

Agrammatic Broca's Aphasia Is Not Associated with a Single Pattern of Comprehension Performance

Alfonso Caramazza,* Erminio Capitani,† Arnaud Rey,* and Rita S. Berndt‡

**Harvard University*; †*Universita di Milano, Milan, Italy*; and ‡*The University of Maryland*

One influential hypothesis posits that the brain regions implicated in Broca's aphasia are responsible for specific syntactic operations that are necessary for the comprehension and production of sentences (Grodzinsky, 1986, 1990, in press). The empirical basis of this hypothesis is the claim that Broca's aphasics have no difficulty understanding sentences in the active voice (and other "canonical" sentence types, such as subject relatives and clefts with agentive predicates), but perform at chance level with passive voice constructions (and other "non-canonical" sentences such as object-gap relatives and object clefts). In the face of well-established results indicating that Broca's aphasics can exhibit several different performance patterns on these sentence types, Grodzinsky, Piñango, Zurif, and Drai (1999) argued that these conflicting results do not challenge the theory when the data are analyzed appropriately. They carried out a creative statistical analysis of the comprehension performance of published cases of Broca's aphasia and concluded that all of these cases are in agreement with the predicted pattern: chance on passives and 100% correct on actives. Here we show that the statistical reasoning adopted by Grodzinsky et al. (1999) is flawed. We also show that the comprehension performance of a substantial number of the Broca's aphasics in their own sample does not conform to the pattern required. Rather, contrary to these authors' claim, Broca's aphasia is not associated with a consistent pattern of sentence comprehension performance, but allows for a number of distinct patterns in different patients. © 2001 Academic Press

Scientific study of the relationship between brain regions and language functions has been based almost entirely on data from the performance of adult patients with focal brain damage. Research studying the performance of aphasic patients has a long and interesting history and has yielded a general framework for mapping language functions such as comprehension and production to relatively delimited regions of the left hemisphere (Geschwind, 1965). More recently, as cognitive and linguistic models of language functions have increasingly guided patient testing and as methods for imaging structural brain lesions have improved, hypotheses have become much more specific about the neural substrates that support discrete language processes. One particularly influential hypothesis posits that the brain regions implicated in causing Broca's aphasia are responsible for specific syntactic operations that are necessary for the comprehension and production of sentences (Grodzinsky, 1986, 1990, in

The work reported here was supported in part by NIH Grants NS22201 to Harvard University and DC00262 to the University of Maryland School of Medicine.

Correspondence and reprint requests regarding this article may be addressed to Alfonso Caramazza, Cognitive Neuropsychology Laboratory, William James Hall, Harvard University, 33 Kirkland St., Cambridge, MA 02138. E-mail: caram@wjh.harvard.edu.

press). As this hypothesis has developed and evolved over time, it has generated much interest from linguists and others who attempt to understand language/brain relations and to test specific linguistic theories (see special issues of *Brain and Language*, 1993, 1995). Most of this effort has focused on the comprehension of distinct sentence structures by patients clinically classified as Broca's aphasics. In the most recent formulation, the hypothesis (as it concerns comprehension) states that the brain region responsible for Broca's aphasia "is neural home to receptive mechanisms involved in the computation of the relation between transformationally moved phrasal constituents and their extraction sites" (Grodzinsky, in press, abstract).

The empirical basis of this hypothesis is the claim that Broca's aphasics have no difficulty understanding sentences in the active voice (and other "canonical" sentence types such as subject relatives and clefts with agentive predicates), but perform at chance level with passive voice constructions (and other "noncanonical" sentences, such as object-gap relatives and object clefts). This pattern is said to result because canonical sentences do not involve moved phrasal constituents; therefore, they are understood without difficulty. The passive voice and other noncanonical sentence types require transformational movement and thus cannot be understood because the source of the moved constituent cannot be traced.

