On the Informativity of Different Measures of Linguistic Acceptability

Thomas Weskott and Gisbert Fanselow

University of Potsdam

weskott@uni-potsdam.de

Address: University of Potsdam, Dept. of Linguistics, Komplex II, Karl-Liebknecht-Str. 24-25, Haus 35, D-14476 Golm, Germany

On the Informativity of Different Measures of Linguistic Acceptability

Abstract

The paper deals with the claim that the magnitude estimation (ME) method of gathering acceptability judgments produces data which are more informative for linguists than binary or n-point scale judgments. We performed three acceptability rating experiments which directly compared ME data to binary and 7-point scale data. The results clearly falsify the hypothesis that data gathered by the ME method carry a larger amount of information about the acceptability of a given linguistic phenomenon. The three measures are largely equivalent with respect to informativity. Moreover, ME judgments are shown to be more liable to producing spurious variance under certain circumstances.*

Keywords: acceptability judgments, empirical syntax, magnitude estimation, informativity

*Parts of the contents of this article were presented at the Linguistic Evidence 2008 Conference at Tübingen University and at the WCCFL 28 Special Session on Experimental Syntax, Semantics and Pragmatics at UCLA. The authors want to thank the following people for comments and discussion: Markus Bader, Sam Featherston, Lyn Frazier, Robin Hörnig, Konstantin Mergenthaler, and Eike Richter. Special thanks to our student assistants Jutta Boethke, Pavel Logačev, Anselm Metzger, Mareike Orschinski, and Nikolaus Werner for collecting and typing in the data. Furthermore, we thank the editors and two anonymous referees of Language for their valuable suggestions how to improve the manuscript. Any remaining errors are our own. The work reported here was carried out in the project *Theoretische und methodische Fundierung von gradierten Akzeptabilitätsurteilen*, funded by the DFG, grant no. FA 255/6-1.

1. Introduction. The recent years have seen the common practice of linguists to base their theories on intuitive judgments of grammaticality or acceptability come under attack. This attack was spear-headed by a number of publications, one of which was a highly influential article by Bard, Robertson and Sorace published in 1996 in Language. In the first part of that article, the authors followed up on the criticism brought forward by Schütze (1996) that the informal use of intuitive judgments poses a severe problem for the empirical assessment of theoretical claims in linguistics. Part of this problem, according to Schütze, can be attributed to the reluctance of theoretical linguists to apply the empirical standards of hypothesis testing that are common in, for example, psycholinguistics. These standards include the testing of multiple lexical instantiations of a given structure, as well as testing multiple naïve participants, and applying the statistical machinery of hypothesis testing to the results. Bard et al. (1996) agreed with Schütze's critical assessment of the state of hypothesis testing in current theoretical linguistics, claiming that the type of data gained by introspective judgments of acceptability severely underdetermines the theoretical possibilities of linguistic theories. However, they also went on to propose a solution to that problem, a methodological technique called 'magnitude estimation' (ME henceforth), which originally was used in psychophysics, i.e. that part of psychology that deals with the perception of the magnitude of physical stimuli such as length, brightness, loudness, etc.

Bard et al. (1996), and in their succession, Cowart (1997), argued that ME can be successfully applied to the study of linguistic acceptability, and that it is not only superior to the standard practice of introspective assessment of acceptability common in most theoretical linguistic work, but also to other measures of acceptability like binary judgments (categorizing linguistic expressions as either 'acceptable' or 'unacceptable'), and judgments on an n-point scale (like, e.g., judging acceptability on a 5- or 7-point scale).

In this paper, we want to deal with the claim that data from controlled experiments on acceptability using the ME methodology are to be preferred over data from controlled experiments of acceptability that use measures that are of a lower scale type like binary judgments, and 7-point judgments. This is to say that we fully subscribe to the view that linguists collecting acceptability data should adhere to the experimental standards alluded to above. However, we do not agree with the reason given for the claim that ME should be preferred over other experimental methods of eliciting acceptability judgments, namely an alleged advantage in informativity of ME over binary or n-point judgments of acceptability. In dealing with this issue, we will first provide a comparison of ME with these two other types of acceptability measures with respect to the requirements of the experimental task (section 2.1), and with respect to the resulting data and their statistical properties (2.2). In section 2.3, we will be in a position to make more precise the claim that ME is superior to other types of acceptability measures in terms of informativity. For the sake of argument, we will adopt the hypothesis put forward in Bard et al. (1996) and Keller and Sorace (2005) that the data from ME experiments of linguistic acceptability are more informative for linguistic theory than data from other types of measures like binary and 7-point judgments. This hypothesis will then be put to test in three experimental studies in which we directly compare the different measures with respect to their informativity when testing a particular phenomenon, namely word order variation in German (section 3). Without entering into details here, we can state that our results clearly speak against this hypothesis. The bottom line of our discussion will be that, for the types of experimental designs we investigate here, there is no difference in the amount of information linguists can draw out of an experiment using ME judgments versus using binary or 7-point judgments.

2 Comparing Different Measures of Acceptability. Until the 1970s, the experimental study of attitudes toward perceptual, social or, for that matter, linguistic stimuli, was mostly confined to having participants give the assessment of their attitude by means of nominal or ordinal scales, i.e. scales that categorize perceived properties of the stimuli into equivalence classes, or, in the case of ordinal scales, into total orders of discrete categories. An example for the former is the categorization of the acceptability of a linguistic structure into being acceptable vs. being unacceptable; an example for the latter is constituted by judging the acceptability of a structure on a 5 or 7-point scale. However, with the advent of ME in psychology (see Stevens (1946), Stevens (1956)), participants' judgment of the perceived intensity of stimuli became possibly much more fine-grained. In ME experiments, participants first have to judge the perceived intensity of a reference stimulus (the so-called 'modulus'), and then judge the perceived intensity of a different stimulus relative to that of the modulus. The judgments themselves could be given by drawing lines the length of which would correspond to the estimated magnitude, or by means of numerical judgments. This method was successfully applied to a vast number of psychophysical and perceptual domains in psychology, showing that participants in ME experiments are able to estimate the intensity of a stimulus in direct proportion to the actual intensity the stimulus exhibits. Stevens (1975) was the first to suggest an extension of this method to domains outside of perceptual psychology. Lodge (1981) took up this suggestion and showed that the perception of social constructs such as e.g. prestige and merit could be tested using ME, and found the estimates of such constructs to correlate significantly with underlying metrical scales as e.g. income. While the use of ME in the social sciences did not remain undisputed (see Wegener (1983), Schaffer and Bradburn (1989)), it represents until now a

methodological option for the study of subjective attitudes in the social sciences.

In linguistics, ME received attention with the publication of Bard et al. (1996) and Cowart (1997). Bard et al. (1996) presented a thorough investigation of the various drawbacks of the practice, then common in most parts of theoretical linguistics, to base claims solely on an introspective and mostly uncontrolled assessment of the acceptability of a given structure. Among these drawbacks are the property of such judgments to be prone to theoretical bias, their lack of generalizability across lexical instantiations of a given structure, and their inherent nonreplicability; more generally speaking: the lack of experimental control this informal method is beset with. They went on to suggest a solution to this problem by advertising ME as a method that not only adhered to the standards of empirical hypothesis testing, but in addition admitted participants in judgment experiments to give much more fine-grained estimates of the perceived acceptability of stimuli than is permitted by a predefined scale like e.g. a 7-point scale. The authors showed how especially this last property could be used to detect very fine-grained differentiations of grammatical properties of stimuli such as e.g. the influence of verb class on auxiliary selection in Italian. Cowart (1997) showed how ME can be used to validate theoretical claims about the status of the that-trace effect in English. Since then, a large number of rating studies has made use of the ME methodology, covering a wide range of linguistic phenomena (see Keller (2000), Keller (2003), and Keller and Sorace (2005) for an overview).

From the outset, the proponents of ME made it clear that there is a problem with the transfer of the methodology from psychophysics, where a physical magnitude such as e.g. the objectively measurable brightness of a flash serves as the stimulus, to the domain of linguistic acceptability, which obviously cannot be measured objectively. However, they argued that the methodological

tool of cross-modal matching (see Lodge (1981), Bard et al. (1996)) can remedy this problem. We will not deal with the issue of the objective correlate of linguistic acceptability here.

Recently, however, ME has incurred criticism concerning some of its inherent properties, among them the problematic aspects of the ME task (see Featherston (2009), Weskott and Fanselow (2009)), and the variability of effects inherent in ME data (Sprouse (2007a), Sprouse (2007b), Weskott and Fanselow (2008)). Sprouse (2008) argues convincingly that the choice of the modulus severely affects the assumption of colinearity of effects in ME experiments on acceptability: the choice of the modulus, Sprouse argues, does not affect acceptability patterns in a linear fashion, as it should if colinearity were to hold. Rather, some acceptability patterns react more extremely to a change of the modulus than others. This may, according to Sprouse, partly be attributed to the fact that the constant repetition of the modulus sentence may exert a priming effect on the participants' processing mechanism. Furthermore, a recent article by Bader and Häussler (2009) directly addresses the question we are partly concerned with here, too, namely that of the comparison of binary judgments with ME judgments. We will discuss the issues brought up by these papers successively in the next sections.

