# Variance and Informativity in Different Measures of Linguistic Acceptability

**Thomas Weskott & Gisbert Fanselow**
**Department of Linguistics, University of Potsdam**

## 1. Introduction

In this paper, we deal with the issue of variability in different measures of linguistic acceptability. It has been argued that acceptability, when measured with the magnitude estimation method (ME), reveals the underlying gradience of linguistic judgments, while other measures, like Likert-scale judgments (e.g. on a 7-point scale) do not provide the same amount of information about the gradient basis of linguistic judgments. We will question this assumption by claiming that not only do Likert-scale judgments provide the same amount of information about a given empirical hypothesis, but also that the inherent variability of ME judgments makes them more susceptible to the production of spurious variance. We will back up these claims by reporting data from a study which compares three different measures of linguistic acceptability, categorical, Likert-scale, and ME. This study was concerned with a phenomenon which has attracted much attention in the literature on linguistic judgments, namely word order variation in German. For the empirical hypothesis we investigated, all three measures of acceptability contain the same amount of information relevant to hypothesis testing. In addition, we show that ME judgments contain more spurious variance and hence are more vulnerable to a reduction in statistical power. The next section deals with the issues of gradience and variance in different measures of linguistic acceptability. Section 3 reports on the empirical study. The final section discusses the findings and relates them back to the question of informativity and variability of different acceptability measures.

## 2. Gradience, Variance, and Informativity in Different Measures of Linguistic Acceptability

The practice of using introspective judgments as the main—and sometimes the sole—source of empirical validation of linguistic theories has been under attack now for more than a decade (see Featherston (2007) and the open peer commentary therein, for a reopening of the case). The charge was initiated by three publications, two books (Schütze (1996); and Cowart (1997), and an article (Bard et al. (1996)). The main argument in all three cases was that as an empirical basis for linguistic theory, acceptability judgments as used by theoretical linguists lack the degree of experimental control that is assumed to be standard in the social sciences. In the latter two publications, this charge was coupled to the plea for a methodological alternative: the use of magnitude estimation (ME) judgments. Bard et al. (1996), as well as Cowart (1997), suggested that the ME method was the cure to the problem that the malpractice of introspective judgments posed, because it adhered to the common standards for empirical hypothesis testing, and it allowed for the treatment of grammaticality as a gradient phenomenon. Later, this claim was made more precise by Keller (2000), Keller (2003), and Keller & Sorace (2005). Although other methods of empirical validation of linguistic hypotheses are mentioned for example by Keller & Sorace (2005, p.1501f.), magnitude estimation is propagated as the method of choice for its ability to "[...] provide fine-grained measurements of linguistic acceptability, which are robust enough to yield statistically significant results, while being highly replicable both within and across speakers." (Bard et al. (1996, p.1500)). In the same vein, it was claimed that ME judgments were methodologically superior to more common judgment elicitation methods such as categorial judgments (i.e., acceptable/unacceptable decisions) and Likert-scale judgments (for example, an acceptability rating on a 7-point scale). The following quote from Bard et al. (1996, p.40) illustrates this claim: "[...] these scales [e.g. ordinal, TW & GF] are too low in the series either to capture the information that could be made available or to serve the current needs of linguistic theories.". Similarly, Keller & Sorace (2005, p.1500) state that "[u]nlike the five- or seven-point scale conventionally employed in the study


Cascadilla Proceedings Project
Completed August 8, 2008

of psychological intuition, ME allows us to treat linguistic acceptability as a continuum and directly measure acceptability differences between stimuli." Inherent in these quotes is the supposition that only the ME method meets the standards of empirical hypothesis testing, and only ME data supply the kind of empirical information that can serve to validate the empirical claims of linguistic theories. Since we want to argue that these claims fairly overshoot the mark, we will first restate them in order to then refute them both by theoretical and empirical arguments.

> *Claim C1: ME is the most informative measure for the statistical testing of acceptability differences.* Only ME adheres to the standards of empirical hypothesis testing. Other measures like ordinal seven-point, or, for that matter, categorical 0/1-judgments, are not appropriate to establish acceptability differences between stimuli and to test the statistical significance of these differences.
>
> *Claim C2: Only ME is able to provide data that represent gradient acceptability.* Other measures like ordinal seven-point, or, for that matter, categorical 0/1-judgments, do not allow one to treat linguistic acceptability as a continuum.