One of the most important aspects of this argument is that it is stated, not as a vague association of symptoms, but as a testable hypothesis about *necessary* performance patterns. The hypothesis is that *all* patients exhibiting the symptoms of Broca's aphasia will show this particular comprehension impairment. Thus, Grodzinsky's thesis is subject to falsifiability by nonconforming cases. The principle of falsifiability is widely assumed to be the bedrock of the scientific method. For example, the eminent philosopher of science Sir Karl Popper stated this principle unequivocally: "a statement (a theory, a conjecture) has the status of belonging to the empirical sciences if and only if it is falsifiable" (1983, p. xix). Grodzinsky (in press) clearly recognized the importance of this aspect of his thesis when discussing reports of nonconforming data about the comprehension patterns found among Broca's aphasia: "[these] presumed inconsistencies must be taken very seriously: an unstable syndrome . . . is the wrong object of inquiry; likewise, a false hypothesis is, most likely, the wrong one to follow. It must be revised, perhaps even abandoned, when confronted with data that cannot be accounted for."

Despite these assurances, Grodzinsky and his colleagues have not proved willing to revise the hypothesis to accommodate the occurrence of nonconforming data. In the face of well-established results indicating that Broca's aphasics can exhibit several different performance patterns on these sentence types, Grodzinsky, Piñango, Zurif, and Drai (1999) argued that these conflicting results do not challenge the theory when the data are analyzed appropriately. They carried out a creative statistical analysis of the comprehension performance of published cases of Broca's aphasia and concluded that all of these cases are in agreement with the predicted pattern: chance on passives and 100% correct on actives. Here we show that the statistical reasoning adopted by Grodzinsky et al. (1999) is flawed. We also show that the comprehension performance of a substantial number of the Broca's aphasics in their own sample does not conform to the pattern required. Rather, contrary to these authors' claim, Broca's aphasia is not associated with a consistent pattern of sentence comprehension performance, but allows for a number of distinct patterns in different patients.

1. THE BACKGROUND

1.1 *Early Evidence of Sentence Comprehension Deficits in Some Broca's Aphasics*

Almost 25 years ago, Caramazza and Zurif (1976) challenged the view then prevalent that Broca's aphasia is associated with "normal" comprehension. They found that a group of Broca's aphasics performed poorly in comprehending certain types of sentences. They compared patients' performance on sentences whose meaning could be inferred from the meanings of the major lexical items alone (semantically nonreversible sentences: "The apple that the boy is eating is red") and sentences whose accurate interpretation depends on the correct analysis of their syntactic structure (semantically reversible sentences: "The boy that the girl is chasing is tall"). The Broca's aphasics tested by Caramazza and Zurif performed quite well on the semantically nonreversible sentences but very poorly on the semantically reversible sentences. This pattern of performance has been referred to as "asyntactic comprehension." The authors interpreted these results as indicating that Broca's aphasia is the result of damage to syntactic processing mechanisms that are used both in sentence comprehension and in sentence production. On this view, the agrammatic production of Broca's aphasics necessarily co-occurs with asyntactic comprehension because the two deficits are the result of damage to common mechanisms that are used in the two tasks.

Although various other investigations of the comprehension performance of Broca's aphasics initially seemed to confirm a link between agrammatic production and asyntactic comprehension (e.g., Caramazza, Berndt, Basili, & Koller, 1981; Heilman & Scholes, 1976; Schwartz, Saffran, & Marin, 1980), it soon became apparent that the relationship did not hold for all agrammatic patients (see Berndt, 1991, for review). In fact, a number of Broca patients were described with flawless sentence comprehension on complex reversible sentences (e.g., Miceli, Mazzuchi, Menn, & Goodglass, 1983; Kolk, VanGrunsven, & Keyser, 1985). These results undermine the claim that Broca's aphasics show a single pattern of impairment on these sentences, and they support the conclusion that the underlying cause of agrammatic production (as clinically defined) is not the same as the underlying cause of asyntactic comprehension.¹ The results also reveal that membership in clinical categories like Broca's aphasia does not uniquely determine the nature of the underlying deficit in the patients included in those categories.²