2.1 Task-related Issues**.** The task of the participant in typical experiment on ME of linguistic acceptability, after having assigned a numerical value (from the set of positive rational numbers) to the modulus at the outset of the experiment, is to assign a numerical value to each of the stimuli relative to the value she has assigned to the modulus.[1] A perspicuous difference between the ME procedure and the one employed in customary n-point judgment tasks is that the former involves relational reasoning in an explicit manner. The participant is instructed to come up with the numerical value expressing her assessment of the relative acceptability of the stimulus by an

arithmetic operation on the value assigned to the modulus. This arithmetic operation is exemplified in the instructions given to the participant at the outset of the ME experiment, typically containing a statement of the following form: if you have assigned the modulus the value '25', and you consider the stimulus sentence to be twice as acceptable as the modulus, then you should assign the stimulus sentence the value '50'. We note in passing that this formulation of the instructions implicitly assigns acceptability the status of a ratio scale, since an operation like multiplication is defined on this scale type only.

The mental arithmetic involved in the ME task, although warranting an explicitly relational judgment of the stimulus, poses a possible problem for both the participant in the experimental situation, as well as the researcher in his or her interpretation of the data. Evidence from a psychophysics ME experiment by Ellermeier and Faulhammer (2000) suggests that certain mathematical properties of the measurement model on which the ME methodology rests (see Narens (1996)) are violated by participants in ME experiments. We will shortly digress to present the results of these authors, since their data nicely exemplify the task-related problems of the ME methodology, albeit in a non-linguistic domain. The authors performed a loudness production experiment where participants are asked to assess their estimate of the stimulus intensity (the loudness of a tone) by producing a tone by, loosely speaking, turning the volume knob. For example, a participant would be presented with a stimulus with a loudness of 30dB and would be asked to produce a tone that was three times as loud. Ellermeier and Faulhammer tested the hypothesis, adopted from Narens (1996), that the properties of commutativity and multiplicativity should hold for the participants' responses. Commutativity would hold if the resulting tone c from the two following conditions would be identical: in the first condition, a stimulus a is given and the participant is asked to first produce a tone b that is twice as loud as a,

and then produce a tone c that is three times as loud as b. In the second condition, b would have to be three times as loud as a, and c twice as loud as b (obviously, (2×a)×3=(3×a)×2). Multiplicativity predicts that the resulting tone c under the instruction to make b twice as loud as the stimulus a, and c three times as loud as b should be the same as under the instruction to make c six times as loud as a ((2×a)×3=6×a). Whereas the property of commutativity was fulfilled by the loudness production data in Ellermeier and Faulhammer's study, multiplicativity failed to hold (see Augustin and Maier, 2008, for similar results in the visual domain). Not only was the resulting tone c not identical under the two instructions. The mean loudness of the resulting tone c produced by the participants under the instruction 'Make b twice as loud as a, and then make c three times as loud as b! ' was actually lower than that produced under the instruction 'Make c five times as loud as a! ', which casts considerable doubt onto the assumption that participants are capable of performing the kind of mental arithmetic that the ME task demands. Note that this even holds for a task domain like loudness which can be supposed to exhibit a much less complex underlying scale than the domain of linguistic acceptability. We refer the interested reader to Sprouse (ms.), where an attempt is made to test commutativity and multiplicativity for linguistic ME judgments.

On the other hand, an obvious difference between the different types of measures speaking in favor of ME concerns the granularity of individual judgments. Whereas in the n-point judgment task, the participant may, for lack of differentiation possibility, find herself forced to squeeze her judgment of two or even more stimuli into one of the n categories, the ME task grants her the freedom of a possibly infinitely fine-grained differentiation, either by making use of the open-endedness of the scale, or the possibility to use infinitely small positive numbers. From the point of view of participants, this is clearly an advantage of the ME procedure. However, it may at the

same time be a cause for concern on the part of the experimenter. The freedom given to the participant to express her subjective judgment of the acceptability of some linguistic structure may turn into a source of interindividual variance that does not contribute to the explanation of variance between experimental condition means in the statistical analysis. If the freedom granted to participants in ME experiments is a source of variance, we can reasonably ask whether this surplus variance is informative (i.e., it tells us something about the factors we manipulated), or whether it is simply spurious. We will take this issue up in section 2.3, where we will relate the question of variability of judgments to that of the informativity of different measures.

2.2 Variability and Gradience. In this section we want to deal with the question how the variability of different measures of linguistic acceptability relates to the notion of gradient grammaticality. As we noted in the preceding section, ME data exhibit a much greater amount of variability, simply because the individual judgments do. Accordingly, judgments from a 7-point scale will exhibit more variability than those from a 5-point scale, which in turn will be more variable than data from a binary judgment experiment. It has been argued that this property of the larger variability of ME data makes them a more adequate measure of acceptability, since grammaticality (as one ingredient of acceptability) has recently been increasingly often been argued to be gradient (see the contributions in Fanselow, Féry, Vogel, and Schlesewsky (2006) for an overview). However, since we assume that the standards of experimental hypothesis testing are adhered to in a given experiment (at least four items per condition etc.), even the resulting means of the binary and, accordingly, the n-point measures exhibit variability to a certain degree: a mean value of four binary judgments can take on five different possible values (0, .25, .50, .75 and 1). The mean value of i judgments on an n-point scale judgments can take on

($\underline{i}\times\underline{n}$) -$\underline{i}$ values; in the case of four 7-point judgements, this comes down to (4×7)-4=24 different possible values for the mean to take on. Thus, even fixed scale judgments, depending on the number of observations gathered, can exhibit a remarkable range of variability, and are not per se less suited to represent gradient acceptability.

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) for an overview) make strong claims about the probabilistic nature of linguistic phenomena, including the grammaticality of sentences. We do not want to deny that linguistic acceptability is a graded phenomenon. What we take to be a theoretically and empirically open issue, however, 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 irreconcilable 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. We want to illustrate the latter claim by 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. Although it cannot replace a thorough meta-analysis of the two data sets, which is beyond the scope of the present paper, we still think the graph in Figure 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 (binary, $\underline{n}$-point, ME) is and should be independent of the theoretical position the experimenter takes on the gradience issue.

--- INSERT FIGURE 1 ABOUT HERE ---

More generally, we want to point out that the gradience of a dependent variable does not in itself imply the gradience (or non-gradience) of the underlying mental representation which the dependent variable is related to. An analogy from phonological perception may illustrate that point: McMurray et al. (2002) showed that participants are sensitive to gradient effects in phoneme detection, depending on the task they are instructed to perform. Nonetheless, no one would doubt that phoneme perception in everyday speech comprehension constitutes a paradigm case of categorical, i.e. non-gradient perception.

By way of concluding, we want to call into question the theoretical significance of the prima facie close match between gradient ME judgment data and gradient conceptions of grammaticality. We think it should not be taken to speak unequivocally in favor of a gradient notion of acceptability, or even grammaticality—whether or not grammaticality should be conceived of as a continuous, rather than a binary notion, should be regarded as an open empirical and theoretical issue. Nor do we take it to be the case that the apparently larger variability of gradient ME judgment data make them the sole candidate for the study of continuous linguistic phenomena.

2.3 Comparing the Informativity of Different Measures. In this section, we want to compare the three types of measures—binary, n-point, and ME—with respect to their informativity in a hypothetical experiment employing all three types of measures. For the comparison to make sense, we presuppose that everything apart from the different dependent variables is equal (i.e., the circumstances of data gathering, the participants and items etc.). Our level of comparison is not the individual data point of the judgment of one item by one participant, but rather the means

of the ratings of a group of participants for a group of items. More specifically, we assume that the data come from a repeated measures design (e.g., a latin square design, see e.g. Cowart (1997)), and that we compare the mean ratings of two linguistic properties of a stimulus, A and B.

To compare the three types of measures with respect to informativity, let us first look at the possible outcomes of the hypothetical experiment. Given our empirical hypothesis, we will want to make sure that the means for the conditions instantiating property A and the ones instantiating property B are different, and that this difference is statistically reliable; that is, the error probability α that the difference we have found in our experimental sample is not present in the population from which the sample was drawn should be 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.

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 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 rather 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 eta-squared (see Cowart 1996: 123-125). It can take on values between 0 and 1. An eta-squared value 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 an eta-squared value 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-squared value 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. If we have no hint at what might be responsible for the additional variance, we have to consider this variance as spurious. Eta-squared, which is sometimes also referred to as 'estimate of effect size', is exactly the measure that connects the two issues raised above: the informativity of a given measure, and the differences in variability in a given data set. In order to assess a difference in informativity between two measures of linguistic acceptability, we will have to compare the eta-squared 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 eta-squared value of the experimental factor for each of the two measures to determine to which proportion the variance in the two data sets 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

eta-squared in the analysis of the data obtained with $\underline{m}_2$ to be lower than the eta-squared of the analysis of the data obtained with $\underline{m}_1$. The difference in informativity between the two measures can thus be determined by assessing the amount of spurious variance (1−eta-squared), that is, the amount of 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.

