MAGNITUDE ESTIMATION OF LINGUISTIC ACCEPTABILITY

ELLEN GURMAN BARD, DAN ROBERTSON, ANTONELLA SORACE

University of Edinburgh

Judgments of linguistic acceptability constitute an important source of evidence for theoretical and applied linguistics, but are typically elicited and represented in ways which limit their utility. This paper describes how MAGNITUDE ESTIMATION, a technique used in psychophysics, can be adapted for eliciting acceptability judgments. Magnitude estimation of linguistic acceptability is shown to solve the measurement scale problems which plague conventional techniques; to provide data which make fine distinctions robustly enough to yield statistically significant results of linguistic interest; to be usable in a consistent way by linguistically naive speaker-hearers, and to allow replication across groups of subjects. Methodological pitfalls are discussed and suggestions are offered for new approaches to the analysis and measurement of linguistic acceptability.*

1. GRAMMATICALITY INTUITIONS AS EVIDENCE.

1.1. INTRODUCTION. For many linguists, intuitions about the grammaticality of sentences comprise the primary source of evidence for and against their hypotheses. Typically provided by the linguist or by close associates, the intuitions are reported in a variety of terms—acceptable, marginally acceptable, unacceptable, good, terrible, etc.—and coded with such symbols as ?, *, **. Although this system has supported a research program of considerable accomplishment over several decades, it presents difficulties that are widely, if informally, recognized, and seldom confronted (for exceptions, see Newmeyer 1983; Sorace 1988, 1990). The purpose of this paper is to characterize some of the major difficulties inherent in current methods of judging grammaticality and to propose a better way to elicit intuitions.

To do this we will treat judgments about the grammatical status of sentences as psychological evidence, that is, like judgments about other phenomena on which human perception and cognition work. We will discuss the difficulties of eliciting reliable, consistent judgments revealing subjects' true powers of discrimination. Over the last century, psychophysics has dealt with problems surprisingly similar to those presented by linguistic judgments, and we will apply some of the lessons of psychophysics to the present problem. We will show how the customary measurement of judged grammaticality loses information and makes it difficult to test hypotheses of current linguistic interest. We attribute these problems to the use of the wrong kind of measurement scale, and we describe what the right kind of scale would be.

In the remainder of the paper we present a technique called MAGNITUDE ESTIMATION, developed by psychophysicists to make maximal use of subjects' ability to make fine judgments about physical stimuli, and we describe how it can

* This work was supported by ESRC Project Grant R000233965 to the authors, whose names are listed in alphabetical order. The authors are grateful to J. Levy and C. Theobald for their advice, to E. Engdahl, S. Garrod, and two anonymous reviewers for comments, and to the subjects for their participation. A preliminary version of this paper was presented at the Spring 1993 meeting of the Linguistics Association of Great Britain.

be adapted to the elicitation of judgments about the grammaticality of sentences. We show that it is easy to operate informally, that it can support statistically more robust distinctions than more familiar techniques when applied to a question of linguistic interest, and that it elicits judgments that are consistent within and between subjects. Finally we discuss how magnitude estimation may be applied to make better use of our capacity to make judgments about sentences.

1.2. SOME DIFFICULTIES. By performing the small and imperfect experiment of judging those sentences that are critical to linguistic theories, linguists intend to assess grammaticality, that is, compatibility with the grammar of a particular language, or well-formedness under the assumptions about linguistic competence used to build the grammar. By asking speakers of the language to make judgments about sample strings, linguists test the hypothesis that speakers' views and linguists' proposals for the grammar match. Yet eliciting those views does not give direct access to speakers' linguistic competence. What is observed instead is a particular kind of linguistic behavior, an overt response to the subjects' opinion about characteristics of the sentence. Thus we can make a three-way distinction among GRAMMATICALITY, a characteristic of the linguistic stimulus itself, ACCEPTABILITY, a characteristic of the stimulus as perceived by a speaker, and the ACCEPTABILITY JUDGMENT which is the speaker's response to the linguist's inquiries. The fact that the subject offering the opinion and the linguist generating the proposals are often the same person does not change the fact that the impression on offer is an acceptability judgment, behavioral evidence around which the theory develops.

The distinction between acceptability and grammaticality unveils a further distinction between RELATIVE GRAMMATICALITY, which is an inherent feature of the grammar, and RELATIVE ACCEPTABILITY, which is perceived by the subject. Insofar as judgments about acceptability represent effects of the grammar, the overt manifestation of both relative grammaticality and relative acceptability is gradience in acceptability judgments. While the existence of relative acceptability is easily accepted (cf. for instance Newmeyer 1983, Rizzi 1990), inherent gradience within the grammar has a more controversial status, since it appears to be difficult to accommodate within formal linguistic theories (McCarthy & Prince 1993; Sorace 1995).[1] Nonetheless, the possibility remains that acceptability is graded because grammaticality is.

Of course, acceptability judgments, like other manifestations of linguistic performance, need not be one-to-one reflections of grammaticality. First, it is always possible that the subject is not reporting directly on grammaticality but is responding to any number of other features of the stimulus (Botha 1973, Quirk & Greenbaum 1970). Impressions of acceptability may be based, for example, on estimated frequency of usage, on conformity to a prescriptive norm or a prestigious register, or on degree of semantic or pragmatic plausibility.

[1] While syntactic theorizing in different frameworks is assigning growing importance to the notion of comparison, the consensus is that there is only one optimal output that best satisfies the system of interacting constraints and therefore receives a grammatical interpretation (see Chomsky 1991).

Second, where the linguist acts at the same time as the theoretician and the source of the data (Labov 1970), results may be subject to bias, however unconscious, towards an outcome concordant with the judge's vested interests. Even judges with no direct knowledge of the field can be biased in another way, by the context in which the judgment is made, and in particular by repeated exposure to sentences of particular kinds (Levelt 1972). Finally, details of extralinguistic context may have consistent effects on judgments, which may tell us as much about the process of introspection as about linguistic abilities (Carroll et al. 1981, Nagata 1987a, 1987b, 1988, 1989).

Although these difficulties may obscure the primary data linguists need, we seldom react, as cognitive psychologists would, by attempting to develop methods of minimizing the artifacts. Instead, a 'small is beautiful' principle seems to operate: the empirical damage is limited by dependence on striking rather than exhaustive examples and judgments are made by a small community of subjects who share an agreed definition of acceptability.

Whether or not the small-is-beautiful approach solves the problems of interpretation and of bias remains to be seen. Certainly it does little to mitigate an even greater difficulty, the inherent inadequacy of the measuring instrument used for linguistic acceptability judgments. One symptom of the problem is the fact that symbols used for categorizing example sentences tend to vary in application even within the work of a single author. Consider, for example, the following items drawn from a textbook on GB theory (Haegeman 1991):

(1) a. *Which man did Bill go to Rome to visit?* (H:35a, p. 500)
b. ?*Which man do you wonder when to meet?* (H:44a, p. 502)
c. ?*This is a paper that we need someone who understands.* (H:50a, p. 505)
d. ?*Which car did John announce a plan to steal tonight?* (H:53a, p. 506)
e. **Whom do you know the date when Mary invited?* (H:31a, p. 495)
f. **Where did Bill go to Rome to work?* (H:35b, p. 500)
g. **This is a book which reading would be fun.* (H:38a, p. 500)
h. **With which pen do you wonder what to write?* (H:44b, p. 502)
i. **This is a paper that we need someone that we can intimidate with.* (H:50b, p. 505)
j. ***This is a pen with which writing would be fun.* (H:38b, p. 500)

Most of these examples are cited in pairs in their original source. In each case we are invited to note that the second member of the pair is less acceptable than the first. If an impression of relative acceptability were the only goal of acceptability judgments, however, the symbol > would always be adequate to express the critical data. By reassembling the original pairings of the examples in 1, the reader can demonstrate that the implied relative judgments are usually easy to reproduce. The use of the 0–?–*–** scale (where 0 indicates an acceptable sentence) indicates something more, however. As is normal practice, Haegeman is attempting to indicate the absolute acceptability of these sentences. This is where the problem arises.

Even though absolute acceptability is usually not of primary interest, the 0–?–*–** scale ought to facilitate building extended linguistic arguments on the basis of acceptability judgments. To deliver this, however, the symbols recording judgments should be capable of consistent application over a few pages of text. Thus, if the scale allowed reasonable representation of both relative and absolute acceptability, then sentences marked with the same symbol should be roughly comparable in acceptability and any sentence marked ** should be worse than any marked *, which should in turn be worse than any marked ?, and all of these should be recognizably less acceptable than an unlabelled acceptable sentence. This condition does not appear to hold in 1. Example 1c seems less acceptable than 1b, for instance, though both are labelled ?, and 1j, marked **, does not seem to be markedly worse than 1e, marked *. That this can be true even when we agree with the original relative judgments means that something is amiss with the scale in which they are represented.

The fault is certainly not Haegeman's. It derives instead from the disproportion between the fineness of judgments people can make and the symbol set available for recording them. Each of the symbols in the 0–?–*–** scale appears to cover a range of acceptability levels. That is, if the sentences in 1 are ranked by acceptability without regard to the grammaticality annotations, and if the annotation scale is adequate, we should find that only four different degrees of acceptability are discriminable. The greater the number of discriminable ranks beyond four, the more information the four-point scale must be hiding from us.

The same argument applies to extended scales like 0–?–??–?*–*–** or to the five-point scale often used in empirical studies or indeed to any other scale that predetermines the number of distinctions subjects may use. There is no way of knowing in advance if our sensitivities are limited to a five-way distinction any more than a four-way distinction. It is instructive to illustrate the problem with a five-point numerical scale, which can be used both carefully and to good effect in many domains. In this domain, the relative and the absolute uses of the scale can conflict. Imagine that a sentence Sa and the appreciably less acceptable corresponding sentence Sb might both fall within a carefully defined '3' category. Imagine that Sa and its corresponding Sc differ more than Sa and Sb, but still reside within that part of the range labelled 3. Neither difference will be recordable on this scale, for all three examples will be coded 3. Nor will there be any legitimate way to report that one difference is perceived to be larger than the other. Now imagine that the a v. b pair for another sentence, Z, differ in acceptability as noticeably as Sa and Sb, but this time the difference genuinely crosses the carefully marked 2/3 boundary. Now of two equal differences, Sa v. Sb and Za v. Zb, one is lost to view. Of two unequal differences, the smaller, Za v. Zb, can be reported, while the larger, Sa v. Sc, does not register. The only way to get around these difficulties without expanding the scale is to pervert it, that is, to move the boundaries between numbers in order to reflect perceived differences. Thus the first sentence less acceptable than a genuine 3 will be labelled 2 and any subsequent even less acceptable sentence then has to be labelled 1. Confusion between genuine or absolute 2 and forced

or relative 2 will then arise. In effect, the subject has to choose between being less than informative and being less than consistent.

Working linguists know very well, of course, that each symbol covers a range of judged degrees of acceptability, and that in practice the ranges covered by different symbols will often overlap. Sensibly enough, linguists rely more heavily on the ability of the symbols to express relative acceptability and make less direct use of their dubious relationship to absolute acceptability.

If the field has progressed using limited annotations, can it be important that they tend to lose information subjects might be able to provide? For some time it was arguable that the loss was harmless, because the generalizations of interest were fairly broad. More recently, however, the scale for measuring linguistic acceptability has begun to curtail the utility of the elicited judgments. Consider two examples.

One is well known. It deals with the relative effects of Subjacency and ECP violations (Chomsky 1986, Rizzi 1990). To support the proposal that one of these principles is 'stronger' than the other, it is necessary to elicit intuitions about the relative unacceptability of strings that violate them. Subjects must therefore judge whether a sentence that violates the Subjacency Principle (like 2b and 3b) or a sentence that violates the ECP (like 2c and 3c) is less acceptable.

(2) a. *John announced a plan to steal Bill's car late tomorrow.*
b. *?Which car did John announce a plan to steal late tomorrow?*
c. **When tomorrow did John announce a plan to steal Bill's car?*

(3) a. *I wonder whether John can solve the problem.*
b. *?Which problem do you wonder whether John can solve?*
c. **Who do you wonder whether can solve the problem?*

Here it is not necessary to show that all (c) examples are equally undesirable. The hypothesis about the relative importance of Subjacency and ECP does not preclude the possibility that some (a) sentences will be less acceptable than others or that our judgments may be subject to adventitious effects of the lexical, propositional, or pragmatic contents of each set of sentences. Instead it is important to determine whether, despite all these factors, an ECP violation generally reduces perceived acceptability more than a Subjacency violation in whatever sentence structures they may be instantiated. What is needed is a comparable effect of a violation over a number of sentence structures. This is rather more demanding than finding that for every acceptable (a) example, the (c) version is worse than the (b) version. It requires that, for example, 2b should be less acceptable than 2a to roughly the same degree as 3b is worse than 3a, and that each of these reductions in acceptability should be smaller than the one created by the violations in 2c and 3c. In other words, testing the hypothesis requires comparing differences in acceptability.