¹ Grodzinsky, Zurif, and Swinney (1985) acknowledged that the existence of these cases presents a serious challenge to their claims about Broca's aphasia. However, they pointed out that the performance criteria used to categorize patients are ". . . too crude to insure comparability in patient selection . . . in terms of principled cognitive distinctions." It is not implausible to argue that putative cases of Broca's aphasia that do not show the predicted pattern of comprehension performance might have been miscategorized as Broca's aphasics and therefore would be irrelevant to issues about comprehension performance in this type of aphasia. However, by adopting the "miscategorization argument," Grodzinsky et al. (1985) implicitly concede that clinical classification does not provide an adequate basis for the identification of the *functional* lesions in a brain-damaged patient. Furthermore, if this argument were to have any force it would have to be accompanied by *explicit, independent* criteria for distinguishing the "real" cases of Broca's aphasia from superficially similar cases whose problematic performance could safely be ignored. Otherwise we could always appeal to the "miscategorization argument" when confronted with nonconforming data, thereby insulating the theory from falsification. We will return to this point in the Discussion.

² Grodzinsky et al. (1999) correctly point out that it has been argued that, since membership in a clinical category does not provide an adequate basis for determining the functional lesion(s) in a patient, the only valid method for investigating the nature of cognitive deficits is through single-patient analyses (Caramazza, 1986). However, from this observation, they attribute to proponents of this position the claim that ". . . linguistic behavior, once examined at a sufficient level of detail, reveals vast inter-

Despite the evidence that there are a variety of comprehension patterns among Broca's aphasic patients, Grodzinsky and his collaborators (1986, 1990, in press; Grodzinsky, Pearce, & Malakovitz, 1991) have continued to maintain that Broca's aphasia is associated with a single pattern of comprehension performance: chance performance on sentences with noncanonical structure and normal performance on sentences with canonical structure.³ We will refer to this pattern of performance as "agrammatic comprehension."

1.2 The "Same" Database but Opposite Conclusions

Berndt, Mitchum, and Haendiges (1996) and Grodzinsky et al. (1999) recently carried out reviews of a large number of studies dealing with the comprehension of active and passive voice sentences in Broca's aphasia. The aim of these reviews was to decide whether Broca's aphasia is *systematically* associated with a specific pattern of sentence comprehension impairment. The two reviews used a similar (and partially overlapping) database but performed different analyses and reached different conclusions. The database consists of the comprehension performance of Broca's aphasics in sentence/picture matching or verification tasks. Patients hear a sentence in the active voice ("the boy chases the dog") or in the passive voice ("the dog is chased by the boy") and must either choose between a correct depiction of the sentence and a distracter picture showing a reversal of roles of the nouns in the sentence (a dog chasing a boy) (Schwartz et al., 1980) or judge whether the meaning of the sentence is accurately portrayed by a single picture.

In an analysis of the individual performance of 42 Broca's aphasics, Berndt et al. (1996) found that a number of different patterns were exhibited. Using a similar database of published studies, Grodzinsky et al. (1999) claimed that only one pattern of comprehension performance was consistently observed: chance performance on passives and flawless performance on actives. How is it possible for two analyses of essentially the same data to reach such discrepant conclusions about the nature of the comprehension performance in Broca's aphasics? Are our methods of analysis so flexible that they allow us to legitimately reach diametrically opposed conclusions? Or is one of the two methods of analysis flawed and only one of the two conclusions correct? Let us consider in more detail the analyses carried out in these two studies.

Berndt et al. (1996) used the following criteria in selecting studies for inclusion in their review: (1) the patients were described in the studies as nonfluent agrammatic aphasics (i.e., Broca's aphasics); (2) active and passive voice sentences were tested; and (3) a sentence/picture matching or verification paradigm was used in which the probability of performing correctly by chance was .50. Fifteen studies met these crite-

patient variation that defies generalization, or inference to a theory" (p. 135). This attribution is incorrect, and is essentially the opposite of what has actually been claimed. Proponents of single-patient studies have argued that the detailed analysis of individual patients' performance is *necessary* to constrain hypotheses about functional relationships among symptoms. Although comparison of symptom patterns in different patients may be very useful, the variability across patients in the same clinical category is irrelevant.