We take it that this way of assessing the possible differences in informativity between different measures of acceptability is the correct one, and, moreover, is more accurate than the one used in Bader and Häussler (2008), where the lack of difference in informativity is established by computing a correlation over the subject means, item means, and grand means for the conditions employed in their experiment. However, these authors in their comparison of binary and ME judgments partly found effects that were significant for one of the two measures, but failed to reach significance for the other measure. Although the authors do not comment on this fact, we take it as a strong indication of a difference in informativity of the two measures. In our own experiments, we will test the much stronger assumption that all the experimental effects (i.e., the effects of the experimental manipulations) all reach the level of significance, and concentrate our analysis on the comparison of the amount of variance accounted for by these effects.

We conclude this section by stating that the issue of whether the data gathered by means of an ME experiment on linguistic acceptability are more informative than the data from a binary or an $\underline{n}$-point judgment experiment is essentially an open empirical question. Proponents of the ME method have argued that ME data are indeed more informative than those of the other two measures, as the following quote illustrates: '[…] 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.' (Bard et al. 1996:38). In what follows, we will, for the sake of argument, adopt the hypothesis inherent in this quote, namely that the informativity of ME data is superior to that of binary or 7-point judgment data. If this hypothesis is correct, we should be able to establish the difference in informativity by a direct comparison of ME data with the two other data types in experiments in which the same participants have to judge the same set of items by means of the ME method vs. the two other types of measures.

3 Empirical Evidence. In this section, we will present three experiments that directly compare the informativity of three measures of linguistic acceptability: binary judgments (acceptable/unacceptable), judgments on a 7-point scale, and ME judgments. The experiments are concerned with a phenomenon that has received much attention in the literature on ME judgments, namely word order variation in German (see Pechmann et al. (1994), Keller (2000), Keller (2003), Bader and Häussler (2008), among many others). Experiment 1 and 2 each tested a two-level word order factor. Experiment 1 is concerned with the comparison of the unmarked order subject-before-direct object (SO) vs. the marked OS order. Experiment 2 compares SO vs. OS with indirect objects, that is, Scrambling of datives across the subject. Experiment 3 combines the two previous experiments by testing a three-level word order factor, namely Scrambling in ditransitives in German.

In all three experiments, we expect the markedness of the deviant word order to yield a decrease in acceptability, i.e. a markedness effect. This is to say that the marked variants (e.g., (2) below), although fully grammatical in German, are usually judged as less acceptable by participants than their unmarked counterparts, if presented out of context. However, we will not be concerned here with this markedness hypothesis, since our aim is to compare the amount of

information that the different measures carry about that effect. The hypothesis throughout the three experiments will be that if ME data are indeed more informative than the other two measures, then there should be a difference in the size of the markedness effect that analyses of the data from the three measures exhibit: ME data should show larger effect sizes, that is, a higher amount of variance accounted for, than the data from binary and 7-point judgments. It is this hypothesis that our experiments aim to falsify.

3.1 Experiments 1 and 2 — Method. Since Experiment 1 and 2 were carried out as subexperiments of the same larger study, we collapse the description of the materials, participants and procedure for these two experiments.

Materials for Experiment 1. Experiment 1 was concerned with direct object Scrambling in German (see Haider and Rosengren (1998), Fanselow (2001)). The unmarked word order where the subject precedes the direct object was compared to its marked variant, where the order of the arguments is reversed. The clauses with the scrambled argument order were embedded into contexts like 'Peter hat erzählt . . . ' (Peter has reported . . .). All arguments had masculine gender, which renders case and number marked unambiguously. The two variants of a sample item are given in 1 and 2.

(1) . . . dass der Präsident den Scheich empfangen hat.
. . . that the$_{NOM}$ president the$_{ACC}$ sheik received has.
'. . . that the president has received the sheik.'

(2) . . . dass den Scheich der Präsident empfangen hat.
. . . that the$_{ACC}$ sheik the$_{NOM}$ president received has.
'. . . that the president has received the sheik.'

There were eight items (i.e., lexical variants) overall, four items instantiating the SO and the OS order, respectively. The items were distributed onto two different lists such that the first list contained items 1-4 in the SO and items 5-8 in the OS condition, while the reverse was true for the second list. These two lists were then included into the lists containing the other 102 items of the subexperiments of the larger experiment, which served as fillers for the items of Experiment 1. These filler items stemmed from a wide range of different constructions and instantiated different degrees of acceptability, among them the benchmark items described in Weskott and Fanselow (2009). This procedure yielded an overall of 110 items per list. These large lists were then pseudo-randomized such that each of pair of instances of items from Experiment 1 was separated by at least two filler items.

Materials for Experiment 2. Experiment 2 was concerned with the Scrambling of indirect objects in German. As in Experiment 1 above, we compared the unmarked SO order to the marked OS order, where the indirect object precedes the subject. Although this is not our main concern here, we want to point out that the Scrambling of indirect objects (4) usually yields a smaller decrease in acceptability ratings than the Scrambling of direct objects employed in Experiment 1 (cf. (1) above). An example of the two variants of an experimental item is given in 3 and 4.

(3) . . . dass der Mönch dem Jäger geholfen hat.
. . . that the$_{NOM}$ monk the$_{DAT}$ hunter helped has.
'. . . that the monk has helped the hunter.'

(4) . . . dass dem Jäger der Mönch geholfen hat.
. . . that the$_{DAT}$ hunter the$_{NOM}$ monk helped has.
'. . . that the monk has helped the hunter.'

As in Experiment 1, there were eight items overall. Again, these items were distributed onto two lists such that each list contained four items in each of the two conditions. The procedure of including these two lists into the larger experimental lists was identical to the one described for Experiment 1.

Participants. The participants of Experiments 1 and 2 were 48 students of the University of Potsdam (23 of them female, 25 male; age range 20-34, mean 22.8). They were native speakers of German from the Berlin-Brandenburg area and did their studies in different departments of the social sciences faculty. Students of linguistics could only take part if they had not gone further in their study than the third semester (enhanced undergraduate level). This guaranteed that all the participants were sufficiently naïve with respect to the linguistic structures tested in the experiments. All participants received course credits or cash remuneration (6,– EUR, approx. $8,50) for their participation.

Procedure. Experiments 1 and 2 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 binary judgment task in one experimental session (which we will refer to as ‘binary1’), 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 binary judgment task on one occasion (‘binary2’ henceforth), 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 binary task first, and then the other task. The other half had the reverse order of task assignment. This manipulation entered into the statistical analysis as the factor task order.

The binary task was a non-speeded forced-choice acceptability judgment. Below each sentence, there were two boxes, one labelled with ‘akzeptabel’ (acceptable), the other with ‘nicht akzeptabel’ (unacceptable). In the instructions, an example for a judgment of an ungrammatical sentence as unacceptable was given, followed by a judgment of a grammatical sentence as acceptable.

In the 7-point scale task, there were seven small circles in a row displayed below each sentence. The outermost two circles were labelled with the words ‘völlig akzeptabel’ (totally acceptable) and ‘völlig inakzeptabel’ (totally unacceptable). In the instructions, an example of a perfectly acceptable sentence, judged as ‘totally acceptable’, and an example of a judgment of an ungrammatical sentence as ‘totally unacceptable’ were given. In this task, participants were instructed to use the full range of the scale for the judgments.

For the ME task, we followed the procedure standardly applied by authors using this technique. First, there was a short training phase which served to familiarize the participants with the ME task. In the first part of this training phase, they were asked to estimate the length of a line relative to the length of a reference line (the ‘modulus’), to which they had to assign an arbitrary value (> 0). They were instructed that if they thought that the ‘stimulus’ line was twice as long as the reference line, then they should assign the stimulus line twice the value of the modulus. Accordingly, the instruction told them that if they thought that the stimulus was half as long as the modulus, they should assign the stimulus half the value of the modulus. In the second part of the training phase, the participants were then familiarized with estimating the relative acceptability of sentences. A sentence of medium acceptability, given in (5) below, was chosen as modulus. They were told that if they thought that the stimulus (which was grammatical) was

ten times more acceptable than the modulus (a sentence of medium acceptability), then they should assign the stimulus ten times the value of the modulus.

(5) Man bekommt zu selten von ihr zugelächelt.
One gets too seldom by her smiled-at
'One is smiled at by her too seldom.'

Participants were encouraged to use arbitrarily high numbers, as well as arbitrarily fine differentiations and gradations. They were then asked to perform this task on two training items before the experiment proper would start. Throughout the experiment, the modulus sentence, which was grammatical, though mildly deviant, was held constant (cf. Bard et al. (1996), Keller (2000)).

In all three judgment tasks, the participants were instructed to judge each sentence carefully, but speedily, and not to skip any sentences. Furthermore, they were instructed not to consult any normative knowledge of grammar which they might have from school, but to simply rely on their intuition.

