

A Maximum Likelihood Procedure for the Analysis of Group and Individual Data in Aphasia Research

ELIZABETH BATES

University of California, San Diego

JANET McDONALD

Louisiana State University

BRIAN MACWHINNEY

Carnegie-Mellon University

AND

MARK APPELBAUM

Vanderbilt University

The limitations inherent in group versus case studies appear to lie in a complementary distribution, underscoring the importance of combining both strategies within a single research program. However, this compromise approach requires analytic tools that permit us to combine and evaluate individual and group data in a common format. Maximum likelihood estimation (MLE) belongs to a family of procedures for determining goodness of fit. MLE can be used in conjunction with a linear or nonlinear model of the way that sources of information combine to determine a given behavioral outcome; such models can be used to estimate

This research was supported by a PHS-NIDCD Grant R01-DC00216-07 to Elizabeth Bates. We are also grateful to Angela Friederici, Gabriele Miceli, Beverly Wulfeck, and Edgar Zurif for their assistance in collecting the data on which we based the formal models illustrated here, and to George Carnevale and Norman Anderson for advice on different aspects of the analysis. The paper is an extended version of a presentation to the Academy of Aphasia in October, 1986; selected portions of the MLE analysis are also discussed in McDonald and MacWhinney (1989). Address all correspondence to Elizabeth Bates, Department of Psychology C-0109, University of California, San Diego, La Jolla, California, 92093.

the distance between two groups, the degree to which an individual case deviates from a given empirically or theoretically defined group profile, and the degree to which one individual case resembles another. We offer a demonstration of how MLE can be used to evaluate group and individual profiles, in a cross-linguistic study of sentence comprehension in nonfluent aphasic speakers of English, Italian, and German. This includes a demonstration in which the MLE models for each language are "lesioned" to simulate several competing accounts of receptive agrammatism. © 1991 Academic Press, Inc.

An old debate in aphasia research has flared up once again in the last few years, revolving around the relative importance and legitimacy of group versus case studies. On one side, Alfonso Caramazza and his colleagues have presented a series of logical arguments against a widely practiced research strategy in neuropsychology and cognitive neuroscience, in which the results obtained with individual patients are collapsed and analyzed according to diagnostic groups, based on some *a priori* set of behavioral criteria (e.g., Caramazza, 1986). A spirited defense of research with patient groups has been provided by Caplan (1986, 1988), and by Zurif, Gardner, and Brownell (1989). These investigators acknowledge some of the problems raised by the Caramazza group, but they point out some equally serious flaws in the single-case method (see also Shallice, 1988a; Grodzinsky, 1989).

To escape from the heat and profit from the light generated by this debate, most aphasiologists and neuropsychologists would prefer to occupy some middle ground (e.g., Bub & Bub, 1988; Newcombe & Marshall, 1988; Whitaker & Slotnick, 1988). Because the limitations inherent in group versus case studies appear to lie in a complementary distribution (as discussed in more detail below), many investigators recommend combining these two strategies within a single research program. If group and case studies are compared side by side, there are sound reasons to believe that their methodological flaws will cancel out to yield a deeper truth.

This compromise approach is appealing, but it is also difficult to implement. Case and group studies typically involve different laboratory procedures, presented at radically different levels of detail, analyzed with different inferential and descriptive statistics. When we set these studies side by side, we often find ourselves in the position of comparing apples and oranges (or, perhaps, comparing forests and trees). In this paper, we will illustrate a set of analytic tools that can facilitate the comparison of group and individual data in aphasia research. With these procedures, we can cast individual and group data into a common format, and determine the extent to which individuals deviate from group results, from other individuals, and from the parameters that are predicted by different theories of aphasia.

To set the framework for this effort, we first need to consider some of

the major measurement problems associated with group and case studies, respectively.

ON PROBLEMS OF MEASUREMENT IN GROUP AND CASE STUDIES

Caramazza and his colleagues have presented their arguments against group studies in several different papers (Badecker & Caramazza, 1985, 1986; Caramazza, 1986; Caramazza & McCloskey, 1988; Miceli, Silveri, Romani, & Caramazza, 1989). Although these arguments are elegantly wrought, in unprecedented detail, they boil down (in our view) to two serious but long-recognized problems of measurement in research based on diagnostic groups.

(1) Nonrepresentativeness of the mean: When data for individual subjects are collapsed, the resulting group means may represent no one at all. No matter how many statistics we manage to gather, it remains true that no Vassar graduate has 2.5 children and 75% of a dog. In the same vein, it may be that no individual brain-damaged patient fits the syndrome profiles obtained by investigators working with group data.

(2) The importance of minority patterns: Even if the group mean does faithfully represent performance by a majority of patients, summary statistics can still hide information about extreme scores and other minority patterns that are of great theoretical importance. We know, for example, that the average annual temperature in Chicago differs little from the average annual temperature in San Diego, however, the temperature extremes encountered by residents in these two cities are very different indeed. In making decisions about where to spend a full calendar year, information about temperature extremes on rare but terrible days has greater implications for one's quality of life than information about the mean. In the same vein, information about extreme scores and/or minority patterns may be more important than information about the mean in research on the brain bases of language and other cognitive functions (e.g., patients with expressive agrammatism who show no signs of a receptive agrammatism).

As Zurif et al. (1989) point out, inferential statistics are designed to take facts like these into account. But this does not free us from the responsibility of explaining stable exceptions to the group trend. These well-known problems are all the more pressing in research on brain-damaged patients, where within-group variance may be considerably greater than within-group variance in the normal population (e.g., Miceli et al., 1989).

In their defense of group studies, Zurif et al. and Caplan (1986) go on to describe some equally serious flaws in the single-case method. Like the original criticisms by Caramazza and colleagues, these arguments are remarkable for their eloquence and detail. However, the main problems with single-case studies are well known to most aphasiologists, and can

be summarized in two parts: problems in the interpretation of idiosyncratic patterns, and measurement problems inherent in the study of individuals.

The Questionable Relevance of Idiosyncratic Findings

Unusual patterns of association and dissociation in an individual patient may occur for idiosyncratic reasons that have no direct bearing on the organization of the brain and/or the organization of a normal language processor. The most obvious examples come from the psychiatric literature, where we find a vast array of peculiar and often unique obsessions, compulsions, phobias, and/or fixations in patients who are (as far as we know) neurologically intact. Most of us would agree that strong conclusions about the architecture of brain and mind would be unwise in such cases. And yet, when unusual deficits appear in a patient with focal brain damage (e.g., a selective impairment in naming fruits and vegetables—Hart, Berndt, & Caramazza, 1985), explanations of this sort are not uncommon. We certainly are not arguing for the equivalence of psychiatric and neuropsychological symptoms. But it is worth considering why we are so often willing to assume a transparent relationship between neural architecture and symptom patterns in one case, but not in the other.

One might argue that the transparency assumption is on firmer ground when patients who were perfectly normal prior to their illness display unique symptoms following focal brain damage. And yet, might we not expect a pathological adaptation to a pathological condition, even in patients who were well adjusted in their premorbid life? A bored or frightened patient has many hours to worry and to analyze his disease, developing his own particular self-fulfilling theory of those things that are “hard” or “easy” for him; hence unusual symptom patterns could in principle reflect an adaptation (positive or negative) that is related only indirectly to the disease process itself (Kolk, 1985; Kolk & van Grunsven, 1985; Heesch, 1985). The notion of adaptation is rendered all the more plausible by increasing evidence that aphasic patients retain knowledge of linguistic structure, and (in many cases) the ability to reflect consciously on that structure (e.g., recent studies documenting preservation of grammaticality judgments in Broca’s aphasics; see especially Linebarger, Schwartz, & Saffran, 1983; Wulfeck, 1988; Shankweiler, Crain, Gorrell, & Tuller, 1989). If a patient is capable of metalinguistic reflection, then it follows that he is also capable of reflection upon his own symptom patterns. This fact forces us to recognize the further possibility that individual patients may shift their behavior in individual directions (adaptive or maladaptive) as they gain experience with the effects of their disease. We cannot assume a transparent relationship between symptom patterns and the architecture of either brain or mind.

A particularly clear example of such an adaptation comes from a recent study by Gandour, Marshall and Windsor (1989) of a patient with selective

apraxia of phonation. This patient (D.T.) demonstrated no linguistic deficits in reading, auditory comprehension, or writing, and his gestural performance was also intact. Nevertheless, as originally described by Marshall, Gandour, and Windsor (1988), D.T. produced a large number of syntactically ill-formed utterances in which the verb is shifted to the immediate right of the following constituent within a sentence. For example, "The boy hit the girl" was transformed into "Boy girl hit," "It's hard" was expressed as "Hard be," and "They will make many friends" appeared as "Many friends make." Transpositions of this type account for approximately half of the sentences produced by D.T. Because his other sentences preserve the normal syntactic structure of English, the transposition strategy cannot be viewed as an "automatic" consequence of laryngeal apraxia, nor is it possible to conclude that the patient has somehow "lost" his knowledge of or ability to apply English phrase structure rules in spontaneous speech. In an ingenious analysis of the relationship between prosody and syntax in D.T.'s speech, Gandour et al. were able to show that his deviant syntactic pattern reflects an interaction between intonational limitations and the number of syllables contained within a syntactic phrase, resulting in a coterminous relationship between tone groups and phrase length. In other words, this *syntactic* pattern constitutes a selective adaptation to the patient's *phonational* deficit. The Gandour et al. paper illustrates two points that are important for our purposes here: (1) the study of individual cases can be very useful, but (2) such cases may reflect specific and idiosyncratic adaptations that bear a very indirect relationship to the mental and neural architecture that supports language under normal conditions.

Caramazza and his colleagues have argued against the adaptation argument, on two grounds. First, they claim that no coherent theory of adaptation has yet been proposed. Second, they conclude that neuropsychology would be impossible, and nothing could be learned, if it were true that patients are capable of developing or modifying symptoms through some process of adaptation. The validity of the first criticism turns in part on one's generosity in using the term "coherent." For example, one might argue that any general theory of learning could form the basis for a specific theory of adaptation in aphasia (see, for example, Shallice, 1988b; Seidenberg, 1988; Seidenberg, McClelland, & Patterson, 1987). The second criticism is also questionable: the coherent study of change and adaptation in aphasia may be possible after all, if researchers decide to turn their energies in that direction (e.g., Basso, Capitani, & Vignolo, 1979; Basso, Capitani, Laiacona, & Luzzatti, 1980; Holland, 1986; Kolk & van Grunsven, 1985; Byng, 1988). In either case, scientific policy cannot be determined on the basis of scientific convenience; if idiosyncratic adaptations are a real phenomenon, then we have to take them into account in research and theory construction. For precisely this

reason, aphasiologists have sought patterns of brain–behavior correlation that are relatively robust, associations and dissociations that generalize across individual conditions. This brings us to the next point.