The difficulty is that we have no obvious way of estimating such differences. Even if it included a different symbol for every sentence, a scale like ?–*–** would not allow us to subtract the acceptability of 2b from the acceptability of 2a and compare the result with the outcome of the parallel operation in 2c and 2a, 3b and 3a, and so forth. Because there is no scale on which the difference

between * and ? can be represented, the notion comparable effect will not find an easy definition. The five-point scale is a tempting alternative here, because operations like subtracting 2 from 4 would seem to allow the necessary comparisons. As we have just seen, however, it might be impossible to perform the arithmetic accurately without assurance that we had encountered the genuine rather than the relative 2.

A second example is drawn from Sorace 1992, 1993a, 1993b, 1996 (to which we will return in §5). Here the issue is whether our knowledge of a syntactic generalization is equally secure throughout its domain of application. Sorace proposes that Italian native speakers' knowledge about the restrictions on combinations of auxiliaries and lexical verbs is not equally determinate for all pertinent verbs. Although, for example, speakers of Italian agree that unaccusative verbs select the auxiliary *essere* 'be' and unergative verbs select *avere* 'have' (Perlmutter 1978, 1989, Burzio 1986), they should not agree to the same extent for all unaccusative and all unergative verbs. For cases toward the core of the system, it is predicted that speakers should very clearly accept the canonical auxiliary and reject the alternative, while for other, more peripheral cases, they should be progressively less definite in their views. In this case, the hypothesis finds at least two natural interpretations. Unfortunately, neither is currently easy to apply.

First, indeterminacy might be reflected in differences between the acceptability of canonical (a) and alternate (b) auxiliaries with particular verbs. The two might differ greatly in acceptability in the case of core examples like those in 4 below or relatively little in peripheral instances like 6. For this approach we once again need to be able to subtract the acceptability of the dispreferred form from the preferred, and as we have seen, the scales in use offer no such facility.

(4) a. *Maria è andata in ufficio a piedi.*
Maria.FEM.SG is gone.FEM.SG to office on foot
b. **Maria ha andato in ufficio a piedi.*
Maria.FEM.SG has gone.MASC.SG to office on foot
c. 'Maria went to the office on foot.'

(5) a. *Paolo è rimasto a letto tutto il giorno.*
Paolo.MASC.SG is stayed.MASC.SG in bed all the day
b. **Paolo ha rimasto a letto tutto il giorno.*
Paolo.MASC.SG has stayed.MASC.SG in bed all the day
c. 'Paolo stayed in bed all day.'

(6) a. *Gli unicorni non sono mai esistiti.*
Unicorns.MASC.PL not are never existed.MASC.PL
b. **Gli unicorni non hanno mai esistito.*
Unicorns.MASC.PL not have never existed.MASC.SG
c. 'Unicorns never existed.'

The second interpretation is more direct, but even more problematical: indeterminacy might be reflected in the variability of an individual's or a group's judgments, whether from verb to verb at a particular position between core and periphery, from trial to trial on the same verb, or from subject to subject.

For example, judgments on 4b should be more consistent than those on 5b or 6b, even for subjects who are very secure in the belief that all the (b) examples are less acceptable than the corresponding (a) examples.

Yet variability of judgments will be difficult to assess using terms like ?, *, and **. Although we might score each dispreferred example for the number of times it was given each annotation, we will still suffer from the use of a scale that can lose distinctions in apparent acceptablity. Deciding whether results are due to genuine confusion about the status of examples or to an inadequate and confusing set of symbols may be more trouble than it is worth. Using more common measures of variability, like the standard deviation, for instance, is simply out of the question with annotations that preclude simple arithmetic.

To give such examples the serious study they merit, we need a better way of measuring acceptability. In both cases we need a measure of perceived acceptability so sensitive that we can use all the judgments our subjects produce and so structured that we can at least make simple arithmetic estimates of differences in perceptions.

Readers unfamiliar with the history of experimental psychology may feel at this point that we are trying to replace a simple and well-practiced technique with an alternative of unknown and unnecessary complexity. Readers familiar with this history, on the other hand, may recognize the sorts of problems that inspired the development of measurement theory and of a phalanx of judgment-elicitation techniques in experimental psychology. The purpose of this paper is to bring to the service of linguistic investigations one such method originally developed by psychophysicists to elicit subjects' impressions of various physical phenomena and subsequently adapted for use with a number of psychosocial domains. With the proper application of this method, many of the difficulties outlined here can be overcome. Insofar as linguistics is a branch of psychology that studies a specialized kind of human perception, it is a sister field to psychophysics, the study of relationships between human sensations and the physical universe. Transfer of techniques is more than appropriate.

2. A MEASUREMENT SCALE FOR ACCEPTABILITY. To understand what is at stake here, it will be helpful to recast the problem in terms of the kinds of measurement scales involved. Measurement is often defined as the 'assignment of numbers to things according to rule' (after Stevens 1946:667). Four types of scale are commonly distinguished: nominal, ordinal, interval and ratio (Stevens 1946). They are ordered in terms of their formal properties, the kinds of information they use, and, consequently, the kind of mathematical operations that can be performed on the measurements (Stevens 1951; see Michell 1990 for a discussion). Because the scales are effectively ordered in the precision with which they use available information, any type of data will be most adequately measured on the highest applicable scale. Our introductory examples illustrated two claims: first, that the scales on which acceptability has heretofore been measured appear to be too condensed to reflect our intuitions accurately and, second, that whatever their length, these scales 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. We will illustrate this claim as we set out the characteristics of the different sorts of scales.

The simplest measurements are via NOMINAL SCALES. These are easiest to view as a set of labels assigned according to rule, like *apple, banana, orange.* Nominal scales have one formal property, the property of equality: if it is meaningful to say of two objects A and B that they are either equal or not equal with respect to some attribute or property, then that attribute or property can be measured on a nominal scale. Items measured on a nominal scale can be categorized but not ordered in any way. No mathematical operations can be performed on these measures other than counting the items in each category and comparing the totals.

Some kinds of data may be perfectly well measured via a nominal scale. The fruit example is typical. There is no inherent order among the apple, banana, and orange categories. There are no intermediate cases. There is no notion like 'average fruit' which we are prevented from expressing because we cannot add apples and oranges or divide by bananas. In many views, grammatical and ungrammatical form an exhaustive nominal scale. This scale will not measure relative ungrammaticality, however, because points intermediate between the two categories will be as impossible to reflect as fruits that are a bit more apple than banana. Even if a nominal scale were expanded to include a doubtful category, it would not order this category between grammatical and ungrammatical any more than it could order pears between apples and bananas.

Once order is introduced, the scale is an ORDINAL SCALE. These have two formal properties: equivalence and order. If two objects are the same with respect to a particular property, while each has more of that property than a third, then the property we are dealing with can be measured on an ordinal scale. An ordinal scale rank orders scale points but makes no commitment to any other kind of difference between them. If we were to put ordinal scale points on one axis of a graph, we would have to assume that the axis was elastic, for the distance between successive points is both unknown and undependable. For this reason, we can count the number of items at each rank or groups of ranks, but we would have difficulty performing arithmetic across them.

In §1 we suggested that the system of symbols {0, ?, *, **} comprises such a scale: each member indicates less acceptability than the previous one. We showed, however, that this scale is often applied in such a way as to violate both the equivalence and the order conditions. Even if the scale were appropriately applied, the mathematical limitations of ordinal scales, not to mention the nonnumerical symbols used to measure acceptability, would stand in the way of testing linguistic hypotheses dependent on notions like 'comparable difference'. The difference between successive ranks, 0 and ?, for example, is not only an odd concept, but also one that cannot be predicted to be equal to another successive-ranks difference, * and **, or less than a nonsuccessive difference, ? and **.

INTERVAL SCALES allow us to measure difference. To equality and order, interval scales add regular difference between successive pairs of measurements.

A property is measurable on an interval scale if we can meaningfully compare the differences between pairs of objects with respect to that property. Once we can do this, various useful mathematical operations become available. Skirt length in inches below the knee is an interval scale: a skirt 2 inches below the knee is longer than one 1 inch below the knee by as much as a skirt 4 inches below is than another 3 inches below. Because interval scales have measurement points at equal intervals, they support subtraction.

It may seem strange to think of linguistic acceptability as an interval scale, but we contend that only historical accident and the basic nature of early linguistic hypotheses originally led to the use of nominal and ordinal scales of measurement rather than interval. Once it is proposed, as the ECP/Subjacency discussion does, that we can reliably judge the difference in acceptability between one principle-respecting sentence and its principle-violating mate as greater than the corresponding difference for a pair respecting and violating another principle, then linguistic theory has outgrown simpler measurement scales.

If interval scales can be applied, our analytic tools multiply. Although there are descriptive and inferential statistics for nominal and ordinal scales, much greater variety is available for interval data.[2] So long as judgments actually take on the necessary characteristics, we should be able to pursue the psycholinguistics of intuitions in detail comparable to what is available in other branches of perceptual and cognitive psychology.

More informative still are RATIO SCALES. A property is measurable on a ratio scale if it satisfies the criteria for an interval scale and the additional condition that the ratios between measurements can be discovered. To make this possible, the distance of each item from a common 0-point must be known. For skirt lengths, measurement from the knee gives interval measurement of the differences between lengths, but will not allow us to say that one skirt is 1.5 times as long as another. To do this we need to measure the skirts from their waistband origin, so that we can determine that one skirt is 33 inches long and the other 22 inches long from waist to hem.

It may stretch the imagination to suppose that ratio scale measurement is appropriate for judgments of acceptability. If the principled arguments are less compelling here than in the case of interval scales, principally because it is unclear what a string with 0 acceptability would be like, the two scales are linked by a judgment elicitation technique called MAGNITUDE ESTIMATION: providing that subjects' abilities are as great as we have supposed, attempts to say which sentence is 1.5 times as acceptable as another, and which .6 times as acceptable, and so forth, can at least give us the interval scales that we need.

3. MAGNITUDE ESTIMATION: ESTABLISHING THE SUBJECTS' SCALE. Magnitude estimation was developed to provide better than ordinal scales for measuring impressions of physical continua (Stevens 1956). As originally applied to the direct estimation of brightness or loudness, magnitude estimation in its simplest

[2] There is considerable disagreement about the use of parametric statistics with ordinal measurement. For discussions, see Gaito 1980, Townsend & Ashby 1984, Michell 1986.

version requires the subject to associate a numerical judgment with a physical stimulus (see Stevens 1975 for a review). Once the initial stimulus, or modulus, is presented and a number associated with it by experimenter or subject, the subject assigns to each successive stimulus a number reflecting the relationship between that stimulus and the modulus. Subjects are explicitly instructed to reflect perceived ratios in their judgments: a stimulus that appears to be 10 times as bright as the first is to be given a number 10 times the original number; one that seems one-third as bright is given a number one third the size. However bizarre they may find the task at first, normal adults can reliably perform it for a large number of physical continua.

Magnitude estimation fills exactly the needs which we have been discussing. First, it does not restrict the number of values which can be used to measure the property of interest. Subjects decide whether each stimulus should be assigned the same number as another stimulus or a different number, and they have complete freedom about which of the infinite set of numbers to use. Accordingly both the range of responses and the distribution of individual responses within that range are informative. Second, because ratio-scale judgments subsume an interval scale, it is possible to subtract the number assigned to one stimulus from the number given to another and produce meaningful differences which directly reflect differences in impressions. By the same token it is also possible to calculate the mean and the variance for multiple judgments on a particular type of stimulus.

Most important for psychophysics, magnitude estimation provides measurements of impressions on a numerical scale which can be plotted against the objective measure of the physical stimuli giving rise to the impressions. As a result, psychophysical relationships can be viewed as a set of mathematical functions. Although there is dispute about the generality of the finding (see Poulton 1986, 1989 for a critique), when the subject's estimates of magnitude (or group geometric mean estimates or medians) are plotted in log-log coordinates against the physical dimension, the points tend to follow a straight line with a slope characteristic of the physical property being assessed. The straight line in log-log coordinates means that equal ratios on the physical dimension give rise to equal ratios of judgments. In judgments of brightness, for example, every time the stimulus energy doubles, the subjective brightness becomes 1.5 times larger. In judgments of line length, on the other hand, the function is steeper: doubling physical line length doubles subjective line length as well. The characteristic relationship is reflected in the value of this slope, called b or B.[3]

[3] Psychophysical relationships of this kind are expressed in the form of equations called power laws with the alternative forms below (Stevens 1957):

(i) $\psi = R = kS^b$ or $log\ R = log\ k + b\ log\ S$

Here ψ is the subjective magnitude of the stimulus, R is the response estimating that magnitude, S is the physical magnitude of the stimulus itself, k is a constant, and b is the exponent that is characteristic of the S (Lodge 1981:13). In its log form, b gives the slope of the straight line function in log-log coordinates and log k the intercept. Thus, the variable b is what characterizes a sensory domain: in brightness estimation, as we indicated, b is .5, while in line-length estimation it is 1.0.