All four experimental sessions (binary1, binary2, 7-point and ME) took place in a lecture room at the University of Potsdam. Each participant was handed one of the 24 versions of a booklet which contained all the experimental items and fillers. The versions consisted of different randomizations. Participants sitting next to each other got different versions of the booklet. The experimenters took care that participants did not communicate or copy judgments. As soon as they had performed the second session, participants received their course credits or remuneration. By this course of action we were able to minimize the rate of drop-outs for the parallel sessions.

Data Analysis. For each of the subexperiments, binary judgments were coded as '0' (for unacceptable) and '1' (for acceptable). These judgments were then aggregated over items and participants and arcsine-transformed. The purpose of this transformation is to equalize the variance of proportions; it improves the equality of variance in the angles of the distribution (but see Jaeger (2008) for a different treatment of binary data). 7-point judgments were left as is, and also aggregated over items and participants. Following the practice in the literature (but see Featherston (2005) for an alternative data treatment), ME judgment values were divided by the modulus value and log-transformed. For reasons of comparability across the two comparisons, which involved different participants, we will only report item analyses.

Design and Predictions. We will report two analyses. Analysis 1 was a 2×2 repeated measures design in which we tested the factors word order (SO vs. OS) and judgetype (binary1 vs. 7-point/binary2 vs. ME). In this analysis, we tested whether there are any differences between pairs of measures in terms of a main effect of the factor judgetype, or an interaction of judgetype and word order. If there is a difference in informativity for the different measures, then the comparison of two measures should show a main effect, or an interaction, or both.

In Analysis 2, we directly compared the effect sizes of the word order factor for the different measures stemming from a one-way ANOVA. Again, a difference in informatitivity between the three types of measures would predict that there are substantive differences in the effect sizes for the word order effect. More specifically, if ME judgments contain a higher amount of information, this difference in informativity should be visible in our results in terms of the ME data exhibiting substantially higher eta-squared values than the other two measures. Apart from the hypotheses stated above that pertain directly to our research question about the informativity of different measures of acceptability, the experiments reported here also tested linguistic

hypotheses. Since both experiments dealt with the markedness of word order variants, we hypothesized that in both experiments, the marked word order should get a higher judgment score than the marked one, respectively the highest one in Experiment 3. In addition, we can put up the hypothesis that the markedness effect for the marked word order OS should be larger for scrambling a direct object over the subject than for scrambling an indirect object (see Pechmann et al. (1994), Keller (2000)). That is, the difference between the SO and OS condition should be larger in Experiment 1 than in Experiment 2.

Finally, we included the factor task order into our overall analysis to check for possible effects of the order of task assignments. We predicted that this factor should not show a significant main effect, nor should it enter into an interaction with any of the other experimental factors.

3.2 Results for Experiments 1 and 2. Before going into the details of the subexperiments, we first report on the analysis of the task order factor which we computed to check for possible effects of task assignment (binary first, 7-point/ME second, and the reverse). Task assignment did not have a significant main effect (all *F*s <1) in any of the experiments, nor did the task order factor enter into any of the interactions with the other experimental factors, word order and judgetype (all *F*s <1). We can conclude from this that it did not influence the judgments of our participants whether the binary task, or the 7-point/ME task was administered first.

Results for Experiment 1, Analysis 1. The results for Analysis 1 of Experiment 1 are summarized in Table 1 and 2 below. Table 1 gives the descriptive statistics for the two word order conditions for all four measures. All values are based on the untransformed scores. Although the effect of the word order factor per se is not of interest here, we note that Table 1 shows that all four

measures, though on different scales, exhibit a clear markedness effect: on average, the OS structures are always judged to be less acceptable than the SO structures.

--- INSERT TABLE 1 ABOUT HERE ---

More important for the purpose at hand are the results of the two-way ANOVA presented in Table 2.

--- INSERT TABLE 2 ABOUT HERE ---

The first and second column of Table 2 present the comparison of the binary judgments with the 7-point judgments; the third and fourth column give the comparison of the binary to the ME judgments. For each of these two data sets, we performed an ANOVA with the factors word order (SO vs. OS) and judgetype (binary1 vs. 7-point, and binary2 vs. ME). The first thing to note with respect to Table 2 is that for neither of the comparisons is there a significant effect of the factor judgetype; nor does judgetype enter into an interaction with the word order factor—all F-values are smaller than one. Since we can conceive of the factor judgetype as an indicator of possible differences between the four measures, we can conclude from this result that there was no difference between the measures with respect to their sensitivity to the word order manipulation, and thus no difference in informativity between the binary1 and the 7-point judgments on the one hand, and the binary2 and the ME judgments on the other hand. To get a clearer picture of the differences between the variance in the measures and how much of it we can account for, we established the possible differences in effect size between the individual measures in Analysis 2.

Results for Experiment 1, Analysis 2. The effect sizes (variance accounted for) were computed from four separate one-way ANOVAs with word order as the only factor.[2]

--- INSERT TABLE 3 ABOUT HERE ---

As a glance at Table 3 reveals, the measures do not exhibit any differences with respect to effect sizes. The percentage of variance accounted for by the experimental factor word order is exceedingly high for all four of the measures, ranging between 95 and 98 %. Before we discuss these results in more detail, let us first turn to the results of Experiment 2.

Results for Experiment 2, Analysis 1. The results for Analysis 1 of Experiment 2 are summarized in Table 4 and 5 below. Table 4 presents the descriptive statistics for the two word order conditions for all four measures. As before, all values (apart from the ME data) are based on the untransformed scores, and all four measures exhibit a clear markedness effect: the judgment means for the OS structures are always lower than those of the SO structures. Note, however, that the markedness effect for the fronted indirect objects in Experiment 2 is smaller than that of the fronted direct objects in Experiment 1. This is in line with our prediction which was based on earlier studies on the acceptability of Scrambling in German.

--- INSERT TABLE 4 ABOUT HERE ---

Table 5 presents the results of the two way ANOVAs with the word order and the judgetype factor. As in Experiment 1, we predicted that there is a difference between ME and the other two measures in terms of informativity, there should be a significant effect for the factor judgetype, or an interaction of judgetype and word order, or both.

--- INSERT TABLE 5 ABOUT HERE ---

As before, the factor judgetype showed no significant main effect, nor did it enter into a statistically reliable interaction with the factor word order. Thus, Experiment 2 also shows no difference in informativity for the three measures investigated.

Results for Experiment 2, Analysis 2. The direct comparison of the effect sizes of the word order effect in Experiment 2 is given in Table 6 below. As in Experiment 1, these effect sizes

were computed from four separate one-way ANOVAs with word order as the only factor.

--- INSERT TABLE 6 ABOUT HERE ---

Table 6 clearly shows that there is only one measure whose effect sizes deviate from the others, namely binary2. The effect sizes of the other three measures are all in the same range: the word order factor accounts for 91 to 96% of the variance in the samples for the binary1, 7-point and ME judgments. The binary2 judgments show a drop of the eta-squared value to 83%, which is still high, but considerably lower than the other three values. We attribute this to the fact that the high degree of variance in the SO condition for this measure (cf. the standard deviations for this condition in Table 5 above), which is not readily interpreted. We conjecture that the binary judgments show a greater sensitivity to random variance in the slightly more subtle manipulation in Experiment 2. We will return to this point in the discussion below. Focussing on the two other measures, 7-point and ME, we can still conclude that it is not the case that the 7-point judgment data carry substantially less information about the word order effect than the ME judgment data.

3.3 Discussion. First of all, the results of Experiments 1 and 2 both clearly speak against the assumption that the three types of measures we employed carry different amounts of information about the experimental factors we investigated, namely two-level word order manipulations in German. We found significant effects of the word order manipulation for all three types of measures, and the sizes of these effects as given by the eta-squared values were within the same range.

If there were a difference in informativity between the three types of measures, we should have seen effects of the judgetype factor in both experiments, or an interaction of the word order manipulation with the judgetype factor. However, this was not borne out by Analysis 1, which

we take as a clear indication that there is no reliable difference in informativity between the three measures.

This interpretation is further backed up by the results of Analysis 2 for the two experiments. The comparison of the effect sizes in Experiment 1 showed that there are no substantive differences in the variance accounted for. This means that, for all three measures employed in Experiment 1, the variance generated by the experimental factor word order is in the same range for all three measures, namely between 95 and 98%. In Experiment 2, the effect sizes for the binary1, 7-point and ME variables, although being somewhat lower overall, were also in the same range (between 91 and 96 %). Although there is, to our knowledge, no mathematical procedure to directly compare eta-squared values, we take our results to show that all three measures are equally capable of capturing the effect that the word order manipulation has on the judgments of our participants. In other words: given the results of Experiments 1 and 2, we are forced to reject the assumption that the three measures are different with respect to their informativity. More specifically, we are forced to interpret the data reported so far as rejecting the hypothesis that ME data are more informative than either binary or 7-point data.

