A TEST OF THE COGNITIVE ASSUMPTIONS OF MAGNITUDE ESTIMATION: COMMUTATIVITY DOES NOT HOLD FOR ACCEPTABILITY JUDGMENTS

JON SPROUSE

University of California, Irvine

The introduction of the psychophysical technique of MAGNITUDE ESTIMATION to the study of acceptability judgments (Bard et al. 1996) has led to a surge of interest in formal acceptability-judgment experiments over the past fifteen years. One of the primary reasons for its popularity is that it was developed as a tool to measure actual units of perception, offering the possibility of data that is inherently more informative than previous scaling tasks. However, there are several untested cognitive assumptions that must hold in order for ME to be the perceptual measurement test that it is purported to be. Building on the recent formalization of these assumptions in the psychophysics literature (Narens 1996, Luce 2002), this article presents two experiments designed to test whether these assumptions hold for acceptability-judgment experiments. The results suggest that the cognitive assumptions of magnitude estimation do not hold for participants in acceptability-judgment experiments, eliminating any reason to believe that ME could deliver inherently more meaningful data than other acceptability-judgment tasks.*

Keywords: magnitude estimation, acceptability judgments, experimental syntax, commutativity, multiplicativity

1. INTRODUCTION. The past fifteen years have seen a steady increase in the use of formal experiments for the collection of acceptability judgments as opposed to the traditional informal experiments that have characterized much of (generative) syntactic literature. The recent surge in popularity of formal experiments is due in no small part to the introduction of a psychophysical task known as MAGNITUDE ESTIMATION to the field of syntax by Bard, Robertson, and Sorace (1996), and the initial studies that adopted the task (e.g. Cowart 1997, Keller 2000, 2003, Featherston 2005a,b). Because of its unique development as a tool within psychophysics, magnitude estimation is a fundamentally different cognitive task from the other acceptability-judgment tasks: in magnitude estimation, participants are asked to estimate the acceptability of a target sentence by using the acceptability of a different sentence as a unit of measure. This fundamentally different measurement procedure has been argued to lend magnitude estimation a type of measurement reliability that other scaling tasks cannot achieve (Stevens 1956, 1957, Bard et al. 1996, Cowart 1997, Keller 2000). Because of claims such as these, magnitude estimation has in many respects become a 'gold standard' in the acceptability-judgment literature today.

Given the ascendency of magnitude estimation in linguistics, it is perhaps surprising to note that in the years since Bard and colleagues (1996) first adapted magnitude estimation for use in syntax, the field of psychophysics has systematically questioned whether participants can actually perform the cognitive task asked of them by the magnitude-estimation procedure. Psychophysicists have formalized the set of cognitive assumptions underlying the magnitude-estimation task, and have also developed a set of procedures for empirically verifying whether those assumptions are met by participants in psychophysical magnitude-estimation experiments (Narens 1996, Luce 2002). Recent experiments have suggested that participants meet only one of the two critical cognitive assumptions, at least for magnitude estimation of loudness (Ellermeier &

* This research was supported in part by National Science Foundation grant BCS-0843896 to JS. I would like to thank Diogo Almeida, as well as three anonymous referees, for helpful comments on previous versions of this work. All errors remain my own.

Faulhammer 2000, Zimmer 2005). Though several early adopters of magnitude estimation in linguistics have independently expressed similar skepticism of participants' ability to perform magnitude estimation of acceptability (e.g. Sprouse 2007, Featherston 2008, Weskott & Fanselow 2011 (this issue)), critical differences between the stimuli used in psychophysics (physical stimuli) and the stimuli used in syntax (sentences) have thus far prevented a direct investigation of whether the cognitive assumptions of magnitude estimation hold for acceptability experiments. This article introduces a novel methodology for adapting the psychophysical procedures for use with acceptability judgments, and presents two experiments that employ this methodology to finally determine whether the cognitive assumptions of magnitude estimation are met during acceptability experiments.

The ability to test the fundamental cognitive assumptions of magnitude estimation is particularly relevant given the recent discussion about the statistical power of magnitude estimation (compared to the statistical power of other tasks) in the literature (Featherston 2008, Myers 2009, Bader & Häussler 2010, Weskott & Fanselow 2011). These discussions are direct consequences of the early claims that magnitude estimation is a fundamentally different type of measurement task. That claim is in turn predicated upon the currently untested assumption that the cognitive assumptions of magnitude estimation hold for participants in acceptability-judgment tasks. If it indeed turns out that the assumptions of magnitude estimation do not hold, then there would no longer be any reason to believe that magnitude estimation could yield data that is superior to other judgment tasks.

Several papers in the literature have attempted to address this question from a different direction by directly comparing the results of magnitude-estimation experiments with the results of other tasks (e.g. Featherston 2008, Myers 2009, Bader & Häussler 2010, Weskott & Fanselow 2011). However, because such comparisons are dependent on the sentence types chosen, the sample sizes tested, and the statistical tests employed, it is logically possible that future experiments could reveal that magnitude estimation does indeed yield superior data. A direct test of the cognitive assumptions of magnitude estimation avoids this problem altogether: barring any major technical problems with the experiments, there is no logical way for the cognitive assumptions of the task to change based on sentence types, sample sizes, or statistical tests. This means that syntacticians can determine once and for all whether there is any reason to believe that magnitude estimation offers a type of measurement that is distinct from, and superior to, the other scaling tasks. As becomes clear below, the results of the experiments presented here suggest that the cognitive assumptions of magnitude estimation do not hold with respect to acceptability judgments. This suggests that there is no way for magnitude estimation to be the distinct (and superior) cognitive measurement task that it has been purported to be, and therefore magnitude estimation has no inherent claim to the mantle of 'gold standard' among acceptability-judgment tasks.

2. THE PURPORTED PROPERTIES OF MAGNITUDE ESTIMATION. Before discussing the formal cognitive assumptions of magnitude estimation (ME), it seems important to establish the basic properties that have been attributed to it. The most appropriate way to do this is to compare it directly to standard scaling tasks, as this was the original motivation for the development of ME (Stevens 1956) and the importation of ME to syntax (Bard et al. 1996). Syntacticians generally agree that acceptability is a continuum (although the boundaries of that continuum are not always obvious), and that acceptability-judgment experiments can help quantify the position of a given sentence along that continuum.