4. THE CASE FOR MAGNITUDE ESTIMATION OF LINGUISTIC ACCEPTABILITY.

4.1. MAGNITUDE ESTIMATION FOR LINGUISTIC INTUITIONS. Magnitude estimation has often been applied to linguistic stimuli with properties for which some objective interval scale is available: speech rate (Grosjean 1977, Grosjean & Lass 1977, Green 1987), vowel roughness (Toner & Emanuel 1989), similarity of syllables from different languages (Takefuta et al. 1986), quality of synthesized speech (Pavlovic et al. 1990), and speech intelligibility (Fucci et al. 1990).

Acceptability differs qualitatively from these examples, however. Unlike apparent vowel roughness, which can be plotted against amplitude of aperiodic energy, linguistic acceptability has no obvious physical continuum to compare with the subjects' impressions. In theory, it is the predictions of the grammar that should replace objective physical descriptions here: psycholinguistic and psychophysical relationships should be analogous. They are not, because linguistic theory does not make predictions in the same measurement scales as physics does. Even though we need interval scales to test linguistic theories, the theories themselves do not predict precise intervals. At best, the kinds of predictions we have been describing deal in orders of results: they predict, for example, that one error should be worse than another but not how much worse or how absolutely bad. So although we could select a pair of stimuli for a typical psychophysical experiment in such a way that one is twice as bright as the other, or twice as long, or twice as red, we cannot find a sentence that linguistic theory designates as twice as grammatical as some other.

In psychophysics, the utility of magnitude estimation is demonstrated when an orderly psychophysical function emerges. Without a suitable 'physical' scale for acceptability, we are unable to make such a simple argument here. Instead, we have had to take a multipronged approach to discovering whether magnitude estimation can serve the needs of linguistics.

First, in §4.2, we demonstrate that the technique is easy to apply informally with naive or experienced judges and that it produces data with a prima facie resemblance to familiar acceptability judgments. Second, in §5, we demonstrate the ability of magnitude estimation to reveal distinctions of linguistic interest in a statistically robust way. To do this we summarize selected results from a large-scale study (Sorace 1992) which elicited the views of native and non-native speakers of Italian. Third, we address the issue of the missing axis. We have borrowed an extension of magnitude estimation, called CROSS-MODALITY MATCHING, which also originated in psychophysics (J. C. Stevens et al. 1960) but which is widely used in studies of psychosocial domains where the physical axis is missing (see Dawson 1974 and Lodge 1981 for reviews). It validates magnitude estimation not with reference to a physical continuum, but in terms of self-consistency. Section 6 reports a successful validation study of this kind. Finally, we turn to straightforward replication to show that magnitude estimation results will generalize across subjects. Section 7 shows that our subjects in the validation study (§6) replicated the results reported by Sorace (§5).

4.2. AN ILLUSTRATION. This study was designed in the manner of a prepilot, an exercise to catch difficulties and reveal whether the technique or the materi-

als were worth continued study. Compared to the larger scale studies we have done, this one is something of a classroom exercise. We offer it here because if magnitude estimation comes to be used regularly by linguists, it is most likely to be applied in this relaxed fashion without large groups of subjects, major training, or complicated apparatus. The study shows how readily the technique can be applied, both to linguists and to naive adult native speakers of English. We have used a set of materials based on the items in 1 to offer further comments on our discussion in §1.2.

Although it is not necessary to allow subjects extensive practice with physical magnitude estimation, we used both line and sentence stimuli in this study. For both, the initial stimulus, the modulus, was from the middle of the range of stimuli. The 12 horizontal lines ranged in length from 2mm to 98mm. Each line was displayed horizontally in the middle of a separate page of a small booklet. The lines were used to illustrate the method of magnitude estimation, under the supposition that our subjects would find length estimation easy to understand. The linguistic materials, also displayed one per page, were the 16 sentences below in the order given. We include their original grammaticality/acceptability markings, and their page and number references in Haegeman 1991 for convenience here, though these were not offered to subjects.

(7) a. ?*Which man do you wonder when to meet?* (H:44a, p. 502)
b. *Which book would you recommend reading?* (H:41a, p. 501)
c. **When does John like the plan to steal the crown jewels?* (H:53b′)
d. ***When do you know the man whom Mary invited?* (H:31b, p. 495)
e. **With which pen do you wonder what to write?* (H:44b, p. 502)
f. **Whom do you know the date when Mary invited?* (H:31a, p. 495)
g. **When did John announce a plan to steal Bill's car?* (H:53b, p. 506)
h. **This is a book which reading would be fun.* (H:38a, p. 500)
i. ?*Which car did John announce a plan to steal tonight?* (H:53a, p. 506)
j. ?*Who did Bill buy the car to please?* (H:35a′)
k. **Where did Bill go to Rome to work?* (H:35b, p. 500)
l. ?*This is a paper that we need someone who understands.* (H:50a, p. 505)
m. ***This is a pen with which writing would be fun.* (H:38b, p. 500)
n. **This is a paper that we need someone that we can intimidate with.* (H:50b, p. 505)
o. *Who did John invite?* (H:22, p. 489)
p. **Where did Bill buy the car to drive?* (H:35b′)

Some comments on the set of materials are in order. Most of the sentences were taken directly from Haegeman. Several (7j (=35a′), 7p (=35b′), and 7c (=53b′)) were alternative lexicalizations. They were composed to help subjects arrive at the interpretation intended by Haegeman in the absence of explicit instructions as to how sentences should be construed. To invite particular interpretations in materials for psycholinguistic experiments which do not allow

stage directions for each example (see Trueswell & Tanenhaus 1991 for examples), lexical content of examples is often manipulated. Here, item 7g, Haegeman's original example 53b, allows the interpretation that *when* originates in the IP containing *announce,* though it is intended to represent a sentence in which *when* is an adjunct from the IP containing *steal.* Example 7c uses a verb in the upper IP which is harder to construe with *when.* Examples 7j and 7p are analogues of Haegeman's 35a and 35b (the latter reproduced in 7k) which seemed easier to judge as Haegeman intended, perhaps because they contain less common sequences *buy the X to please* t_i and *buy the car to drive* t_j instead of the more common *go to X to visit* t_i and *go to X to work* t_j. One of the uses of a prepilot is to check the efficacy of such alternate lexicalizations.

In all, we tested four undergraduate anatomy students, none of whom had ever judged acceptability before, and nine experienced linguists. Because so few inexperienced subjects were readily available, we selected the four most experienced linguists as a comparison group. Data from a fifth linguist are examined separately because her use of the scale was unique in this group. We tested subjects in groups in their own departments during a coffee break. We asked all subjects to assign a number to the first line to represent its length and then assign a number to each subsequent line to reflect its length relative to the first, doubling the first number if the second appeared twice as long, dividing it in three if it seemed a third as long. After they had judged the lines, we asked subjects to make analogous numerical estimates of the acceptability of the sentences, again judging each subsequent example relative to the first. We told subjects to judge acceptability of construction rather than meaning, assigning higher numbers to better sentences and lower numbers to worse. We reminded them that there is no limit to the set of positive numbers, that fractions are legitimate numbers, and that all multiples and fractions of any positive number assigned to the modulus would have to be greater than zero.[4]

Table 1 shows summary figures for each subject. Much of our argument thus far depends on the possibility that more degrees of acceptability are distinguishable than the usual symbolic scale reflects. The table shows that all the subjects used more than 4 different numbers to express their acceptability judgments. Without further validation, this limited set of data cannot be conclusive, but, compared to more restricted responses, it encourages the belief that subjects genuinely find more than four different levels of acceptability respresented in the stimuli. The table also shows that subjects used a wide range of numerical estimates, a result consistent with the view that these sentences represented very different degrees of acceptability. The ratio of the largest to the smallest estimated magnitude for a subject, what we call the max/min ratio, varied from 5 to 500.

Finally, a cursory glance at the table shows that although subjects chose quite different moduli, there is no obvious relationship between the modulus and the max/min ratio. There is always some worry that numerical estimates may be distorted by subjects' unwillingness to use large numbers or to calculate

[4] Full instructions are available from the authors on request.

SUBJECTS	NUMBER OF LEVELS EXPRESSED	MODULUS	MAX/MIN RATIO
Linguists			
A	8	4	48
B	9	10	500
C	8	1	40
D	8	9	10
E	5	3	5
Anatomists			
A	7	4	5
B	11	10	150
C	6	3	12
D	8	5	200

TABLE 1. Use of magnitude estimates of acceptability in an informal exercise by experienced linguists and inexperienced undergraduates.

estimates to several decimal places in small ones.[5] These results indicate no major effects of this kind.

Characteristics of magnitude estimates can be highlighted by comparing a subject who appeared to follow the instructions, Linguist A (Fig. 1) with the unusual subject, Linguist E, who effectively rejected magnitude estimation, using only the integers of a five-point scale, and telling us subsequently that she could not imagine acceptability being judged in any other way. Although subjects may not be the best judges of their behavior in these tasks, this one seems to have been correct. Note that both linguists put the acceptable sentences 7b and 7o at the top of their respective ranges and sentences originally marked * or ** at the bottom, so we can suppose that both were assigning larger numbers to better sentences. But note also that neither follows the implications of the original classifications on 0–?–*–** scale, for neither maintains a discrete range of numbers for all the items originally marked with a given symbol.

Several dissimilarities are clear. The first, unsurprisingly, is the number of values used: Linguist A used more different values in attempting to estimate the acceptability of the 16 stimuli than Linguist E and, accordingly, produced fewer ties. Second, Linguist A not only produced different estimates of acceptability for items directly compared in Haegeman's original presentation (for example, 7h v. 7m, 7a v. 7e, 7l v. 7n) but also produced differences of different sizes: 7h and 7m show a small difference in estimated acceptability while the

[5] It is well known that departures from linearity often occur at the lower end of the psychophysical scales where a preference for integer responses raises estimates that ought to be less than 1.00 to 1.00, since, as the informants have been told, they must not be 0. Though the integer bias may work over the whole range of responses, it produces most distortion at the lower end, where smaller numerical differences represent larger proportional differences. To avoid this problem, researchers often choose a modulus in the middle of the testable range, in the hope that it will inspire a conveniently large intial estimate. It is also customary to advise subjects to assign the modulus a number that is easy to work with in multiplication or division. Although in the last analysis subjects will do what they please, most have the sense not to start with values like 0.73.

