# Magnitude Estimation and the Non-Linearity of Acceptability Judgments

**Jon Sprouse**
**University of California, Irvine**

## 1. Magnitude estimation in linguistics and psychophysics

The term *experimental syntax* – the use of psycholinguistic methodologies for the collection of acceptability judgments – can cover any number of designs, tasks, and statistical analyses (Cowart 1997, Schütze 1996). Over the past decade, one task in particular, the magnitude estimation task, has received significant attention for its alleged ability to provide more accurate data, almost to the point of becoming a 'gold standard' among judgment collection techniques (Bard et al. 1996, Keller 2000, and Keller 2003). While magnitude estimation has been a staple task of psychophysics for over 50 years (Stevens 1957), it has only become a part of linguistic methodology thanks to the demonstration by Bard, Robertson, and Sorace (1996) that it could be profitably adapted for the collection of acceptability judgments. In the ensuing decade, magnitude estimation has been applied to a number of areas of syntactic research with exciting results (e.g., Featherston 2005a, 2005b, Sorace and Keller 2005); however, since that seminal paper there has been relatively little research into the task itself. For example, given that magnitude estimation was originally developed to measure the perception of physical stimuli, there may be certain assumptions built into the magnitude estimation task that may not be compatible with the perception of linguistic acceptability. This paper investigates one such assumption: that participants are able to use the modulus to estimate acceptability along a linear scale. As we shall see shortly, there are two manners in which the magnitude estimation task could be used to estimate acceptability along a linear scale, yet the four experiments presented in this paper suggest that neither manner is actually adopted by the participants. These results indicate that linguistic magnitude estimation differs significantly from psychophysical magnitude estimation, which may necessitate a reconsideration of one of the methodological advantages that have been offered for the widespread adoption of linguistic magnitude estimation by syntacticians: that the use of a modulus stimulus leads to more accurate measures of acceptability.

Before beginning the investigation of the linguistic magnitude estimation task, it seems worthwhile to give a brief overview of the magnitude estimation task itself. The magnitude estimation task was originally developed to investigate humans' perception of physical stimuli. For instance, if the brightness of a light source is doubled, is it the case that we perceive the light as twice as bright as the original? While a priori you may be inclined to answer 'Yes, of course,' psychophysicists in the middle of 20$^{th}$ century used the magnitude estimation task to determine that, in fact, you would perceive the light as only 1.4 times as bright. The task itself was straightforward. Participants were presented with a single example of the stimulus, for example a light source, and told that its magnitude (in this case brightness) is 100 units. They were then presented with other examples of the stimulus, and asked to estimate the magnitude of the new stimuli based on the original. If they perceived the second light as twice as bright, they would report a brightness of 200; if they perceived the second light as half as bright, they would report a brightness of 50. The reference stimulus (called the modulus) is kept constant throughout the experiment so that every experimental item is estimated using the same reference. In this way, the experimenters could compare participants' reported perceptions with the actual physical measurements of the stimulus to determine the nature of human perception for various stimuli (brightness, volume, heat, length, and hundreds of others). As it turns

 Proceedings of the 27th West Coast Conference on Formal Linguistics, ed. Natasha Abner and Jason Bishop, 397-403. Somerville, MA: Cascadilla Proceedings Project.

out, the perception of each stimulus has a characteristic relationship to its physical value, a property captured in Stevens' (1957) Psychophysical Power Law.

Bard et al. 1996 demonstrated that the magnitude estimation (ME) task could be straightforwardly adapted to the measurement of the acceptability of sentences, with the resulting task looking nearly identical to the psychophysical task. Participants are presented with an example sentence and told that its acceptability is 100 units. They are then asked to estimate the acceptability of subsequent sentences based on the initial sentence. For instance, if they feel the second sentence is twice as acceptable as the first, they will respond 200; if the feel the second sentence is only half as acceptable as the first, they will respond 50. Again, the reference (or modulus) sentence is kept constant throughout the experiment. However, because there is no independent (physical) measure of acceptability, there is nothing to compare participants' responses to. Instead, the participants' responses themselves are taken as the actual measurement (as is standard in linguistics). Because the task itself mirrors the psychophysical version in all respects except final data interpretation, it has been assumed that the properties and benefits of psychophysical ME transfer to linguistic ME as well.