Prior to Bard et al. 1996, the most common methodology for this was to define a set of equally spaced points along the continuum, usually five or seven points, and ask participants to choose the point that is closest to the position of a given sentence. This task is sometimes called the seven-point scale task, or a (type of) Likert scale task; for ease of exposition I call it the *n*-point scale task. If acceptability is indeed a continuous measure, then the *n*-point scale task introduces two possible distortions to the reported judgments. First, the limited number of points along the scale means that participants can maximally distinguish *n*-levels of acceptability intentionally. This raises the possibility that a participant may wish to distinguish more levels of acceptability than the scale allows. Whether this is a problem in practice is still a matter of debate given the standard practice of averaging across samples of participants, though several recent studies have demonstrated that there is no obvious sensitivity difference between ME and other tasks, at least for the constructions and sample sizes investigated (Myers 2009, Bader & Häussler 2010, Weskott & Fanselow 2011).

The second distortion introduced by *n*-point scale tasks arises because of the nature of the *n*-point scaling task itself. Participants are asked to optimally assign a set of experimental items to a discrete (and finite) scale. The numerals used to define locations on this scale define regular intervals: the interval between 1 and 2 is one unit, the interval between 2 and 3 is one unit, and so on. The psychological units that are represented by each of these numerically defined units can vary, however: the psychological unit between 1 and 2 is defined by the psychological distance between the items assigned to 1 and 2, the psychological unit between 2 and 3 is defined by the psychological distance between the items assigned to 2 and 3, and so forth. In this way, there is no guarantee that the distances between all successive units (the intervals) are stable—it depends on multiple decisions by the participant. In short, the fact that the *n*-point task is a scaling task introduces at least two types of distortion into the ratings, and perhaps worse, the possibility of losing data (or statistical power) that might be significant for grammatical theories (these effects are sometimes known as SCALING EFFECTS; see also Bard et al. 1996, Schütze 1996, and Cowart 1997).

ME was originally developed by the psychophysicist Stanley Smith Stevens (1956, 1957), building on previous work by Merkel (1888) and Richardson (1929), to explicitly overcome the possible problems caused by scaling effects in psychophysical experiments. Stevens was interested in investigating how the human perceptual system represents the properties of physical sensory stimuli such as the brightness of light and the loudness of sound. In order to do so, Stevens needed a measure of the PERCEPTION of physical stimuli that was as accurate and precise as possible. Because there was (and is) no device that can measure the PERCEIVED VALUES of physical stimuli in a participant's brain, psychophysicists were forced to rely on REPORTED VALUES, which, like all types of reported data, had to be reported along some sort of scale. Stevens was keenly aware of the limitations of scaling tasks such as the *n*-point scale (he was also the one who developed the theory of types of statistical data taught to undergraduates today: nominal, ordinal, interval, and ratio), and as such, was a forceful proponent of using the magnitude-estimation task to overcome those limitations.

Stevens's psychophysical magnitude-estimation (ME_P) task works as follows. Participants are presented with a physical stimulus, such as a light source set at a prespecified brightness by the experimenter (Stevens 1956 suggests that the magnitude of the stimulus be in the middle of the comfortable range of perception). This physical stimulus is known as the STANDARD. The standard is paired with a numerical value, which is called the MODULUS (Stevens 1956 suggests that the modulus be a relatively large, easily divisible number, such as 100). The participants are told that the brightness of the light

source is 100, and that they are to use that value to estimate the brightness of other light sources. They are then presented with a series of light sources with different brightnesses, and are asked to write down their estimates for the values of these light sources. For example, if the participant believes that a given light source is one half the brightness of the standard, they would give it a value that is one half of the modulus, in this case, 50. If the participant believes that a given light source is twice as bright as the standard, they would give it a value that is twice the modulus, in this case, 200. The standard remains visible throughout the experiment.

There are two innovations in the ME_P task that Stevens argued make it superior to the *n*-point scale task with respect to the scaling effects mentioned previously. The first innovation is what makes ME a fundamentally different cognitive task: the standard acts as a unit of measure. Whereas the units in an *n*-point scale task (the distance between two points on the scale) can vary from participant to participant, and can even vary between points for a single participant, the numeric intervals in the ME task are stably defined as the magnitude of the standard. Every item in ME is judged in relation to the standard such that one numeric unit is always equal to the psychological magnitude of the standard. This makes ME a more accurate measurement task because the standard acts like a stable unit of measure (a perceptual 'inch'). The second innovation is that the response scale is based on the (theoretically infinite) positive number line. This response scale better reflects the continuous nature of the stimuli under investigation, allowing participants to indicate any distinction that they feel is psychologically relevant. It should be noted that the choice of response scale is technically not unique to the ME task, and indeed, could be independently added to any task (e.g. the 'thermometer task' proposed in Featherston 2008). It just so happens that Stevens introduced both innovations simultaneously. For obvious reasons, the focus of this article is on the cognitive assumptions of the ME task itself, not on the use of an infinite response scale.

Given the potential benefits of ME_P over the *n*-point scale task in psychophysics, Bard and colleagues (1996) proposed a straightforward methodology for a type of magnitude estimation of acceptability that I call SYNTACTIC MAGNITUDE ESTIMATION (ME_S) when it is necessary to distinguish it from ME_P. In ME_S, participants are presented with a sentence (the standard) and a numeric value representing its acceptability (the modulus); an example is given in Figure 1. They are instructed to indicate the acceptability of all subsequent sentences using the acceptability of the standard as a unit of measure.

Standard:	Who thinks that my brother was kept tabs on by the FBI?	100
Item:	What did Lisa meet the man that bought?	___

FIGURE 1. An example of syntactic magnitude estimation.

As in ME_P, the standard in ME_S remains visible throughout the experiment so that it can act as a stable unit of measure.