The Reliability and Validity of Single-Case Analyses

Let us assume for the moment that there is a systematic (if not transparent) relationship between mental/neural architecture and the idiosyncratic symptoms displayed by an individual patient (as in the Gandour et al. case). How do we go about evaluating the reliability and validity of this symptom profile? In fact, case studies pose a number of very serious measurement problems.

First, as Zurif et al. have shown, the two principal measurement problems encountered in group studies are also encountered in the study of single cases. To characterize the level of performance that is typical for an individual patient, we typically have to sum across different items, different trials, and often different tests. This means that we lose information about extreme scores and other minority patterns, and we may also arrive at a mean score that is not representative of the patient's "true" performance at any individual moment in language or cognitive processing (i.e., nonrepresentativeness of the mean).

Once again, inferential statistics are designed to take this kind of variation into account. However, the assumptions that underlie inferential statistics (parametric and nonparametric) are difficult to meet in single-case studies, unless we have substantial background information about every measure. This information must include **base rate performance** in the appropriate control populations (to serve as the expected values in assessments of the degree to which a individual patient deviates from normal), **reliability** (to determine the probability that a given profile could be obtained by chance), and the **underlying shape of the distribution** (to guard against spurious dissociations). In the absence of such information, we are forced to assume that our measures are perfect, and no interesting distribution of symptoms can be presumed to occur by chance, or that our measures are imperfect but so well understood that we can distinguish intuitively between "interesting" deviations and random deviations. Let us consider a few brief examples.

Base rates and expected values. Claims about single dissociations within a single patient, and double dissociations within two patients, typically rest on a nonparametric comparison of error rates on two or more tests. Such comparisons rest, in turn, on two crucial assumptions that are rarely met in single-patient studies: first, that the different errors produced by that patient are independent entities; second, that the expected distributions should be equivalent for both measures (i.e., both measures are equally difficult for the population at large). In fact, it is quite unlikely that the error produced by our patient at Moment A is independent of

the error produced at Moment B.¹ Even more important, we cannot always assume that the expected distribution is identical for both measures. For example, using a very large sample of speech error data from normal adults, Dell (1988) reports that errors in inflectional morphology (though rare) are more common than errors in derivational morphology. This fact comes to light only when the normal sample is extremely large; in a smaller sample of normal performance, we rarely observe errors of either type. Nevertheless, the "true" base rate difference revealed in a larger sample of normals may become quite important when we observe errors in inflectional versus derivational morphology in aphasic patients.

For example, Miceli and Caramazza (1988) report a case study of an aphasic patient who shows a supposed selective impairment between two aspects of lexical/grammatical morphology; this patient produces many errors of inflectional morphology (e.g., noun and verb suffixes), but virtually no errors of the derivational type (e.g., substituting "-tion" for "ment" in words like "government" or "management"). Their conclusion that this profile represents a selective and theoretically important deficit is necessarily based upon the implicit assumption that inflectional and derivational errors should be equally likely when a patient starts to make mistakes. Dell's results regarding baseline error rates in the normal population suggest that this may not be the case. Indeed, we should expect to find that inflectional morphology is selectively vulnerable under a range of different conditions. This prediction is supported in a series of cross-linguistic studies by Bates and her colleagues (for reviews, see Bates & Wulfeck, 1989a,b). In richly inflected languages with a great deal of inflectional morphology, supposed dissociations of the kind demonstrated by Miceli and Caramazza (1988) are quite common; they occur in receptive and expressive language, in several different forms of aphasia. This could occur because the two forms of morphology are handled by different processors; but it might also reflect quantitative differences between the two morpheme types within a single processor (in which, for some reason we do not yet understand, derivational morphemes are more "robust").

Reliability. Spurious single and double dissociations can also arise from differences in the internal reliability of measures. Suppose, for example, that we find two patients with the following complementary profiles: Smith scores at 80% on Task A, with a 45% score on Task B; Jones scores at 20% on Task A, with a 55% score on Task B. At face value, it seems fair to conclude from these patterns that Task A and Task B tap into mechanisms that can be doubly dissociated. However, there are alternative explanations for the same distribution. Figure 1 illustrates the "true"

¹ As one reviewer has pointed out, there are legitimate means for establishing the statistical independence of observations within a single patient, using autocorrelations borrowed from time series analysis.

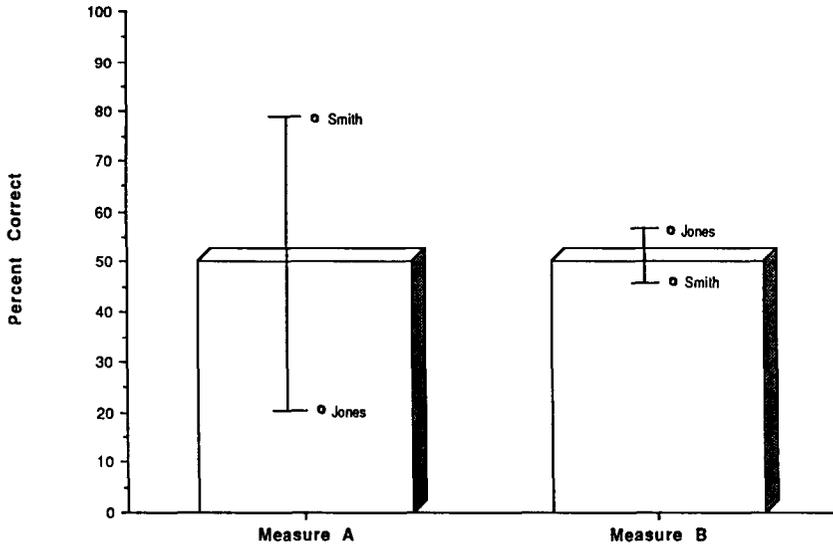


FIG. 1. Hypothetical distribution for two patients against measures with markedly different standard deviations.

distributions that underlie our two hypothetical tasks. As we can see, although the two tests have similar means, they differ greatly in their internal reliability. Task A is extremely noisy, with a huge standard deviation; indeed, a single standard deviation can span the distance from 20 to 80%. Task B is much more reliable, with a single standard deviation that spans the distance between 45 and 55%. For present purposes, our point is that the double dissociation observed in Smith and Jones could have been obtained entirely by chance; in both cases, on both measures, the patients are within one standard deviation from the mean (see Miceli et al., 1989, for some claims about double dissociation based on distributions of this sort).

Researchers committed to the single-case method might dismiss these arguments by pointing out that normals make no errors at all on Tasks A and B. Therefore the “true” mean in the population at large is 100%. However, it is possible to obtain ceiling effects on two measures as different in their internal reliability as the hypothetical tests described here. Differences in the fidelity or noise inherent in two instruments may be evident only at certain points in the distribution, e.g., in patients who are operating with global but severe resource limitations. When the entire system is subjected to stress or noise, we start to find “double dissociations” of the sort just described—patterns that may tell us nothing at all about the neurological underpinnings of cognitive processing.

Shape of the underlying distribution. Finally, Shallice (1988a,b) points

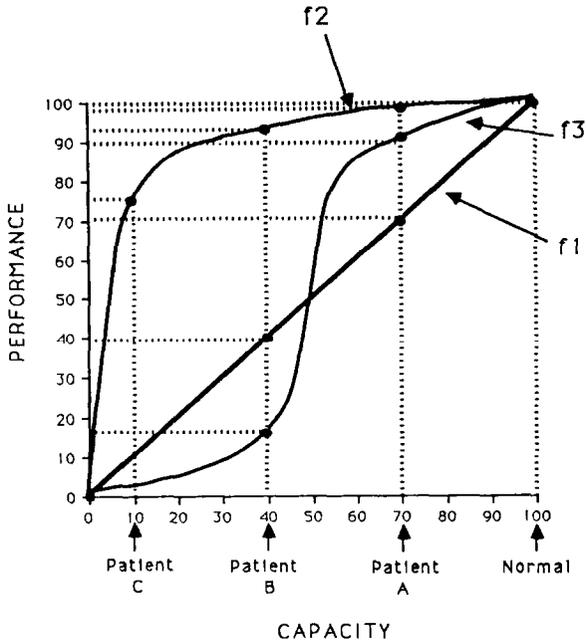


FIG. 2. Hypothetical linear and nonlinear distributions between performance and capacity.

out some related problems revolving around the assumptions of linearity and additivity that underlie so much research in this field, leading investigators to spurious conclusions about the dissociations responsible for contrasting performance profiles. To illustrate, consider the three performance functions illustrated in Fig. 2 (an adaptation and elaboration of a figure in Shallice, 1988a, p. 234). Each function represents a relationship between performance on a given measure and the amount of capacity or resources required for various levels of performance (we will use the terms “capacity” and “resource” interchangeably in this discussion). Function 1 (f1) illustrates the kind of linear performance/capacity function that we typically assume (implicitly or explicitly) in drawing conclusions about single and double dissociations. Function 2 (f2) illustrates a nonlinear function that rises sharply from baseline, levels off around 90%, and rises slowly toward asymptote after that point. Function 3 illustrates another nonlinear relationship between capacity and performance (f3), the kind of S-shaped function that is in fact very common in the literature on attention and performance. If we assume a linear function when the task in question is really governed by a nonlinear distribution, there are at least three ways that we can be led to erroneous conclusions about differential patient profiles.