This conclusion, however, is confronted with at least two problems. The first qualification we have to make concerns the obvious limitation of the conclusions we can draw to the kind of design we have tested so far. It is not unconceivable that in an experiment with, for example, a design with a three-level word order factor, the binary and 7-point scale judgments would fare worse than the ME judgments with respect to informativity. Although it is far beyond the scope and aims of this paper to formulate a claim with respect to the informativity of measures of acceptability which could be generalized (inductively) across all possible kinds of experimental designs, we nevertheless have to show that the lack of informativity differences is not limited to

the kind of one-factorial two-level design that we have tested here. We will refer to this problem of the generalizability of the assumption of equal informativity of the measures across experimental designs as Problem 1.[3]

The second qualification we have to add to our conclusions drawn from the results reported so far is related to the difference in robustness of the effects from Experiment 1 and 2. We will call this problem, which concerns the assumption of equal sensitivity of the measures to effects differing in robustness, Problem 2 in what follows. While the scrambling of direct objects tested in Experiment 1 showed a robust effect on all three types of judgments with exceedingly high eta-square values, the effect of the scrambling of indirect objects tested in Experiment 2 was a bit less robust, as witnessed by the slightly lower values of eta-squared. This finding, we argued, is in line with earlier findings in which these two types of scrambling were compared (Pechmann et al. (1994), Keller (2000), Keller (2003)). We take it that the connection between the robustness of an effect and the informativity of a dependent variable used to measure that effect is crucial to our line of argument. That is, if two measures m_1 and m_2 are supposed to carry the same amount of information about two effects X and Y, one of which (say, Y) is less robust than the other, then m_1 is as informative as m_2 if and only if it is as informative as m_1 about X and Y. Although we have tested two effects with different robustness in Experiment 1 and 2, we have not tested them with the same set of materials. Thus, the difference we have obtained might be attributed to an effect of the item sets used. To obtain a direct comparison of the informativity of different acceptability measures about the difference in judging e.g. accusative vs. dative scrambling, it is imperative to use the same materials in testing the two effects.

We will try to deal with the problems raised above in a further experiment with a three-level word order factor (handling Problem 1) with a new set of participants in which we directly

compared the effects of direct vs. indirect object scrambling (taking care of Problem 2), comparing .

4 Experiment 3

Materials of Experiment 3. In this experiment, we compared to each other the three word order variants resulting from Scrambling the arguments of a ditransitive verb in German. Note in passing that in contrast to experiments 1 and 2, in this experiment, the clauses in which Scrambling took place were main clauses with the preverbal position filled by a temporal adverbial. From the possible 6 permutations, we chose to compare the following ('IO' and 'DO' stand for 'indirect object' and 'direct object' respectively): (a) S≺IO≺DO; (b) IO≺S≺DO, and (c) DO≺S≺IO. All arguments had masculine gender, hence case marking was overt. Examples of the three orders are given below.

(6) a. Dann hat der Lehrer dem Schüler den Beweis erklärt.
Then has the$_{NOM}$ teacher the$_{DAT}$ student the$_{ACC}$ proof explained.
'Then the teacher explained the proof to the student.'

(6) b. Dann hat dem Schüler der Lehrer den Beweis erklärt.
Then has the$_{DAT}$ student the$_{NOM}$ teacher the$_{ACC}$ proof explained.
'Then the teacher explained the proof to the student.'

(6) c. Dann hat den Beweis der Lehrer dem Schüler erklärt.
Then has the$_{ACC}$ proof the$_{NOM}$ teacher the$_{DAT}$ student explained.
'Then the teacher explained the proof to the student.'

There were 24 items overall, 8 per condition. These items were distributed across three lists such that each list contained 8 items in one of the three conditions. These lists were then

randomly combined with the 62 other items of the larger study which Experiment 3 was a part of. These other items served as pseudo-fillers and came from a variety of different constructions and instantiated different degrees of acceptability. This combination yielded an overall of 86 items per list; these lists were then pseudo-randomized for each participant separately such that each pair of instances of an experimental item from Experiment 3 was separated by at least 3 fillers.

Participants. 24 students of the University of Potsdam took part (10 female, 14 male; age range 21-31, mean 23.4). As before, they all were native speakers of German coming from the Berlin-Brandenburg area, and did their studies on different subjects within the social sciences. Linguistics students could only take part before having reached the third semester. None of the participants of Experiment 3 had taken part in Experiment 1 or 2. All participants received cash remuneration (6.- EUR) or course credits for their participation.

Procedure. Experiment 3 was part of a different larger study than Experiments 1 and 2, and compared only 7-point and ME judgments. Apart from this difference, the logic of the task assignment was the same as in Experiments 1 and 2: one half of the 24 participants were asked to perform the 7-point task first, and then performed the ME task on the same set of materials in a second session two weeks later. For the other half of the participants, task assignment was reversed. This manipulation entered into the statistical analyses as the factor task order.

Both experimental sessions took place in the PC pool of the University of Potsdam. Participants were seated individually in front of a PC. All other details of the procedure (labelling of the 7-point scale, training of the ME task etc.) were identical to the one in Experiments 1 and 2. One session lasted approximately 25 minutes. As soon as the second session was completed,

participants received their course credits or remuneration. As in Experiments 1 and 2, this course of action guaranteed that the rate of drop-outs for the parallel sessions was small.

Data Analysis. The data analysis for the 7-point and ME data was identical to the one in Experiments 1 and 2.

Design and Predictions. As before, we will report two analyses. Analysis 1 was a 2×3 (judgetype × word order) repeated measures design; the word order factor had the levels S-IO-DO (for 'subject-indirect object-direct object'), DO-S-IO, and IO-S-DO; judgetype was realized this time as 7-point vs. ME. This analysis tested for differences between the two measures in terms of a main effect of the factor judgetype, or an interaction of judgetype and word order. A difference in informativity for the different measures predicts that the comparison of the two measures should show a main effect, or an interaction, or both.

Analysis 2 directly compared the effect sizes of the word order factor for the different measures stemming from a one-way ANOVA. Again, a difference in informativity between the two measures would predict there to be substantive differences in the effect sizes for the word order effect. More specifically, if ME data are more informative than 7-point data, the effect sizes of the latter should be lower than that of the former.

In order to clarify the problems raised above in the discussion of Experiments 1 and 2, we recast these hypotheses with respect to these two problems. In order to show that the measures are equally informative in two-factorial and in three-factorial designs (Problem 1 above), we hypothesize that the measures show the same sensitivity to the effects in Experiment 3 as they had done in Experiments 1 and 2. This is to say that we predict that in Experiment 3, we should see the same behavior of the judgetype factor in Analysis 1 (no main effect, no interactions); and, for Analysis 2, we have to predict that the effect sizes for the two measures lie within the same

range, as they did in Experiments 1 and 2. Similarly with respect to Problem 2—our failure to show that the measures are equally sensitive to effects with differing robustness with the same set of items. Here, we have to predict that the difference in effect size that we found between the word order effects in Experiment 1 and 2 also shows up when the two types of word order manipulation show up in one Experiment. This comes down to the prediction that the difference between the S-IO-SO and the IO-S-SO condition should be smaller than the one between S-IO-DO and DO-S-IO, and that the effect sizes of the two measures should reflect this difference in more or less the same way: the effect sizes should lie within the same range for both measures irrespective of the robustness of the effect.

As in Experiment 1 and 2, we also hypothesized the markedness differences to pan out in the judgment data from Experiment 3. That is, there should be an overall markedness effect, and both a significant difference in the judgments for subject-initial vs. indirect-object-initial (4.b above) structures, as well as a significant difference between the former and the judgments for direct-object-initial structures (4.c). As stated before, this prediction is in line with the results of similar studies of the acceptability of German word order variation. Moreover, the markedness effect for the structures with scrambled indirect objects should be smaller than that for the structures with scrambled direct objects.

We also again included the factor task order into the overall analysis to check for possible effects of the order of task assignments. As above, we predicted that this factor should not show a significant main effect, nor that it should enter into an interaction with any of the other experimental factors.

4.1 Results. As in Experiments 1 and 2, the factor task order had no significant main effect on

the judgments of our participants, nor did it enter into an interaction with any of the other experimental factors (all Fs < 1).

Concerning the factor word order, the descriptive results given in Table 7 show the predicted pattern: for both measures, the subject-initial structures are judged best, the structures with a scrambled indirect object receive intermediate judgments, while the structures with a scrambled direct object are judged to be least acceptable. Furthermore, the descriptive difference between mean judgments for the subject-initial and the indirect-object-initial structures is smaller than the one between subject-initial and direct-object initial structures.

--- INSERT TABLE 7 ABOUT HERE ---

Experiment 3, Analysis 1. Following the logic of the analyses for Experiments 1 and 2, these data were entered into a two-way ANOVA with the two-level factor judgetype and the three-level factor word order. The results of the overall analysis are given in Table 8. Contrary to Experiments 1 and 2, where two samples of participants were involved in comparing the different judgment measures, in Experiment 3 there was only one sample of participants. Thus, it was licit to compute a participant analysis. Accordingly, F-statistics of the participant analysis (F_1) are given below, too.