³ Grodzinsky et al. (1999) contrast this claim of a unique pattern of comprehension failure in Broca's aphasia with the view that "... the comprehension of actives and passives in agrammatic Broca's aphasics varies *randomly* across patients" (p. 135; emphasis added). They attribute the latter position to those researchers who have argued that functional lesions cannot be inferred reliably on the basis of clinical categories. To our knowledge, such a claim has not been made. Rejection of the claim that "Broca's aphasia is associated with a unique pattern of comprehension failure" does not imply that comprehension of actives and passives should vary *randomly*. The pattern of comprehension performance in an individual Broca's aphasic will depend on the specific cognitive mechanisms that are damaged in *that* patient. This claim is silent on the distribution of comprehension patterns across such patients.

ria. The comprehension performance of each of the 42 patients from these studies was analyzed individually. A Binomial Test was used to determine for each patient whether his/her performance on active and passive sentences differed from what would be expected by chance. They found that only about one third of the patients demonstrated the comprehension profile predicted by Grodzinsky and his collaborators. The remaining two thirds of the cases were approximately equally divided between patients who were impaired in comprehending both active and passive sentences and those who performed well on both sentence types. These authors argued, therefore, that the claim that Broca's aphasia is associated with a consistent pattern of comprehension performance is empirically unfounded.

Grodzinsky et al. (1999) criticized Berndt et al.'s review on methodological and statistical grounds. Their principal criticism focused on the review's failure to combine the data from the individual patients and to carry out an analysis of the group data. They contend that since Berndt et al. reported analyses of the performance of individual patients, no meaningful conclusion could be drawn. The basis for this contention is their claim that when a hypothesis predicts chance level outcomes, only the performance of groups of patients can provide a fair test of the hypothesis.

Grodzinsky et al. (1999) also criticized Berndt et al.'s review on three specific points. First, they argued that the patient selection used by Berndt et al. did not match the standard criteria for classifying Broca's aphasics (see Berndt & Caramazza, 1999, and section 3 for a discussion of this issue). However, as will be demonstrated later, similar results and conclusions are obtained if one uses the classification criteria and the data from the Broca's aphasics in Grodzinsky et al.'s (1999) review. Second, they argued that all patients were inappropriately given the same weight, so that patients who were tested with only 8 sentences were considered to be as representative of the putative comprehension deficit in Broca's aphasia as patients who were tested with 48 sentences.⁴ The objection here concerns the number of trials and the reliability of data obtained in experiments that use small numbers of sentences. We will deal with this issue in some detail below, since it is obviously of crucial importance in evaluating the results obtained in single-patient studies. However, concerning the Berndt et al. analysis, this objection is not valid because those authors used a binomial test to investigate each individual patient's performance. A key feature of the binomial test is that it takes sample size into account. The probability of rejecting the chance hypothesis is a function of the number of sentences included in the task. The binomial test allows us to determine (probabilistically) whether the score obtained with a given number of sentences belongs or does not belong to the chance distribution.⁵ Once this probability has been determined for each patient's score, then the probabilities for all patients will have the same weight. The last criticism concerned the fact that different independent testing sessions with the same patient were presented (and counted) separately. The separate presentation of data from the same patient across different testing sessions (and often across many years) indicated the stability of the patterns exhibited. Beyond this, however, this methodological decision

⁴ Surprisingly, a few lines after commenting on the importance of sample size (p. 139, footnote 3), Grodzinsky et al. (1999) provide a figure displaying the distribution of patient performances on active and passive sentences in which each patient is given the same weight, in violation of their earlier warning (p. 140, Fig. 1). Similarly, later in the text, they demonstrate numerically that *mean* performance on passives and actives is significantly different (using a *t* test) with a type of test that assumes that the patients contributing to the passive and active distributions have exactly the same weight (p. 140).