First, suppose we observe three patients (A, B, and C) on a single task, and obtain the following scores: 98% correct (2% error), 93% correct (7% error), and 75% correct (25% error). We typically conclude that Patients A and B show very little impairment on the task, and that the corresponding difference between these two patients in underlying capacity is very small; by contrast, we conclude that Patient C is significantly impaired on this task, and that his underlying deficit is significantly greater than the deficits uncovered in Patients A and B (i.e., 3–10 times as severe). This assumption is warranted if there is a linear relationship between performance and severity of impairment (i.e., absolute amount of an underlying resource). However, suppose the relationship between resources and performance is multiplicative rather than additive, as illustrated by the curvilinear function (f2) in Fig. 2? This function represents a skill system that can withstand resource deficits fairly well up to a point (maintaining a level of performance above 90% despite severe losses in capacity), but falls from asymptote quite rapidly when a certain resource threshold is reached. If the underlying system follows a function of this kind, then conclusions based on absolute error rates can be quite misleading. In fact, as we can see by tracing these three performance scores (vertical axis) down to their corresponding capacity levels (horizontal axis), the amount of impairment necessary to drive error rates from 0 to 7% (i.e., a 60% loss of capacity) is actually *twice* as great as the amount of impairment required for a further increase in error rates from 7 to 25% (i.e., an additional 30% loss in capacity).

Now consider a case in which we are trying to establish a dissociation between two measures within a single patient (take Patient B). Unbeknownst to us, Task 1 follows the linear function (f1); Task 2 has a distribution closer to the nonlinear function (f2). Our patient displays a 60% error rate on Task 1, but commits errors only 7% of the time on Task 2. It is tempting to conclude that this patient suffers a selective deficit in the domain measured by Task 1. However, upon closer examination we can see that the deficits underlying this profile (i.e., absolute amounts of resource reduction) are equivalent for the two tasks (i.e., a 60% loss in capacity). In other words, there is no selective dissociation at the resource level. Presumably, the resource level is the one that is most relevant to the claim that processing mechanisms can be dissociated in brain-damaged patients.

Finally, as Shallice notes, differences in the performance/capacity functions that underlie two tasks can also create spurious double dissociations across patients, if these patients differ markedly along some global dimension of resource or capacity. Suppose we observe Patients A and B on Tasks 1 and 3. We are unaware of the fact that Task 1 follows a linear resource/performance function (f1), while Task 3 follows an S-shaped resource/performance function (f3). Patient A is functioning at a fairly

high level overall, placing him in the 70% range on our hypothetical axis of resources; Patient B is more impaired across the board, falling at the 40% resource level on both tasks. Because of the underlying difference in resource/performance functions in Task 1 and Task 3, Patient A will perform better on Task 3 (90% correct) than on Task 1 (70% correct); for the same reasons, Patient B will display the reverse pattern, performing better on Task 1 (40% correct) than on Task 3 (16% correct). It looks very much as though there is a double dissociation between the processing mechanisms or knowledge systems that underlie each task; in this case, however, the supposed dissociation is actually produced by differences along a continuous dimension of capacity.

The only way to avoid such pitfalls is to obtain exhaustive information about the reliability of our measures and the shape of their distribution: across many different individuals, in many different situations, under different levels of severity and/or stress. This means that group studies are required for the proper interpretation of individual cases, and vice versa; the two methods must go hand in hand. However, to meet this goal we need a set of measurement tools that facilitate "eyeball" comparison of group and individual patterns. In the remaining pages, we present an example of one analytic tool (maximum likelihood estimation) that permits simultaneous examination of group and individual data, in one experimental procedure (sentence comprehension). Although the example we have chosen is quite limited, the modeling procedure that we will apply belongs to a larger class of mathematical tools for determining goodness-of-fit, all of which hold considerable promise for neuropsychological research. This particular example will also permit us to do two things: (1) we can compare the outcomes obtained with the outcomes predicted by several "standard" theories of aphasia, and (2) we can illustrate in more detail how the results obtained with brain-damaged patients are affected by assuming that performance is controlled by a nonlinear function, in contrast with the linear assumptions that are implicit in many neuropsychological studies.

We begin by describing a recent cross-linguistic study of sentence comprehension in aphasia, a group study designed and analyzed in a standard analysis of variance format. Then we will introduce a procedure for maximum likelihood estimation implemented in a program called STEPIT (Chandler, 1969) and demonstrate its utility in comparing group and individual results under a range of theoretical assumptions.

SENTENCE COMPREHENSION IN NORMAL AND APHASIC SPEAKERS OF ENGLISH, ITALIAN, AND GERMAN

The results that we use to illustrate maximum likelihood estimation are taken from a cross-linguistic study by Bates, Friederici, and Wulfeck (1987), examining factors that contribute to sentence comprehension in

three groups of speakers (Broca's aphasics, Wernicke's aphasics, and normal controls), in three different languages (English, Italian, and German). In this experiment, subjects were asked to respond nonverbally to simple declarative sentences, acting them out with small props representing the animate and inanimate nouns used in each sentence. This "enactment" or "whodunit" paradigm has been used extensively with normal adults and children in a wide range of different languages (for reviews, see Bates & MacWhinney, 1987; MacWhinney & Bates, 1989). It has also been used successfully in several other studies of sentence comprehension in aphasia (e.g., Caplan & Hildebrandt, 1988; Smith & Mimica, 1984; Smith & Bates, 1987).

Bates et al. (1987) used this paradigm with a sample of 25 Broca's aphasics (8 English, 7 German, and 10 Italian), 15 Wernicke's aphasics (9 German and 6 Italian), and a series of different control groups that varied in size and composition within each language (i.e., college students versus age-matched normals in all three languages; patient and nonpatient controls in German and Italian; orthopedic patients versus neurological patients in Italian only). Because the details of patient selection are not germane to our arguments here, the reader is referred to the original paper for further information.² For present purposes, we will concentrate on the cross-linguistic sample of Broca's aphasics compared with college controls.

The materials used in this particular study consisted of 54 sentences in each language, representing orthogonal combinations of three different sources of information about sentence meaning (i.e., who did what to whom): these included three levels of word order (Noun-Verb-Noun; Noun-Noun-Verb; Verb-Noun-Noun), three levels of noun animacy (first noun animate and second inanimate; first noun inanimate and second animate; both nouns animate), and three levels of subject-verb agreement (first noun agrees with the verb in person and number; second noun agrees with the verb; both nouns agree with the verb). When all competing and converging combinations of information are represented, it is possible to

² Briefly summarized, patient groups were defined within each language on behavioral grounds. We asked neurologists and speech pathologists in each language community to provide us with patients who, in their experience working within that language, provide the best fit to relatively conservative behavioral definitions of Broca's aphasia (i.e., nonfluent speech characterized by reductions in syntactic complexity, with a tendency toward omission of grammatical elements, in patients whose comprehension appears to be intact at a clinical level) and Wernicke's aphasia (i.e., patients with moderate to severe problems in comprehension, exhibiting speech that is fluent with a well-preserved melodic line despite word finding problems and/or a tendency toward paraphasias and neologisms). The particular diagnostic tools used to derive these classifications necessarily vary from one language to another. For a complete list of exclusionary criteria that are also applied in our cross-linguistic studies, see Bates et al., 1987.

determine with some precision which cues (syntactic, semantic, and/or morphological) play the largest role in sentence interpretation. With this procedure, Bates et al. set out to study the relative strength and weakness of word order, semantics, and grammatical morphology in fluent and nonfluent aphasic patients, to determine whether and to what extent focal brain damage can affect the role that each of these sources of information plays in auditory sentence comprehension.

The cross-linguistic design was justified by results obtained in cross-linguistic studies of normal children and adults, showing that these "cue weightings" differ dramatically from one language to another, reflecting realistic differences in "cue validity" or information value in languages that reflect word order variability and/or amount of inflectional marking. To the extent that these strong cross-linguistic differences are preserved in aphasia, they constrain our interpretations of the functional lesions responsible for receptive agrammatism.

For example, normal English-speaking adults base their interpretations almost exclusively on word order: they choose the first noun in NVN strings (reflecting the high-probability Subject-Verb-Noun structure of their language), but they also have a very strong tendency to choose the second noun in VNN and NNV items (reflecting a bias toward the sentence orders VOS and OSV). These strong word order biases persist even when they are placed in competition with semantic and/or morphological factors. Given a sentence like "Are hitting the cows the pencil," most English-speaking subjects insist on choosing the second noun (i.e., the pencil) as the "one who did the action" despite contrary information from animacy and subject-verb agreement (see also Bates, McNew, MacWhinney, Devescovi, & Smith, 1982; MacWhinney, Bates, & Kliegl, 1984).

Native speakers of Italian behave quite differently: they rely on word order information to interpret simple sentences only if no other source of information is available. If unambiguous subject-verb agreement is available, they use this cue regardless of competition from word order and/or semantics. If morphology is ambiguous, they rely heavily on semantic information (e.g., choosing the cow as the subject/actor in "The pencil hits the cow"). When semantics and morphology are both ambiguous, Italians demonstrate a bias toward SVO on NVN strings, but their performance on NNV and VNN is essentially random.

Native speakers of German perform somewhere between the English and Italian extremes, although they are closer to the Italian end of the spectrum: subject-verb agreement is a very strong cue to meaning, and animacy is certainly stronger than it is in English, but word order strategies (in particular, a bias toward SVO, SOV, and VSO) put up a much stronger competition against semantics or morphology in German than in Italian.

Extending these findings to research on aphasia, Bates et al. reason as follows. On the one hand, if these ordinal differences in cue weighting

can be destroyed in aphasia, they would provide evidence in favor of the notion that grammar as a whole (word order and morphology) or closed-class elements in particular (e.g., subject-verb agreement) can be destroyed or disconnected as a result of focal brain damage. On the other hand, if fluent and/or nonfluent patients retain a hierarchy of cues to sentence interpretation that is typical for their language, then we must conclude that language-specific information is retained in these forms of aphasia. Clearly, Bates et al. base these predictions on the usual logic used in group studies: a robust pattern that holds for a well-selected group of patients can be generalized to other patients with a similar pattern of symptoms.

In fact, their findings at a group level suggest a compromise between these two extremes. Significant cross-linguistic differences were obtained even for the so-called agrammatic Broca's aphasics: English Broca's aphasics made more use of word order than Germans, who in turn made more consistent use of word order than Italian patients; Italian Broca's aphasics made significantly more use of subject-verb agreement than German patients, who in turn used agreement more extensively than their English counterparts. The expected cross-linguistic differences were also obtained in use of semantic cues (i.e., more use of semantics by Italians and Germans). These patterns suggest that language-specific information (qualitative and quantitative) is indeed preserved in Broca's aphasics. However, there were also consistent effects associated with focal brain damage in all three languages: significantly less use of subject-verb agreement in Broca's aphasics than in normal controls, in every language; slightly less use of the nonstandard VOS and OSV strategies in English Broca's, but slightly more use of word order cues in German Broca's than in normals, and no evidence of any kind for an impairment in the use of canonical word order in any language.