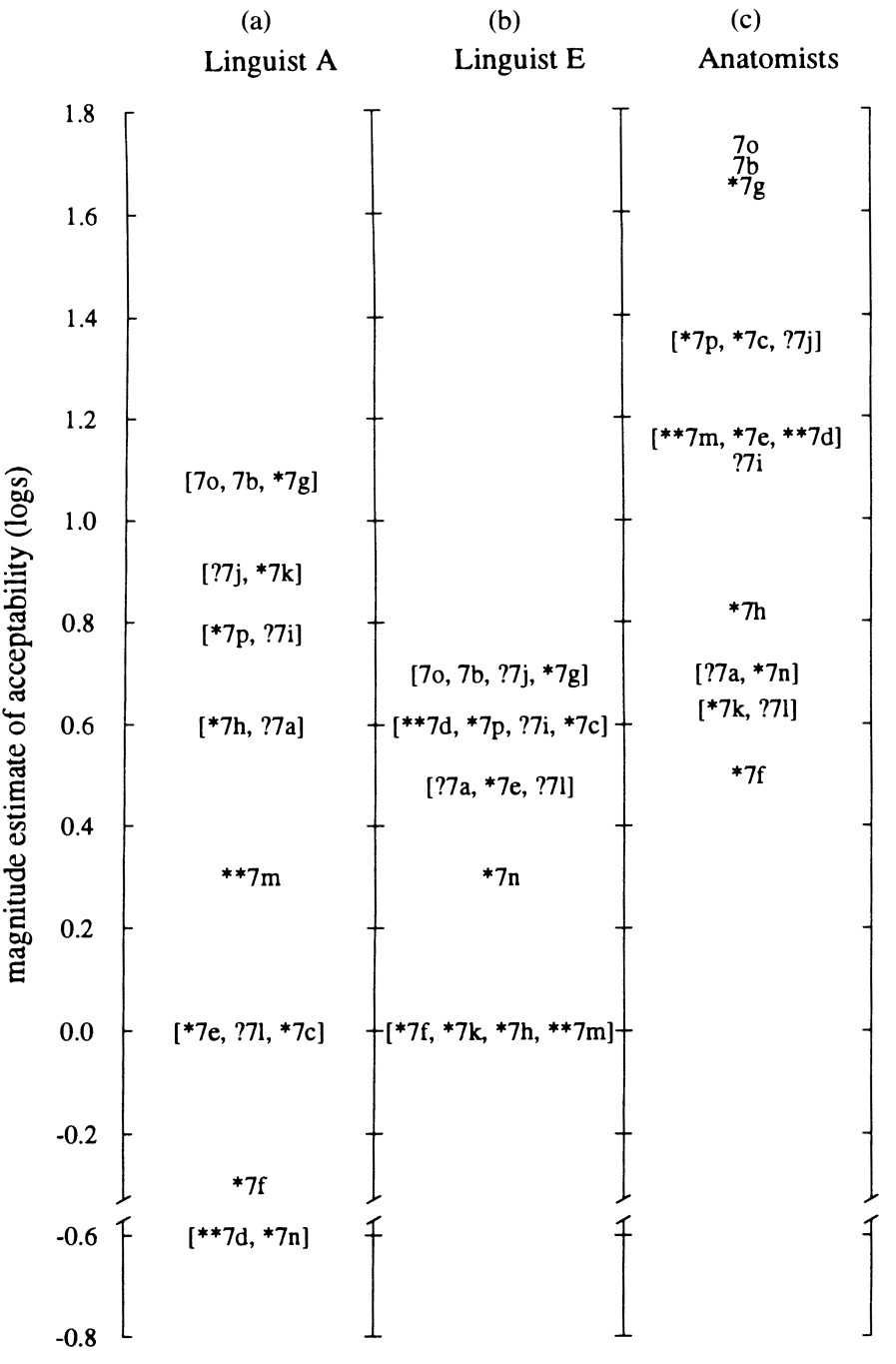
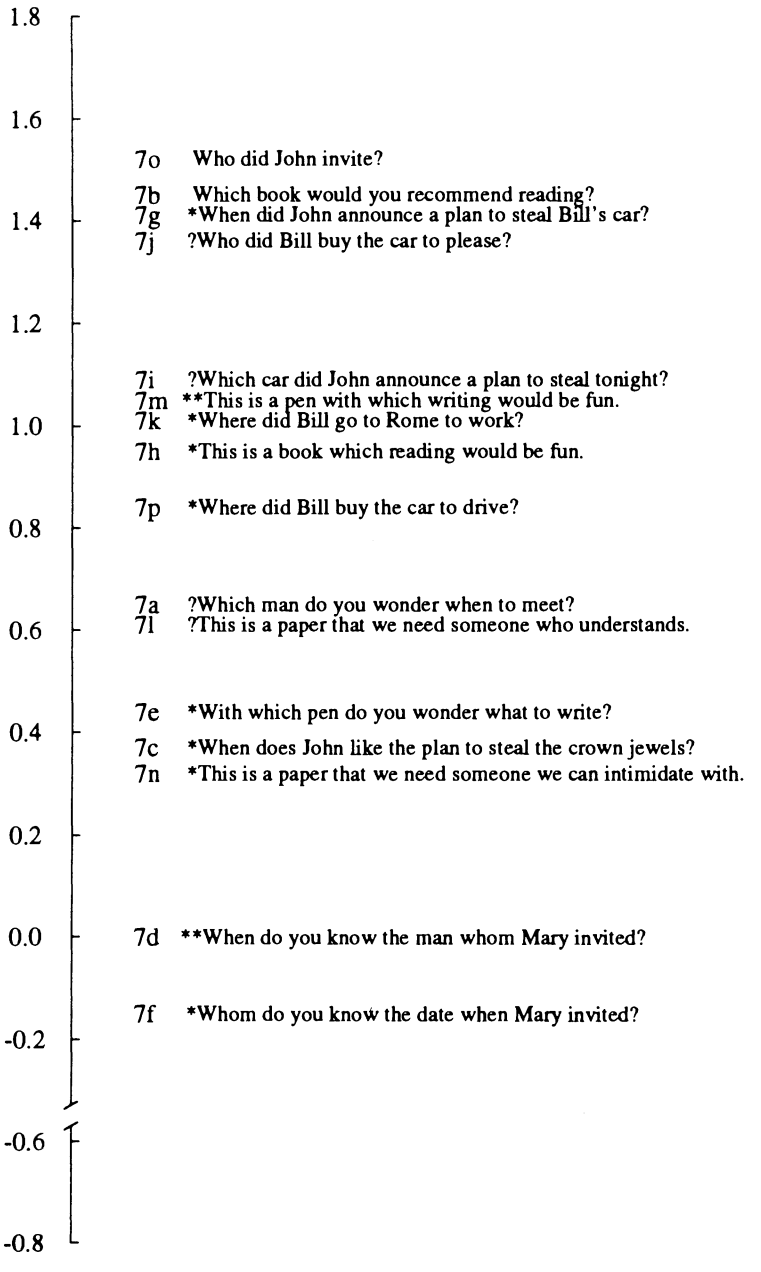


FIGURE 1. Results of an informal study showing (a) one linguist estimating magnitude; (b) one linguist using a 5-point scale; (c) averaged results for 4 naive subjects; (d) averaged results for 4 experienced linguists.

(d)

Linguists


1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.6
-0.8
7o Who did John invite?
7b Which book would you recommend reading?
7g *When did John announce a plan to steal Bill's car?
7j ?Who did Bill buy the car to please?
7i ?Which car did John announce a plan to steal tonight?
7m **This is a pen with which writing would be fun.
7k *Where did Bill go to Rome to work?
7h *This is a book which reading would be fun.
7p *Where did Bill buy the car to drive?
7a ?Which man do you wonder when to meet?
7l ?This is a paper that we need someone who understands.
7e *With which pen do you wonder what to write?
7c *When does John like the plan to steal the crown jewels?
7n *This is a paper that we need someone we can intimidate with.
7d **When do you know the man whom Mary invited?
7f *Whom do you know the date when Mary invited?

differences between 7a and 7e and between 7l and 7n are larger. If these results are reliable, Linguist A may be expressing an intuition that there is no uniform relationship between the effect of subjacency violations and the effect of ECP violations. Linguist E, on the other hand, produced the kind of pattern used to exemplify the problems with short scales. She rated 7l more acceptable than 7n, making a distinction only where Linguist A had recorded a large difference. Linguist A's equally large difference for 7a v. 7e corresponds to E's tie, as does A's smaller difference between 7h and 7m. As we suggested earlier, it is not clear whether E reports fewer perceived differences because fewer are perceptible or simply because fewer are reportable.

For other results we turn to Figures 1c and 1d, which average respectively over the estimates of Anatomists A–D and of Linguists A–D. Again both groups put the unexceptionable sentences at the top of their ranges and an unacceptable sentence at the bottom. Both fail to reserve distinct ranges for items given different symbols in the 0–?–*–** scale.

The averaged results suggest that our alternate lexicalizations may be useful. For example, the four linguists found 7i, Haegeman's 53a (an object extraction without a subjacency violation) less acceptable than 7g, (Haegeman's 53b, which violates ECP under the intended construal), but more acceptable than the alternate lexicalization of the ECP violation, 7c, which discouraged the unintended construal of the adjunct *when*. They also found 7p worse than its alternate lexicalization 7k (both committing weak subjacency and ECP violations), though this time both were less acceptable than 7j. If subjects' assessments are to be relied on, then developing suitable versions of sentence types, and even averaging over different lexicalizations may prove necessary.

For insight into the effects of experience, refer to Table 1 and Figs. 1c and 1d. We noted that none of the anatomy students balked at magnitude estimating acceptability or produced results that were radically out of line with the linguists'. Although the groups did disagree on the relative acceptability of some items, especially in the case of those (b) examples with more acceptable competing interpretations, results were substantially alike: linguists' and anatomists' geometric means, using the average of estimates over any alternate lexicalizations, show a strong positive correlation ($r = .81$, $p < .001$).

5. ROBUSTNESS AND DELICACY. What we have just described is a mere exercise. To show substantial and replicable effects, it is often necessary to invest in larger scale studies. Sorace (1992, 1993) has successfully used magnitude estimation to test her hypotheses with respect to the use and acquisition of the Italian auxiliaries *avere* 'have' and *essere* 'be' and their syntactic and semantic properties. Our purpose in citing this work here is limited in two ways. First, we still cannot plot a psychophysical function, for we still have a single axis, the judgments themselves, to examine. All we can do is test the statistical robustness of the differences among judgments for different classes of verbs. Second, we do not attempt to defend or even relate the full set of linguistic arguments to which Sorace recruits these results. Not only are those claims too broad in scope for the present paper, but making such claims without inde-

pendently establishing the reliability of the technique would bend our argument into a neat circle. Instead we cite examples from Sorace's data to show that magnitude estimation compares well with more familiar techniques in revealing delicacy of judgment and in supporting robust statistical effects.

Briefly summarized, Sorace's position is that a purely syntactic account of unaccusativity is insufficient to capture the systematic variation exhibited in the use of Italian auxiliary verbs. Instead, she suggests that the unmarked selection of *essere* with unaccusatives and of *avere* with unergatives in compound tenses is sensitive not only to a hierarchy of syntactic configurations (as assumed by the Government-Binding version of the Unaccusativity Hypothesis) but also to lexical-semantic hierarchies that subdivide the range of unaccusative and unergative verbs along gradable dimensions such as CONCRETE/ABSTRACT, DYNAMIC/STATIC, and TELIC/ATELIC, referring to the type of event denoted by the verb. These hierarchies distinguish core or prototypical types of verbs from peripheral ones, and therefore account for the well-recognized fact that some verbs are 'more unaccusative' than others, that is, they behave more naturally in particular diagnostics of unaccusativity (cf. Levin & Rappaport Hovav 1994, 1995).[6] Conversely, auxiliary selection in syntactically marked 'restructuring' constructions (Rizzi 1982, Burzio 1986) induced by certain Raising and Control verbs rests exclusively on the unaccusative syntactic configuration. Sorace predicted that the interaction between syntactic and semantic constraints would give rise to systematic variability in native speakers' linguistic intuitions, manifested in consistent and determinate acceptability judgments on core types of verbs, and variable and indeterminate judgments on peripheral types of verbs. Moreover, if the terms core and periphery have any general meaning, then learners of Italian as a foreign language should acquire the distinction starting from the core verbs. It follows from this view that advanced learners of Italian, even those who make no production errors in the language and share many intuitions with native speakers, will have their less nativelike intuitions in the periphery of the system, where native speaker judgments are most indeterminate.

The study we cite here belonged to a set of three subexperiments, each defined by materials making a pertinent linguistic contrast, distinguishing (a) unergative from unaccusative verbs by means of *ne* cliticization; (b) different lexical-semantic types of unergative and unaccusative verbs; (c) syntactically marked restructuring phenomena (optional 'transmission' of auxiliary *essere* from an embedded to a matrix verb; obligatory auxiliary change from *avere* to *essere* under Clitic-Climbing; ungrammaticality of clefting in restructured constructions). For purposes of exemplification, we will restrict ourselves to the results

[6] Sorace (1996) argues that the lexical-semantic representations identified by the hierarchies belong to a potentially universal 'semantic space'. What varies from language to language is the mapping of these representations onto positions in argument structure, which in turn determine the unaccusative or unergative syntactic status of a verb. Linking rules, which govern the assignment of lexical-semantic categories onto argument structure positions, are the main locus of cross-linguistic variation within this account.

of the unaccusative subexperiment. Here the prediction was that (a) paired unaccusatives, which have a transitive alternant, would be less unacceptable when conjugated with *avere* than unpaired unaccusatives, and (b) within the category of unpaired unaccusatives, motion verbs would be perceived as more core *essere* cases than verbs denoting the continuation or the existence of a state.

These materials were assessed via several techniques by 36 native speakers of Italian and by non-native learners of the language at various levels of proficiency.[7] The results of the experiments were largely consistent with the predictions. To a conventional level of significance, the judgments of native Italians were sensitive to lexical-semantic hierarchies of unaccusative and unergative verbs: judgments on both auxiliary selection and *ne* cliticization were more consistent and determinate for core verbs than for peripheral verbs. Among learners, auxiliary selection was acquired earlier with core verbs than with peripheral verbs.

Figure 2 indicates the kinds of discriminations the technique revealed. It shows subjects' strength of preference for the grammatical auxiliary (*essere*) over the ungrammatical (*avere*) with different subclasses of unaccusative verbs.