⁵ Although the binomial test takes into consideration sample size, it is obvious that studies based on large samples provide more reliable evidence of a patient's true performance. The binomial test is simply a way to treat properly the available data given the variations in sample sizes.

had no implications for the analyses conducted, since the data were not combined for analysis. Moreover, we will demonstrate that this aspect of Berndt et al.'s analysis does not change the results obtained in their study.⁶

Grodzinsky et al. (1999) adopted a very different approach in the analysis of patients' performance. As already noted, they argue that only the performance of groups of patients can provide a test of their claim that Broca's aphasia is systematically associated with a specific pattern of comprehension failure. On this view, finding (one or many) individual patients whose comprehension performance does not conform to the expected pattern—say near-perfect comprehension of passives—does not falsify their claim. Grodzinsky et al. (1999) argue that we cannot know with confidence whether these patients are performing at chance. This is because statistical variation allows the existence of such cases even if their "true" performance with passives is at chance level. If we assume a chance process for comprehension of passive sentences, it is indeed statistically *possible* that some patients would obtain flawless performance in a test session even though they are truly impaired in sentence comprehension.⁷ Based on this reasoning, Grodzinsky et al. (1999) go on to argue that the score of an individual patient is therefore meaningless if it is not combined with the performance of other patients.

Grodzinsky et al. (1999) carried out two types of analyses on the comprehension performance of 42 patients who met their criteria for classification as Broca's aphasics: (1) an analysis of the *average* performance across sentence types and (2) an analysis of the *distribution* of correct performance for different sentence types. The first analysis involved the use of traditional inferential statistical tests (χ^2 and *t* tests). These are straightforward analyses even though, as noted above, all patients were inappropriately given the same weight (see footnote 4). The percentage of correct responses was used in these analyses, neglecting the heterogeneity of sample sizes. Using χ^2 tests, Grodzinsky et al. (1999) assessed whether the average performance for active sentences and the average performance for passive sentences were different from chance. The results of these tests showed that the mean performance on active sentences (86% correct) was significantly different from chance, but that the average performance on passive sentences (55.3% correct) did not differ from chance.⁸ Furthermore, the mean correct performance on passives was significantly different (by repeated measures *t* test) from the mean correct performance on actives. Assuming that an appropriate calculation weighting the contribution of each subject by the sample size would lead to the same results, such tests tell us only that there is a general tendency for Broca's aphasics to perform poorly with passive sentences. We cannot conclude from this that there is a *systematic* link between this clinical syndrome and

⁶ In fact, this is relatively easy to calculate from the data presented in that study, since all of the patients with repeated testing were clearly marked in the Appendix. If each patient is counted only once (and the total number of correct responses across all trials is submitted to the Binomial Test), the results for the 42 patients differ little from those presented by Berndt et al. for 64 data samples: 16 patients show better than chance performance on both structures; 10 patients show chance performance on both structures; and 16 patients show the pattern predicted by Grodzinsky et al. ($p < .05$, one tailed).

⁷ Of course, the probability of such events varies as a function of the number of observations we make (sample size). This issue is discussed in detail below. We will show that even though Grodzinsky et al.'s (1999) reasoning about *possible* outcomes is unimpeachable, it is the *probability* of the outcome that is the relevant measure of the validity of a theoretical claim.

⁸ As far as we can determine, Grodzinsky et al. (1999) performed a χ^2 test using the average percentage correct on a 1×2 contingency table, with an expected value of 50% for both the observed average correct performance and $(1 - \text{average correct})$ performance. However, the use of a χ^2 test is not appropriate in this context. This test cannot be used because the mean of a set of percentages (each with a different *N*) is not a frequency.

a deficit in sentence comprehension. Indeed, as will be shown later, some patients “hidden in the mean” do not show the comprehension pattern claimed by Grodzinsky.

In the second analysis, Grodzinsky et al. (1999) considered the distribution of patient scores in evaluating their hypothesis. They compared the distribution of the passive scores obtained by the patients in their sample to a binomial-like distribution generated by a computer simulation. The two distributions looked similar (e.g., they are described as “virtually identical” in the paper’s abstract), and Grodzinsky et al. (1999) concluded that this similarity indicates that Broca’s aphasics’ comprehension of the passive is at chance. However, from this analysis, all we can learn is that the distribution of comprehension scores on passives obtained by a group of Broca’s aphasics is grossly “similar” to an artificial and heterogeneous binomial-like distribution. The analysis is silent on whether there are patients in the group whose comprehension performance is not at chance.