These two claims have different implications for the way empirical hypotheses and experimental designs can be derived from linguistic theories. While C2 concerns the relation of assumptions about grammar formalisms and issues in measurement, claim C1 only concerns the latter. Since we will spend more time with claim C1, we will deal with C2 first.

### *2.1. Variance and Gradience*

It is especially one property of ME data that at first blush makes them attractive to a gradient conception of grammar: the individual judgments from ME experiments are made on an possibly open-ended, possibly infinitely fine-grained numerical scale (the set of positive rational numbers). Participants are free to assign a stimulus sentence whichever value they find suitable to express the relative acceptability of the stimulus sentence relative to the modulus (a reference sentence). From the perspective of the participant in an acceptability judgment experiment, it is this property that most clearly differentiates ME from other methods of judgment elicitation such as categorical or 7-point judgments, which grant the participant less freedom in the expression of relative acceptability. Still, it should be noted, that, contrary to what sometimes seems to be assumed by proponents of ME, even categorical judgments of acceptability are *relative* judgments, because every sentence is judged in relation to the sentences that preceded it. The main difference lies in the explicit nature of the relative judgment in the ME rating task: every sentence has to be judged in relation to the same modulus sentence. Sprouse (2007) investigated the influence of the modulus sentence on ME judgments and not only found that the relative judgments vary with different moduli (which is expected), but that they do so in a non-systematic fashion—the choice of modulus affects the rank ordering of the experimental conditions. This result seems fit to question the practice of ME researchers to stick to just one modulus sentence. In the original psychophysic ME experiments, every experimental item serves as the modulus for every other experimental item once (see Lodge (1982)). Not least because this practice is probably not possible to be adopted in linguistics, it should be noted that Sprouse's findings pose a severe problem the measurement of acceptability with the ME method as it is currently employed.

Returning to the issue of variance in ME judgments, we acknowlegde that the judgment of individual items grants the participant more freedom than the categorical or 7-point rating task. However, since the standards of experimental hypothesis testing prescribe that each experimental condition is tested with at least 4 items, the resulting means of the categorical and 7-point measures also exhibit some amount of variability: a mean value of four categorical judgments can take on four different possible values (0, .25, .50, .75 and 1). The mean value of $i$ n-point scale judgments can take on $(i \times n) - i$ values; in the case of four 7-point judgements, this comes down to $(4 \times 7) - 4 = 24$ different possible values. Thus, even fixed scale judgments exhibit a remarkable range of possible variability, and are not per se less suited to represent gradient acceptability. Apart from the freedom granted for the individual judgment, there seems to be no principled reason to prefer one measure over the other. It may even be argued, and Hubert Haider (Haider (2007, p.393)) has recently done so, that the inherent variability of the ME measure may make it more susceptible to spurious variance: "The variance in the aggregate data, however, should in