--- INSERT TABLE 8 ABOUT HERE ---

There was a significant overall effect of word order (independent of the measure employed) on the judgments of our participants. The judgetype factor showed no significant main effect, nor did it enter into an interaction with the word order factor. Although the F-value for the interaction of word order and judgetype was not smaller than 1 in the participant analysis, as it was in the item analysis, it did not approach marginal significance. Still, we will return to this point in the discussion. Conceiving of the judgetype factor as an indicator of possible differences

between the two measures, we can conclude from this result that 7-point and ME judgments do not differ with respect to the word order factor and thus are equally informative with respect to the question pursued in this experiment. This finding will be relevant for the discussion of Problem 1 raised above, which we will pick up in the discussion section.

To look closer into the differences between the levels of the factor word order, we computed simple contrasts between the first and the second level, and the first and the third level of the word order factor. That is, we compared the judgments for the S-IO-DO word order to those for the IO-S-DO order, and S-IO-DO to DO-S-IO. For each of these contrasts, the main effect of word order, as well as the word order× judgetype interaction were computed. The results are given in Table 9.

--- INSERT TABLE 9 ABOUT HERE ---

Most important for our hypothesis concering possible informativity differences are the last two rows of Table 9: there, we can read off possible interactions of the judgetype factor with the single contrasts of the word order factor. None of these interactions reached the level of significance. However, there was a marginally significant interaction in the participant analysis for the constrast between subject-initial and direct object-initial structures (cf. the last row). We will return to this point in Analysis 2 below, as well as in the discussion.

To conclude: the results of Analysis 1 did not show any significant differences between the 7-point and ME judgments with respect to the three-level word order factor. We interpret this finding as a further piece of evidence against the assumption of an informativity difference between the measures employed. More specifically, this finding speaks against the assumption that ME judgments carry a larger amount of information about the three-level word order factor than the 7-point judgments.

Experiment 3, Analysis 2. In order to directly compare the effect sizes for the 7-point and ME judgments for the three-level word order factor, we computed two one-way ANOVAs for the two measures separately. The results are given in Table 10; as in Analysis 1 above, both participants and items were treated as a random factor in these analyses.

--- INSERT TABLE 10 ABOUT HERE ---

Table 10 reveals at a glance that the participant and the item analysis yielded different results. While in the item analysis, the F-values and eta-squared values for the two measures are almost identical, the participant analysis shows a proportion of explained variance that is still high (79.3%) for the 7-point judgments, but, for the ME judgments, effect size has dropped by almost 25 % to 54.4% as compared to the 7-point judgments. Thus, while the item analysis shows no difference in sensitivity to the word order factor, and thus no difference in informativity between the two measures, the results of the participants analysis force us to conclude that there actually is a difference. Treating participants as a random variable, the ME judgments show a substantial drop of variance accounted for as compared to the 7-point judgments, and thus appear to carry less information about that factor, then the 7-point judgments.

In order to bring more light into this pattern of results, let us again look at the simple contrasts of the three levels of the word order factor. By this analysis, we are able to establish the strength of the markedness effect of the two noncanonical structures, i.e. the structure with a scrambled indirect object, and the one with a scrambled direct object. Recall that we hypothesized on the basis of the findings of Experiments 1 and 2 that the markedness effect of the latter (DO-S-IO) should be stronger than the one for the former (IO-S-DO), as compared to the unmarked S-IO-DO. Moreover, given the drop in the variance accounted for in the participant analysis of the ANOVA for the overall design, the more detailed analysis with the simple contrasts might shed

some light on the question as to where this drop in the effect size in the participant analysis came from. The results of this analysis are given in Table 11.

--- INSERT TABLE 11 ABOUT HERE ---

First of all, all the markedness contrasts reported in Table 11 are highly significant. However, a comparison of the eta-squared values for the first vs. the second contrast reveals that the markedness effect of scrambling a direct object is stronger than that of scrambling an indirect object. This holds for both measures. Note that the overall relatively low eta-squared values have to be attributed to the fact that the simple contrasts we are reporting here only constitute one half of the overall design. Thus, the contribution of the effect of the simple contrasts to the overall effect size is smaller than the overall effect size; in fact, the partial eta-squared values for the simple contrasts are additive with respect to the overall effect.

Comparing the two measures with respect to the first contrast (S-IO-DO vs. IO-S-DO), Table 11 shows that there are no substantive differences in the effect sizes between the two measures. In the participant analysis, the eta-squared values range between 21 and 26 %, and in the item analysis, between 32 and 36 %. Thus, the two measures show almost no difference in the sensitivity to the difference between these two levels of the word order factor. However, as the second comparison (S-IO-DO vs. DO-S-IO) in the bottom row of Table 11 reveals, the participant analysis apparently again shows a drop of about 20 % of the eta-squared values of the ME judgments as compared to the 7-point judgments. In the item analysis, there is no difference in the variance accounted for, the eta-squared values ranging between 57 and 60 %.

To conclude: whereas the item analysis shows no difference in the sensitivity to the word order for the two measures, the substantive difference between the effect sizes for the two measures in the participant analysis for the second contrast supplies us with much stronger

evidence to reject the assumption that the ME judgments are more informative than the 7-point judgments. Whereas in the previous experiments, the ME judgments were equally informative as the other measures, in Experiment 3, we have found evidence that the ME judgments seem to be less informative about the effect of the word order manipulation than the 7-point judgments.

4.2 Discussion. The interpretation of the results of Experiment 3 is less straightforward than that of Experiments 1 and 2. Although Analysis 1 of Experiment 3 showed essentially the same picture as the results of Experiments 1 and 2, the results of the second analysis deviated to some degree from the earlier findings. Before discussing these deviations, we will first integrate the findings from Analysis 1 into the larger picture. Then we will turn to the differences between Experiment 3 and the previous experiments and clarify what the evidence from Experiment 3 tells us with respect to the two problems brought up in the discussion of Experiments 1 and 2 in section 3.3.

As before in the two previous experiments, our main hypothesis for Experiment 3 was couched in conditional terms: if it is the case that ME judgments are more informative than 7-point judgments, then we should be able to read this difference off a main effect or an interaction of the judgetype factor with word order in Analysis 1. We found none of these in Experiment 3, although the word order × judgetype interaction (cf. the last row of Table 9) was marginally significant in the participant analysis. Given that only the participant analysis yielded this result and that said interaction failed to reach the level of significance, we would not have to worry about it too much. But since this finding corresponds to the one of Analysis 2 below, we will pick up this issue in the discussion of these results.

Setting aside this one observation for the moment, the results we obtained for Analysis 1 of

Experiment 3 are in agreement with the ones of the two earlier experiments and disconfirm the hypothesis that ME judgments are more informative than 7-point judgments for the three-level word order factor we have employed—looking at the ANOVA for the overall 2×2-design, the two measures are equally informative.

Contrary to all the results reported so far, the results of Analysis 2 of Experiment 3 showed a difference between the results for the 7-point and the ME judgments. While the overall effect of the word order factor was the same for both measures in the item analysis, the two measures showed a difference in sensitivity to that factor in the participant analysis: the ME judgments showed a substantive drop (-25%) of variance accounted for as compared to the 7-point judgments. This, however, is strong evidence against the hypothesis that ME judgments are more informative than 7-point judgments. Rather, at least for the measurement of acceptability of Scrambling in Ditransitives in German, it seems that the reverse is true. This interpretation is further backed up by the closer look we took at the contrast between the three levels of the word order factor. The drop in the eta-squared value for the ME judgments was confined to the comparison of the unmarked subject-initial to the strongly marked direct object-initial structure. Furthermore, this equips us with a tentative explanation for the marginally significant judgetype×word order interaction found in Analysis 1. Given the findings of Analysis 2, it must be attributed to the fact that the two measures did not show the same amount of informativity for the comparison of the judgments for the S-IO-DO vs. the DO-S-IO structure. Rather, ME judgments exhibit a weaker sensitivity to this comparison than the 7-point judgments.

What does this finding tell us? First of all, that in comparing how the two measures respond to the S-IO-DO vs. DO-S-IO contrast, the participants produced a larger amount of variance we cannot account for in the ME task than they did in the 7-point task. Wherever this variance has

come from, it was not generated by the other independent variables we have controlled for. That is, it was neither generated by the word order factor, nor the influence of the experimental items (recall that the item analysis did not show any differences in variance accounted for), nor the task order factor. By way of exclusion, we are left with the explanation that it was the individual variance produced by our participants in the ME task that we have to blame for the drop in the eta-squared values in Analysis 2 of Experiment 3. That is, the condition means of the individual participants differed among each other for the ME judgments to a considerably larger extent than the individual condition means of the 7-point judgments did. Although we are aware that we are engaging in speculation at this point, we can conjecture that the explanation for this difference between the measures in the variance of the individual condition means lies in the difference in variability that the two experimental tasks grant the participant. While the 7-point judgment task restricts participants to express their relative assessment of the acceptability differences between an item in the unmarked (cf. (5.a) above) and the marked condition (5.c) to the range between the values '1' and '7', in the ME task, the participants are free to use whatever value they want to judge (5.a) (relative to the modulus) and (5.c) (also in relation to the modulus). Given the findings reported in Weskott and Fanselow (2009), where ME judgments were also shown to be less informative than binary and 7-point judgments for a factor that introduced a large amount of interindividual variability, it does not seem unconceivable that it is the relative freedom granted by the ME task that is likely to be held responsible for the drop in informativity that the ME judgments exhibit in the contrast of S-IO-DO vs. DO-S-IO reported above. While it is, from the participants' perspective, a desirable feature of the ME judgment task that they have the freedom to assign a sentence whatever value they find appropriate, this desirable feature may come, from the researcher's perspective, at the quite high and undesirable cost of being more prone to

produce variance which cannot be accounted for and hence has to be considered as spurious. The results of Experiment 3 lend some credibility to a conjecture put forward recently by Haider (2007) who suspected that the inherent variability of the ME measure may make it more susceptible to generating spurious variance: 'The variance in the aggregate data [, however,] should in 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.' (Haider 2007:393).