3. THE COGNITIVE ASSUMPTIONS OF MAGNITUDE ESTIMATION. Narens (1996) and Luce (2002) argue that there are two fundamental cognitive assumptions of the ME task that must hold in order for the responses to represent true perceptual magnitude estimates.

(1) a. Participants must have the ability to make ratio judgments (in the given domain).

b. The number words (sometimes called NUMERALS) that participants use must represent the mathematical numbers (sometimes called NUMBERS) that they denote.

These two assumptions seem straightforward enough, but Narens (1996) and Luce (2002) argue that it is by no means safe to assume that they hold. To test the validity of these assumptions, Narens (1996) defined the empirical conditions given in 2 to test each assumption respectively.

(2) a. COMMUTATIVITY: magnitude assessments are commutative if the order in which successive adjustments (symbolized by ★) are made is irrelevant: $p \star (q \star \mathbf{X}) \approx q \star (p \star \mathbf{X})$.

b. MULTIPLICATIVITY: magnitude assessments are multiplicative if the result of two successive adjustments matches the result of a single adjustment that is the numeric equivalent of the product of the two successive adjustments: $p \star (q \star \mathbf{X}) \approx r \star \mathbf{X}$, when $p \cdot q = r$.

As the use of the word 'adjustment' above suggests, these empirical conditions were not designed to be tested with magnitude-estimation tasks, but rather with magnitude-PRODUCTION tasks. The following subsections introduce the magnitude-production task, commutativity, and multiplicativity in detail, as well as review the results of two attempts to test these conditions empirically in the psychophysical literature (Ellermeier & Faulhammer 2000, Zimmer 2005).

3.1. COMMUTATIVITY IN MAGNITUDE PRODUCTION. Magnitude production (MP) is in many ways the complement of magnitude estimation. In an MP task, participants are presented with a standard, such as a light source with a prespecified brightness, just as in ME. However, in MP the task is not to estimate the brightness of a second light source; instead, participants are asked to PRODUCE a second light source that has a given ratio to the standard light source (they are provided with a second light source that they can control). For example, they may be asked to create a second light source that is one half as bright as the standard light source; or they may be asked to create a second light source that is twice as bright as the standard light source. Crucially, MP is well suited to successive adjustments: the participant could be asked to produce a second light source that is one half as bright as the first, and then create a third light source that is one fourth as bright as the second. It is these successive adjustments that are exploited in the definitions of commutativity and multiplicativity.

If commutativity holds for a given type of stimulus, the order of successive adjustments is irrelevant. For example, let's say that a participant is presented with a 1 kHz tone at 82 dB (let's call this tone **X**), and asked to create a second tone that is one half as loud by adjusting a dial. The resulting tone is an adjustment of **X**, which we can call $q \star \mathbf{X}$, where ★ represents the adjustment procedure. The participant is then asked to take the resulting tone, which can be labeled $(q \star \mathbf{X})$, and adjust it to be one fourth as loud. The resulting tone can be labeled $(p \star (q \star \mathbf{X}))$, or simply $p \star q \star \mathbf{X}$. Then the procedure can be repeated from the beginning, but with the order of adjustments reversed: the 82 dB tone can be adjusted to be one fourth as loud, resulting in $(p \star \mathbf{X})$, and the resulting tone $(p \star \mathbf{X})$ can then be adjusted to be one half as loud, resulting in $(q \star (p \star \mathbf{X}))$, or simply $q \star p \star \mathbf{X}$. Then we can ask the question: Are these two final tones approximately equal in loudness ($p \star q \star \mathbf{X} \approx q \star p \star \mathbf{X}$)? If so, then commutativity holds, because the order of the adjustments is irrelevant to the final outcome.

It should be noted that the definition of commutativity makes no reference to the values of the adjustments. We are not interested in the actual loudness of the final tones in

dB. The only question is whether the two tones are equal, whatever their value may be. In this way, commutativity is independent of the meaning of the number words (numerals) to the participant (for example, the numeral ½ could actually represent the mathematical number ¾ to the participant, and commutativity would still hold). Ellermeier and Faulhammer (2000) and Zimmer (2005) each tested commutativity with respect to MP of loudness. Ellermeier and Faulhammer (2000) tested integer adjustments, whereas Zimmer (2005) tested fraction adjustments. Both studies found that commutativity holds for the majority of participants (e.g. seven out of eight participants tested in Zimmer 2005 demonstrated commutative adjustments). This suggests that at least for loudness, participants are able to make meaningful ratio judgments.

3.2. MULTIPLICATIVITY IN MAGNITUDE PRODUCTION. If multiplicativity holds for a given type of stimulus, the result of two successive adjustments should match the result of a single adjustment that is numerically equivalent to the product of the numerals that represent the two successive adjustments. For example, let's again say that a participant is presented with a 1 kHz tone at 82 dB (let's call this tone **X**), and asked to create a second tone that is one half as loud by adjusting a dial. The resulting tone is ½ ★ **X**. If the participant is then asked to adjust the resulting tone (½ ★ **X**) to be one fourth as loud, the resulting tone is (¼ ★ (½ ★ **X**)). We can then ask the participant to make a single adjustment to the original tone **X** that would make the tone one eighth as loud. If the resulting tone from this single adjustment (⅛ ★ **X**) is equal in loudness to the resulting tone from the two successive adjustments (¼ ★ (½ ★ **X**)), then multiplicativity holds, because ¼ · ½ = ⅛. Crucially, multiplicativity can only hold when (i) ratio judgments are possible (i.e. commutativity holds), and (ii) the number words (numerals) as used by the participant represent the actual mathematical numbers that they are intended to denote.