## 2. The role of the modulus and two types of linearity

From the brief description of the ME task above, it should be clear that the primary innovation of the ME task is the comparison of target stimuli to the modulus stimulus, and that this innovation is the motivating force behind claims that ME data is more accurate than other types of data (e.g., n-point scale tasks). Crucially, these claims assume that the modulus sentence is playing a role in the measurement of the target sentences along a linear scale of acceptability, with two logically possible roles for the modulus to play:

i. The modulus could be serving as a uniform unit of measure against which the acceptability of the target sentences are estimated

ii. The modulus could be serving as a single point of reference along the linear scale of acceptability

The first potential role of the modulus, that of a unit of measure, is a logical possibility given the nature of the ME task itself: participants are instructed to report the magnitude of each target stimulus by comparing it to the magnitude of the modulus stimulus. In psychophysical ME, this has the effect of converting the modulus stimulus into a unit of measure with which to estimate the magnitude of subsequent stimuli. This becomes readily apparent when the stimulus in question is the physical length of lines: if the modulus stimulus is a line whose length is assigned the number 100, then reporting the length of a target line as 200 would be equivalent to estimating its length as 2 modulus-units. The appeal of a unit of measure for acceptability is obvious: acceptability is a psychological property with no physical measurable equivalent, therefore there has never been any independent way to ensure that two speakers are using the same scale (i.e., the same units) to report their perceptions of acceptability. For example, prior to the introduction of ME, the most sophisticated acceptability judgment experiments employed a Likert-style n-point scale task. However, in n-point scale tasks each participant is free to determine the meaning of the points on the scale independently of the other participants – just think about the meaning of letter grades in classes taught by different teachers (Lodge 1981). By using the modulus as a psychologically defined unit of measure, and assuming that this unit of measure is stable across participants, the ME task standardizes the scale across participants by providing a meaningful measure of the distance between stimuli on the scale of acceptability. Theoretically, this means that ME data is not only more accurate than n-point scale data, but also amenable to a wider, and more sensitive, array of statistical analyses.[1]

[1] For the statistically minded, this is just a long-winded way of saying that ME yields ratio data while n-point scales yields ordinal data.

The second potential role of the modulus, that of a reference point along a linear scale, would suggest that participants in fact ignore the explicit instructions of the linguistic ME task, a position expressly advocated in Featherston 2007. Featherston argues that instead of a unit of measure, participants interpret the modulus a single reference point along a linear scale of acceptability and that they then compare each subsequent sentence to this reference point, assigning it a numeric distance from the reference as they see fit. Crucially, he argues that participants are consistent in assigning distances between the target sentences and the reference point: if sentence A and sentence B are both psychologically equidistant from the reference for a given participant, they will be reported as numerically equidistant by that participant as well.[2] Again, if the distances between sentences rated in an ME task are indeed meaningful, then ME data is technically more accurate than n-point scale data, as is amenable to a wider, and more sensitive, array of statistically tests.

## 3. The Experiments

Four experiments serve as the empirical basis for the investigation of the two possible roles of the modulus in linguistic ME. The four experiments are identical in every respect, except that the modulus sentence is different in each experiment. The modulus sentences and the number of participants in each experiment are presented in Table 1 below. All participants were self-reported native speakers of English with no formal training in linguistics.

The body of the experiments consisted of 5 blocks of 10 sentences, for a total of 50 sentences. Each block contained 1 token of each of 8 violation types, and 2 grammatical sentences, therefore each participant saw 5 tokens of each violation type. Examples of each of the violations, all of which are constraints on wh-movement, are presented in Table 2 below. All items were matched for length in clauses (2) and in words (9). Five orders of the blocks were derived using a Latin-Square design, and items within each block were randomized. Responses were divided by the value of the modulus sentence (100) prior to analysis, and the 5 tokens of each violation per participant were averaged prior to statistical tests..