Furthermore, the results from Experiment 3 bring us into a position to deal with the two problems raised in connection with the discussion of Experiments 1 and 2 above. Problem 1 concerned the question wether the lack of an informativity difference between binary and 7-point judgments on the one hand, and ME judgments on the other hand, carries over from two-level to three-level repeated measure designs. Unfortunately, the answer is 'yes' and 'no': yes, because there we did not find an influence of the judgetype factor on our results in Experiment 3, which is in accordance with the previous results from Experiments 1 and 2. 'No', because the effect sizes, as revealed by Analysis 2, do differ between the two measures in Experiment 3, which they did not do in the previous experiments. However, the difference in informativity we found was not in favor of ME judgments being more informative than the 7-point judgments (as the hypothesis we followed throughout would have it), but rather the other way round. That is, ME judgments showed a drop in informativity in the three-level design, which they had not shown in the two-level design; the 7-point judgments behaved essentially the same in all three experiments.

Similarly for Problem 2: it consists of the generalizability of the approximately equal informativitiy of our measures with respect to the result that the Scrambling of direct objects yields a stronger markedness effect than the Scrambling of indirect objects. Recall that in order

to show that this is really the case, we had to show that this holds when the acceptability of the two types of Scrambling structures is tested with the same set of materials. In Experiment 3, we did just this. The general result of Experiment 3 is that both measures show that the markedness effect of the DO-S-IO structure (compared to the unmarked S-IO-DO) is stronger than that of the IO-S-DO structure; both measures yielded larger effect sizes for the former contrast than for the latter. This is true despite the fact that the absolute effect sizes for the two markedness effects were overall lower in Experiment 3 than in Experiment 1 and 2, respectively—this absolute difference is easily explained by the presence of an additional level of the word order factor in Experiment 3, lowering the overall contribution of the single contrasts. However, taking the difference in effect size obtained in the participant analysis of Experiment 3 into account, we cannot draw the conclusion that the two measures employed in Experiment 3 show the same sensitivity to the different robustness of the two markedness effects. Rather, it seems that, while the 7-point judgments show the same sensitivity to the difference in robustness of the two markedness effects as they did in Experiments 1 and 2, the sensitivity of the ME judgments was affected by the greater variability in the S-IO-DO vs. DO-S-IO contrast in Experiment 3, which we did not observe in the comparison of Experiments 1 and 2. The drop in variance accounted for we observed in the participant analysis jeopardizes the assumption that the ME data carry the same amount of information about the robustness of the two markedness effects.

5 Conclusions.

In our experiments, we aimed at an empirical assessment of the hypothesis put forward by proponents of the ME method that ME judgments as a measure of linguistic acceptability are more informative than binary or 7-point judgments. Given our results, we can outright reject this hypothesis. Not only did we not find any indication of an informativity

difference in terms of the judgetype factor in the three experiments reported. We also did not find an effect for which our measure of effect size, the eta-squared value, was substantially higher for the ME data than for the other two types of data. In fact, the opposite seems to be true at least for one contrast employed in Experiment 3, where the effect size of the ME data dropped to considerable degree in the participant analysis, while the eta-squared value of the 7-point judgments did not show such a drop.

These results clearly disconfirm the hypothesis that ME is more informative than binary and 7-point measures for the kind of structures we have investigated. Together with the problems discussed in section 2, we think that there are a number of reasons to doubt that ME constitutes the method of choice for the investigation of linguistic acceptability.

First, the problems raised in section 2.1 in connection to the task demands of ME are in need of a solution. Given the results of Ellermeier and Faulhammer (2000) and Augustin and Maier (2008) from psychophysics, as well as their recent extension to linguistic acceptability by Sprouse (ms) it appear far from clear what participants are actually doing when performing the mental arithmetic involved in an ME judgment. If the mathematical assumptions of the underlying measurement model would be fulfilled, this would not be a problem. But these assumptions are violated by the interpretation given to the numerical judgments by the participants: they do not interpret the numbers in the scientific sense, and hence their judgments violate against the assumptions about the multiplicativity property, and hence ultimately against assumptions about the scale type of ME.

Second, we have found the inherent variability of ME judgments to have a detrimental effect on the amount of variance explained in our data. Although we are confident that these findings are not limited to the design and type of manipulation employed in the studies reported here (s.

Footnote 3), we are aware that our findings are in need of further clarification and should be put under scrutiny in more complex designs and with other types of linguistic manipulations. Since a lot of empirical syntax studies involve the prediction of interactions (as opposed to main effects), looking at the effect strengths of different measures in, for example, a 2×2 design is an obvious next step. We leave it to future research to show whether there are informativity differences between ME and other acceptability measures, and if there are, what they look like.

Thirdly, and more generally, we think that in the empirical study of what people do when they give linguistic acceptability judgments, the questions still vastly outnumber the answers. In our search for the right questions and their answers, we need all the information we can get, which means that confining ourselves to one methodological option like ME does not seem to be good advice. Converging evidence from different methods has always been the gold standard of psycholinguistics. Accordingly, we think that empirical linguistics, instead of narrowing down its methodological options, should strive to accumulate evidence coming from a plurality of different methodological tools.

References

Augustin, Thomas, and Katja Maier. 2008. Empirical evaluation of the axioms of multiplicativity, commutativity, and monotonicity in ratio production of area. *Acta Psychologica* 129.1, 208-216.

Bader, Markus, and Häussler, Jana. 2009. Towards a Model of Grammaticality Judgments. *Journal of Linguistics* 46.2, 273-330.

Bard, Ellen Gurman, Robertson, Dan, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72.1, 32–68.

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

Ellermeier, Wolfgang, and Günther Faulhammer. 2000. Empirical evaluation of axioms of fundamental to Stevens's ratio-scaling approach: I. Loudness production. *Perception & Psychophysics* 62.8, 1505–1511.

Fanselow, Gisbert. 2001. Features, theta-roles, and free constituent order. *Linguistic Inquiry* 32, 405–437.

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

Featherston, Sam. 2005. Magnitude estimation and what it can do for your syntax: some wh-constraints in German. *Lingua* 115.11, 1425–1440.

Featherston, Sam. 2009. A scale for measuring well-formedness: Why syntax needs boiling and freezing points. *The Fruits of Empirical Linguistics Volume 1: Process,* ed. by Sam Featherston and Susanne Winkler. Berlin, New York: Mouton De Gruyter, 47-73.

Haider, Hubert. 2007. As a matter of facts: comments on Featherston's sticks and carrots. *Theoretical Linguistics* 33.3, 381–394.

Haider, Hubert, and Inger Rosengren. 1998. Scrambling. *Sprache and Pragmatik*, 49, Lund.

Jaeger, T. Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59, 434-446.

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. *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, ed. by Richard Alterman, and David Kirsh. Boston, pp. 652–657.

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

Lodge, Milton. 1981. *Magnitude Scaling. Quantitative Measurement of Opinions*. Newbury Park, CA: Sage Publications.

Manning, Chris D., 2003. Probabilistic Syntax. *Probabilistic Linguistics,* ed. by: Rens Bod, Jennifer Hay, and Stefanie Jannedy. Cambridge, MA: MIT Press.

McMurray, Brian, Tanenhaus, Michael K., Richard N. Aslin. 2002. Gradient effects of within-category phonetic variation on lexical access. *Cognition* 86.2, B33–B42.

Narens, Louis. 1996. A theory of ratio magnitude estimation. *Journal of Mathematical Psychology* 40, 109–129.

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

Peirce, Charles A., Block, Richard A., and Herman Aguinis. 2004. Cautionary Note on Reporting Eta-Squared Values From Multi-Factorial ANOVA Designs. *Educational and Psychological Measurement* 64.6, 916–924.

Schaffer, Nora Cate, and Norman M. Bradburn. 1989. Respondent Behavior in Magnitude Estimation. *Journal of the American Statistical Association* 84, 402–413.

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

Sorace, Antonella. 2000. Gradients in auxiliary selection with intransitive verbs. *Language* 76.4, 859–890.

Sprouse, Jon. 2007a. Experimental Syntax: what does it get you? San Diego, CA, Talk presented at the 20th Annual CUNY Conference on Human Sentence Processing.