Ellermeier and Faulhammer (2000) and Zimmer (2005) each tested multiplicativity with respect to MP of loudness. Both studies found that multiplicativity does NOT hold for a majority of participants (e.g. six out of seven participants in Zimmer 2005 for whom commutativity held failed to demonstrate multiplicative adjustments). This suggests that at least for loudness, the numerals that participants use do not represent the actual mathematical numbers (in other words, the numerals used by participants cannot be taken at face value). In developing an extension of Narens's (1996) model, Luce (2002) points out that it is logically possible for the numerals used by participants to be related to mathematical numbers via a nontrivial function other than the identity function, though such a scenario would be disconcerting to theories of number cognition. Zimmer (2005) tested this explicitly for two functions suggested as possibilities in Luce 2002, and still found that multiplicativity was violated. This suggests either that multiplicativity does not hold for loudness, or that the complex function relating numerals to numbers is still unknown.

Taken as a whole, the Ellermeier & Faulhammer 2000 and Zimmer 2005 results for ME_P suggest that participants are able to provide meaningful ratio judgments of loudness, but are unable to report those judgments in a mathematically meaningful way. This is a particularly troubling result: the data is there, but psychophysicists cannot access it through the reports of the participants. This leads to the obvious question of whether the same is true for syntactic magnitude estimation, a topic taken up in the next section.

4. TESTING THE COMMUTATIVITY ASSUMPTION IN ME_S. The Ellermeier & Faulhammer 2000 and Zimmer 2005 method for testing commutativity and multiplicativity in ME_P is crucially tied to the magnitude-production (MP) task. Unfortunately, MP is fundamentally incompatible with sentence acceptability since it is unlikely that participants

would be able to construct novel sentences that represent an intended level of acceptability. In fact, it seems eminently plausible that even professional syntacticians would find such an acceptability production task extremely difficult. Therefore in order to test the cognitive assumptions of ME, the psychophysical methodologies must be modified to be compatible with the exigent circumstances of sentence acceptability. In practice this means adapting the commutativity test to magnitude estimation itself.

4.1. MODIFYING THE COMMUTATIVITY TEST. In theory, commutativity can be tested with three successive magnitude-estimation experiments as follows (though the logic is complex). Let the standard used for experiment 1 serve the place of **X** in the definition of commutativity. From the target items in experiment 1, two target items (Y and Z) with different acceptability ratings can then be chosen to serve as the equivalent of ($p \star \mathbf{X}$) and ($q \star \mathbf{X}$) in the definition of commutativity. Crucially, this means that their ratings relative to the modulus of experiment 1 will also serve as the p and q adjustment factors (respectively) in the definition of commutativity. The identical set of experimental materials can then be tested two more times using item Y (equivalent to ($p \star \mathbf{X}$)) and item Z (equivalent to ($q \star \mathbf{X}$)) as the standards in each of the two subsequent experiments respectively. One can then perform a search for an experimental item in the Y-standard experiment with a rating of ($q \star p \star \mathbf{X}$), and perform a second search for an item in the Z-standard experiment with a rating of ($p \star q \star \mathbf{X}$). If those two searches yield the same experimental item, then commutativity holds. If these two searches yield different items, or if no items with the appropriate ratings can be found, then commutativity does not hold. To simplify the searches, the same modulus can be used in all three experiments (e.g. 100), which means that the p and q adjustment values do not need to be independently calculated. Instead, the raw ratings of item Y and item Z in experiment 1 can be used directly in the search: an item in the Y-standard experiment with a rating that is equal to the rating of item Z in experiment 1 is equivalent to ($q \star p \star \mathbf{X}$), and an item in the Z-standard experiment with a rating that is equal to the rating of item Y in experiment 1 is equivalent to ($p \star q \star \mathbf{X}$).

Though the logic in the preceding paragraph demonstrates that commutativity can IN THEORY be tested with just three experiments, IN PRACTICE the situation is much more complicated. For one, there is no guarantee that the values p and q will be present in the Z-standard and Y-standard experiments respectively. One would likely need to run several sets of experiments before finding the right set of materials. Perhaps more importantly, presenting the same set of experimental materials two (or three) times to a single participant could lead to various types of repetition effects that may obscure commutativity, or even superficially simulate commutativity. One way around these problems would be to use different samples for each experiment; however, commutativity is defined as a cognitive property of an individual, not a property of sample means.

To overcome these problems, the following design features were incorporated in the experiments. First, in order to decrease the likelihood that the results obtained were due to the specific syntactic properties of the sentence types tested, two separate experiments were conducted. Only two conditions occurred in both experiments, meaning fourteen separate sentence types were tested. Second, in order to increase the likelihood that participants would use the full range of acceptability in their judgments, the eight conditions in each experiment were chosen from two previously conducted large-scale acceptability-judgment experiments (one with 120 participants (unpublished), one with 173 participants (Sprouse et al. 2011)). Third, in order to increase the likelihood of finding viable p and q values, each experiment was divided into eight blocks, each of which

contained the same set of eight items. In this way, each block can be viewed as a mini-replication of the previous block. By using a different condition for the standard in each block, these eight mini-experiments stand in for the three experiments required by the logic discussed above. Since eight is greater than three, this design significantly increases the likelihood of finding viable *p* and *q* values. Finally, in order to decrease the likelihood of repetition effects, each experiment was limited to eight blocks of eight conditions (one standard and seven target items), thereby limiting the total number of items seen by each participant to fifty-six.

4.2. EXPERIMENTS 1 AND 2. The motivation for two experiments is to decrease the likelihood that the results obtained are due to the specific syntactic properties of the sentence types tested. Therefore each experiment tested eight different conditions, and used a different sample of twenty-four participants. In all other respects the experiments were identical; therefore they are discussed in parallel.

PARTICIPANTS. Experiment 1 consisted of twenty-four participants, all self-reported monolingual native speakers of English. Eight of the participants were UCI undergraduates who participated for course credit or $5 (their choice). Sixteen participants were recruited using the Amazon Mechanical Turk online marketplace, and were paid $2 for their participation (Sprouse 2011). The two groups were used to compare the results of laboratory-based participation (UCI undergraduates) and online-based participation, since Sprouse 2011 has suggested that there is little or no difference in the data collected using the two approaches. The laboratory-based participants are reported as participants 1–8, and the online participants are reported as 9–24.

