation. Participants were recruited via postings to cinema related discussion forums on the internet.

Comparison of response rates from these participants confirmed that face-validity of magnitude estimation disproportionately deters prospective respondents. Significantly higher numbers of judgements were returned per prospective participant assigned the Likert scale.

Aggregate judgements for each sentence stimulus over participant responses were also compared. A correlation of the Likert and magnitude estimation replications to the original experiment showed that they converged similarly, and this result could not be accounted for by the lower number of judgements gathered among participants assigned the magnitude estimation scale.

Finally, all three data sets were evaluated in terms of how strongly they confirmed five experimental hypotheses that are both predicted by relevant literature, and which received support from the original experiment (roughly in order of increasing subtlety): that authentic sentences are more acceptable than their constructed alternation counterparts; that the dative alternation is more successful if its indirect object is animate, and it is given in the discourse; and that the passive alternation is more successful if its subject (logical direct object) is animate, and its by-object (logical subject) is new to the discourse. Both the pairwise comparison and Likert data provided equal or stronger support for three of these hypotheses than that provided by the magnitude estimation data, regardless of any parametric assumptions and after outlier magnitude estimates had been manually removed. Again, this could not be accounted for by differences in response rates.

While different results might have been obtained in a more controlled laboratory environment (where motivation is higher, participants can be carefully screened for numeracy, and individual training in the methodology can be carried out), this suggests that magnitude estimation does not provide clear benefits as a general methodology for gathering judgements of linguistic acceptability.

LE-References

Bard, E., D. Robertson, and A. Sorace: 1996, ‘Magnitude estimation of linguistic acceptability’. Language 72(1), 32–68.

Brennan, E. M., E. B. Ryan, and W. E. Dawson: 1975, ‘Scaling of apparent accentedness by magnitude estimation and sensory modality matching’. Journal of Psycholinguistic Research 4(1), 27–36.

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

Kaplan, R. M., J. W. Bush, and C. C. Berry: 1979, ‘Health Status Index: Category Rating versus Magnitude Estimation for Measuring Levels of Well-Being’. Medical Care 17(5), 501–525.

Keller, F. and A. Asudeh: 2001, ‘Constraints on Linguistic Coreference: Structural vs. Pragmatic Factors’. In: J. D. Moore and K. Stenning (eds.): Proceedings of the 23rd Annual Conference of the Cognitive Science Society. Mahawah: Lawrence Erlbaum, pp. 483–488.

Lawless, H. T.: 1989, ‘Logarithmic Transformation of Magnitude Estimation Data and Comparisons of Scaling Methods’. Journal of Sensory Studies 4(2), 75–86.

Myers, J.: 2007, ‘MiniJudge: Software for small-scale experimental syntax’. International Journal of Computational Linguistics and Chinese Language Processing 12(2), 175–194.

Orth, B. and B. Wegener: 2006, ‘Scaling occupational prestige by magnitude estimation and category rating methods: A comparison with the sensory domain’. European Journal of Social Psychology 13(4), 417–431.

Russell, M., M. Corley, and R. J. Lickley: 2005, ‘Magnitude estimation of disfluency by stutterers and nonstutterers’. Phonological Encoding and Monitoring in Normal and Pathological Speech 1(5), 248–260.

Shapiro, M.: 1997, ‘Style Shifting: The perception and production of formality in English’. Ph.D. thesis, Department of Linguistics, University of Texas at Austin.

Southwood, M. H. and J. E. Flege: 1999, ‘Scaling foreign accent: direct magnitude estimation versus interval scaling’. Clinical Linguistics and Phonetics 13(5), 335–349.

Stevens, S. S.: 1946, ‘On the Theory of Scales of Measurement’. Science 103, 677–680.

Stevens, S. S.: 1975, Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects. New York: John Wiley & Sons.