The first thing to notice about this graph is the dependent variable described on the vertical axis: the strength of preference for one form over another. Sorace made use of the interval scale measurement present in magnitude estimates by subtracting the log of a subject's estimate for the acceptability of the dispreferred *avere* version of a given sentence from the log of his or her estimate for the preferred *essere* version of the same sentence. Strength of preference can be assessed in this way even for sentences never juxtaposed for direct comparison. Figure 2 portrays the arithmetic mean of these between-auxiliary differences averaged over a group of subjects and a group of grammatically equivalent sentences. Logs are used both to keep the scale manageable in the presence of the very large numbers some subjects used and also to provide a straightforward way of dealing with judgments of proportions: when exponentiated, the difference between log estimates provides the ratio of the acceptability of the two versions of the sentence.

The second thing to note here is that verb subclasses are arranged along the horizontal axis with core unaccusatives, change-of-location verbs (as in 4) on the left, followed by increasingly peripheral subclasses as we move rightwards. Continuation and existence-of-state verbs are exemplified in examples 5 and 6 above, while 8 and 9 below illustrate unaccusative verbs with transitive and unergative alternants respectively.

(8) a. *Le tasse sono aumentate del 20%.*
The taxes.FEM.PL are increased.FEM.PL by 20%
b. **Le tasse hanno aumentato del 20%.*
The taxes.FEM.PL have increased.MASC.SG by 20%
c. 'Taxes have gone up by twenty percent.'

[7] See Sorace 1992 for details.

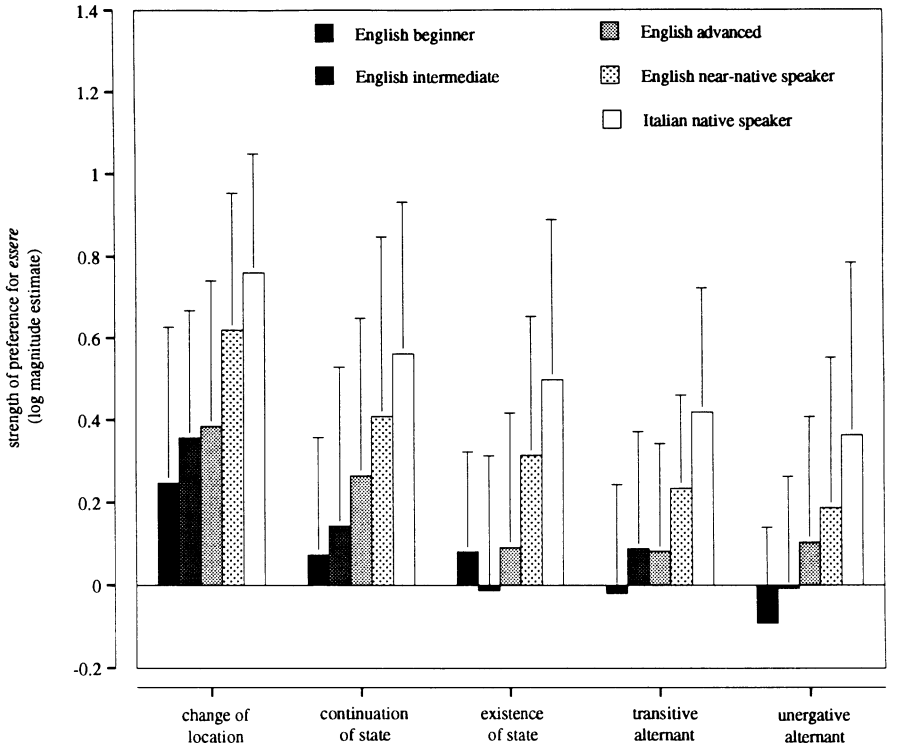


FIGURE 2. Strength of preference for grammatical over ungrammatical auxiliaries with Italian unaccusative verbs of various subclasses as expressed via magnitude estimation of linguistic acceptability by learners (beginning, intermediate, and advanced), near native speakers and native speakers of Italian (from Sorace 1992).

(9) a. *Paola è corsa in farmacia.*
Paola.FEM.SG is run.FEM.SG to pharmacy
b. **Paola ha corso in farmacia.*
Paola.FEM.SG has run.MASC.SG to pharmacy
c. 'Paola ran into the pharmacy.'

Subjects had more decisive views about core unaccusatives than about peripheral items. Preferences get significantly weaker as distance from the core increases ($p < .001$) for all classes of subjects: Spearman's ρ for native speakers of Italian is -0.363, $df = 178$; for near-native speakers -0.406, $df = 118$; for advanced, intermediate, and beginning learners -0.315, $df = 158$, -0.322, $df = 178$, -0.340, $df = 158$.[8]

[8] The reader should recall that it is not possible to space the verb categories along the abscissa of this graph according to the predictions of a linguistic theory that comments only on their order. For this reason, Sorace could not indulge in the kind of statistics psychophysicists use and which we shall attempt to return to in §6. The closest we can come to the kinds of correlations discussed elsewhere in this paper is correlate rankings of all the judgments to the rankings of all the categories.

The figures represented by the patterned columns are means, however. The spread around each, represented by error bars, can be considerable. Sorace used Analysis of Variance, which treats verb category as a nominal scale, to establish that variation between verb categories exceeded variation within them. Either considering the behavior of individual subjects or considering the responses to individual sentences, the intercategory variation is greater than intracategory to a degree that is unlikely to be the result of chance responding.

To see how the effects of verb category work in detail, we can examine the results for different groups of judges, shown with least advanced learners on the left of each group of patterned bars and native speakers on the right. As Sorace predicted, the increasing length of bars within verb categories shows that intuitions about the unaccusativity hierarchy become more determinate with increasing proficiency. In fact, by Tukey test, a sequel to ANOVA, the English beginner and intermediate groups show only a single significant difference each: between the mean for the core verb type (change of location) and the mean for the verb type hypothesized to be the most peripheral (unergative alternant). Sensitivity to the unaccusativity hierarchy is more clearly evident in the judgments of native speakers of Italian, of near-native speakers, and of advanced learners: their mean preference scores for core verbs are significantly different from their scores for each of the three most peripheral verbs (existence of state, transitive alternant, unergative alternant).

Compared to the examples in 7, which involved major differences in type and location of constituent, the sentences of the unaccusative study, like 8a and 8b, differed minimally. Yet in Sorace's studies, both native and non-native speakers not only produced significant effects, but also judged acceptability via magnitude estimation with at least as much delicacy as they did via a rank-ordering task. For instance magnitude estimation judgments—but not rank ordering responses—distinguished natives from near-natives whose speech and writing were virtually undistinguishable from natives': near-natives produced variable judgments about some sentences which elicited determinate judgments in native speakers. Even with modest differences in materials, magnitude estimation appears to be the tool of choice for distinguishing among subject groups.

6. Validation studies.

6.1. Cross-modality matching. Validation is the process of establishing that a response measure reflects what it is supposed to reflect. With no continuous measure of the stimulus to plot against subjects' impressions, we might be at a loss to determine what it is that controls magnitude estimates of acceptability (Stevens 1966). Linguistics shares this difficulty with the social sciences, which have nonetheless made good use of magnitude estimation in providing interval scale judgments of such diverse properties as prestige of occupations (Kuennapas & Wikstroem 1963, Dawson & Brinker 1971), support for political policies (Lodge et al. 1976), moral judgments (Ekman 1962), and the stressfulness of events (Zautra et al. 1986; see Lodge 1981 for a more extensive list). In each case, as in the study by Sorace discussed above, the estimates for different stimuli have proved informative in their own right.

To solve this problem, social psychologists have borrowed the CROSS-MODALITY MATCHING technique from psychophysics. In the psychophysical version of cross-modality matching, subjects use one sensory modality to estimate magnitudes presented in another. For example, brightness might be estimated by adjusting the length of a line to correspond to the perceived brightness of a light. If the subject thinks that the second light stimulus is twice as bright as the first, she or he draws a line that appears twice as long as the line drawn (however arbitrarily) to represent the brightness of the first stimulus. Two psychophysical functions contribute to the results, the function for brightness perception and the function for line-length perception. The plot of judgment (log of line length) against stimulus (log light energy) is characterized by a straight line, representing a power function with a predictable slope in log-log coordinates. If subjects are using their abilities to judge brightness and line length normally in this unusual situation, then the slope of the cross-modal function should be equal to the characteristic slope for numerical magnitude estimation of the stimulus (for example, .5 for estimated brightness of a point source) divided by the characteristic slope for numerical magnitude estimation of the response (1 for line-length estimation).[9]

The psychosocial application of cross-modality matching makes use of this regularity. When two familiar modalities are used to express judgments of dimensions which have no objective physical points of comparison (Cross 1974, Hamblin 1974, Stevens 1969, 1975, Lodge et al. 1976, Lodge 1981), the cross-modality plot of judgments against judgments will still approximate to the predicted slope, that is, to the ratio of the two psychophysical slopes, as long as subjects are able to use the modalities to estimate the new continuum consistently. Thus the appearance of the expected line in a cross-modality plot becomes a test of validity of judgments.

As usually applied, the cross-modality technique has two phases. The CALIBRATION PHASE is cross-modality matching in the psychophysical sense, with each modality used to judge stimuli in the other. The proportions holding among stimuli in each modality are the same: subjects are judging the same proportions in two ways. This phase helps familiarize subjects with the concept of proportionality, which underlies the technique of magnitude estimation, and is used

[9] The reasoning runs as follows. In the cross-modality experiment, there is an appropriate psychophysical function (i) for the stimulus modality and (ii) for the response modality:

(i) $log\ R_1 = log\ a_1 + b_1\ log\ S_1$

(ii) $log\ R_2 = log\ a_2 + b_2\ log\ S_2$

Here, R_1 and R_2 represent the responses for the two modalities and S_1 and S_2 represent the stimuli. In the case cited, the subject is attending to S_1, a bright light stimulus, and matching to it S_2, a line stimulus. In effect, S_2 is made to equal S_1, at least insofar as each will have exactly the same relationship with all the other members of their respective series. Since $S_1 = S_2$, we can derive the following equation by substitution:

(iii) $(log\ R_1 - log\ a_1)/b_1 = (log\ R_2 - log\ a_2)/b_2$

From which it follows that

(iv) $log\ R_1 = (log\ a_1 - b_1/b_2\ log\ a_2) + b_1/b_2\ log\ R_2$

Hence the slope of the cross-modal function is b_1/b_2.

to assess subjects' basic self-consistency in well-understood domains. That is, this exercise should produce a cross-modality (psycho-psychological rather than psychophysical) plot with the slope predicted from classical psychophysics. In the VALIDATION PHASE the same two modalities are used to judge a single set of nonmetric stimuli, in our case, to judge the acceptability of the same sentence types. The question is whether they act as if they are judging the same proportions in two ways. If they are, whatever slope the cross-modal plot had in the calibration phase, it should have here.

Cross-modality matching does not invent the linguistic scale of a psycholinguistic plot. On the other hand it does go some way toward confirming that, however unprincipled it seems to them, the spacing of judgments by our subjects is no matter of whim, but a reflection of intuitions on which they can draw repeatedly. Lodge 1981 gives a complete and clear account of the experimental and statistical procedures required to test the hypothesis that subjects' estimates are operating in the same way on the physical and the 'social' stimuli. Here we offer only an abbreviated report of a pair of studies following Lodge's procedures (for a full description, see Bard et al. 1994), to support our claim that the magnitude estimation technique elicits consistent expressions of opinion.

6.2. CROSS-MODALITY MATCHING OF ACCEPTABILITY JUDGMENTS.

6.2.1. METHOD. Each of our two studies used a separate group of 32 young adult native speakers of Italian, all residents of Italy and all visiting Edinburgh for a brief course in English. They included professional working people, university students and teachers. None were linguistics students and none had participated in any of the other studies described here.

CALIBRATION PHASE. To introduce the calibration phase, we demonstrated the notion of simple proportion and allowed subjects some initial practice with judging one of two physical continua. Subjects then performed two psychophysical magnitude estimation tasks that normal adults are known to execute accurately. They gave numerical magnitude estimates of the lengths of 48 horizontal lines where those lengths were distributed more or less evenly over the width of a PC screen. They also used lines to express magnitude estimates of the size of 48 numbers, matched pairwise to the line stimuli so that both represented the same set of ratios. For example, the ratio of largest to smallest stimulus was 17.5 in both sets. Because judgments in each dimension were expressed by manipulations of the other, subjects were always using their feel for both number and length. All that differed was which dimension was stimulus and which was response modality.

VALIDATION PHASE. When they had finished judging numbers and line lengths, we showed subjects that linguistic acceptability could be assessed in the same way, giving them examples of more and less acceptable sentences and allowing them to practice on sentences which varied considerably in acceptability and in source of unacceptability.[10] The first group of subjects received no explicit

[10] Copies and, where appropriate, translations of the materials will be found in Bard et al. 1994.