Table 1: Modulus Sentences and Sample Sizes

| Experiment | Type | Modulus | Sample |
|---|---|---|---|
| 1 | IF-island | What do you wonder if Larry bought? | 22 |
| 2 | CSC-violation | What do you think that Larry bought a shirt and? | 24 |
| 3 | NC-island | What did you start the rumor that Larry bought? | 23 |
| 4 | RC-island | What did Larry help the customer who bought? | 23 |

NC = Noun Complement, CSC = Coordinate Structure Constraint, RC = Relative Clause

Table 2: Sentence Types

| Sentence Type | | Example |
|---|---|---|
| (G) | Grammatical | What does Bill think that you are cooking tonight? |
| (Adj) | Adjunct island | Who did Mary hide her face because she recognized? |
| (CSC) | Coordinate Structure | What does Jane think that you should eat carrots and? |
| (FSS) | Finite Sentential Subject | Who did that Frank danced with shock the guests? |
| (ISS) | Infinitival Sent. Subject | What can to see be scary for a child? |
| (LBC) | Left Branch Condition | Whose did John think that you saw father yesterday? |
| (NC) | Noun Complement | What did you doubt the claim that Jessica invented? |
| (RC) | Relative Clause | Who does Erin trust the nurse who cared for? |
| (WH) | Whether island | Who do you wonder whether Mike met on vacation? |

[2] Presumably the actual numerical value of the intervals may vary from participant to participant. As we shall see in section 5, there is no evidence of this type of variation in the ME data collected in these experiments. However, if such variation were to occur, it could be trivially normalized with something like a z-score transformation.

## 4. The modulus as a unit of measure

If the modulus is indeed treated like a unit of measure, then there should be a fixed relationship between the acceptability value of the modulus and the acceptability values of all of target structures measured with that modulus. For example, if one target structure is measure as 2 modulus-units, and a second is measured as .5 modulus-units, then these two structures should be in a fixed 4:1 relationship regardless of the experimental context. Therefore, if the first target structure were then made the modulus in a subsequent experiment, and therefore assigned a value of 1 modulus-unit, the second target structure should still be measured at the same ratio of 4:1, or in other words, it should be .25 modulus-units in the new experiment. The four experiments build on this logic directly: the first experiment establishes a relationship among 9 target sentence types, and then tests those relationships in three follow-up experiments. In each follow-up experiment, a different sentence structure from the original 9 is given the role of modulus (equivalent to 1 modulus-unit) and the experimental acceptability values of the other 8 sentence structures is compared to the predicted value based on the relationships from the first experiment. If the modulus is indeed acting as a unit of measure, we would expect a nearly perfect convergence between the predicted values and the experimental values.

Responses to the ME task were divided by the value of the modulus sentence (100) prior to analysis. Predictions for each of the subsequent experiments were calculated in the following way: the mean for the appropriate modulus condition (NC-island, CSC-island, and RC-island) was taken from the IF-reference results. That value was then set equal to 1 in order to obtain a prediction factor (Mean x Prediction Factor =1), and the prediction factor was then applied to the other conditions in the IF-reference experiment to obtain the predicted values for the new experiment. The results from each condition in the three experiments were compared to the predicted values using a one-sample t-Test to determine if they differed significantly. The results are summarized in the three graphs in Figure 2, with significant results indicated by asterisks (* = p < .05, ** = p < .01, *** = p < .001).