In the next section, we provide a general demonstration that by analyzing the performance of individual patients with the proper tests, we can reject with a high level of confidence Grodzinsky et al.’s (1999) claim that Broca’s aphasics systematically perform at chance on passive sentences. We first show that the chance performance hypothesis can be tested (and in this case rejected) very easily by conducting simple binomial tests. We also illustrate the well-known fact that the power of this test strongly depends on the number of trials included in the task. If a sufficient number of trials is used, then the probability that a patient with a good score on that task belongs to the chance distribution is very low. In light of these elementary statistical facts, we show that the performance of individual Broca’s aphasics can be used to test the hypothesis that these patients’ comprehension of passives is at chance. We then critically evaluate the analyses used by Grodzinsky et al. (1999) to demonstrate that the performance of the Broca’s aphasics included in the review shows that their comprehension of passives is at chance. We show that the analysis of distributions of scores carried out by Grodzinsky et al. (1999) is statistically flawed. We then show that although their group-mean analyses could have been consistent with their claims had they been carried out appropriately, they would also have been consistent with many very different claims about the nature of comprehension impairment in Broca’s aphasia. We consider here two of these alternative claims and demonstrate the theoretical weaknesses and limits of such broad group analyses.

2. CHOOSING THE PROPER TEST

2.1 Testing the Chance Performance Hypothesis with the Binomial Test

How can we test Grodzinsky et al.’s (1999) hypothesis that Broca’s aphasics are systematically at chance in comprehending passive sentences? If patients effectively respond at chance, then each patient can be thought of as flipping a coin and using it as a guide for responding. In this situation, the expectation is that in the long run we would get about 50% heads and 50% tails. However, for any finite number (N) of coin tosses the actual number of heads and tails is unlikely to result in exactly $N/2$ heads and $N/2$ tails. The ratio of heads to tails can vary considerably. The actual values for any finite series will be distributed binomially with a mean of about .50. So, for example, if we were to flip a coin 10 times we might get 8 heads and 2 tails. If we were to do it again, we might get 4 heads and 6 tails, or 5 heads and 5 tails, or 3 heads and 7 tails, and so on. If we kept doing this for many many trials, we would end up with a distribution of heads and tails that approximates the binomial distribution. Figure 1 shows the (theoretical) binomial distribution of heads or tails

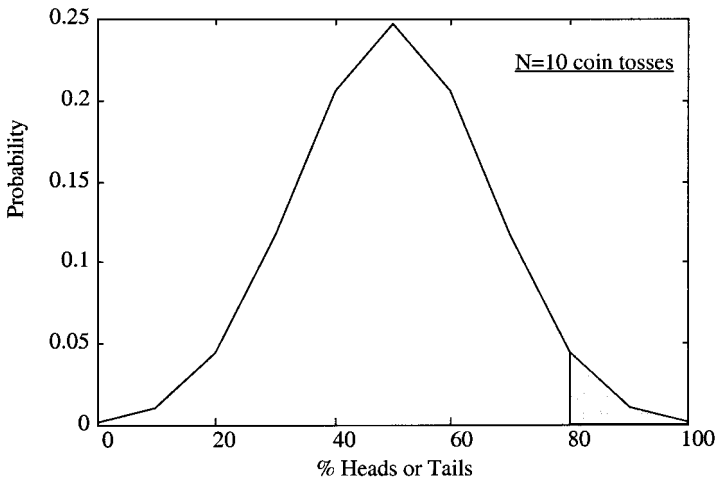


FIG. 1. Binomial distribution for $N = 10$ coin tosses (bin size, 10%).