SUBEXPERIMENT	VARIABLES		
	VERB CATEGORY	WORD ORDER	AUXILIARY
Unergatives	+motional (*camminare* 'walk') −motional (*dormire* 'sleep') +unaccusative alternant (*correre*, 'run')	basic *ne* cliticized	*essere* *avere*
	VERB CATEGORY	WORD ORDER	AUXILIARY
Unaccusatives	+motional (*arrivare* 'arrive') continuative (*rimanere* 'stay') existential (*esistere* 'exist') +transitive alternant (*aumentare* 'increase') +unergative alternant (*volare* 'fly')	basic *ne* cliticized	*essere* *avere*
	VERB CATEGORY	WORD ORDER	FORM
Restructuring verbs	raising verbs (*dovere* 'have to') control verbs (*volere* 'want')	basic clefted +clitic climbing −clitic climbing	+restructured −restructured

TABLE 2. Designs of three subexperiments on auxiliary choice with Italian verbs: variables and levels of variables (Sorace 1992, and Robertson, Sorace, and Bard 1993).

instructions as to what numbers they should use in their numerical estimates. The second group was asked not to restrict their responses to the 10-point academic marking scale used in Italy. There were no other differences between the methods for the two studies.

The linguistic materials in this phase of the studies were 192 Italian sentences presented visually. All were drawn from the materials devised by Sorace (1992). These covered three subexperiments on factors controlling auxiliary choice in Italian: the Unaccusative subexperiment discussed in §5 above, an Unergative subexperiment, and a Restructuring subexperiment. Table 2 outlines the three subdesigns.

Each subexperiment had its own factorial design: each alternative value of each variable was combined with each value of every other, producing 48 basic item types, half using *essere*, half *avere*. Half of the 48 should be fully grammatical, the other half ungrammatical with varying levels of unacceptability. To make it possible to separate the effect of a syntactic manipulation from the effect of the particular lexical items in a sentence, 4 distinct lexicalizations were devised for each item type. For example, the 4 unergative [+motional], basic, *avere* lexicalizations were:

(10) a. *Maria ha nuotato tutti i giorni quest'estate.*
'Maria swam every day this summer.'

b. *Mia zia ha viaggiato molto da giovane.*
'My aunt traveled a lot when she was young.'

c. *Carla ha passeggiato nel parco per un'ora.*
'Carla strolled in the park for an hour.'

d. *Paola ha camminato in campagna per tre ore.*
'Paola walked in the countryside for three hours.'

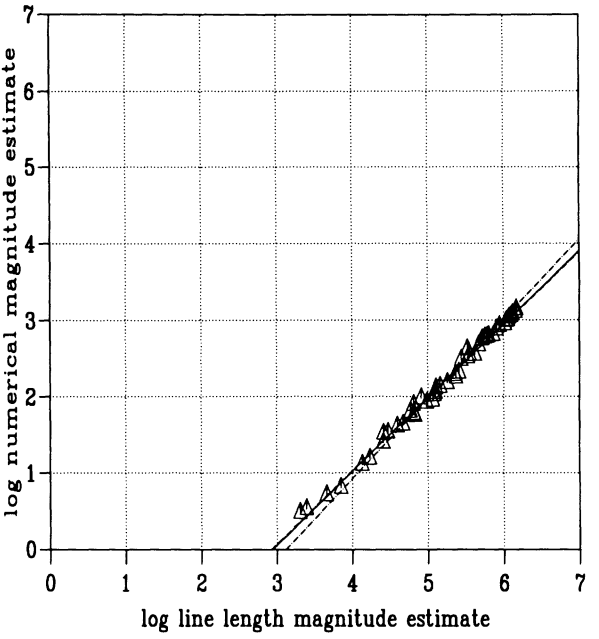


FIGURE 3. Study 1. Cross-modality plot for metric stimuli: mean numerical and mean line-length magnitude estimates of stimulus sets representing the same relative magnitudes (solid line = observed regression line, $B = 0.96$; broken lines = 95% confidence intervals for population β).

The resulting 192 sentences were divided into four groups, each containing one lexicalization from each of the 48 original types. Each subject encountered two groups of sentences, judging one via line lengths and the other via numerical estimates on their first presentation, and then reversing the combination of item and modality in a second session three or four days later. Each sentence was judged by 16 subjects per study. Each study represented all of the materials with the same counter-balancing for first combination of lexicalization and magnitude estimation technique, the order of techniques in the first session, and the ordering change between sessions. To free averaged results from order-based bias (Levelt 1972), 8 different random orders of sentences were used.

In all three subexperiments, Sorace's subjects had performed according to predictions based on the general stance set out in §5.[11] Our immediate purpose here was not to retest Sorace's hypotheses, but to allow the new subjects to

[11] In the Unergative subexperiment, subjects judged 'paired' unergatives, those with unaccusative counterparts, as more peripheral, that is less unacceptable when conjugated with *essere*, than 'unpaired' unergatives (+ or −motional). Among unpaired unergative verbs, they treated the −motional items as more core *avere* cases than + motional verbs. For Restructuring materials, native speakers were able to discriminate categorically between possible and impossible, optional and obligatory auxiliary change in restructuring sentences, although these sentences on the whole elicited lower acceptability values than those assigned to core sentences.

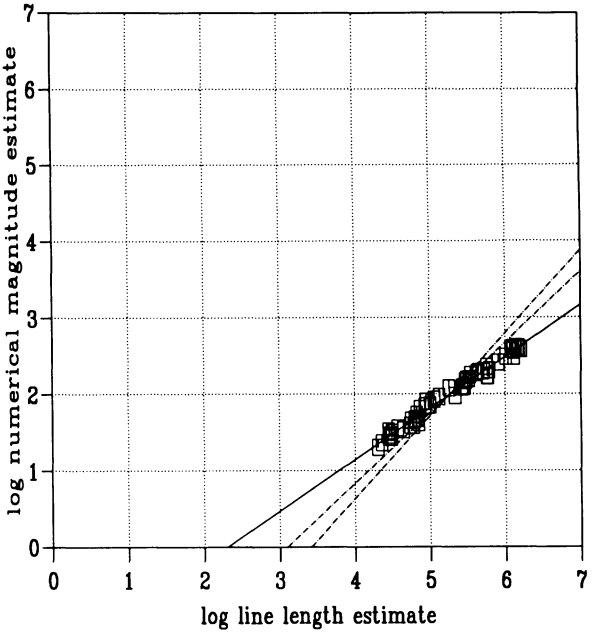


FIGURE 4. Study 2. Cross-modality plot for metric stimuli: mean numerical and mean line-length magnitude estimates of stimulus sets representing the same relative magnitudes (solid line = observed regression line, $B = 0.88$; broken lines = 95% confidence intervals for population β).

judge materials that should differ markedly in acceptability. In the validation study, therefore, we did not subdivide the linguistic materials in any way.

6.2.2. RESULTS.

CALIBRATION. Figures 3 and 4 show that our subjects, like other people who have contributed to psychophysical experiments, can estimate line length and numerical magnitude, and that exchanging response and stimulus modalities did not interfere with their ability to make such judgments. The figures show the averaged log judgments of length and numerical magnitude plotted against each other, Fig. 3 for the first group, Fig. 4 for the second. In both studies, because the correlations between line length and numerical estimates of the same ratios were effectively perfect ($r = 1.00$ in Study 1 and 0.99 in Study 2), the points cluster around a straight line.

Were our subjects good, self-consistent judges of length and number? The slopes of these lines tell us that they are quite good, but not perfectly self-consistent. Because both the psychophysical functions on which these plots are based should have slopes of 1 in the coordinates used here, the cross-modal plots in these figures should approach a regression line with a slope of 1/1 or 1. Our first group was conservative in their use of measurements, however, particularly in their use of numbers to estimate line lengths: the ratio of the largest to the smallest numerical estimate was only 14.25, though the ratio of

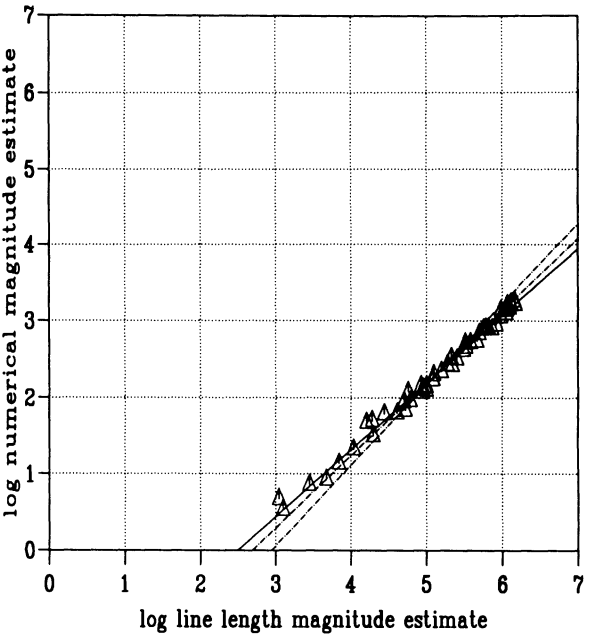


FIGURE 5. Study 1. Cross-modality plot for linguistic stimuli: mean numerical and mean line-length magnitude estimates of the acceptability of the same Italian sentences. (Each of the 48 data points represents 32 numerical and 32 line length estimates on items in a cell of the factorial linguistic design of Table 1; solid line = observed regression line, $B = 0.67$; broken lines = 95% confidence intervals for population β.)

largest to smallest true length was 17.5. Nonetheless, the cross-modal plot in Figure 3 shows a slope of 0.96, a value close to the ratio of the two actual psychophysical slopes (0.98), as it should be, and close to the usual 'theoretical' slope of 1.[12] In the second study, subjects were even less dependable on length and number. The regression line in Fig. 4 has a slope of 0.88, significantly shallower than the predicted value of 1.00, because these subjects slightly underestimated numbers via line lengths (max/min ratio 15), but overestimated lengths via numbers (max/min ratio 23). The difference between the two studies shows the kind of variation we should expect to find in ability or interpretation of a constant set of instructions.

VALIDATION. Figures 5 and 6 show the cross-modal plot for estimates of the acceptability of the same sentences expressed by the same subjects in the form of line lengths and numbers. As in Figs. 3 and 4, the figures include only averaged logs of estimates. Both cross-modal plots show nearly perfect correlations

[12] This cross-modal slope is based on an errors-in-both-variables regression model (Cross 1974, Lodge 1981) which minimizes the *perpendicular* distances from plot points to the regression line (see also Cross 1982). We use 'close to' here as an informal paraphrase of 'contains within its 95% confidence limits', that is, this is a likely outcome of sampling from a population similar to the one characterized by the theoretical value.

across modalities: subjects were self-consistent when rejudging the same sentences via different magnitude estimation modalities, just as they had been when rejudging the same physical and numerical proportions in the calibration phase. The slopes of the lines along which the judgments cluster differ between studies, however.

Were our subjects good and consistent judges of acceptability? The subjects in study 1 were not: the shallow slope (B = .67) in Fig. 5 falls far short of what theory or the calibration results predict (between 0.96 and 1.00). In study 1, where no explicit instructions were given about avoiding familiar numerical scales, subjects gave more restricted estimates of acceptability ratios when responding in numbers than when responding via lines: the ratio of the highest to the lowest geometric means is only 3.71 for numerical responses, while it is 6.68 for line responses. With only one or two exceptions, this group of subjects used numbers in the range from 2 or 3 to 10, the scale used in the Italian school system for assessment purposes. This result might have indicated an inability to make fine numerical estimates of acceptability, or it might merely be a case of defaulting to a familiar numerical scale in the absence of any suggestions to the contrary.

To discover which, we instructed subjects in study 2 not only to use appropriate numbers but also to avoid restricting their choices to the numbers from 1

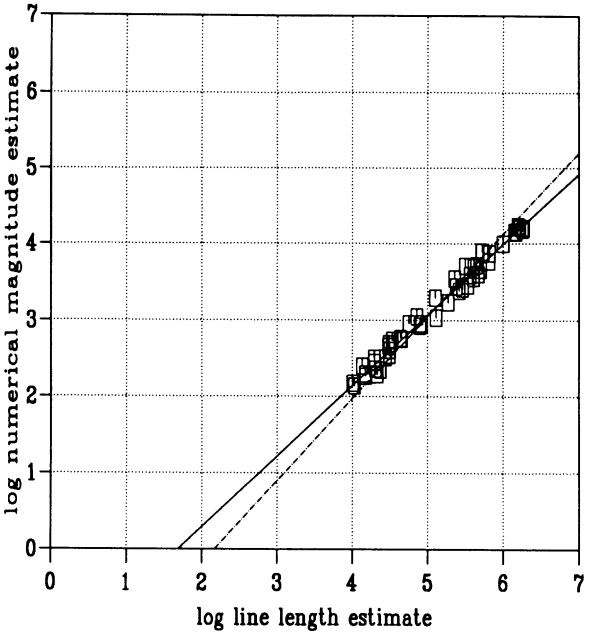


FIGURE 6. Study 2. Cross-modality plot for linguistic stimuli: mean numerical and mean line-length magnitude estimates of the acceptability of the same Italian sentences. (Each of the 48 data points represents 32 numerical and 32 line length estimates on items in a cell of the factorial linguistic design of Table 1; solid line = observed regression line, B = 0.93; broken lines = 95% confidence intervals for population β.)

to 10. Subjects now judged the acceptability of sentences consistently across modalities. Figure 6 meets the psychophysical predictions. Data points closely approximate ($r = 0.99$) a linear function with a slope ($B = 0.93$) close to the predicted slope of 1.00. The change gives every appearance of being a direct result of the change in instructions. The range of numerical estimates of acceptability increased very markedly between studies, much more so than the range of line-length judgments: the ratio of highest to lowest mean line length in study 2 was 9.9, 50% larger than the corresponding ratio in the earlier study (6.7), while the ratio for numerical responses (8.4) was more than double the ratio produced under the old instructions (3.7).