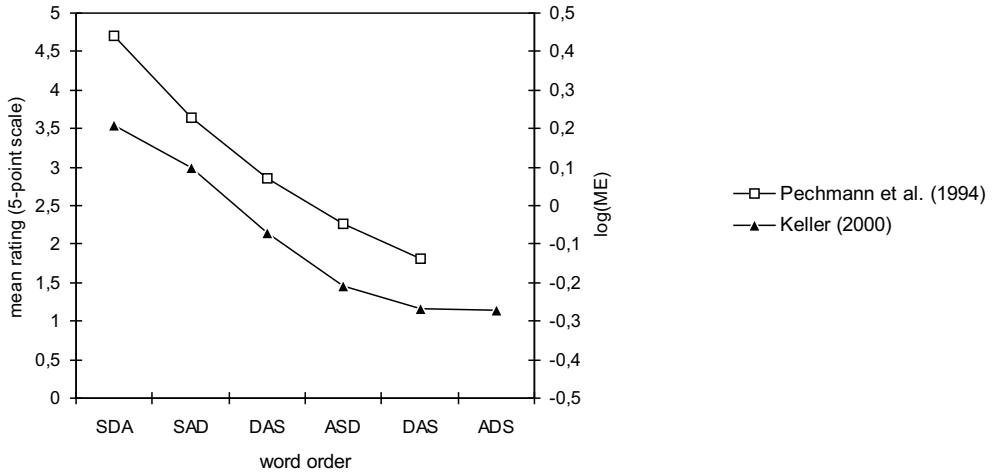


**Figure 1:** Comparison of the data on word order variation in German ditransitives from Pechmann et al. (1994) and Keller (2000). The left vertical axis plots the Pechmann et al. (1994) mean ratings from a 5-point scale, the right vertical axis plots the logarithmized ME rating values from Keller (2000). On the horizontal axis, the word order conditions are plotted; 'S' stands for 'subject', 'D' for 'dative', and 'A' for 'accusative'. The ADS condition was not tested by Pechmann et al. (1994).

my opinion not be mistaken as an indicator of grammatical variance. It should be acknowledged as what it seems to be, namely a test artefact.". We will come back to this strong claim regarding the variability of ME data when we discuss the results of our own experiments.

Just as the choice for one measure of acceptability, the choice for a grammar formalism with a gradient vs. non-gradient conception of grammaticality seems to us to be underdetermined by the empirical facts. Gradient theories of Grammar such as Probabilistic Grammars, Stochastic OT and Linear OT (see Manning (2003), and the contributions in Fanselow et al. (2006) for an overview) make strong claims about the probabilistic nature of linguistic phenomena, including the grammaticality of sentences. We take it as undisputable that linguistic acceptability is a graded phenomenon. What we however take as an theoretically and empirically open issue is whether *grammaticality* is a property that exhibits gradience. The experimental data that have been adduced in favor of gradient grammaticality (as e.g. the data in Keller (2000) and Keller (2003)) are neither unreconcilable with a dichotomous notion of grammatical well-formedness, nor do they in themselves show a gradience that goes beyond what is known from other rating studies. Evidence for the latter claim comes, for example, from a comparison of the data on word order variation in German ditransitive structures that Keller (2000) gathered with the ME method and the data that Pechmann et al. (1994) collected with a 5-point rating scale. Figure 1 shows how closely the results from the two studies match each other. For lack of space, we can not provide a thorough meta-analysis of the data from the two studies. Still, we think the graph in Fig.1 clearly shows that the gradience in markedness of the word order patterns is equally well represented by the 5-point scale data as by the ME data, and that the choice of the measure (categorial, n-point, ME) is and should be independent of the theoretical position the experimenter takes on the gradience issue.

Taking these considerations one step further, we think that just as data collected with different measures are able to represent gradient phenomena, so can gradient data patterns be obtained in domains that do not per se exhibit gradience. An analogy stemming from a different domain of cognitive science may help to clarify this point: Armstrong et al. (1983) conducted experiments that showed that prototypicality effects can occur in domains that do not exhibit the gradient property of typicality. The most impressive and oft-cited example of their results is that people had no problem rating odd numbers as more or less typical. Thus, people had no problem imposing a gradient on a dichotomous concept. Similarly, McMurray et al. (2002) showed that participants are sensitive to gradient effects in phoneme detection, depending on the task they are instructed to perform, although phoneme perception was long considered to be a paradigm case of categorial perception. We take these examples to show

that gradience as a property of behavorial data does not in itself imply the gradience (or non-gradience) of the underlying mental representations.

To conclude, we note that it is neither the case that the close match between gradient ME judgment data and gradient conceptions of grammaticality unequivocally speeks in favor of gradience—whether or not grammaticality should be conceived of as a continuous rather than a categorial notion, should be regarded as an open empirical and theoretical issue. Nor do we take it to be the case that the apparently larger variance of gradient ME judgment data make them the sole candidate for the study of continuous linguistic phenomena. To conclude, we take these considerations to refute claim C2. Moreover, as we will show in section 3, the inherent variability of ME judgments may even be harmful for the their informativity. But before returning to this point, we will turn to claim C1 and discuss how ME fares in comparison to other measures of acceptability in the next section.

### *2.2. Informativity*

In this section we will put to closer scrutiny the claim—claim C1 above—that data from categorial and n-point judgments contain less information than those from ME judgments. In order to be able to compare the informativity of two data sets in a meaningful way, we have to make a few assumptions, which, however, are uncontroversial in experimental psycholinguistics and completetly innocent with respect to the issue at hand.