Putting these group findings together, Bates et al. suggest that there is cross-linguistic evidence in favor of the notion that closed-class morphemes are particularly vulnerable in aphasia—although the level of impairment depends upon the strength of morphological cues in the patient's pre-morbid language. They conclude that grammatical *knowledge* is retained in Broca's aphasia (in line with independent results on metalinguistic judgments, cited above), but patients suffer from problems of *access* during sentence comprehension that are particularly severe for closed-class morphemes.³

³ The comfort these data provide for champions of agrammatism is tempered by the fact that Wernicke's aphasics showed patterns of receptive sparing and impairment on this task that were almost identical to the patterns displayed by Broca's aphasics. Indeed, evidence for the selective vulnerability of morphology was also found in anomics, and in some nonaphasic patient controls; healthy age-matched controls, on the other hand, displayed

To the extent that these conclusions are based entirely on group data, they are subject to the same criticisms described above. How representative are these group effects? Are there patients within each language who present a true disconnection syndrome, i.e., a total loss of the ability to use grammatical information in general, and/or a complete loss of morphological cues in particular? If so, then perhaps these patients are the real agrammatics, the patients who should inform our theory of the neurological structures that subserve grammar (Grodzinsky, 1984). There might also be other individual patterns hidden behind the group statistics (c.f. Caplan & Hildebrandt, 1988), patterns that would constrain the interpretation suggested by group effects.

In their original cross-linguistic study, Bates et al. did acknowledge the importance of considering individual variation within groups, and they present individual data for some of their main effects. In no case did they find a clear reversal of the group pattern within any particular language and/or patient group. However, the complexity of their design makes it virtually impossible to present all or even a significant portion of the data for every subject. With three levels each for three within-subject variables (word order, semantics, and morphology), there are 27 cells for every single patient and control; it is simply not possible for a human observer to compare two 27-cell distributions side by side.

This situation is not unusual in neuropsychological research, particularly if the investigator is informed by a theory of normal language processing of some complexity, i.e., a model that predicts specific patterns of interaction among a series of within-subject variables. The Competition Model developed in our laboratories is one such model (Bates & MacWhinney, 1987, 1989; MacWhinney, 1987; McDonald & MacWhinney, 1989), but many other candidates come to mind (e.g., the Cohort Model—Marslen-Wilson & Tyler, 1980). What we need is a method that permits us to evaluate the entire pattern of main effects and interactions for each individual, comparing it with the pattern displayed by normal controls and by other patients. Maximum Likelihood Estimation (MLE) is one of several related procedures that can be used for this purpose, providing a single goodness-of-fit statistic that tells us at a glance the degree to which a given individual or group deviates from a complex pattern of interactions that has been determined independently on theoretical and/or empirical grounds.

the same hierarchy and strength of cues observed in college students within each language. Hence the selective vulnerability of closed-class elements in receptive language processing appears to be induced under a range of different pathological conditions—underscoring our earlier point about the importance of extensive control information in the interpretation of data from a particular individual or patient group.

Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) belongs to a family of procedures for estimating parameters of a model, a family that includes path analysis, structural equation modeling, and multiple regression (among others). When parameters of a model have been estimated through MLE (or alternative procedures), general methods for testing goodness-of-fit can be applied to evaluate the degree to which the observations that we have obtained correspond to or deviate from this model. Within the framework of a larger cross-linguistic study of normal and abnormal sentence processing (MacWhinney & Bates, 1989), we have recently made progress in the application of MLE to the sentence comprehension profiles obtained with native speakers of a range of structurally distinct language types, including normal monolingual adults (usually college students), children acquiring their language for the first time, and adult bilinguals at different levels of fluency in their second language (McDonald & MacWhinney, 1989; Sokolov, 1989; Kilborn, 1987).

The MLE implementation that we have used in our work to date utilizes a program called STEPIT (Chandler, 1969; for some applications, see Massaro & Oden, 1980; Oden & Massaro, 1978). This procedure is particularly well suited to the underlying theory and experimental design used by Bates et al., and it is the only procedure (among several which exist) that we will use in the illustrations that follow. However, the general point that we are trying to make for the analysis of group and individual data in aphasia research should generalize to other data sets, analyzed with alternative techniques for determining goodness-of-fit.

This MLE procedure assumes a situation in which we have n sources of information or *cues* available to predict a choice between two or more outcomes. For example, Massaro and his colleagues have used this approach to model the effect of different phonetic cues on phoneme perception (e.g., the conditions under which a listener hears BA versus GA). The procedure is used in conjunction with a linear (additive) or nonlinear (multiplicative) mathematical model. The equation that we will use in the simulations described below is an adaptation of the multiplicative model used by Massaro and his associates. The probability that a subject will choose the first noun on a given sentence is given

$$\% \text{ first noun choice} = \frac{\prod S_i}{[\prod S_i + \prod S_j]}$$

where S is the cue strength, i is the cue favoring the first noun, and j is the cue favoring the second noun. This formula can be paraphrased as follows:

The product of the weights for order type, verb agreement, and animacy in favor of choosing the first noun as agent, divided by that same number plus the product

of the weights for order type, verb agreement, and animacy favoring the second noun as agent.

STEPIT is applied iteratively, using this formula to locate that set of input strengths or “cue weights” that provides the best possible fit to the choice behaviors that are observed when the cues are combined in different ways. The procedure seeks one absolute weight for each cue; this value is entered into the equation to predict behavior in every cell of the design in which that cue is present. On each try, the program generates a goodness-of-fit statistic to compare the observed values with the values predicted by that particular set of cue weights, until the smallest possible deviation or discrepancy between the two distributions is obtained. If the final fit is good, then we may conclude that our model predicts the data very well. If the fit is bad, we must assume either that the model is bad (in which case, we may seek a different mathematical model, and/or a different set of predictors) or we may conclude that the data are too noisy and un-systematic to be fit by any model (in which case, we may add subjects to increase the reliability of a group mean or add data points to increase the reliability of an individual distribution).

Several different goodness-of-fit statistics are available for this purpose, each with slightly different theoretical consequences. These include **rho** (which estimates the linear correlation between the 27 observed cells and the 27 expected cells, regardless of their absolute magnitude) and **root mean square deviation** or **RMSD** (which tries to keep the absolute deviation low across all cells in the design). Both of these techniques give equal importance to every cell; hence in theory it is possible to obtain a relatively good overall fit despite very bad fits between observed and expected values in a few outlier cells. If those few cells are important for our theory, we may want to minimize outliers by choosing a higher polynomial for our goodness-of-fit statistic (e.g., root mean cubed deviation); in principle (but rarely in practice), we could also design a model in which cells that are particularly important on theoretical grounds are given greater weight. Although these alternatives are attractive, the most conservative and widely used technique for evaluating goodness-of-fit is RMSD; this is the statistic that we will adopt in the present study. RMSD can vary between zero (representing a perfect fit) and 1 (representing a maximally poor fit). Hence we are seeking models that will generate a low RMSD (i.e., the smaller the better).

Exactly what constitutes a “good” fit, and how can we decide whether one model fits our data better than another? Suppose one model fits the data with an RMSD of .2500 while another yields an RMSD of .2480. Do we really want to conclude that the second model is better? The major disadvantage of MLE, compared with traditional parametric statistics, is that it offers no probability value to assess the reliability of the difference

between two models. This is the case because the model compares absolute numbers, without taking into consideration the amount of variation that may be present around those numbers. If the numbers themselves are stable (for example, if they represent reliable group means or stable performance estimates from one patient), then the data will be more systematic and it should be possible to obtain a good fit with the "right" model; if the numbers themselves are unreliable, then it will be difficult to obtain a low RMSD with any set of weights.

On the other hand, the fact that MLE deals with absolute numbers also permits us to carry out comparisons that are not possible with parametric statistics. We can, for example, use the same goodness-of-fit statistics to test the degree to which a given individual fits the distributions obtained for one or more groups (e.g., does this bilingual speaker behave more like a modal Italian or a modal German? Does this aphasic patient look more like a Broca or a Wernicke?). We can also use the goodness-of-fit statistic to compare the abstract distance separating two or more groups (e.g., is the German distribution more similar to English or to Italian?). Furthermore, if the data set is highly reliable, we can generate a meaningful comparison of the degree to which one individual patient fits the profile produced by another single case. Finally, as we shall see below, we can use MLE to create a series of theoretically specified models (e.g., different models of receptive agrammatism) and test the fit of each patient group and each individual patient to the patterns of performance predicted by each theory.

The nonlinear (multiplicative) model that we have adopted here makes predictions that are in some cases quite different from the predictions that would apply if we assumed a linear (additive) model. On the basis of our previous work (McDonald and MacWhinney, 1989) we know that the multiplicative model generally gives a much better fit to the sentence comprehension data that we have obtained with normal English, Italian, and German speakers, compared with the results that are obtained when an additive model is applied to the same data (i.e., an equation in which we divided the *sums* of the weights of the cues for the first noun by the *sums* of the weights of the cues for both the first and second noun). However, as we shall see, this nonlinear model does yield some rather surprising results when it is applied to the data for aphasic patients.