Sprouse, Jon. 2007b. A program for experimental syntax: Finding the relationship between acceptability and grammatical knowledge. Ph.D. thesis, University of Maryland.

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

Sprouse, Jon. Ms. Evaluating the Assumptions of Magnitude Estimation of Linguistic Acceptability. University of California at Irvine.

Stevens, Stanley Smith. 1946. On the theory of scales of measurement. *Science* 103, 677–680.

Stevens, Stanley Smith. 1956. The direct estimation of sensory magnitude – loudness. American *Journal of Psychology* 69, 1–15.

Stevens, Stanley Smith. 1975. Psychophysics. Introduction to its perceptual, neural, and social prospects. Wiley, New York.

Wegener, Bernd. 1983. Category-Rating and Magnitude Estimation Scaling Techniques: An Empirical Comparison. Sociological Methods and Research 22(1), 31–75.

Weskott, Thomas, and Gisbert Fanselow. 2008. Variance and Informativity in Different Measures of Linguistic Acceptability. In: Abner, Natasha, and Jason Bishop (eds.). Proceedings of the 27th West Coast Conference on Formal Linguistics. Somerville, MA: Cascadilla Press, 431-439.

Weskott, Thomas, and Gisbert Fanselow. 2009. Scaling Issues in the Measurement of Linguistic Acceptability. In: Featherston, Sam, and Susanne Winkler (eds)., 231-245.

Notes

[1] It should be noted that in the original psychophysics ME experiments, each stimulus has to serve as a modulus in the judgment of all the other stimuli, that is, these studies employ complete permutations of stimulus-modulus pairs. This procedure, although it may solve the problem of effect variability that we mention below, as well as that of the lack of colinearity noticed by Sprouse (2008), is probably not feasible in experiments on linguistic acceptability, if only for the sheer amount of items each participant would have to judge in order to reach a sufficient statistical power.

[2] We want to point out that there is a difference between 'classical' eta-squared and the partial eta-squared reported here. While classical eta-squared is the ratio of the variance induced 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-squared values are not necessarily additive, and 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 Analysis 2, we do not differentiate here between classical and partial eta-squared.

[3] We have shown that the claims we make about the informativity of acceptability measures can be generalized beyond experiments concerned with two- and three-level word order manipulations in Weskott and Fanselow (2009), where we report on results similar to the ones obtained in the current study. These experiments, however, differ from the ones reported here in two respects: the employ a four-level factor, and they make use of both syntactic and semantic

violations. We take these results to show that the results reported here can be generalised beyond two-level factors, as well as beyond the type of linguistic phenomena tested here.

Tables

	SO		OS	
binary1	.94	(.24)	.54	(.50)
binary2	.83	(.37)	.41	(.49)
7-point	6.33	(1.44)	4.14	(1.97)
ME	1.35	(.84)	.51	(.87)

Table 1

Condition means and standard deviations (in brackets) for the word order factor in Experiment 1

	7-point vs. binary1		ME vs. binary2	
	$F_2(1,7)$	p	$F_2(1,7)$	p
wo	197.63	<.001	397.04	<.001
jt	<1	>.10	<1	>.10
wo×jt	<1	>.10	<1	>.10

Table 2

Results for the two-way ANOVA (word order (wo) × judgetype (jt)) for Experiment 1

	F_2 (1,7)	p	eta-squared
binary1	140.41	<.001	.953
binary2	141.87	<.001	.953
7-point	159.57	<.001	.958
ME	359.72	<.001	.981

Table 3

Results of the one-way ANOVA with the word order factor only for Experiment 1

	SO		OS	
binary1	.93	(.26)	.62	(.49)
binary2	.80	(.40)	.52	(.50)
7-point	6.25	(1.55)	4.33	(1.88)
ME	1.25	(.84)	.64	(.87)

Table 4

Condition means and standard deviations (in brackets) for the word order factor in Experiment 2

	7-point vs. binary1		ME vs. binary2	
	$F_2(1,7)$	p	$F_2(1,7)$	p
wo	72.57	<.001	73.11	<.001
jt	<1	>.10	<1	>.10
wo×jt	<1	>.10	<1	>.10

Table 5

Results for the two-way ANOVA (word order (wo) × judgetype (jt)) for Experiment 2

	F_2 (1,7)	p	eta-squared
binary1	70.43	<.001	.910
binary2	34.44	<.001	.831
7-point	84.71	<.001	.924
ME	163.93	<.001	.959

Table 6

Results of the one-way ANOVA with the word order factor only for Experiment 2

	S-IO-DO	IO-S-DO	DO-S-IO
7-point	6.38 (1.05)	4.06 (1.68)	3.15 (1.67)
ME	1.00 (.82)	-.02 (.87)	-.28 (.93)

Table 7

Condition means and standard deviations (in brackets) for the word order factor in Experiment 3

	7-point vs. ME			
	$F_1(2,46)$	p	$F_2(2,46)$	p
wo	78.04	<.001	450.44	<.001
jt	<1	>.10	<1	>.10
wo×jt	2.40	>.10	<1	>.10

Table 8

Results for the two-way ANOVA (word order (wo) × judgetype (jt)) for Experiment 3, participant (F_1)and item analysis (F_2)

	7-point vs. ME			
	$F_1(2,46)$	p	$F_2(2,46)$	p
wo				
s-io-do vs. io-s-do	54.90	<.001	926.87	<.001
s-io-do vs. do-s-io	100.97	<.001	584.92	<.001
wo×jt				
s-io-do vs. io-s-do	<1	>.10	<1	>.10
s-io-do vs. io-s-do	3.14	.09	<1	>.10

Table 9

Results for the simple contrasts for two-way ANOVA (word order (wo) × judgetype (jt)) for Experiment 3, participant (F_1)and item analysis (F_2)

	participant analysis			item analysis		
	$F_1(2,46)$	p	eta-squared	$F_2(2,46)$	p	eta-squared
7-point	88.34	<.001	.793	169.30	<.001	.880
ME	27.41	<.001	.544	166.66	<.001	.879

Table 10

Results of the one-way ANOVA with the word order factor for Experiment 3

	participant analysis			item analysis		
	$F_1(2,46)$	p	eta-squared	$F_2(2,46)$	p	eta-squared
s-io-do vs. io-s-do						
7-point	62.83	<.001	.257	583.30	<.001	.320
ME	21.81	<.001	.210	273.43	<.001	.358
s-io-do vs. do-s-io						
7-point	122.18	<.001	.545	218.42	<.001	.603
ME	33.13	<.001	.335	327.63	<.001	.572

Table 11

Results of the simple contrasts of the one-way ANOVA with the word order factor for Experiment 3

Figures

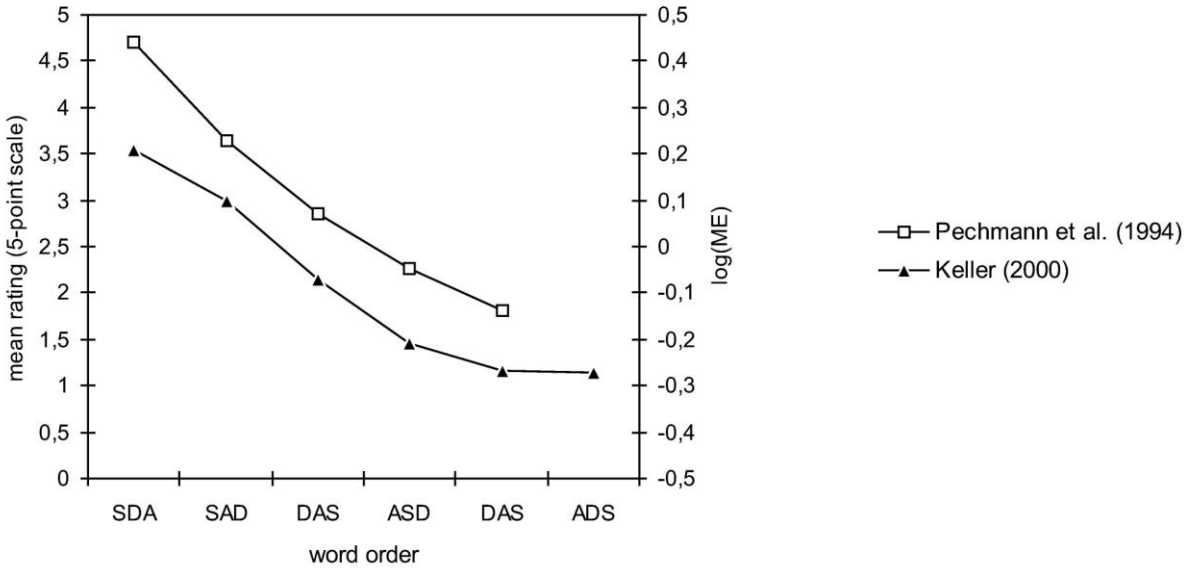

Figure 1: Comparison of the data on word order variation in German ditransitives from Pechmann et al. (1994) and Keller (2000). The left vertical axis plots the Pechmann et al. (1994) mean ratings from a 5-point scale, the right vertical axis plots the logarithmized ME rating values from Keller (2000). 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).