Experiment 2 also consisted of twenty-four participants, all self-reported monolingual native speakers of English. All twenty-four were recruited using Amazon Mechanical Turk, and were paid $2 for their participation, since the results of experiment 1 suggest no difference between laboratory-based and online-based data collection.

MATERIALS. Based on the pretest experiment, the eight conditions used in experiment 1 were those given in Table 1.

CONDITION	EXAMPLE
Left branch extraction	Whose did John think you saw father yesterday?
Double center embedding	The ancient manuscript that the grad student who the new card catalog had confused a great deal was studying in the library was missing a page.
Whether island with d-linked WH-phrase	Which necklace does the detective wonder whether Paul took?
Subject relatives	The accountant that insulted the robber read the newspaper article about the fire.
Object relatives	The banker that the teacher instructed approved the loan after asking a few questions.
Long distance WH-question with *claim*	What did the reporter claim that you saw?
WH-questions with complex NP	Who made the claim that Amy stole the pizza?
WH-questions with embedded *that*-clause	Who thinks that Walter likes hockey?

TABLE 1. The conditions of experiment 1 with example sentences.

Based on the pretest experiment, the eight conditions used in experiment 2 were those given in Table 2.

CONDITION	EXAMPLE
Adjunct island	What do you worry if the lawyer forgets at the office?
Whether island	What does the detective wonder whether Paul took?
Agreement violation	The slogan on the poster unsurprisingly were designed to get attention.
Agreement violation with a foil	The citation on the notecards unfortunately were quite difficult to track down using the library's limited resources.
WH-question with embedded conditional	Who sighs if the bride and groom neglect to send a thank you note promptly?
Long-distance WH-question with *think*	What does the professor think that Walter likes?
WH-questions with complex NP	Who heard the statement that Jeff baked a pie?
WH-questions with embedded *that*-clause	Who thinks that Jessica submitted the report?

TABLE 2. The conditions of experiment 2 with example sentences.

Eight lexicalizations were created for each condition. The center-embedded, relative clause, and agreement-violation conditions were taken from the published materials of Gibson and Thomas (1999), King and Just (1991), and Wagers and colleagues (2009) respectively. The lexicalizations were distributed among eight sublists to form the eight blocks for each experiment. One condition from each sublist was selected as the standard such that each condition served as the standard once. Each sublist was randomized in a different order from the other sublists. Each participant saw all eight lists in the same order (a total of fifty-six items). The full experiments in the order of presentation are available on the author's website (currently: www.socsci.uci.edu/~jsprouse).

PRESENTATION. Each experiment began with a practice phase during which participants estimated the lengths of seven lines using another line as a standard set to a modulus of 100. This practice phase ensured that participants understood the concept of magnitude estimation. During the main phase of the experiment, the standard was also always set to a modulus of 100 (as has been standard in the psychophysics literature since Stevens 1956). Participants were instructed that the reference sentence would change with each block, but the number would remain the same, and that they should change their ratings accordingly.

For the in-laboratory portion of experiment 1, the experiment was presented on a computer screen using a custom-made PHP-based presentation program. All eight sentences constituting a single block appeared simultaneously on a single screen. The standard appeared at the top of the screen in bold. Participants indicated their judgment by typing numbers into a response field next to the target sentence and clicking a button labeled 'submit'. Participants were under no time constraints. After each block, the experimenter cued up the next block and reminded the participant that the reference sentence had changed.

For the online portion of experiments 1 and 2, the experiment was presented using the HTML system of Amazon Mechanical Turk. The HTML form was designed such that all eight sentences constituting a single block appeared simultaneously on a single screen. Furthermore, all eight items constituting a single block were surrounded by a colored box. Because the experimenter could not be present to verbally remind participants that the reference sentence had changed, each block received its own unique colored box to serve as a visual cue that the reference sentence had changed. The standard appeared at the top of the screen in bold. Participants indicated their judgment by typing numbers into a response field next to the target sentence.

4.3. ANALYSIS PROCEDURES. The first step in the analysis of the results of these experiments requires following the logic of Figure 2 below for every possible unique combination of X, Y, and Z given the eight conditions in each experiment. Because each block contains one standard and seven target conditions, the number of unique Y-Z pairs is 7 choose 2, which is twenty-one. Because there are eight potential Xs, each with twenty-one unique Y-Z pairs, there are 168 unique X, Y, Z triplets to be checked for matching conditions. To avoid human error, a script was written using the R language (R Core Development Team 2009) to automatically perform this search procedure. The performance of the script was validated against human (by-hand) searches of eight participants.

The second step in the analysis of the results requires determining whether the number of matching conditions returned by the search (recall that matches are indicative of commutativity) is greater than the number expected by chance. A priori, there is no way of knowing exactly how many matches are expected by chance. We do know, however, the two factors that should contribute to the number of matches returned by chance. The first factor is the number of searches that must be performed. The value of the first factor is 168 searches, and it is constant for each participant since it is a function of the design of the experiment. In this case, the experiment was designed to increase the likelihood of finding matches (eight blocks instead of the minimum three required), which is why the number of searches is so high. The second factor is the number of conditions that are returned that match the p value and the q value respectively for each search. This second factor will vary from participant to participant, and from search to search, because it is completely dependent on the ratings that the participant gives to each condition in each block. To make this discussion concrete, imagine that one search returned four conditions that match value p, and two conditions that match value q. The intersection of those two sets yields one condition. This one condition appears to have been rated commutatively. The question we need to answer is whether one commutative match is greater than or less than we would expect by chance given four conditions in the p set and two conditions in the q set. And then we want to scale that question up to encompass all 168 searches for a given participant.

Though we do not know the probability of getting matches a priori, we can use the results of each participant, specifically how many conditions were returned as part of the p set and how many were returned as part of the q set for each of the 168 searches, to simulate a random assignment of conditions to the p and q sets. We can then ask how many matches were found between the random p and q sets. By repeating that random simulation 10,000 times, we can derive a distribution of the number of matches that we would expect by chance. We can then use that distribution to estimate the likelihood that the actual result occurred by chance. In other words, we can perform a type of randomization test to estimate a p-value for each participant (Onghena & Edgington 2007). An R script was written to perform the randomization test.