We will consider a situation where the same acceptability rating experiment has been conducted under three different task assignments resulting in data coming from three different measures: categorial judgments (acceptable vs. unacceptable), judgments on a 7-point scale with labelled extremes, and a ME judgment task. The empirical hypothesis (the $H_1$) we want to test is a difference hypothesis of the kind "The judgments of $A$s are different than the judgments of the $B$s.", where $A$ and $B$ are any linguistic factors, be they phonological, syntactic, semantic, pragmatic or a (controlled) combination thereof. We assume that the standards of good experimentation are adhered to and multiple lexical variants are used to instantiate the conditions $A$ and $B$, and that everything else that covaries with $A$ and $B$ is controlled for in the experimental materials. Furthermore, we assume that there is a sufficiently large number of participants, and that every participant is confronted with a sufficiently large number of instances of the conditions. That is, in short, we assume a standard repeated measures design as, for example, a latin square design.

In order to compare the three types of measures with respect to informativity, let us first look at the possible outcomes of the hypothetical experiments employing the three types of measures. The results of experiments with a repeated measures design will be condition means of categorial judgments (ranging from 0 to 1), or condition means of n-point judgments (ranging from 1 to $n$), or condition means of ME judgments (with no predefined range). Given our empirical hypothesis, we will want to make sure that the means for the conditions which instantiate property $A$ and the ones that instantiate property $B$ are different, and that this difference is statistically reliable; that is, that the error probability $\alpha$ that the difference we have found in our experimental sample is not present in the population from which the sample was drawn is equal to or smaller than 5 %. Let us further assume that our empirical hypothesis states that the condition instantiating property $A$ is fully acceptable, while the condition instantiating $B$ is marginal, or mildly unacceptable. What does our hypothesis state with regard to our three different measures?

For the categorial judgments, the hypothesis states that the means of categorial judgments for condition $A$ is higher than the condition means for $B$ (both when aggregating over participants and items). In order for this to be true, the number of participants rejecting an item as unacceptable should be smaller in the $A$ condition than in the $B$ condition. That is, our hypothesis is now couched in terms of a difference in frequencies of rejections depending on condition. Turning to the n-point judgments, the hypothesis states that the mean value assigned to condition $A$ should be higher than the mean value assigned to condition $B$. In order for this to be the case, the values assigned to condition $A$ items should on average stem from the upper end of the n-point scale rather than from the lower end, while the reverse should be true for condition $B$ cases. Our hypothesis is now formulated in terms of a difference in mean numerical values depending on condition. And similarly for the ME judgments: in this case, the hypothesis also states that the mean numerical values should be different for the two conditions,