Subjects in the second study did not behave as if the instructions forced them to expand what they actually viewed as a severely restricted continuum. Had they been unable to discriminate more finely than the seven usable points of the academic scale allowed, study 2 should have been characterized by the signs of guessing: unsystematic use of numbers to reflect subjects' attempts to make impossible distinctions, poor agreement between numerical and line responses, and poor match in proportions across modalities. Figure 6 gives no such impression. Instead it displays orderly and consistent use of eightfold differences in judged acceptability.

6.2.3. DISCUSSION. Under suitable instructions even quite naive subjects appear to give self-consistent magnitude estimates of physical dimensions and of linguistic acceptability. In fact, when warned against falling back on familiar assessment scales, subjects were more consistent in maintaining their judgments of relative acceptability of sentences ($B = 0.93$) than they were at maintaining judgments of relative physical magnitude in well-trodden psychophysical domains ($B = 0.88$). Whatever subjects do when magnitude-estimating linguistic acceptability, and however odd they find the whole process at first, they clearly have this ability in their psychological repertoire, just as they have the ability to give proportionate judgments of brightness or prestige. If any of these subjective characteristics of the world were only bi-valued, the kind of results we report here would be difficult to produce.

At the same time, the artifact in the numerical judgments of study 1 reminds us that the psychophysical tool must be applied carefully. Most of us have copious experience with scales that fail to reflect our full powers of discrimination in many areas, and we succumb to their limitations without complaint. An advantage of magnitude estimation is that it gives us the freedom to express as many distinctions as we can make. It would appear that we have to be explicitly released from our habits to use this freedom. In time, and particularly with subjects whose arithmetic skills are questionable, it may be wiser to use unfamiliar judgment modalities like line length, to avoid the artifacts of our individual relationships with numerical scales.

7. RELIABILITY. For any single study to offer generalizable results, a method of accessing human judgments must be reliable not only within but across subjects. The validation studies just described demonstrate within-subject reliability. In this section, we test for consistent results between groups of subjects:

we compare the numerical estimates made by the subjects in our successful validation study, study 2, with those offered for the same sentence stimuli by the subjects in Sorace's original experiments (1992, 1993a,b).

Both groups of subjects were native speakers of Italian living in Scotland. Certain other details differed. In Sorace's study, the 36 subjects were, on average, longer term residents of the U.K. and slightly older than the present group of 32. In Sorace's study, sentences were presented individually by overhead projector, timing was controlled by the experimenter, and subjects were tested in small groups, using pencil and paper to record their responses. Numerical magnitude estimation was only one of the techniques they used. In the current studies, subjects worked individually at PCs, controlled the time they took to respond, and performed magnitude estimation in and on several modalities. Replication despite these differences will indicate that magnitude estimation results are stable over some degree of methodological variation.

To make the necessary comparison, we applied the appropriate statistical tools to answer two questions. First, we needed to know whether the different groups of subjects produced the same relative acceptability judgments for the same sentences. If they did, we would find a high positive correlation between the numerical magnitude estimates of acceptability by the two groups. Second, we needed to know whether the significant results of Sorace's study were replicated, that is whether magnitude estimation supports delicate discriminations or is just incidental but loud noise. Effects which are significant at the .05 level may, after all, occur by chance once in 20 tests on a population for which the effect is not generally true. If Sorace's results were adventitious events in an essentially random process of assigning numbers to impressions, then a replication of an original study might not produce the same results.

7.1. AGREEMENT AMONG ESTIMATES OF ACCEPTABILITY. In the test for agreement between studies, the present work was represented by the numerical magnitude estimates of acceptability from study 2.[13] To make comparisons on maximally similar conditions, only first presentations from study 2 were used. Figure 7 plots the averaged logs of estimates from study 2 against those from Sorace 1992 sentence by sentence. The plot shows significantly close agreement: $r = 0.89$, $t_{46} = 13.08$, $p < 0.001$. The two groups of subjects gave very similar estimates of the relative acceptability of stimulus sentences. As the regression line on Fig. 7 illustrates ($B = 1.09$), however, and as we might have expected from the special instructions they were given, study 2 subjects used a somewhat wider range of numerical estimates.

7.2. REPLICATION OF SIGNIFICANT EFFECTS. Although the relative locations of sentences on the acceptability scale appear to be constant across studies, it could still be the case that clear differences between the cells of Sorace's 1992, 1993 design might fail to reappear. As Table 2 shows, Sorace's study is actually

[13] Study 1 was also compared with Sorace 1992. In general, agreement was even stronger than it was with study 2, perhaps because only study 2 included a warning against limiting the range of numerical estimates.

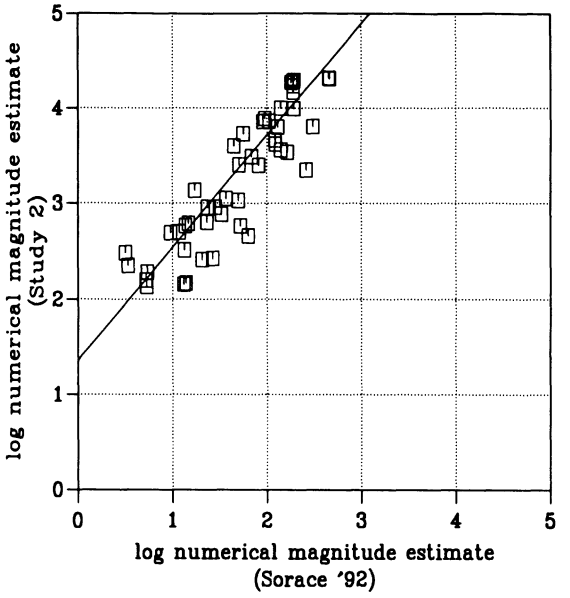


FIGURE 7. Cross-experiment plot for linguistic stimuli: mean numerical magnitude estimates of the acceptability of the same Italian sentences. (Each of the 48 data points represents 36 numerical judgments from Sorace 1992 and 32 numerical judgments from study 2, first presentation only, on the sentences in a cell of the factorial linguistic design of Table 1, $B = 1.09$.)

composed of three experiments, each of which presents all possible combinations of all the levels of several variables. Only a few of the comparisons among cells or groups of cells were critical to Sorace's theory and these were the basis for a comparison between studies.

Sorace's conclusions were supported by Analyses of Variance, which determine whether the differences among cell means outweigh background noise, the incidental differences among items in the same cell. The form of ANOVA Sorace used makes all the comparisons possible in an experimental design, even those that are not directly pertinent to the linguistic issues around which the experiment was designed.

For the subexperiments on unergative and unaccusative verbs, it was the interaction between the category of the sentence's main verb and the auxiliary used with the verb that provided the important effects on acceptability, since these interactions compared the acceptability of the preferred and the dispreferred auxiliaries with verbs of each kind from core to periphery. Because an overall observed preference for the correct auxiliary must be found before the change in auxiliary preference can be interpreted, the main effect of auxiliary is critical, too. For the restructuring verbs, it was the interaction of word order and form that should have an effect: in clitic climbing sentences, [+restructuring] versions (with *essere*) should be preferred to [−restructuring] versions (with *avere*); in cleft sentences the reverse preferance should hold, while the

other two word orders should allow either version. By examining these effects closely, we could determine how far study 2 and Sorace's study would give us different impressions of the effects of certain linguistic factors on acceptability.

To find out how far results differed between experiments, we first ran ANOVAs, by subjects and by materials, for each of the three subexperiments which included data from both Sorace's study and study 2. These analyses followed the original designs set out in Table 2 with an additional variable for experiment. Significant interactions with this variable indicated dependable differences between the outcomes of these studies. For each subexperiment we also ran separate ANOVAs with the designs in Table 2. Whenever the crucial effects, or related effects, showed an interaction with experiment, the analyses for the individual studies could be consulted to determine the nature of the difference.[14] In particular, it was important to determine whether any differences could be damaging to Sorace's conclusions, either because acceptability ordering reversed between studies or because the effect, though numerically present in both studies, was significant only in Sorace's.

The only important effect to produce a significant interaction with experiment was the auxiliary effect within unaccusative verbs (by materials, $F = 4.89$, $df = 1, 30$, $p < 0.035$; by subjects, $F = 1.99$, $df = 1, 66$, $p > .10$). The difference fell in neither damaging category, however, for both Sorace's subjects and the present group strongly preferred *essere* (for Sorace 1992: $F_2 = 232.86$, $df = 1, 15$, $p < 0.0001$; for study 2: $F_2 = 243.68$, $df = 1, 15$, $p < 0.0001$). The only difference was that Sorace's subjects showed a stronger preference than the present group. This difference is typical of the two experiments: although none of the other critical results differed significantly, study 2 generally produced less marked contrasts than Sorace's study.

8. CONCLUSIONS. The empirical work we have reported suggests that magnitude estimation can be a useful tool in the study of linguistic acceptability judgments. The technique is easy to use informally, but warrants the additional effort needed to mount full-scale experimental studies, for it delivers delicate and robust distinctions among linguistic categories. The cross-modal validation studies indicate that magnitude estimation can be applied to linguistic acceptability in much the same way as to typical psychosocial continua: its validity comes from intrasubject consistency, which was easily achieved with instructions that encourage subjects to make full use of the numerical scale in expressing their impressions. The reliability study demonstrates that the technique gives intersubject consistency as well, despite modifications of procedure.

We are unwilling to claim that magnitude estimation of linguistic acceptability is the philosopher's stone. Instead, we see it as a useful tool. Certainly this method should allow us to overcome the problems outlined in §1 of this paper.

[14] A more conventional treatment might have been to use post hoc tests on the combined data, but these tests are so stringent that, when used with the combined variance levels, they sometimes fail to reveal the significance of those effects whose replication was in question.

Measurement can now be as fine as subjects' capacities allow. With no preemptive limitation of the measurement scale, the tension between relative and absolute measurement is lost as subjects build a whole scale by means of relative judgments. Gradience of grammaticality/acceptability can be captured empirically. Estimates of acceptability can be made consistent across large sets of examples without direct pairwise comparison, as the validation and calibration studies show. Estimates of differences in acceptability and of variation of acceptability judgment can not only be calculated straightforwardly but also produce statistically significant results, as the example from Sorace's work demonstrated. The lessons of psychophysics can profit us as they have profited social scientists for several decades.

Magnitude estimation per se will not do away with the artifacts that plague judgment techniques. In fact, much of the benefit of the technique should be felt in experimental research, where it can provide scope to deploy well-known design strategies against artifacts. For example, effects of context, in particular, of order of stimulus presentation, are as well known in psychophysics as in the study of linguistic acceptability. In magnitude estimation experiments, as in nonjudgment techniques, stimuli are presented in different orders to different subjects, or to the same subjects at different points in time. Because judgments can be affected by the modulus, a different modulus may be chosen for comparison on different trials. Averaged results sample all levels of the artifact in all critical conditions. If the result of interest is larger in scale than the variation induced by the artifact, results are still visible.

Again, magnitude estimation will not change the fact that different kinds of subjects may perform differently. It does, however, give us ways of comparing their performance, as Sorace did in the study reported in §5. Here the advantage is that the technique is readily applied without much apparatus and that results may be comparable across studies. In this case, as in others, magnitude estimation is a tool for exploring acceptability judgments as well as insuring against factors that affect them.

With these tools in place, subjects' capacities can more easily become an object of study. Since it is now easy to subtract one estimate of acceptability from another, we should be able to partition acceptability judgments to study their components. As we have seen, Sorace 1992 used difference between estimates for preferred and dispreferred auxiliary to show the effects of changing auxiliary while holding the rest of the sentence constant. Analogously, one might hold all of a judged sentence constant but vary its context, comparing presentation in isolation with presentation in a short text to determine how much judged acceptability differs in the two cases. Or one might track the difference between a preferred and a dispreferred form as the two are offered in different contexts. Similar manipulations for lexicalizations, social norms, frequent usages, and pragmatic plausibility are imaginable. Experiments of these general designs have certainly been performed. What the new measure of acceptability allows is a more powerful way of integrating the results. A flexible response measure and statistical techniques like linear regression should help us to discover the major factors contributing to acceptability judg-

ments, and to elaborate theories explaining their operation (Robertson et al. 1993).