for 10 coin tosses. As can be seen from the figure, the distribution is symmetric, with a peak at .50. Furthermore, it is quite apparent that the distribution is rather broad, covering a large range of ratios of heads to tails. This indicates that by flipping a coin 10 times we could get as many as 10 heads by chance, even though this is a rather unlikely event. In fact, obtaining by chance a score of N heads of N coin tosses occurs once every 2^N trials. In this particular case, 10 heads of 10 coin tosses occurs by chance one time every 1024 trials; the probability of this event is $2^{-10} = .00098$ (i.e., very low). Although possible, this event is very unlikely and as we increase the number of coin tosses, the probability of obtaining by chance N heads of N coin tosses decreases drastically. For $N = 20$ coin tosses, obtaining by chance 20 heads occurs once every 1,048,576 trials. From this simple numerical example, we can clearly see that the probability of a particular outcome is strongly dependent on the sample size, and we can use this fact to test the hypothesis that some observed event is the result of a chance process.

Similarly, if we consider now a patient score of 80% correct we can calculate the probability that this event occurred by chance if we know the number of trials sampled. In the case of a 10-trial test, the probability of obtaining by chance at least 8 heads (shaded area in Fig. 1) is .0547. It is therefore true, as Grodzinsky et al. (1999) correctly noted, that if we consider an 80% level of performance for a patient tested on a 10-item comprehension test we could not decide conclusively that this patient's comprehension of passives is different from chance level. Indeed, there is approximately a 6% probability that the patient's performance is at chance. From this observation Grodzinsky et al. go on to draw the conclusion that therefore one cannot use the results of single-patient analyses to test their theory of language processing in Broca's aphasia:

... Guessing behavior, which results in "chance" performance, cannot, and should not, be 50% correct per subject. Rather, it should be binomially distributed around the mean of 50% correct level. We can now see why results from multiple subjects are so important in this context: in such a response-type, each subject flips a coin and uses it for responding to each experimental question. A single subject, then, cannot be used to discern the pattern, if there are experimental conditions that might result in chance performance. This is so because the score of this particular subject may be located anywhere on a binomial curve. (p. 137)

This logic is flawed and appears to confuse the N of subjects with the N of trials. It is true that, for tests with a small N of trials, the score of a particular patient may

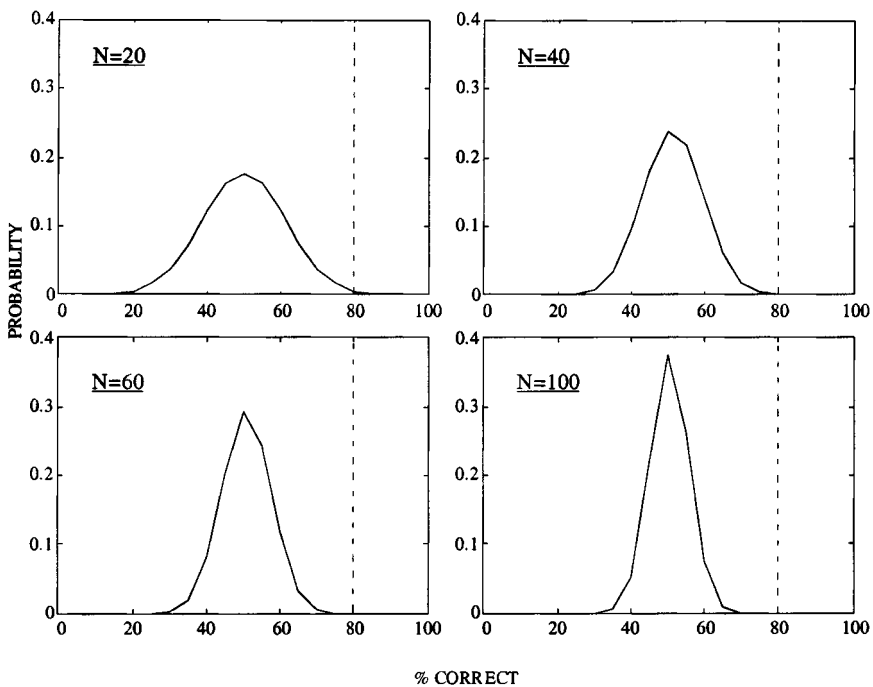


FIG. 2. Binomial distributions for $N = 20, 40, 60,$ and 100 coin tosses (bin size, 5%).