the mean of the values assigned to the stimuli of condition $A$ being higher than the mean of the values assigned to condition $B$ stimuli. As discussed above, an important difference between the three types of measures lies in the difference in variability of individual judgment data points. Trivially, categorial judgments have only two possible outcomes - each individual data point can take on only one of two possible values ("0" and "1"). In contrast, n-point judgments exhibit a larger degree of variability: each individual judgment may take on one of n different values. For ME judgments, the degree of individual variability is even larger, since there is no restriction on the possible values used to express the acceptability judgment. While this property of ME data is taken to be an advantage from the perspective of the participant in a judgment experiment (and rightly so, as noted above), it is far from clear what this larger degree of variability means with respect to the informativity of the condition means that enter into the inferential statistics, i.e. the statistical evaluation of the empirical hypothesis vs. the null hypothesis. From the perspective of inferential statistics, the only point that matters is whether the data help us to reject the null hypothesis (that the means of the $A$ and $B$ conditions are identical), and whether they do so with a sufficient degree of reliability. While in the inferential statistics of, for example, an analysis of variance (ANOVA), the $p$-value informs us about the probability that we have falsely rejected the null hypothesis, the $p$-value tells us nothing about the variance that underlies the pattern we find in the data. In our example above, we may find that for all three measures, there is a statistically significant difference between the condition means for condition $A$ vs. the condition means for condition $B$, and hence that we can reject the null hypothesis that there is no difference between the two conditions with an error probability lower than five percent. But given the different degrees of variability in the individual judgments discussed above, this is not informative with respect to possible differences about how this result of the statistical analysis has come about. We may want to know how much of the actual variance in the data can be accounted for by our experimental factor ($A$ vs. $B$), as opposed to mere random variance or the influence of some other factor which we are ignorant of or did not control in our experiment (as, for example, the shoe size of the participants). In the ANOVA procedure, there is a measure for this proportion of variance accounted for, called partial eta-squared (partial $\eta^2$; s. Cowart 1996: 123-125). It can take on values between 0 and 1. A partial $\eta^2$ of 0 means that none of the variance in the data set can be attributed to the experimental factor (a rather undesirable outcome of an experiment), while a partial $\eta^2$ of 1 would mean that all the variance in the data set is produced by the factor we are investigating (which is rarely the case in the social sciences). An $\eta^2$ of, say, .60 means that 60% of the variance in the sample can be traced back to the experimental factor, while 40% of variance must be attributed to some other factor, be it random, or some other variable. If we have no hint at what might be responsible for the additional variance, we have to consider this variance as spurious. Partial $\eta^2$, which is sometimes also called "estimate of effect size", is exactly the measure that connects the two issues raised above: the issue of differences in informativity of a given measure, and the issue of differences in variability in a data set. In order to assess a difference in informativity between two measures of linguistic acceptability, we will have to compare the partial $\eta^2$-values obtained in the two respective analyses of the two data sets. For illustration, consider the following example: if we investigate a two-level linguistic factor with two measures $m_1$ and $m_2$ under the empirical hypothesis that the mean judgments for the stimuli from the two levels of the factor (call them $A$ and $B$) differ, then we have to compute the partial $\eta^2$-value of the experimental factor for each of the two measures to determine to which proportion the variance in the two samples can be explained by the difference between $A$ vs. $B$. If the two measures differ in informativity such that $m_2$ is less informative than $m_1$, we expect the partial $\eta^2$ in the analysis of the data obtained with $m_2$ to be lower than the partial $\eta^2$ of the analysis of the sample obtained with $m_1$. The difference in informativity between the two measures can thus be determined by assessing the amount of spurious variance, that is, variance which cannot be attributed to the experimental factors we have employed and which, thus, is not relevant to our testing of the empirical hypothesis.[1]

[1]Here, a note is due regarding the difference between $\eta^2$ proper and partial $\eta^2$. While classical eta-squared is the ratio of the variance produced by the experimental manipulation to the total variance, partial eta-squared is calculated by dividing the variance attributable to the manipulation by the sum of effect variance and error variance. That means that partial $\eta^2$ values are not necessarily additive—they can exceed the value of 1, which classical eta-squared, by definition, cannot. However, as noted by Peirce et al. (2004, p.918), classical and partial eta-squared are identical in a design that has only one factor. Since this is the case in the hypothetical example in this section, as well as in the experimental data we report on in the next section, we do not differentiate between classical and

## 3. Empirical Evidence

In this section, we report on a series of experiments which were designed to answer the question whether different measures of linguistic acceptability contain different amounts of information with respect to a given empirical hypothesis. To keep the amount of experimental detail to a minimum, we will only report the results of the comparison of the partial $\eta^2$-values and refer the interested reader to Weskott & Fanselow (in prep.) for a detailed description of the experiments.

### *3.1. Method*

The experiments to be reported below investigated a phenomenon which has received much attention in the literature on gradient linguistic phenomena, namely word order variation in German (see e.g. Keller (2000), Keller (2003)). The first experiment dealt with scrambling of direct objects across the subject (subject$\prec$direct object vs. direct object$\prec$subject). The second experiment employed the same word order manipulation, but with indirect objects (subject$\prec$indirect object vs. indirect object$\prec$subject). All participants in the experiments to be reported were native speakers of German from the Berlin-Brandenburg area and were naïve with respect to the experimental factors investigated. The experiments each consisted of a pairing of two judgment tasks for the same set of participants and the same set of materials. One set of participants (N=48) performed a categorial judgment task in one experimental session, and was then asked to do a 7-point judgment task on the same set of materials two weeks later. The second set of participants (N=48) also performed a categorial judgment task on one occasion, and was then asked to rate the same stimuli with the ME method in a second session two weeks later. To control for possible ordering effects, we split each of these two participants sets in half. One half had to perform the categorial task first, and then the other task. The other half had the reverse order of task assignment. The materials consisted of eight sentences per condition; a sample for each experiment is given below.