Figure 2: NC-Reference - Predicted Means versus Actual Means and t-Test results

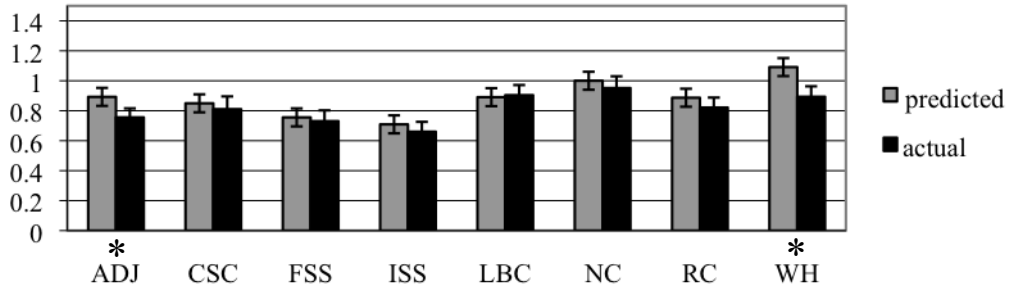



Figure 3: CSC-Reference - Predicted Means versus Actual Means and t-Test results

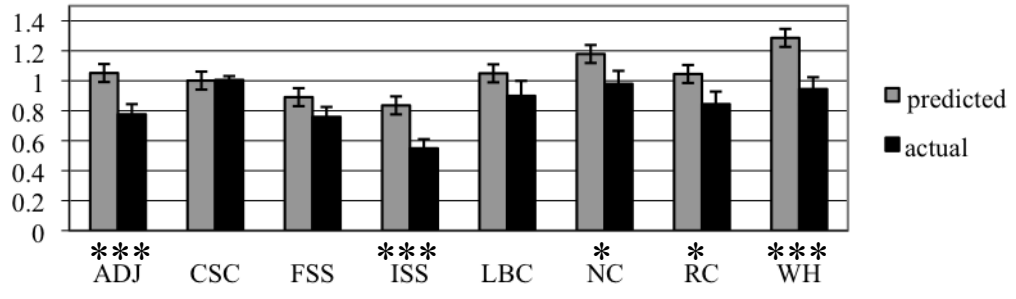
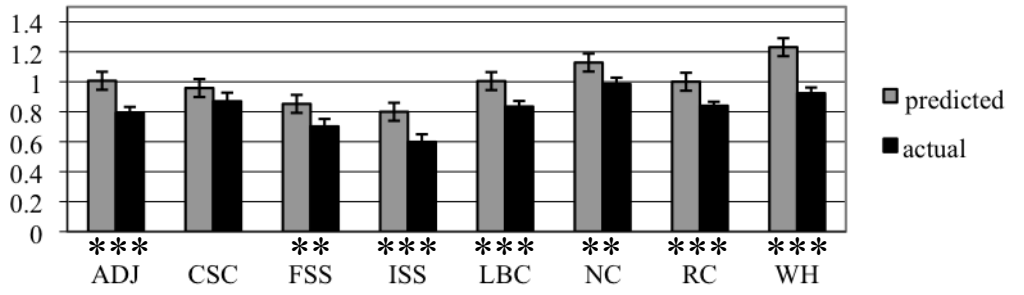


Figure 4: RC-Reference - Predicted Means versus Actual Means and t-Test results




In the NC-reference experiment, two of the conditions differed significantly from the predicted values: ADJ and WH. In the CSC-reference and RC-reference experiments the number of conditions differing significantly from the predictions increases: in the CSC-reference five of the eight conditions differed significantly, and in the RC-reference seven of the eight conditions differed significantly.

The significant differences between predicted and actual values in the follow-up experiments suggest that participants are not using the modulus sentence as a unit of measure. In practical terms, this means that linguistic ME results cannot be interpreted the same way as psychophysical ME: the relationship between stimuli in linguistic ME is dependent on the composition of the experiment itself, and not inherent to the stimuli in question. If linguistic ME is indeed a more precise task for the measurement of acceptability it cannot be because the modulus serves as a unit of measure. These results, however, are not entirely surprising: in order to use the modulus as a unit of measure, there must be a meaningful and stable zero point. In other words, it must be possible to have the absence of that stimulus. Because it is not at all clear what the absence of acceptability would be, it seems unlikely that there is a meaningful zero point for acceptability. Yet despite the lack of a meaningful zero point for acceptability, the instructions of the ME task ask participants to respond as if such a zero point exists. The fact that the modulus ultimately does not act as a unit of measure can be taken as corroborative evidence that there is no meaningful zero point of acceptability.

## 5. The modulus as a reference point along a linear scale

The claim that the modulus acts as a point of reference makes a straightforward prediction for linguistic ME data: the distances between sentence structures should remain constant relative to one another *and shift up and down the numerical scale as the modulus is changed*. We can look for this shifting of the scale in the data from the four experiments presented above. Recall that each of the four experiments tested the same set of sentences, but manipulated the modulus sentence. Because the modulus sentence was always assigned the value 100, which becomes 1 during analysis, we would expect to find the same relative pattern of results in each experiment, but with a different location on the scale depending on which modulus was used: if the modulus was a structure with high acceptability, we would expect the bulk of the sentences to be below 1; if the modulus had low acceptability, we would expect the bulk of the sentences to be above 1.