All these advantages are garnered at considerable empirical cost. How do they relate to small and imperfect exercises in professional judgment by linguists? First, as we have shown, even simple informal exercises in magnitude estimation do yield judgments which are worth pursuing, because we have reason to believe that judges will be self-consistent and will perform like other judges. To take advantage of the technique on an informal basis will not be costly: judging a dozen critical sentences three times in different orders, with a different modulus each time, and then averaging the results should take less than 15 minutes. The effect should be to permit better and more consistent distinctions in conventional cases, and to suggest new data for study. For example, the reader is invited to consider why Linguist A did not maintain a constant effect for all ECP violations.

More important among new objects of consideration is the kind of perceptual ability which underlies the formulation of acceptability judgments. We can consider two possibilities here though the data currently do not decide between them. Acceptability might be composed of psychosocial categories. It would then amount to a binary distinction, analogous to U and non-U, with mid-range judgments indicating only error of measurement. Results with this flavor are predicted if the underlying engine of acceptability is a grammar which makes a binary classification between those strings which are within the language and those which are not. It is clear that linguists do not believe that their judgments of acceptability are binary: hence ?, ??, ?*, **, etc. We have offered their conventional explanation for this fact: other factors interact with purely grammatical intuitions, lowering some judgments and raising others, and giving rise to variable mid-range scores which are the product of interacting sensitivities.

Whether or not other factors participate in acceptability judgments, acceptability might entail quite a different kind of ability, one that underlyingly resembles certain typical continuous psychophysical scales, like apparent brightness or loudness. If so, acceptability judgments should resemble such psychophysical judgments in creating a scale that is genuinely continuous, most orderly in the middle of its range, and most variable at the upper end. The fact that our results show more variance at the lower end of their range may indicate only that subjects were really judging unacceptability. Had we asked them to give the big numbers to the bad examples, the resemblance between our results and the typical sensory measures would have been more striking.

These two models for the underlying ability, the psychosocial categories and the psychophysical continuum, make different predictions about the relative robustness of mid-range and extreme judgments. The present materials do not permit the necessary comparison, for they were not designed to cover the full range of acceptability values from the grotesque to the innocuous. With the right data, we should be able test this and other hypotheses about the sources of this important linguistic behavior. On the basis of the results reported in this paper, however, we do have a tool for discovering what kind of perceptions linguistic intuitions create.

REFERENCES

BARD, ELLEN GURMAN; DAN ROBERTSON; and ANTONELLA SORACE. 1994. Magnitude estimation of linguistic acceptability. Research Paper HCRC/RP-52. Edinburgh: Human Communication Research Centre, University of Edinburgh.

BOTHA, RUDOLPH P. 1973. The justification of linguistic hypotheses: A study of non-demonstrative inference in transformational grammar. The Hague: Mouton.

BURZIO, LUIGI. 1986. Italian syntax: A government-binding approach. Dordrecht: Reidel.

CARROLL, JOHN M.; THOMAS G. BEVER; and CHAVA R. POLLACK. 1981. The nonuniqueness of linguistic intuitions. Language 57.368–81.

CHOMSKY, NOAM. 1986. Barriers. Cambridge, MA: MIT Press.

——. 1991. Some notes on economy of derivation and representation. Principles and parameters in comparative grammar, ed. by Robert Freidin, 417–54. Cambridge, MA: MIT Press.

CROSS, DAVID. 1974. Some technical notes on psychophysical scaling. Sensation and measurement: Papers in honor of S. S. Stevens, ed. by Howard R. Moskowitz, Bertram Scharf, and Joseph C. Stevens, 23–36. Dordrecht: Reidel.

——. 1982. On judgments of magnitude. Social attitudes and psychological measurement, ed. by Bernd Wegener, 73–88. Hillsdale, NJ: Erlbaum.

DAWSON, WILLIAM E. 1974. An assessment of ratio scales of opinion produced by sensory-modality matching. Sensation and measurement: Papers in honor of S. S. Stevens, ed. by Howard R. Moskowitz, Bertram Scharf, and Joseph C. Stevens, 49–59. Dordrecht: Reidel.

——, and RICHARD P. BRINKER. 1971. Validation of ratio scales of opinion by multimodality matching. Perception and Psychophysics 9.413–7.

EKMAN, GÖSTA. 1962. Measurement of moral judgment: A comparison of scaling methods. Perceptual and Motor Skills 15.3–9.

FUCCI, DONALD; LEE ELLIS; and LINDA PETROSINO. 1990. Speech clarity/intelligibility: Test-retest reliability of magnitude estimation scaling. Perceptual and Motor Skills 70.232–4.

GAITO, JOHN. 1980. Measurement scales and statistics: Resurgence of an old misconception. Psychological Bulletin 87.564–7.

GREEN, KERRY P. 1987. The perception of speaking rate using visual information from a talker's face. Perception and Psychophysics 42.587–93.

GROSJEAN, FRANÇOIS. 1977. The perception of rate in spoken and sign languages. Perception and Psychophysics 22.408–13.

——, and NORMAN J. LASS. 1977. Some factors affecting the listener's perception of reading rate in English and French. Language and Speech 20.198–208.

HAEGEMAN, LILIANE. 1991. Introduction to government and binding theory. Oxford: Blackwell.

HAMBLIN, ROBERT L. 1974. Social attitudes: Magnitude measurement and theory. Measurement in the Social Sciences, ed. by Hubert M. Blalock, 61–120. Chicago: Aldine.

KUENNAPAS, TEODOR, and INGER WIKSTROEM. 1963. Measurement of occupational preferences: A comparison of scaling methods. Perceptual and Motor Skills 17.611–24.

LABOV, WILLIAM. 1970. Some principles of linguistic methodology. Language in Society 1.97–120.

LEVELT, WILLEM. 1972. Some psychological aspects of linguistic data. Linguistische Berichte 17.18–30.

LEVIN, BETH, and MALKA RAPPAPORT HOVAV. 1994. A preliminary analysis of causative verbs in English. Lingua 92.35–77.

——, and ——. 1995. Unaccusativity: At the syntax-lexical semantic interface. Cambridge, MA: MIT Press.

LODGE, MILTON. 1981. Magnitude scaling: Quantitative measurement of opinions. Beverly Hills/London: Sage.
——; DAVID CROSS; BERNARD TURSKY; MARY-ANN FOLEY; and H. FOLEY. 1976. The calibration and cross-modal validation of ratio scales of political opinion in survey research. Social Science Research 5.325–47.
MCCARTHY, JOHN, and ALAN PRINCE. 1993. Prosodic morphology: Constraint interaction and satisfaction. Amherst: University of Massachusetts and New Brunswick, NJ: Rutgers University, ms.
MICHELL, JOEL. 1986. Measurement scales and statistics: A clash of paradigms. Psychological Bulletin 100.398–407.
——. 1990. An introduction to the logic of psychological measurement. Hillsdale, NJ: Erlbaum.
NAGATA, HIROSHI. 1987a. Change in the modulus of judgmental scale: An inadequate explanation for the repetition effect in judgments of grammaticality. Perceptual and Motor Skills 65.907–910.
——. 1987b. Long-term effect of repetition on judgments of grammaticality. Perceptual and Motor Skills 65.295–99.
——. 1988. The relativity of linguistic intuition: The effect of repetition on grammaticality judgments. Journal of Psycholinguistic Research 17.1–17.
——. 1989. Effect of repetition on grammaticality judgments under objective and subjective self-awareness conditions. Journal of Psycholinguistic Research 18.255–69.
NEWMEYER, FREDERICK W. 1983. Grammatical theory: Its limits and its possibilities. Chicago: University of Chicago Press.
PAVLOVIC, CHASLAV V.; MARIO ROSSI; and ROBERT ESPESSER. 1990. Use of the magnitude estimation technique for assessing the performance of text-to-speech synthesis systems. Journal of the Acoustical Society of America 87.373–82.
PERLMUTTER, DAVID. 1978. Impersonal passives and the unaccusative hypothesis. Berkeley Linguistics Society 4.157–89.
——. 1989. Multiattachment and the unaccusative hypothesis: The perfect auxiliary in Italian. Probus 1.63–119.
POULTON, EUSTACE C. 1986. The new psychophysics: Six models for magnitude estimation. Psychological Bulletin 69.1–19.
——. 1989. Bias in quantifying judgments. Hove: Lawrence Erlbaum.
QUIRK, RANDOLPH, and SIDNEY GREENBAUM. 1970. Elicitation experiments in English: Linguistic studies in use and attitude. Harlow: Longmans.
RIZZI, LUIGI. 1982. Issues in Italian syntax. Dordrecht: Foris.
——. 1990. Relativized minimality. Cambridge, MA: MIT Press.
ROBERTSON, DAN; ANTONELLA SORACE; and ELLEN GURMAN BARD. 1993. Magnitude estimation of linguistic acceptability as a tool for studying second language acquisition. Paper presented at the International Conference on the Psychology of Language and Communication, University of Glasgow.
SORACE, ANTONELLA. 1988. Linguistic intuitions in interlanguage development: The problem of indeterminacy. Learnability in second languages, ed. by James N. Pankhurst, Michael Sharwood Smith, and Paul van Buren, 167–90. Dordrecht: Foris.
——. 1990. Indeterminacy in first and second languages: Theoretical and methodological issues. Individualising the assessment of language abilities, ed. by John H. A. L. de Jong, and Douglas K. Stevenson, 127–53. Clevedon: Multilingual Matters.
——. 1992. Lexical conditions on syntactic knowledge: Auxiliary selection in native and non-native grammars of Italian. Edinburgh: University of Edinburgh dissertation.
——. 1993a. Incomplete vs. divergent representations of unaccusativity in non-native grammars of Italian. Second Language Research 9.22–47.
——. 1993b. Unaccusativity and auxiliary choice in non-native grammars of Italian and French: Asymmetries and predictable indeterminacy. Journal of French Language Studies 3.71–93.

——. 1995. Optimality, gradients of grammaticality, and interlanguage grammars. Paper presented at the Language Acquisition Research Symposium, Utrecht.
——. 1996. Acquiring argument structures and linking rules in a second language: The unaccusative-unergative distinction. The current state of interlanguage, ed. by Lynn Eubank, Michael Sharwood Smith, and Larry Selinker. Amsterdam: John Benjamins.
STEVENS, JOSEPH C.; JOEL D. MACK; and S. SMITH STEVENS. 1960. Growth of sensation on seven continua as measured by force of handgrip. Journal of Experimental Psychology 59.60–67.
STEVENS, S. SMITH. 1946. On the theory of scales of measurement. Science 103.667–88.
——. 1951. Mathematics, measurement, and psychophysics. Handbook of experimental psychology, ed. by S. S. Stevens, 1–49. New York: Wiley.
——. 1956. The direct estimation of sensory magnitudes—loudness. American Journal of Psychology 69.1–25.
——. 1957. On the psychophysical law. Psychological Review 64.153–81.
——. 1966. A metric for the social consensus. Science 151.530–41.
——. 1969. On predicting exponents for cross-modality matches. Perception and Psychophysics 6.251–6.
——. 1975. Psychophysics: Introduction to its perceptual, neural, and social prospects. New York: John Wiley.
TAKEFUTA, YUKIO; PETER GUBERINA; LUIGI PIZZAMIGLIO; and JOHN W. BLACK. 1986. Cross-lingual measurements of interconsonantal differences. Journal of Psycholinguistic Research 15.489–507.
TONER, MARY ANN, and FLOYD W. EMANUEL. 1989. Direct magnitude estimation and equal appearing interval scaling of vowel roughness. Journal of Speech and Hearing Research 32.78–82.
TOWNSEND, JAMES T., and F. GREGORY ASHBY. 1984. Measurement scales and statistics: The misconception misconceived. Psychological Bulletin 96.394–401.
TRUESWELL, JOHN C., and MICHAEL K. TANENHAUS. 1991. Tense, temporal context and syntactic ambiguity resolution. Language and Cognitive Processes 6.303–38.
ZAUTRA, ALEX J.; CHARLES A. GUARNACCIA; and BRUCE P. DOHRENWEND. 1986. Measuring small life events. American Journal of Community Psychology 14.629–55.

Human Communication Research Center
University of Edinburgh
2 Buccleugh Place
Edinburgh EH8 9LW
United Kingdom

[Received 25 July 1994;
revision received 12 July 1995;
accepted 31 July 1995]