(1) Experiment 1: ACC-scrambling in German (SO vs. OS):

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (1.a) | ... | dass | der | Papst | den | Scheich | eingeladen | hat. |
| | ... | that | the$_{NOM}$ | pope | the$_{ACC}$ | sheik | invited | has. |
| (1.b) | ... | dass | den | Scheich | der | Papst | eingeladen | hat. |
| | ... | that | the$_{ACC}$ | sheik | the$_{NOM}$ | pope | invited | has. |

(2) Experiment 2: DAT-scrambling in German (SO vs. OS):

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (2.a) | ... | dass | der | Mönch | dem | Jäger | geholfen | hat. |
| | ... | that | the$_{NOM}$ | monk | the$_{DAT}$ | hunter | helped | has. |
| (2.b) | ... | dass | dem | Jäger | der | Mönch | geholfen | hat. |
| | ... | that | the$_{DAT}$ | hunter | the$_{NOM}$ | monk | helped | has. |

Each of these conditions was realized in eight different lexical variants, yielding an overall of 16 experimental items per experiment. Each participant saw the full set of items, which were interspersed between 96 other experimental sentences coming from a wide range of different constructions and instantiating different degrees of acceptability. The data were treated as is common in the literature: categorial judgments were aggregated over participants and items and were arcsine-transformed, and ME raw judgments were divided by the modulus value and log-transformed. Since we are interested in the differences in variation that we can account for, we conducted an analysis which allowed us to directly compare the $\eta^2$ values of the word order effects for each of the different measures. Given that we have argued that it is not the case that condition means of ME judgments contain more information than condition means of categorial or 7-point judgments, our hypothesis is the null hypothesis ($H_0$). That is, we predict that all three measures show the same amount of variance accountable for in terms of the factor WORD ORDER. In other words, we predict that there are no substantive differences in the eta-squared values for the three measures.

---

partial eta-squared.

### *3.2. Results*

In Table 1, we give the condition means, standard deviations and eta-squared values (computed by hand from the sum of squares obtained from a one-factorial ANOVA; see Cowart (1997) for the details) for the word order effect in Experiment 1 for the four following measures: Cat1, Cat2, 7-point, and ME. The first two are the ratings from the categorial judgment task, and the second two are the respective 7-point and ME-ratings with which the categorial ones were paired. In order to retain maximal comparability between the two groups (Cat1/7-point and Cat2/ME), we will report eta-squared values from the item analysis only. The Cat1 and Cat2 scores are raw, i.e. untransformed scores, for illustrative purposes.

| | meanSO (StdDev) | meanOS (StdDev) | $\eta^2$ |
|---|---|---|---|
| Cat1 | .94 (.24) | .54 (.50) | .953 |
| Cat2 | .83 (.37) | .41 (.49) | .953 |
| 7-point | 6.33 (1.44) | 4.14 (1.97) | .958 |
| ME | 1.35 (.84) | .52 (.87) | .981 |

**Table 1:** Condition means, standard deviations (in brackets), and eta-squared values for Experiment 1

We note in passing that, although they are not at issue in this paper, all effects of WORD ORDER were significant in this analysis. What is more important and is revealed by Table 1 at a glance is that there is no substantive difference in the amount of variance accounted for by the factor WORD ORDER. The eta-squared values are exceedingly high for all four measures. In fact, the factor explains between 95 and 98% of the variance in the data.