Before we proceed, it may be useful to discuss (briefly) the different philosophy behind MLE and multiple regression, a modeling procedure which may be more familiar to readers. MLE is what we might call a "top down" procedure. An investigator specifies the model that he wants to test in advance; his choice of models is jointly determined by the experimental design (e.g., the $3 \times 3 \times 3$ design of the sentence enactment experiment) and the particular theory that he wants to test. For example, in the present study we have decided in advance to use nine parameters (three levels of word order, three levels of agreement, and three levels

of animacy) and a multiplicative formula for combining those parameters. The nine-parameter approach is relatively conservative (e.g., it allows for the possibility that agreement may have a different effect when it occurs on the first versus the second noun in a sentence). However, it is still a small enough number that the model does not become trivial (e.g., we could predict all 27 cells in the design perfectly if we had 27 parameters to work with). We could have made other assumptions (e.g., in other studies, including some using the same data set, we have tried a smaller set of parameters, and a comparison of multiplicative and additive formulae—see McDonald & MacWhinney, 1989, for a detailed discussion). For present purposes, the point is that these decisions do not emerge from the data themselves (e.g., they are not based on the ANOVA results that we have just described); instead, they are independently motivated and imposed on the data in a top down fashion. Maximum likelihood estimation is a procedure that finds the best possible set of parameter values against which this model can be tested; we are not particularly interested in the parameter values themselves, but in the extent to which a whole theoretically determined model fits this particular sample of reality. By contrast, multiple regression is typically used as a “bottom up” procedure, in which we are trying to determine *a posteriori* which parameters provide the best account for the data in question (e.g., which variables contribute significant and/or unique variance to a given outcome measure). In regression, the parameters themselves are of central interest, as opposed to the way in which those parameters combine. In principle, multiple regression could be used in the “top down” mode we have just described, but it is not particularly well suited for this purpose. First, it is not clear that regression would be very useful for the understanding of individual subject data; in a sense, the advantages of regression over MLE (i.e., availability of significance tests) are part and parcel of its disadvantages (i.e., dependence on derivation from group data with a reasonably well-behaved distribution). Second, the only way to achieve the equivalent of a multiplicative MLE model using regression would be to specify and enter all the possible multiplicative combinations of predictors as variables in the design (i.e., interaction terms like “animacy \times agreement”). This is mathematically feasible, but clumsy and difficult to interpret. Third, the least squares estimation principles of regression will not necessarily yield the same results that are obtained with MLE. In short, MLE is the method of choice for our purposes here. A more complete comparison of these techniques would certainly be useful, but such a discussion would take us far afield from our present goals.

RESULTS AND DISCUSSION

MLE Models of Normal Performance

To understand the statistics typically generated by the STEPIT program, it will be useful to start with STEPIT results for normal speakers of

English, Italian, and German. We begin by modeling the group means for young normal subjects (i.e., college students) within each language.

In these simulations, we will try to fit the 27 cells in our experimental design with a relatively conservative nine-parameter model, in which the cue weights for animacy and for agreement are estimated separately depending on the sentence position in which those cues are found. In other words, this model takes the respective Order \times Animacy and Order \times Agreement interactions into account for each language. The nine cues that are used for this simulation are the following: +NVN, +NNV, +VNN, +first noun animate, +second noun animate, +both nouns animate, +first noun agrees with verb, +second noun agrees with verb, +both nouns agree with verb (see McDonald & MacWhinney, 1989 for very similar results with a less conservative five-parameter model, in which Animacy and Agreement are given only one weight regardless of word order type).

At the group level, the goodness-of-fit obtained within each language was extremely good: an RMSD of .0580 for English, .0749 for German, and .0340 for Italian. These statistics testify to (a) the reliability of our experimental results at the group level, and (b) the excellent match between this nine-parameter multiplicative model and the actual performance observed in normal adult subjects. These figures are compatible with the view that sentence interpretation in normal adults involves a process of cue multiplication that is qualitatively and quantitatively similar to the process assumed by this particular implementation of MLE.

These results also demonstrate how MLE can be used to determine the relative distance between different languages and language types. Table 1 shows that there are systematic differences in the degree to which the three language groups map onto one another. For example, the Italian group data fit the German equation with an RMSD of .2394, compared with an RMSD of .4321 when the same data are fit to the English equation. Similarly, the German group fits the Italian equation at .2337, compared with a fit of .4983 to the English model. However, the English group data are badly fit by both the German equation (.4930) and the Italian equation (.4297). These statistics reflect the fact that German and Italian subjects behave similarly in this experiment, relying heavily on semantic and morphological cues, compared with the strong word order strategies displayed by English listeners.

Table 1 also presents the RMSD statistics obtained for individual English, German, and Italian speakers, fit to the equation for their own language and to the equations for the other two language groups. Not surprisingly, individual subjects vary in their degree of fit to the group statistics for their own language (with RMSDs between .1067 and .4801); in no case do we find a single individual who fits the group equation perfectly—a point that must be kept in mind when we interpret MLE

TABLE 1
FIT OF NORMAL SPEAKERS TO THEIR OWN AND OTHER LANGUAGES

	English	German	Italian
English normals			
1	.2136*	.5387	.4547
2	.4801	.3084	.2261*
3	.2663*	.6296	.5473
4	.2188*	.6230	.5924
5	.1979*	.5977	.5336
6	.2084*	.6590	.6052
7	.2062*	.5778	.5228
8	.2857*	.4991	.4189
9	.3658*	.4010	.3666
10	.2223*	.6890	.6337
11	.2083*	.6772	.6209
12	.3755*	.4548	.4638
Group mean	.2649*	.5547	.4988
Group fit	.0580*	.4930	.4297
German normals			
1	.5417	.2262*	.3242
2	.5410	.2111*	.2305
3	.5053	.3182	.2259*
4	.5412	.1735*	.2858
5	.5435	.1829*	.2228
6	.6538	.3829*	.5690
7	.5366	.3097	.1704*
8	.5191	.2161*	.2215
9	.5636	.1849*	.2790
10	.6738	.4196*	.6114
Group mean	.5620	.2625*	.3141
Group fit	.4983	.0749*	.2337
Italian normals			
1	.4289	.2964*	.3046
2	.4451	.2913	.2230*
3	.5297	.3180	.1447*
4	.4374	.3243	.2755*
5	.5077	.3270	.1453*
6	.5074	.2871	.1067*
7	.4207	.4296	.3286*
8	.5185	.3059	.1842*
9	.5137	.2500	.1212*
10	.4678	.3073	.1180*
Group mean	.4777	.3137	.1952*
Group fit	.4321	.2394	.0340*

Note. Twenty-eight of thirty-two = native model.

* Best fit of three.

results for individual aphasic patients. However, for 28 of the 32 subjects in this study, the equation for their own native language gives a much better fit than the equations for the other two languages. The mismatches obtained with the remaining four "misfits" may reflect nothing more than random error variance (see Seidenberg, 1988). On the other hand, it is interesting that three of the four reversals involve closely related languages (German and Italian). It is also possible that these four subjects have the kind of bilingual history that contributes to such "cross-overs" in our experiments with adult speakers of a second language (for discussions, see McDonald & MacWhinney, 1989; Kilborn & Ito, 1989).

Finally, Table 1 presents the mean RMSD scores within each language (i.e., the average RMSD, summing across individuals). Note that these RMSD averages are considerably higher than the RMSD score we obtain for the group as a whole. This is true for two reasons. First, the set of cue weights that we used to define English, German, and Italian performance were all based on efforts to fit the observed group means across all 27 cells in the design; obviously the model that we finally obtain will provide its best fit to the data from which it was derived. However, it is also true that the group means are more reliable (and hence more predictable) than the data for any individual subject. The mean RMSD scores in Table 1 are sensitive to degrees of individual variation within each group; by contrast, the RMSD score for group performance as a whole represents the central tendency. This contrast will be useful when we consider patterns of individual deviation from the group in our simulations of aphasia data.

Within this framework, we can now proceed to examine results obtained with Broca's aphasics who have presumably veered away from the normal performance profiles displayed by adult speakers in each native language.

MLE Models of Aphasic Patients

Table 2 presents the individual and group statistics for 25 Broca's aphasics (8 English, 7 German, and 10 Italian), fit to the equation for normal speakers of their own language and to the equations for normal speakers in the other two language groups. Even at the group level, it is clear that Broca's aphasics do not behave normally in this sentence interpretation task. English patients fit the English equation with an RMSD of .2393 (with individual fits ranging from .3072 to .4477). These figures are of course much higher (i.e., much worse) than the range of individual fits observed with English normals. At the group level, German patients fit the German model at a group value of .3136, with individual figures ranging from .3107 to .4840. Finally, Italian aphasics fit the Italian equation at .2463, but individual statistics range from .1517 to a high .5243.

It is equally clear, however, that the best fit is still provided by the patients' own native language. At the group level, the RMSD of .2393

TABLE 2
FIT OF APHASIC PATIENTS TO NORMALS IN THEIR OWN AND OTHER LANGUAGES

	English	German	Italian
English patients			
1	.3072*	.5915	.5829
2	.4335	.5274	.5283*
3	.3102*	.5423	.5162
4	.4443*	.4567	.4930
5	.3589*	.5233	.5080
6	.4477*	.5194	.5736
7	.3670*	.5035	.5173
8	.3649*	.5214	.5153
Group mean	.3792*	.5232	.5293
Group fit	.2393*	.4377	.4447
German patients			
1	.5436	.4199*	.4945
2	.5589	.3107*	.4608
3	.3788*	.4840	.5253
4	.5563	.4261*	.5544
5	.5515	.3345*	.4899
6	.6657	.4362*	.5245
7	.6342	.4170*	.5830
Group mean	.5556	.4041*	.4681
Group fit	.4977	.3136*	.5332
Italian patients			
1	.3906*	.4362	.4554
2	.4816	.4205	.4148*
3	.4461	.4725	.4544
4	.4499	.4816	.4306*
5	.4672	.3914	.2712*
6	.4722*	.4907	.5243
7	.4872	.2341*	.2672
8	.5241	.4461*	.5395
9	.5028	.2721	.1517*
10	.4126*	.4231	.4280
Group mean	.4634	.4068	.3937*
Group fit	.3287	.2532	.2463*

* Best fit of three.

provided by the English equation is considerably better than the RMSDs for German (.4377) or Italian (.4447). Furthermore, the English model also provided the best fit for all eight individual Broca's aphasics. In German, the group RMSD of .3136 compares favorably with the models for English (.4977) or Italian (.4681). The German model also provided the best fit of the three for 7 of the 8 individual German patients. The Italian group data are best explained by the Italian equation (.2463), but

the German equation is only slightly worse (.2532) and the English equation provides a surprisingly close fit (.3287). The spread of variation observed in the Italian group is also illustrated by the fact that the Italian equation provides the best fit for only 4 of the 10 individual Italian patients, 2 are better fit by the German model and 4 individual Italians are actually fit best by the English equation. This "slide" away from normalcy and toward other languages in Italian may reflect the fact that use of subject-verb agreement is impaired in many of these patients, bringing them closer to the relatively weak cue weights for subject-verb agreement that typify normal English performance. We will return to this point shortly, when we fit these data to different theoretical models of aphasia. Meanwhile, these data illustrate how group statistics and individual variation can be evaluated simultaneously with the MLE procedure.