The final step in the analysis is to specify what counts as a 'match' between a given p or q value and the rating given by the participant. Though the obvious answer to this question is identity (e.g. a match occurs if the target value is 100 and the rating given by the participant is also 100), given the relatively fine-grained nature of the ME_S response scale (the positive number line), some margin of error may be in order. For example, if the target value is 160, and the participant rates an item as 163, it is plausible that the participant intended the rating to be equivalent to 160, but some sort of noise entered into the judgment (sampling error in the judgment process, or even conscious manipulation of the numbers by the participant). This issue reduces to a question of how participants use numbers in an ME experiment. Though I know of no quantitative studies

explicitly testing this question, anecdotal reports in both the psychophysics (e.g. Stevens 1956) and syntactic literature (Sprouse 2007, Featherston 2008) suggest that while participants may be cavalier with the least significant digit, they intend differences in the second significant digit to represent true perceptual differences. This makes some intuitive sense both in the way that humans use numbers of this magnitude, and given that there would be no motivation to even try ME if participants could not make fine-grained distinctions among stimuli. As such, a margin of ±9 is likely the limit to which the matching definition can be relaxed without a significant risk of returning matches that the participants intended as nonmatches. To be absolutely certain, however, that the results of the analysis were not distorted by the choice of the margin, results are reported for identity, ±9, and ±19.

4.4. RESULTS OF THE COMMUTATIVITY EXPERIMENTS. For psychophysics experiments, it is customary to report results for every participant. Those results can be found in Tables A1 and A2 in the appendix. The graphs in Fig. 2 summarize the results by summing the number of participants that performed above chance. In order to minimize the likelihood that the results are being distorted by performance that is near the chance threshold, two chance thresholds are reported in the graphs: the conventional $p < 0.05$ level, and the less conservative $p < 0.1$. It should be noted that the level of $p < 0.1$ is by convention not considered statistically significant by the APA or the LSA. This level was included simply to reinforce the fact that these results are not due to the choice of the significance level. Furthermore, to decrease the likelihood that the results are being distorted by the choice of match margin, three margins are reported: the obvious identity-based match, the margin of ±9 suggested by the ME_S literature, and a large margin of ±19. It should be noted that the margin of ±19 likely includes matches that the participants intended to be distinct ratings. This margin was included simply to reinforce the fact that these results are not due to the choice of the margin. The level at which commutativity held for ME_P in the Zimmer 2005 study is indicated by the dashed horizontal line in the graphs (i.e. seven out of eight, or an expected twenty-one out of twenty-four participants).

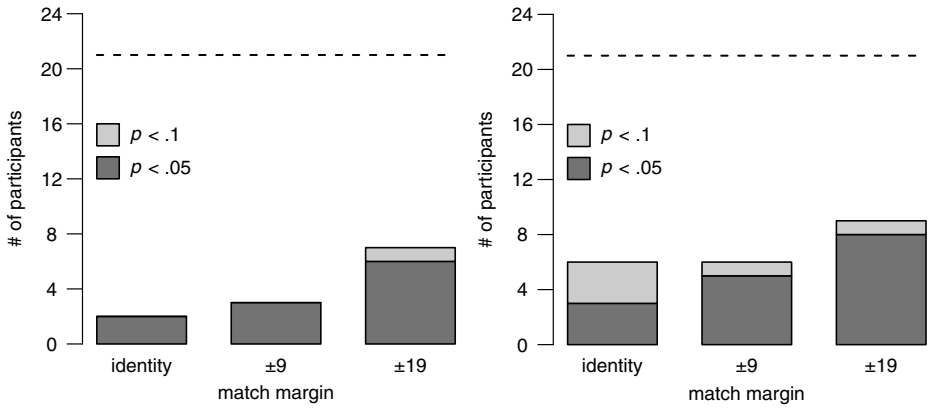

FIGURE 2. The number of participants that reported significantly more matches than would be expected assuming the null hypothesis (i.e. random performance) for three matching margins (identity, ±9, ±19) and two significance levels ($p < 0.05$, $p < 0.1$). The dashed line represents the level expected if commutativity holds (based on the results of Zimmer 2005).

As Fig. 2 makes clear, very few participants revealed significantly more matches than would be expected assuming the null hypothesis (i.e. chance performance) in ei-

ther experiment, at either significance level, and for any of the three margins. At the conventional criterion of $p < 0.05$ and a margin of ±9, commutativity appears to hold for less than 20% of the participants in ME_S. This is far fewer than the nearly 90% of participants that demonstrate commutativity in ME_P. This suggests that commutativity cannot be assumed to hold for ME_S as it does for ME_P.

5. The implications for magnitude estimation of acceptability. The goal of this study was to test the cognitive assumptions of ME with respect to acceptability judgments by adapting psychophysical magnitude-production procedures to syntactic magnitude-estimation procedures. The results of the two experiments reported here suggest that commutativity, though valid for ME_P, does not hold for ME_S. This suggests that participants in ME_S cannot make the ratio judgments that are asked of them by the ME instructions. Crucially, this means that multiplicativity—which represents the ability of participants to accurately report their judgments—cannot hold either, because ratio judgments (commutativity) are necessary to test multiplicativity. These results suggest that ME_S and ME_P are distinct cognitive tasks: participants in an ME_P experiment can perform ratio judgments, but cannot accurately report the results of those judgments (e.g. Zimmer 2005), whereas less than 20% of participants tested in ME_S could make ratio judgments. This implies that more than 80% of the participants in ME_S experiments are not performing the task as instructed by the experimenter.