We see a similar picture in the data of Experiment 2, the Dative scrambling experiment, which are given in Table 2.

| | meanSO (StdDev) | meanOS (StdDev) | $\eta^2$ |
|---|---|---|---|
| Cat1 | .93 (.26) | .62 (.49) | .910 |
| Cat2 | .80 (.40) | .52 (.50) | .831 |
| 7-point | 6.25 (1.55) | 4.33 (1.88) | .924 |
| ME | 1.26 (.85) | .64 (.87) | .959 |

**Table 2:** Condition means, standard deviations (in brackets), and eta-squared values for Experiment 2

In passing, we note that the effects of WORD ORDER, although all highly significant, are somewhat weaker for scrambling the indirect object over the subject than for the direct object scrambling in Experiment 1, which nicely fits with earlier results on the effect that German direct vs. indirect object scrambling has on acceptability (see Pechmann et al. (1994), Keller (2000)). More importantly, the eta-squared values of the Cat1, 7-point and ME judgments show no substantive difference, again ranging between 91 and 96%. Setting aside the slightly deviating result for the Cat2 data in Experiment 2 (which we do not comment upon here for lack of space, but see Weskott & Fanselow (in prep.) for a detailed discussion), our hypothesis finds strong support in the data from both experiments. If we take the eta-squared values as an indication of the amount of information we can extract from a data set given a hypothesis, there is no difference with respect to informativity between the categorial and 7-point judgment data on the one hand, and the ME judgment data on the other.

This result, however, is somewhat unsatisfactory, since we have only argued the null hypothesis so far. That is, we were only able to show that there are no differences in informativity between the three types of acceptability measures. To support our stronger claim made at the outset, namely that the variance in ME data does not necessarily contain more information about the acceptability of a structure, but that this variance may even be harmful to the informativity of the measure, we conducted a further

analysis in which we reduced the number of participants to a half of the sample. That is, we divided the sample of Experiment 1 and 2 in two (participants with an odd vs. an even participant number), and computed the eta-squared values of the WORD ORDER effect for the resulting subsamples. We predicted that, if it is in fact the case that ME judgment data contain more spurious variance than categorial and 7-point data, then we should see a substantive drop in the eta-squared values of the ME measure for the two subsamples compared to the overall analysis, while the two other measures should be robust against the decrease in statistical power and thus show no such drop.

The results of this analysis for Experiment 1 are given in Table 3 below. The drop in variance that we can account for in terms of our independent variable is represented by means of a difference score: the values for the subsamples were subtracted from the values of the total sample.

| | $\eta^2$(total)–$\eta^2$(1st half) | $\eta^2$(total)–$\eta^2$(2nd half) |
|---|---|---|
| Cat1 | -4,6 | 0,5 |
| Cat2 | -36,6 | 0,7 |
| 7-point | $\pm 0$ | -,03 |
| ME | -11,5 | -8,7 |

**Table 3:** Difference score for the eta-squared values of the subsamples (relative to the total) in Experiment 1

Before discussing these results, let us have a look at results of the corresponding analysis for Experiment 2:

| | $\eta^2$(total)–$\eta^2$(1st half) | $\eta^2$(total)–$\eta^2$(2nd half) |
|---|---|---|
| Cat1 | -8,4 | 0,9 |
| Cat2 | -39,5 | 1,7 |
| 7-point | 1,8 | -2,9 |
| ME | -25,8 | -11,4 |

**Table 4:** Difference score for the eta-squared values of the subsamples (relative to the total) in Experiment 2

Again setting aside the results for the Cat2 variable (but see Weskott & Fanselow (in prep.) for details), the measure that shows the largest drop in variance explained are the ME judgment data. The eta-squared values of the Cat1 and 7-point measures do not appear to be affected by the decrease in statistical power: the largest drop is shown by the Cat1 data in Experiment 2, while the 7-point judgments show almost no sensitivity to the decrease at all. On the other hand, the eta-squared of the ME measure seem to be quite sensitive to the split-half manipulation: the drop in eta-squared in Experiment 1 may not be too worrisome, but the drop by 26% and 11% in Experiment 2 is quite substantial. We report a similar result in Weskott & Fanselow (to appear), where variance was induced by including a semantic violation condition. ME showed a larger sensitivity to this increase in variance than did the other two measures. We take the results presented here to indicate two things: firstly, that not only the claim that ME judgment data are more informative than categorial or 7-point judgment data can simply not be upheld. And secondly, that under circumstances were within-group variance matters—as in cases of weak statistical power—the inherent variability of ME data appears to weaken the informativity of this type of measure when compared to categorial or n-point type measures.