As we pointed out earlier, there is no standard way of assessing the significance of a single maximum likelihood model, nor are there inferential statistics for estimating the reliability of the difference between two or more MLE models. We simply have to "eyeball" the RMSD statistics to evaluate goodness-of-fit. However, if we have enough individual cases within each group, we can return to both nonparametric and parametric techniques to evaluate the trends we have just observed. For example, as we just noted, the native-language equation provides a better fit for 18 of 25 patients (compared with 28 of 32 normals). This distribution for aphasic patients is clearly significant by a nonparametric sign test ($p < .01$), although the effect is not as reliable as the distribution observed in normals ($p < .0001$).

Assuming that the scores within any given group are normally distributed, we can also use analysis of variance to test this "native-language trend" at the group level. To illustrate, we carried out an analysis of variance using individual RMSD scores as the dependent variable, in a 3×3 design with three levels of Patient Language (English, Italian, German) as a between-subject variable and three levels of Language Model (English, Italian, German) as a within-subject variable. There was no significant effect of Language Model ($F(2, 44) = 2.40, p < .11$), but there was a significant effect of Patient Language ($F(2, 33) = 5.40, p < .02$), reflecting the fact that Italian Broca's aphasics obtain relatively low RMSD scores in all three languages (i.e., the slide toward English described above). However, the most important effect for our purposes here was a significant Patient Language \times Language Model interaction ($F(4, 44) = 13.20, p < .001$). This interaction provides a statistical confirmation of the pattern illustrated in Table 2: simply put, the English model gives a better fit for English patients, the German model gives a better fit for German patients, and the Italian model provides a better overall fit for Italian patients, despite individual deviations.

This particular use of ANOVA is of course quite different from the

analyses of variance originally used to analyze the Bates et al. experiment; it generalizes only to a rather peculiar world of derived scores (i.e., degrees of deviation from normal), and does not reflect the operation of any single factor in the design. But for supporters of the group method, it at least shows that RMSD scores preserve the stable group trends revealed by more traditional statistical techniques. At the same time, we know from our examination of individual goodness-of-fit scores that this story does not apply to every single patient. Specifically, it does not apply to 7 of the 25 patients in this study, including 6 Italians. And we also know that these patients deviate from normal even if they are closer overall to their native language. This brings us to our final set of exercises, fitting the data for aphasic patients to the outcomes predicted by contrasting theories of aphasia.

Testing Alternative Theories of Agrammatism

So far we have examined the fit of group and individual data to MLE models that are based on empirical observations. It is also possible to use MLE to evaluate the goodness-of-fit of one or more theoretically defined distributions. In this section, we will compare the goodness-of-fit that is obtained with several idealized models of aphasia. (Note: these models represent *simplifications*, for illustrative purposes, of the many subtle and often shifting models of aphasia that are currently under serious consideration in our field. Because we do not want to attribute these simplified models to any of our colleagues in aphasiology, they will be described without citations of the literature. However, the reader will surely find that many of these models have a familiar ring.)

Strong Agrammatism. In its strongest possible form, the doctrine of central agrammatism would predict that agrammatic Broca's aphasics can make absolutely no use of grammatical information during the process of sentence interpretation. These patients are able to mimic normal comprehension in context by relying exclusively on redundant pragmatic and semantic cues to sentence meaning. Using the STEPIT program, we can model the strong agrammatism hypothesis by setting the cue weights for word order and subject-verb agreement at neutral; we can then generate the distribution that would be obtained in each language if patients were relying entirely on semantic information (in this case, the contrast between animate and inanimate nouns), operating at normal levels for each language. The distribution obtained with individual aphasic patients (and with aphasic groups) can then be compared with the hypothetical strong agrammatism distribution, using RMSD scores to evaluate goodness-of-fit.

The strong agrammatism argument could be viewed as a straw man. Few aphasiologists currently accept the view that grammar is lost *in toto*,

and a number of more conservative intermediate views can be derived from the current literature. These might include the following:

Closed-Class Loss. Some investigators have argued that agrammatic Broca's aphasics suffer from a selective impairment in the use of closed-class morphology (i.e., bound inflections and grammatical function words). Stated in its strongest form (i.e., as a complete disconnection of closed-class processing from the rest of grammar and semantics), this model would predict absolutely no use of subject-verb agreement by agrammatic Broca's aphasics. However, word order and lexical semantics may continue to function at normal (and language-specific) levels. We can model this proposal by setting the cue weight for subject-verb agreement at neutral, leaving the weights for word order and animacy at the levels derived for each language in our MLE model of normal speakers. The distribution generated by this model can then be compared with the observed distribution in individual aphasic patients and in the three aphasic groups to assess how well this model fits the data. The RMSD scores obtained with this idealized Closed-Class Loss model can also be compared with the goodness-of-fit scores generated by our model of Strong Agrammatism.

Closed-Class Access. It can be argued that the disconnection assumptions behind these two models are much too strong: agrammatic Broca's aphasics rarely lose every trace of grammatical knowledge; instead, we may argue that these patients are impaired in their ability to *access* closed-class information in real time. This kind of partial impairment can be modeled by setting the weights for subject-verb agreement in each language at half their normal value. Then we can repeat the process of generating RMSD scores for each group and each individual patient, comparing these scores with the goodness-of-fit obtained with the two stronger models. Note that this 50% figure is quite arbitrary; a more realistic approach might require iterative adjustment of impairment levels for every individual (e.g., trying out a 25% cut in agreement weights, 10, 75, and so forth). Such a "hunting expedition" for appropriate levels of noise would also require a good deal of replication (to avoid capitalizing on chance). For present purposes we will restrict ourselves to simulations using the 50% figure, keeping in mind that this level was selected arbitrarily.

Normalcy through Noise. As we argued earlier, in a nonlinear system we can sometimes obtain apparently selective dissociations simply by subjecting the entire system to random levels of noise. In our original study, based on analysis of variance results, we concluded that our patients display a selective impairment in the use of subject-verb agreement—a deficit that is encountered in every language, although between-language differences in the absolute strength of morphological cues are preserved. Hence we accepted some version of the Closed-Class Access model. However, remember that analysis of variance is a linear model; a nonlinear model might therefore yield rather different results. The multiplicative

model of cue strengths used in all these simulations assumes a curvilinear relationship between behavior and degrees of impairment, like function (f_2) in Fig. 2. It is possible that the selective vulnerability of morphology described in our original paper actually reflects differences between morphology and (for example) canonical word order in the absolute cue weights displayed by normal adults. If the absolute weights for morphology are weaker than the absolute weights for canonical word order (as suggested by McDonald and MacWhinney, 1989), then morphological deficits may appear before deficits in canonical word order as a continuous function of degrees of impairment along a curvilinear scale. To test this hypothesis, we have created another model called Normalcy through Noise, in which the weights for all cues are set at 50% of their normal value for each language. (Again, the 50% figure is arbitrary, and it may eventually be necessary to experiment iteratively with different global impairment values.) Results of this model can then be compared with the idealized aphasia models described above, and with the values that would be obtained if we simply assume that these Broca's aphasics are still perfectly normal speakers of their native language (i.e., the Normal model).

Table 3 presents the results obtained with each of the four idealized aphasia models, for each individual patient and for the group means in each language, compared with the fits obtained for everyone if we simply assume that the patient is a normal native speaker of his/her language. At the group level, the Normalcy through Noise model provides an excellent fit to the data for English and Italian, much better than the other three aphasia models and better than the assumption of Normalcy (without noise). Indeed, the Noise model fits these aphasic groups almost as well as the Normalcy model fits the normal adults who furnished those numbers in the first place: an RMSD of .0989 in English, and an RMSD of .0949 in Italian. By contrast, the German data are better fit at the group level by the two closed-class models: an RMSD of .1544 for Loss, and an RMSD of .1618 for Access.

At the individual level, Normalcy through Noise provides the best fit for all 8 individual English patients. This model also provides the best fit for 4 of the 10 Italian patients. Among the other 6 individuals in the Italian group, three are best fit by simply assuming that the patient is normal; with these patients we might conclude that the 50% noise level is much too strong, and that better fits would have been obtained with either Normalcy through Noise or Closed-Class Access if we had assumed a smaller overall deficit level (e.g., a 10–20% decrease in cue strength). The remaining 3 Italians are explained best by Closed-Class Access. The 7 German Broca's aphasics break down as follows: 4 are best fit by Closed-Class Loss, 2 by Closed-Class Access, with a single Broca's aphasic (out of all 25 in the study) receiving his lowest RMSD score under the assumptions of Strong Agrammatism.

Once again, there is no significance test associated with MLE models,

TABLE 3
GOODNESS OF FIT FOR VARIOUS IDEALIZED MODELS OF PATIENT PERFORMANCE

	Competing aphasia models				Normalcy Through Noise
	Normalcy	Agrammatism	Closed-Class Loss	Closed-Class Access	
English Patients					
1	.3072	.4333	.2775	.2840	.2155*
2	.4335	.4086	.4240	.4254	.3013*
3	.3102	.4326	.3112	.3032	.2567*
4	.4443	.4121	.4492	.4467	.1805*
5	.3589	.3432	.3565	.3579	.3255*
6	.4077	.3626	.3821	.3942	.2923*
7	.3470	.3644	.3257	.3330	.2835*
8	.3649	.4444	.3734	.3672	.2517*
Group mean	.3717	.4002	.3625	.3640	.2571*
Group fit	.2393	.2794	.2256	.2283	.0989*
German patients					
1	.4199	.3910	.3392*	.3396	.3439
2	.3107	.3712	.2791	.2537*	.2840
3	.4840	.3648*	.3841	.4033	.3821
4	.4261	.3885	.3194*	.3229	.3531
5	.3345	.4221	.2906	.2677*	.3134
6	.4362	.4287	.2131*	.2527	.3677
7	.4170	.4485	.2658*	.2785	.3650
Group mean	.4040	.4021	.2988*	.3026	.3442
Group fit	.3136	.3074	.1544*	.1618	.2268
Italian patients					
1	.4554	.3550	.3181	.3076*	.3193
2	.4148	.2763	.2819	.2563*	.3122
3	.4544	.3353	.3175	.3078*	.3457
4	.4306	.3805	.3891	.3619	.3398*
5	.2712*	.4902	.4843	.3818	.3302
6	.5243	.4554	.4392	.4486	.4014*
7	.2672*	.4838	.4642	.3625	.3078
8	.5395	.4601	.4505	.4581	.4141*
9	.1517*	.5002	.5028	.3843	.2992
10	.4286	.4425	.4293	.4020	.3339*
Group mean	.3938	.4179	.4077	.3671	.3404*
Group fit	.2463	.2678	.2518	.1738	.0949*

* Best fit of all.