The following pairs of graphs illustrate that this shift along the scale is not present in the data from these four experiments. The graphs on the left are the IF-reference with a horizontal line through the value of the modulus in the experiment in the graph to the right. The horizontal line in the graphs to the right indicates the value 1. If the prediction of this scenario holds, the graphs on the left should look nearly identical to the graphs on the right, with each condition appearing in the same relative position with respect to the horizontal line (in other words, the absolute value of the horizontal line and conditions should shift, but the relative relationship should remain the same). However, as these graphs indicate, the relative positions of the conditions with respect to the horizontal line are not maintained from experiment to experiment:

Figure 5: Dispersion around NC-islands in the If-reference and NC-reference experiments

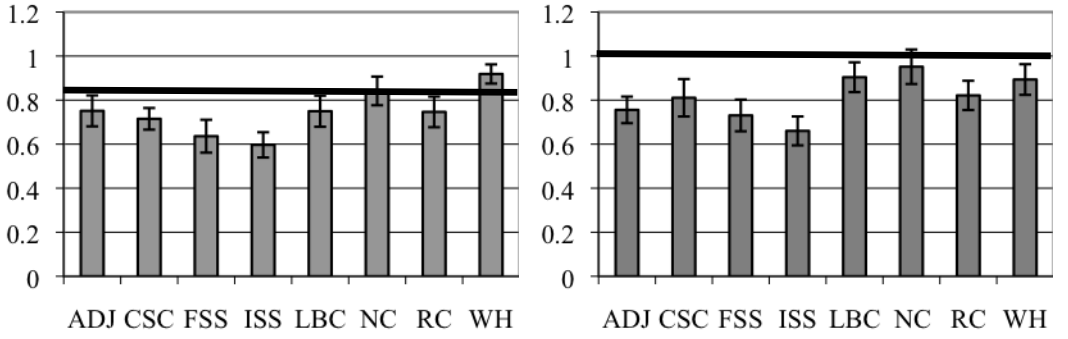



Figure 6: Dispersion around CSC-violations in the If-reference and CSC-reference experiments

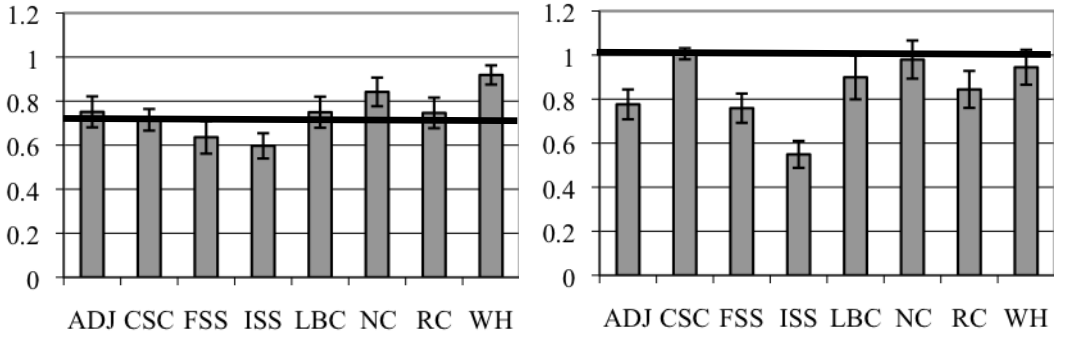
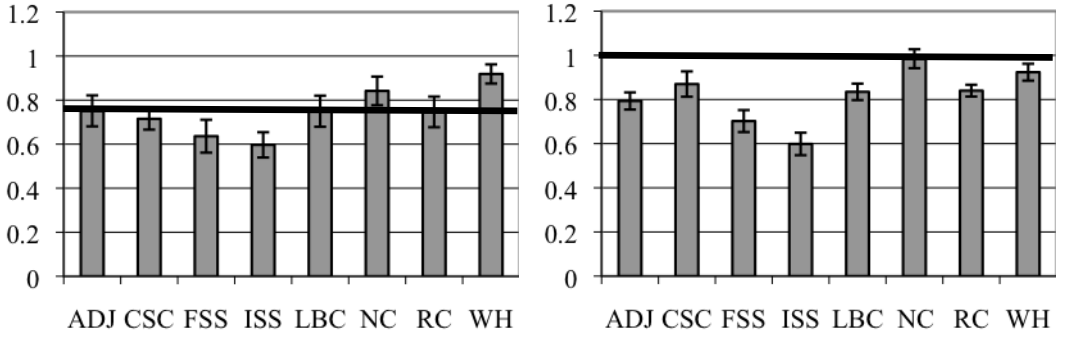