These results accord well with the anecdotal reports of several ME_S researchers that participants may not be performing ME_S correctly, but rather may be reducing the task to a scaling task similar to those using *n*-point scales. These results also accord well with the suspicion of several ME_S researchers that the very nature of acceptability judgments likely makes ratio judgments impossible (Sprouse 2007, Featherston 2008, Weskott & Fanselow 2011): acceptability judgments may not have a true zero point representing the absence of all acceptability the way that physical stimuli such as loudness have a true zero point representing the absence of all sound. True zero points are required for ratio judgments. The fact that participants cannot make ratio judgments in ME_S may indicate that participants cannot even create an artificial zero point to use in place of a true zero point. Finally, these results also offer an explanation for the recent results suggesting that ME_S is no more or less powerful than other acceptability-judgment tasks (e.g. Bader & Häussler 2010, Weskott & Fanselow 2011), since it is now clear that ME_S is not a cognitively distinct task from the other judgment tasks as was originally suggested in the literature.

These results have very practical consequences for syntacticians who must decide which judgment task to use in their research. One of the primary purported benefits of ME_S is that it is a type of true cognitive measurement, free from the confounds inherent in standard scaling tasks, with the potential to yield data that is more reliable than traditional scaling data (Stevens 1956, Bard et al. 1996). Though this may be relatively true for psychophysical stimuli (modulo the multiplicativity problem), the results reported here suggest that the foundational premise of this argument is false with respect to acceptability judgments. ME_S is not the distinct cognitive-measurement task that it was purported to be, and therefore it should not automatically be given privileged position among the variety of acceptability-judgment tasks that are available to syntacticians. Previous empirical results that have compared judgment tasks for certain sentence types, sample sizes, and statistical tests have suggested a similar conclusion; those results, however, leave open the logical possibility that future experiments may one day reveal real differences between the tasks. Barring any major technical problems with

the experiments, the results reported here leave open no such logical possibility, since they directly test the fundamental cognitive assumptions of magnitude estimation that are at the foundation of the claim that ME is a superior task. These results suggest that ME_S is not the distinct cognitive measurement task that it was purported to be, and consequently it should have no inherent claim to the mantle of 'gold standard' among acceptability-judgment tasks.

Appendix: Experimental results

		identity			margin of ±9				margin of ±19			
Partic.	Matches	Sim Mean	Sim SD	*p*	Matches	Sim Mean	Sim SD	*p*	Matches	Sim Mean	Sim SD	*p*
1	0	0.97	0.90	.999	16	12.86	3.03	.190	53	58.71	5.63	.864
2	1	1.71	1.23	.839	4	5.27	2.00	.813	22	23.52	3.68	.703
3	2	2.71	1.36	.808	5	7.13	2.19	.891	19	18.59	3.36	.506
4	10	7.97	2.38	.256	16	13.26	3.02	.231	45	33.13	4.37	*.006
5	15	15.39	3.04	.609	22	22.79	3.62	.632	55	52.31	4.95	.326
6	16	15.31	3.21	.466	16	15.30	3.18	.467	89	71.56	5.60	*.001
7	28	27.09	3.92	.458	32	32.58	4.35	.593	71	68.75	5.64	.378
8	55	61.77	5.11	.923	55	61.80	5.08	.924	55	61.77	5.00	.929
9	1	3.32	1.69	.972	2	4.15	1.90	.928	33	21.93	3.90	*.005
10	1	0.86	0.87	.603	6	5.01	1.96	.386	23	24.95	4.04	.721
11	3	9.92	2.63	.999	9	20.45	3.62	.999	17	34.48	4.55	.999
12	3	5.29	1.98	.933	46	43.67	4.73	.348	208	211.38	7.12	.703
13	4	4.77	2.01	.725	34	25.88	3.89	*.026	112	102.97	5.68	#.065
14	5	5.26	1.99	.630	6	7.01	2.29	.744	10	10.11	2.62	.582
15	5	5.26	1.97	.631	11	8.36	2.41	.183	22	22.31	3.75	.583
16	6	4.77	1.98	.335	20	24.12	4.00	.876	92	114.45	6.85	.999
17	8	5.74	2.07	.195	9	7.13	2.25	.266	38	40.90	4.73	.761
18	9	11.40	2.66	.864	10	12.51	2.78	.860	25	31.27	4.21	.949
19	12	19.03	3.60	.983	12	18.96	3.60	.985	12	19.00	3.58	.984
20	14	18.57	3.35	.937	28	33.38	4.43	.908	223	225.60	7.32	.667
21	15	13.63	2.70	.367	65	60.22	4.34	.164	87	78.11	4.61	*.034
22	16	23.03	3.59	.985	35	40.10	4.72	.884	87	114.38	7.24	.999
23	19	13.44	3.01	*.050	30	22.41	3.68	*.030	84	65.65	5.51	*.001
24	36	23.12	3.75	*.001	38	24.78	3.86	*.001	60	43.45	4.86	*.000

Table A1. Results for experiment 1. Matches is the number of commutative results returned by the participant. Sim Mean and Sim SD report the mean and standard deviation respectively for the 10,000 simulations in the randomization tests. Column *p* reports the likelihood of the participant's result according to the randomization test. Significant results at $p < 0.05$ are indicated with asterisks (*). Significant results at $p < 0.1$ are indicated with hashmarks (#). Participants 1–8 were in-lab. Participants 9–24 were online.

		identity			margin of ±9				margin of ±19			
Partic.	Matches	Sim Mean	Sim SD	*p*	Matches	Sim Mean	Sim SD	*p*	Matches	Sim Mean	Sim SD	*p*
1	0	0.98	0.91	.999	7	5.60	2.08	.321	36	33.58	4.59	.339
2	1	0.87	0.88	.603	6	5.99	2.17	.566	27	31.00	4.45	.847
3	1	1.48	1.13	.800	89	85.10	5.80	.282	227	227.26	6.93	.545
4	1	7.72	2.41	.999	1	7.74	2.40	.999	23	30.99	4.46	.973
5	2	2.10	1.30	.650	9	12.24	3.00	.894	26	32.28	4.42	.938
6	2	1.34	1.08	.396	13	8.82	2.61	#.082	62	48.55	5.11	*.007
7	3	4.57	1.81	.878	21	16.29	3.33	.104	96	78.67	6.09	*.003
8	4	1.98	1.25	.115	24	13.21	2.98	*.001	50	32.08	4.04	*.001
9	6	8.22	2.51	.858	7	8.46	2.57	.772	25	24.47	4.00	.484
10	7	6.72	2.28	.520	10	11.54	2.84	.762	71	64.62	5.45	.143
11	8	9.91	2.38	.845	9	11.25	2.61	.858	55	45.44	4.86	*.031
12	9	6.12	2.23	.141	25	24.25	4.15	.469	116	114.64	7.18	.451