## 4. Conclusions

With the results reported in the last section as a backdrop, we can now turn back to the claim C1 concerning the alleged superiority of the ME method that we set up in section 2. We take our data

to show very clearly that it is not the case that ME judgment data do contain more information than categorial and 7-point scale judgments. As far as categorial data are concerned, there is, apart from our own study, growing evidence that claim C1 cannot be upheld. For example, Murphy & Vogel (2008) and Bader & Häussler (submitted) report results similar to the ones presented here, showing that ME judgment data are not more informative than those from categorial judgments. We think, however, that our analysis of the variability inherent in the different measures takes us even a step further in that it goes beyond arguing the null hypothesis (no difference between the measures). This analysis, showing a pronounced drop in´variance explained for ME data, and no such drop for e.g. 7-point scale data, seems to accord very well with Haider's (2008) conjecture about the variance in ME data being an artefact of the method. Furthermore, it seems fit to cast severe doubts on the assumption that the magnitude estimation measure is superior to categorial or n-point measures of linguistic acceptability. Needless to say that much more experimental and methodologically refined work on this issue has to be done before ultimate conclusions can be drawn. But it seems that linguists, when they want to carry out an acceptability judgment experiment, should not regard magnitude estimation as the single method of choice. Given that the ME method is also beset with problems having to do with task demands and naturalness (see Weskott & Fanselow (to appear)), linguists on the search for empirical validation of their hypotheses, should—depending, of course, on design and hypothesis—also consider its lower-scale siblings, the categorial and n-point judgments.

## References


Armstrong, Sharon L., Leila R. Gleitman & Henry Gleitman (1983). What some concepts might not be. *Cognition* 13(3), pp. 263–308.

Bader, Markus & Jana Häussler (submitted). Towards a Model of Grammaticality Judgments .

Bard, Ellen Gurman, Dan Robertson & Antonella Sorace (1996). Magnitude estimation of linguistic acceptability. *Language* 72(1), pp. 32–68.

Cowart, Wayne (1997). *Experimental Syntax*. Sage Publications, Thousand Oaks.

Fanselow, Gisbert, Caroline Féry, Ralf Vogel & Matthias Schlesewsky (2006). *Gradience in Grammar*. OUP, Oxford.

Featherston, Sam (2007). Data in generative grammar: The stick and the carrot. *Theoretical Linguistics* 33(3), pp. 269–318.

Haider, Hubert (2007). As a matter of facts: comments on Featherston's sticks and carrots. *Theoretical Linguistics* 33(3), pp. 381–394.

Keller, Frank (2000). *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. Ph.D. thesis, University of Edinburgh.

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

Keller, Frank & Antonella Sorace (2005). Gradience in linguistic data. *Lingua* 115, pp. 1497–1524.

Lodge, Milton (1982). *Magnitude Scaling. Quantitative Measurement of Opinions*. Sage Publications, Newbury Park, CA.

Manning, Christopher D. (2003). Probabilistic Syntax. Bod, Rens, Jennifer Hay & Stephanie Jannedy (eds.), *Probabilistic Linguistics*, MIT Press, Cambridge, MA.

McMurray, Bob, Michael K. Tanenhaus & Richard N. Aslin (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition* 86(2), pp. B33–B42.

Murphy, Brian & Carl Vogel (2008). An empirical comparison of measurement scales for judgements of linguistic acceptability. Poster presented at the Linguistic Evidence Conference 2008.

Pechmann, Thomas, Hans Uszkoreit, Johannes Engelkamp & Dieter Zerbst (1994). Word Order in the German Middlefield. *Computerlinguistik an der Univerität des Saarlandes* 43.

Peirce, Charles A., Richard A. Block & Herman Aguinis (2004). Cautionary Note on Reporting Eta-Squared Values From Multi-Factorial ANOVA Designs. *Educational and Psychological Measurement* 64(6), pp. 916–924.

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

Sprouse, Jon (2007). *A Program for Experimental Syntax: Finding the relationship between acceptability and grammatical knowledge*. Ph.D. thesis, University of Maryland.

Weskott, Thomas & Gisbert Fanselow (in prep.). On the informativity of different measures of linguistic acceptability.

Weskott, Thomas & Gisbert Fanselow (to appear). Scaling Issues in the Measurement of of Linguistic Acceptability. Featherston, Sam & Susanne Winkler (eds.), *Fruits: Process and Product in Empirical Linguistics*, Mouton De Gruyter, Berlin, New York.