and it is likely that some of these differences are due to chance variation. However, to derive inferential statistics for this test of the difference between idealized aphasia models, we can subject the RMSD scores for individual patients to a Patient Language by Aphasia model analysis of variance. This analysis yields a significant main effect of Model ($F(4, 88) = 10.37, p < .001$), and a significant Language by Model interaction

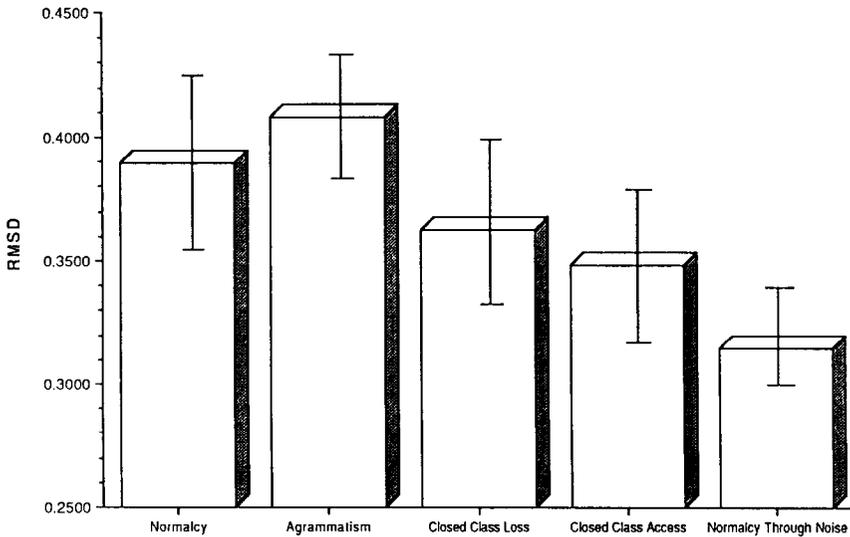


FIG. 3. Main effect for type of aphasia model.

($F(8, 88) = 3.05, p < .01$). The main effect is illustrated in Fig. 3, and the interaction is presented in Fig. 4.

We carried out a series of post hoc comparisons to explore the main effect of Aphasia models (collapsed across language type). These analyses show that the Noise model provides a significantly better fit than any of the four alternative equations. The two closed-class models are statistically identical, but both yield significantly better RMSD scores than we obtain under the assumptions of Strong Agrammatism. The Access model (but not the Loss model) also functions significantly better than the simple assumption that our patients are normal. Finally, Strong Agrammatism provides a significantly worse fit than every model except Normalcy, for which RMSD scores are lower but not reliably different.

In contrast with the main effect of Model, the Language by Model interaction shows that these rank orders vary across language groups. Post hoc analyses of the data in Fig. 4 showed that Normalcy through Noise is indeed the best model for English, differing significantly from the other four models at $p < .01$. The Agrammatism and Loss models are not reliably different from Normalcy or from one another, but the predictions offered by the Access model are slightly better than the fit obtained by assuming that patients are normal English speakers ($p < .05$).

In German, the picture is rather different. Loss, Access, and Noise are all reliably better than Normalcy; only the Agrammatism model fails to improve on the fit obtained by assuming that the patient is unimpaired. However, the four aphasia models break down differently in German:

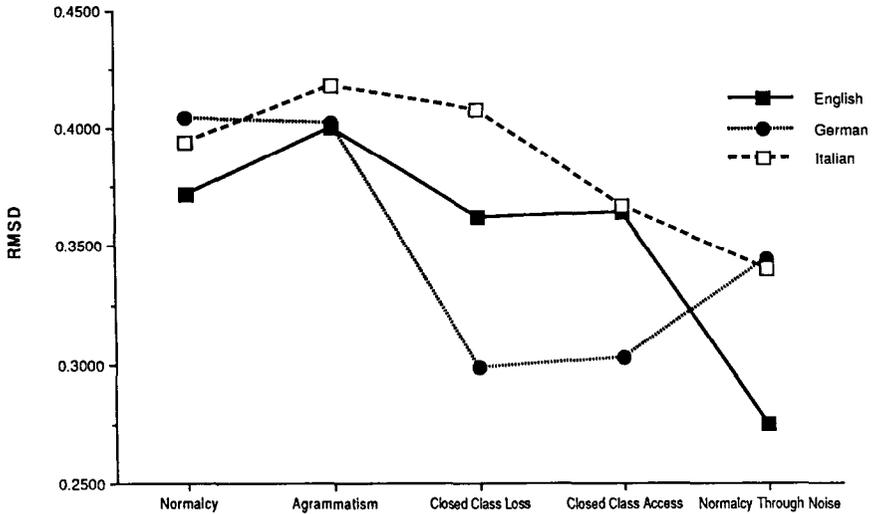


FIG. 4. Patient language by aphasia model interaction.

Loss, Access, and Noise are all significantly better than Agrammatism ($p < .02$), but the two closed-class models are not statistically different, and neither of these models differs reliably from Normalcy through Noise. However, there is a trend in German favoring both the closed-class models over assuming that patients are experiencing global degradation ($p < .10$).

Finally, in the post hoc analyses for Italian none of the individual aphasia models produced reliably better RMSD scores than we obtain by assuming that the patient is normal. However, the Agrammatism model provided significantly worse fits than each of the other three impairment types. Closed-Class Access and Normalcy through Noise both provide better fits than Closed-Class Loss ($p < .05$), but Access and Noise do not differ significantly from one another ($p < .11$, reflecting a slight but nonsignificant advantage for Normalcy through Noise). Note that these results are based on **mean RMSD**, which takes individual variation into account; this contrasts with the relatively good fit that we obtained for Normalcy through Noise when a **single RMSD** was generated for the Italian group average. The Italian group average is a composite, mixing patients who are indistinguishable from normal and patients who have selective impairments in use of the closed class. This difference between MLE and traditional ANOVA procedures testifies to its sensitivity as a tool for reconciling group and individual data in aphasia research.

CONCLUSION

We have now looked at the same data set from several different perspectives: standard analysis of variance (in Bates et al.), individual and

group fits to a variety of empirically and theoretically derived models, and a rather nonstandard application of ANOVA to goodness-of-fit scores for individual aphasics within language groups, under different assumptions of sparing and impairment. The MLE analyses do not contradict the standard ANOVA results reported by Bates et al.: we still see evidence for preservation of cross-linguistic differences in aphasia, and we also see a certain amount of evidence in favor of the conclusion that morphology is selectively vulnerable. But these conclusions are tempered considerably when we are able to examine group and individual data in the same format.

First, the MLE approach has given us much more information about the range of individual differences observed in this experiment, particularly when individual profiles are tested against different models of aphasia. Some patients are indistinguishable from normal in this procedure, and still others look as though a normal processor has been subjected to random noise. We see evidence in favor of Strong Agrammatism in only one case, in one language. A detailed follow-up of this particular patient might well prove interesting, if this Strong Agrammatism profile were displayed reliably in a range of language tasks. However, given the fact that some normals also stray a considerable distance from their group mean, it is probably safer to ascribe this particular case to error variance. In short, there is little support for the assumption that Broca's aphasics experience a complete loss of grammatical information in receptive language processing. However, we did find at least nine cases in which some kind of selective closed-class impairment provides a better fit to the data. We will believe that the Closed-Class theory of agrammatism is viable, for at least some aphasic patients. But these cases may only constitute a small subset of the patients who participated in this cross-linguistic experiment.

Finally, we must also recognize that there are systematic differences among languages in the degree of fit observed for different models of grammatical impairment in aphasia. For all of the eight English patients, we obtain a better fit to the data simply by assuming normal processing under random noise. Evidence in favor of closed-class impairment was obtained only with German and Italian patients, i.e., patients who are or were native speakers of a richly inflected language. Furthermore, impairments in the use of subject-verb agreement seem to be greater for the German Broca's aphasics (four of whom are best fit by assuming closed-class loss); greater sparing of agreement was observed in our Italian patients. How can we explain this interaction between Language and Models?

The answer may lie in the curvilinear shape of the performance/deficit curve within a multiplicative model. The absolute strength of agreement morphology is greatest in Italian, followed by German, with relatively

low cue weights for agreement in English normals. Because these normal weights lie at different points on the curvilinear function, the "same" deficit (e.g., a 50% loss in the strength of agreement) has quantitatively and qualitatively different consequences in each language. Agreement morphology is so weak in English normals that patients will necessarily differ very little from normal in the use of this cue even when its weight is cut in half (hence the good fit to normalcy-through-noise). Agreement morphology is so strong in Italian that a cut in cue strength still leaves some patients close to the normal end of the performance curve. However, for other Italian patients, cuts in the cue weights for morphology push them beyond the "cliff" illustrated by function (2) in Fig. 2, giving them a better fit to the closed-class access model. The German patients are somewhere in between: agreement morphology is strong enough among normals to yield noteworthy deficits when processing is disrupted (e.g., compared with English), but the same cue is still not as resistant in German as it is in Italian. Hence more German patients tend to fall at or below the 50% performance level simulated in our models, resulting in a better fit to Access and/or Loss.⁴

Our point is that there may be a nonlinear relationship between observed symptom patterns and underlying impairment, so that the "same" deficit has different effects depending on the language in question, and on the way that individual patients were "calibrated" prior to their illness. This is an interesting but very troubling result, because it casts doubt on the conclusions reached in many studies analyzed with traditional linear statistics. In particular, it brings many of our claims about qualitative dissociations and double dissociations into question (see also Shallice, 1988a,b; Seidenberg et al., 1987). However, if that is the nature of the reality we face, we have no choice but to explore the use of nonlinear models in research on language breakdown in aphasia.

Returning to the case-versus-group controversy, we believe that goodness-of-fit procedures of the kind illustrated here have something to offer to both sides of the debate.

For proponents of the group approach, we have shown that MLE can be used to carry out analyses of variance at a very abstract level, testing alternative models of aphasia across a set of single goodness-of-fit numbers that represent each patient's performance within a complex design. We

⁴ A recent cross-linguistic study of German- and English-speaking college students by Kilborn (1990) provides additional support for our interpretation of the "normalcy through noise" simulation. Subjects carried out an "on-line" version of our sentence interpretation task under normal acoustic conditions, or under a partial perceptual mask (i.e., "pink noise"). The noise mask had no noticeable effect on the use of word order or animacy information, but it had a selective impact on the ability of normal German listeners to use agreement cues (since agreement effects are extremely small in English even under normal conditions, there was no effect of noise on use of agreement in that language).

can also use procedures of this kind to assess the relative distance between groups in an abstract multidimensional space. Finally, MLE techniques permit us to at least begin exploring the world beyond linear statistics and linear effects. Procedures with this kind of flexibility may help us to close the gap between theory and evidence in aphasia research.