Figure 7: Dispersion around RC-islands in the If-reference and RC-reference experiments




This failure to shift around the modulus as it changes from experiment to experiment suggests that the modulus is not acting as a reference point along a linear scale of acceptability. If anything, the modulus value appears to be acting as an upper bound for the judgments of the ungrammatical sentence types (see also Sprouse *in prep*). This again suggests that the modulus plays no role in the measurement of the other conditions, and therefore cannot be used to argue that linguistic ME provides more accurate or more precise data than other acceptability collection tasks.

## 6. Conclusion

This paper investigated the claim that the linguistic ME task allows participants to more accurately estimate the acceptability of target sentences by using the modulus sentence to locate the acceptability of the target sentence along a linear scale of acceptability. There are two logically possible role for the modulus in such a measurement process: it could either act as a unit of measure for the target sentences, or it could act as a single reference point along the scale around which the target sentences can be located. However, the results from the four experiments presented here suggest that the modulus plays no role in the measurement process at all (and if anything, acts as a numerical upper bound for the ungrammatical target sentences). These results suggest not only a significant difference between linguistic ME and psychophysical ME, but also that linguistic ME may not provide more accurate data than other tasks (e.g., n-point scale tasks) as has been previously suggested (Bard et al 1996, Keller 2000, Featherston 2005, etc.), because previous claims that linguistic ME provides more accurate data have been predicated upon the comparison of the target sentences to the modulus.

## References


Bard, Ellen G., Dan Robertson & Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72, 32-68.

Coulson, Seana, Jonathan King, and Marta Kutas. 1998. Expect the Unexpected: Event-related Brain Response to Morphosyntactic Violations. *Language and Cognitive Processes* 13(1): 21-58.

Cowart, Wayne. 1997. Experimental Syntax: Applying object methods to sentence judgments. Sage.

Featherston, Sam. 2005a. Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua* 115:1525–1550.

Featherston, Sam. 2005b. Universals and grammaticality: wh-constraints in German and English. *Linguistics* 43:667–711.

Featherston, 2007. Data in Generative Grammar: The carrot and the stick. *Theoretical Linguistics* 33(3): 269-318

Keller, Frank. 2000. Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. Edinburgh: University of Edinburgh dissertation.

Keller, Frank. 2003. A psychophysical law for linguistic judgments. In Richard Alterman & David Kirsh (eds.), *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, 652-657.

Lodge, Milton. 1981. Magnitude Scaling: Quantitative measurement of opinions. Sage.

Schütze, Carson. 1996. The empirical base of linguistics: Grammaticality judgments and linguistic methodology. The University of Chicago Press.

Sorace, Antonella, and Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115:1497–1524.

Sprouse, Jon. 2007. A program for experimental syntax. College Park, MD: University of Maryland dissertation.

Sprouse, Jon. *In prep*. Evaluating Linguistic Magnitude Estimation.

Stevens, Stanley Smith. 1957. On the psychophysical law. *Psychological Review* 64:153–181.

# Proceedings of the 27th West Coast Conference on Formal Linguistics

edited by Natasha Abner
and Jason Bishop

**Cascadilla Proceedings Project  Somerville, MA  2008**

## Copyright information



ISBN 978-1-57473-428-7 library binding



## Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

## Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Sprouse, Jon. 2008. Magnitude Estimation and the Non-Linearity of Acceptability Judgments. In *Proceedings of the 27th West Coast Conference on Formal Linguistics*, ed. Natasha Abner and Jason Bishop, 397-403. Somerville, MA: Cascadilla Proceedings Project.

or:

Sprouse, Jon. 2008. Magnitude Estimation and the Non-Linearity of Acceptability Judgments. In *Proceedings of the 27th West Coast Conference on Formal Linguistics*, ed. Natasha Abner and Jason Bishop, 397-403. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #1855.