(*continues*)

		IDENTITY			MARGIN OF ±9				MARGIN OF ±19			
Partic.	Matches	Sim Mean	Sim SD	*p*	Matches	Sim Mean	Sim SD	*p*	Matches	Sim Mean	Sim SD	*p*
13	10	10.93	2.80	.689	31	28.14	4.15	.280	116	128.15	6.88	.970
14	12	11.02	2.73	.421	15	13.86	3.02	.405	42	44.19	4.74	.714
15	13	13.48	2.99	.622	20	17.36	3.33	.261	42	36.78	4.56	.150
16	17	12.01	2.90	#.065	20	14.01	3.06	*.039	91	94.16	6.20	.722
17	17	12.56	2.96	#.096	23	23.19	3.85	.567	61	54.05	5.35	.115
18	22	13.97	2.63	*.002	30	20.79	3.34	*.006	120	93.38	6.13	*.001
19	25	30.21	4.01	.923	34	42.96	4.68	.980	125	136.07	5.93	.974
20	31	22.67	3.60	*.015	31	22.66	3.62	*.017	145	133.96	6.49	#.053
21	33	27.79	3.34	#.078	38	35.38	3.74	.287	115	96.72	5.21	*.001
22	33	28.23	4.30	.158	33	28.28	4.28	.159	54	44.99	5.02	*.049
23	37	55.93	5.13	.999	66	90.85	6.12	.999	130	170.16	7.62	.999
24	51	35.05	4.58	*.001	58	41.72	4.89	*.001	88	68.75	5.74	*.001

TABLE A2. Results for experiment 2. Matches is the number of commutative results returned by the participant. Sim Mean and Sim SD report the mean and standard deviation respectively for the 10,000 simulations in the randomization tests. Column *p* reports the likelihood of the participant's result according to the randomization test. Significant results at $p < 0.05$ are indicated with asterisks (*). Significant results at $p < 0.1$ are indicated with hashmarks (#). All participants were online.

REFERENCES

BADER, MARKUS, and JANA HÄUSSLER. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46.273–330.

BARD, ELLEN GURMAN; DAN ROBERTSON; and ANTONELLA SORACE. 1996. Magnitude estimation of linguistic acceptability. *Language* 72.32–68.

COWART, WAYNE. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage.

ELLERMEIER, WOLFGANG, and GÜNTHER FAULHAMMER. 2000. Empirical evaluation of axioms fundamental to Stevens's ratio-scaling approach: I. Loudness production. *Perception & Psychophysics* 62.1505–11.

FEATHERSTON, SAM. 2005a. Magnitude estimation and what it can do for your syntax: Some WH-constraints in German. *Lingua* 115.1525–50.

FEATHERSTON, SAM. 2005b. Universals and grammaticality: WH-constraints in German and English. *Linguistics* 43.667–711.

FEATHERSTON, SAM. 2008. Thermometer judgments as linguistic evidence. *Was ist linguistische Evidenz?*, ed. by Claudia Maria Riehl and Astrid Rothe, 69–90. Aachen: Shaker.

GIBSON, EDWARD, and JAMES THOMAS. 1999. Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes* 14.225–48.

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. *Proceedings of the 25th annual conference of the Cognitive Science Society*, ed. by Richard Alterman and David Kirsh, 652–57. Mahwah, NJ: Lawrence Erlbaum.

KING, JONATHAN, and MARCEL A. JUST. 1991. Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language* 30.580–602.

LUCE, R. DUNCAN. 2002. A psychophysical theory of intensity proportions, joint presentations, and matches. *Psychological Review* 109.520–32.

MERKEL, JULIUS. 1888. Die Abhängigkeit zwischen Reiz und Empfindung. *Philosophische Studien* 4.541–94.

MYERS, JAMES. 2009. The design and analysis of small-scale syntactic judgment experiments. *Lingua* 119.425–44.

NARENS, LOUIS. 1996. A theory of ratio magnitude estimation. *Journal of Mathematical Psychology* 40.109–29.

ONGHENA, PATRICK, and EUGENE EDGINGTON. 2007. *Randomization tests*. 4th edn. New York: Chapman and Hall.

R DEVELOPMENT CORE TEAM. 2009. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Online: http://www.R-project.org.

RICHARDSON, LEWIS F. 1929. Imagery, conation, and cerebral conductance. *Journal of General Psychology* 2.324–52.
SCHÜTZE, CARSON T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.
SPROUSE, JON. 2007. *A program for experimental syntax: Finding the relationship between acceptability and grammatical knowledge*. College Park: University of Maryland dissertation.
SPROUSE, JON. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43.155–67.
SPROUSE, JON; MATT WAGERS; and COLIN PHILLIPS. 2011. A test of the relation between working memory capacity and syntactic island effects. *Language*, to appear.
STEVENS, STANLEY S. 1956. The direct estimation of sensory magnitudes: Loudness. *American Journal of Psychology* 69.1–25.
STEVENS, STANLEY S. 1957. On the psychophysical law. *Psychological Review* 64.153–81.
WAGERS, MATT; ELLEN LAU; and COLIN PHILLIPS. 2009. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language* 61.206–37.
WESKOTT, THOMAS, and GISBERT FANSELOW. 2011. On the informativity of different measures of linguistic acceptability. *Language* 87.249–73.
ZIMMER, KARIN. 2005. Examining the validity of numerical ratios in loudness fractionation. *Perception & Psychophysics* 67.569–79.

Department of Cognitive Sciences
University of California, Irvine
3151 Social Science Plaza A
Irvine, CA 92697-5100
[jsprouse@uci.edu]

[Received 18 August 2008;
revision invited 6 May 2009;
revision received 7 January 2010;
accepted 1 December 2010]