For proponents of a single-case approach, we have shown how MLE can be used to identify patients with interesting profiles for further study. And within a single study, we can evaluate with some precision the degree to which a given case deviates from patterns that we might have expected on theoretical or empirical grounds. Other possibilities that we have not pursued here include analyses in which the data for one individual patient are compared with the data obtained in a different single-case study.

Above all, we have tried to underscore the value of techniques that facilitate the joint evaluation of individual and group data. The strong experimental design (and strong experimental effects) that we have modeled here were chosen because they are particularly well suited to an illustration of how MLE and related procedures can be used. However, alternative model-fitting procedures could be used with more heterogeneous data sets (e.g., comparing profiles across the subscales of a good standard aphasia test, or multiple indices taken from a free speech sample). The success of a model-fitting enterprise is limited primarily by the validity of our theory and the reliability of our data—constraints that hold in every serious research enterprise. In our view, the progress of our science is best served by combining every reliable source of information at our disposal. To cite C. S. Peirce (1933, p. 56, judiciously cited by Zurif et al., p. 1), “We should block no path to inquiry.” MLE and related modeling procedures may help to light the road.

REFERENCES

- Badecker, W., & Caramazza, A. 1985. On considerations of method and theory governing the use of clinical categories in neurolinguistics and cognitive neuropsychology: The case against agrammatism. *Cognition*, **20**, 97–125.
- Badecker, W., & Caramazza, A. 1986. A final brief in the case against agrammatism: The role of theory in the selection of data. *Cognition*, **24**, 277–282.
- Basso, A., Capitani, E., Laiacona, M., & Luzzatti, C. 1980. Sex differences in recovery from aphasia. *Cortex*, **16**, 631–636.
- Basso, A., Capitani, E., & Vignolo, L. 1979. Influence of rehabilitation on language skills in aphasic patients: A controlled study. *Archives of Neurology*, **36**, 190–196.
- Bates, E., Friederici, A., & Wulfeck, B. 1987. Comprehension in aphasia: A crosslinguistic study. *Brain and Language*, **32**, 19–67.
- Bates, E., McNew, S., MacWhinney, B., Devescovi, A., & Smith, S. 1982. Functional constraints on sentence processing: A cross-linguistic study. *Cognition*, **11**, 245–299.
- Bates, E., & MacWhinney, B. 1987. Competition, variation and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum.
- Bates, E., & MacWhinney, B. 1989. Functionalism and the competition model. In B. MacWhinney & E. Bates (Eds.), *The crosslinguistic study of sentence processing*. New York: Cambridge Univ. Press.

- Bates, E., & Wulfeck, B. 1989a. Comparative aphasiology: A cross-linguistic approach to language breakdown. *Aphasiology*, **3**(2), 111–142.
- Bates, E., & Wulfeck, B. 1989b. Crosslinguistic studies of aphasia. In B. MacWhinney & E. Bates (Eds.), *The crosslinguistic study of sentence processing*. New York: Cambridge Univ. Press.
- Bub, J., & Bub, D. 1988. On the methodology of single-case studies in cognitive neuropsychology. *Cognitive Neuropsychology*, **5**(5), 565–582.
- Byng, S. 1988. Sentence processing deficits: Theory and therapy. *Cognitive Neuropsychology*, **5**(6), 629–676.
- Caplan, D. 1986. In defense of agrammatism. *Cognition*, **24**, 19–67.
- Caplan, D. 1988. On the role of group studies in neuropsychological and pathopsychological research. *Cognitive Neuropsychology*, **5**, 565–582.
- Caplan, D., & Hildebrandt, N. 1988. *Disorders of syntactic comprehension*. Cambridge, MA: MIT Press.
- Caramazza, A. 1986. On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition*, **5**, 41–66.
- Caramazza, A., & McCloskey, M. (in press). The case for single-patient studies. *Cognitive Neuropsychology*.
- Chandler, J. 1969. Subroutine STEPIT finds local minima of a smooth function of several parameters. *Behavioral Science*, **14**, 81–82.
- Dell, G. 1988. *Frame constraints on phonological speech errors*. Paper presented at the First Annual Conference on Language and Connectionism, Rome, December 1988.
- Gandour, J., Marshall, R., & Windsor, J. 1989. Idiosyncratic strategies in sentence production: A case report. *Brain and Language*, **36**, 614–624.
- Grodzinsky, Y. 1984. The syntactic characterization of agrammatism. *Cognition*, **16**, 99–120.
- Grodzinsky, Y. 1989. *The fallacy of the recent fad in neuropsychology*. Unpublished manuscript. Tel Aviv Univ., Tel Aviv.
- Hart, J., Berndt, R., & Caramazza, A. 1985. Category-specific naming deficit following cerebral infarction. *Nature (London)*, **316**, 439–440.
- Heeschen, C. 1985. *The notion of "conventionality" of an expression in cross-linguistic research on aphasia*. Presented at the Royaumont Conference Center, Paris, France.
- Holland, A. 1986. *On recovery from aphasia*. Keynote address to the Academy of Aphasia, Pittsburgh, October 1986.
- Kilborn, K. 1987. *Sentence processing in a second language: Seeking a performance definition of fluency*. PhD thesis, Univ. of California, San Diego.
- Kilborn, K. 1990. *The effects of task and language type on sentence interpretation*. Unpublished manuscript, Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands.
- Kilborn, K., & Ito, T. 1989. Sentence processing strategies in adult bilinguals: Mechanisms of second language acquisition. In B. MacWhinney & E. Bates (Eds.), *The crosslinguistic study of sentence processing*. New York: Cambridge Univ. Press.
- Kolk, H. 1985. *Telegraphic speech and ellipsis*. Presented at the Royaumont Conference Center, Paris, France.
- Kolk, H., & van Grunsven, M. 1985. Agrammatism as a variable phenomenon. *Cognitive Neuropsychology*, **2**(4), 347–384.
- Linebarger, M., Schwartz, M., & Saffran, E. 1983. Sensitivity to grammatical structure in so-called agrammatic aphasics. *Cognition*, **13**, 361–392.
- MacWhinney, B. 1987. The competition model. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum.
- MacWhinney, B., & Bates, E. (Eds.) 1989. *The crosslinguistic study of sentence processing*. New York: Cambridge Univ. Press.

- MacWhinney, B., Bates, E., & Kliegl, R. 1984. Cue validity and sentence interpretation in English, German and Italian. *Journal of Verbal Learning and Verbal Behavior*, **17**, 539–558.
- Marshall, R., Gandour, J., & Windsor, J. 1988. Selective impairment of phonation: A case study. *Brain and Language*, **35**, 313–339.
- Marslen-Wilson, W., & Tyler, L. 1980. The temporal structure of spoken language understanding. *Cognition*, **8**, 1–71.
- Massaro, D., & Oden, G. 1980. Evaluation and integration of acoustic features in speech perception. *Journal of the Acoustical Society*, **67**, 996–1013.
- McDonald, J., & MacWhinney, B. 1989. Maximum likelihood estimation in crosslinguistic studies of sentence processing. In B. MacWhinney and E. Bates (Eds.), *The crosslinguistic study of sentence processing*. New York: Cambridge Univ. Press.
- Miceli, G., & Caramazza, A. 1988. Dissociation of inflectional and derivational morphology. *Brain and Language*, **31**(1), 24–65.
- Miceli, G., Silveri, M. C., Romani, C., & Caramazza, A. 1989. Variation in the pattern of omissions and substitutions of grammatical morphemes in the spontaneous speech of so-called agrammatic patients. *Brain and Language*, **36**, 447–492.
- Newcombe, F., & Marshall, J. 1988. Idealization meets psychometrics: The case for the right groups and the right individuals. *Cognitive Neuropsychology*, **5**(5), 517–528.
- Oden, G., & Massaro, D. 1978. Integration of featural information in speech perception. *Psychological Review*, **85**, 172–191.
- Peirce, C. S. 1933. In C. Hartshorne, P. Weiss, & A. Burks (Eds.), *Collected papers of C. S. Peirce*. (p. 56). Cambridge, MA: Harvard Univ. Press.
- Seidenberg, M. 1988. Cognitive neuropsychology and language: The state of the art. *Cognitive Neuropsychology*, **5**(4), 403–426.
- Seidenberg, M., McClelland, J., & Patterson, K. 1987. *A distributed, developmental model of visual word recognition, naming and dyslexia*. Paper presented at the Experimental Psychological Society (UK) symposium on connectionism, Oxford, July 1987.
- Shallice, T. 1988a. *From neuropsychology to mental structure*. Cambridge: Cambridge Univ. Press.
- Shallice, T. 1988b. *Connectionist modelling of aspects of acquired dyslexia*. Paper presented at the First Annual Conference on Language and Connectionism, Rome, December 1988.
- Shankweiler, D., Crain, S., Gorrell, P., & Tuller, B. 1989. Reception of language in Broca's aphasia. *Language and Cognitive Processes*, **4**(1), 1–33.
- Smith, S., & Bates, E. 1987. Accessibility of case and gender contrasts for agent–object assignment in Broca's aphasics and fluent anomics. *Brain and Language*, **30**, 17–32.
- Smith, S., & Mimica, I. 1984. Agrammatism in a case-inflected language: Comprehension of agent–object relations. *Brain and Language*, **21**, 274–290.
- Sokolov, J. 1989. Cue validity and the acquisition of Hebrew. In B. MacWhinney and E. Bates (Eds.), *The crosslinguistic study of sentence processing*. New York: Cambridge Univ. Press.
- Whitaker, H., & Slotnick, H. 1988. Comments on "The case for single patient studies:" Is (Neuro)psychology possible? *Cognitive Neuropsychology*, **5**(5), 529–534.
- Wulfeck, B. 1988. Grammaticality judgments and sentence comprehension in agrammatic aphasia. *Journal of Speech and Hearing Research*, **31**, 72–81.
- Zurif, E., Gardner, H., & Brownell, H. (1989). The case against the case against group studies. *Brain and Cognition*, **10**(2), 237–255.