be located anywhere on a binomial curve with a fairly reasonable probability. This is true even for extreme scores, therefore rendering uninterpretable any testing done with small numbers of trials. However, for cases with a large numbers of trials, the probability of obtaining by chance an event such as an 80% correct score can be so low that it would be absurd to maintain the chance hypothesis. To illustrate this critical point, let us consider the binomial distributions obtained for different sample sizes displayed in Fig. 2. These distributions show the probability of obtaining a score by chance for different numbers of coin tosses. As can be seen, the larger the number of trials, the thinner the distribution and the smaller the probability of obtaining by chance an 80% score (represented by the dashed line in Fig. 2). For example, we can be very confident ($p < .0001$) that a patient categorized as a Broca's aphasic who obtains a score of 80% correct on passives on a 40-item test did not perform at chance level. We would have even greater confidence in drawing such a conclusion from an 80% correct score if the test had included 60 or 100 items. Furthermore, if such a pattern of performance is observed across a number of different Broca's aphasics, we would have even greater confidence that Grodzinsky et al.'s (1999) claim about the comprehension performance in such patients is false. We will show in section 3 that such patients exist and that they were included in Grodzinsky et al.'s (1999) own review.

These simple numerical and graphical examples illustrate how the binomial test provides a straightforward and unequivocal tool for evaluating the chance hypothesis. On the basis of *individual* patient scores (taking into account the percentage of correct responses and the sample size for each patient), one can confidently classify each patient's data (to whatever level of confidence is desired) as belonging or not belonging to the chance distribution.⁹ Clearly and obviously, increasing the number of trials

⁹ This exercise has attempted to clarify the distinction between uncertainty and probability. Grodzinsky et al. (1999) claim that the uncertainty associated with the scores obtained by individual patients "...

increases the robustness of the results and provides stronger empirical data for evaluating theories of language processing.

2.2 Comparing Distributions

In the preceding section we discussed a simple but efficient method for testing Grodzinsky et al.'s (1999) claim of a systematic link between Broca's aphasia and sentence comprehension. With this method we can *falsify* their claim if there exist patients classified as Broca's aphasics whose comprehension of passives is not at chance level. Grodzinsky et al. (1999) did not use the binomial test with individual patients to evaluate their hypothesis. Instead they carried out a patient group analysis that involved the comparison of two distributions. They reasoned that if their hypothesis were correct, passive scores should be distributed binomially (since patients are guessing the correct response). However, since the number of trials in the comprehension tasks used in the different studies are not equal, it is not possible to carry out a straightforward comparison between the distribution of scores obtained by the *group* of patients and the binomial distribution. Given Grodzinsky et al.'s (1999) assumption that correct performance is generated by a random process, the less likely events should be located in the tails of the distribution, and the more likely events should be closer to the expected value of 0.50. But this is not necessarily true for a distribution of cases with unequal *N*s. This is because a single event located in the tail could be better accommodated by the chance hypothesis than a more central one if the score in the tails derives from a small number of trials and the central score from a large number of trials. To circumvent this problem, Grodzinsky et al. (1999) compared the distribution of the passive scores obtained by the patients in their sample with a computer simulation designed to generate a binomial-like distribution for the particular group of patients included in the study. For each patient, a simulated patient was generated by flipping a coin as many times as that patient's sample size. They then plotted the distribution of scores obtained by the real and the simulated patients and concluded:

... The similarity between the data and the simulation is striking. They are both symmetric with a mean around 50%, and they are unimodal. Most crucially, they are both open to an almost identical extent, that is, the range of possible performances is as broad in both graphs.

Concluding, then, chance performance is equivalent to flipping a coin; a model of the distribution of coin tosses (corrected as our particular case requires) is similar to the actual data from passive; individual variation, thus, is a reflection of this distribution. Broca's aphasics, we can safely conclude, perform at chance levels on comprehension tasks of the passive construction. (pp. 141–142)

The "striking similarity" referred to is shown in Fig. 3 (Fig. 2 in the original Grodzinsky et al., 1999). Grodzinsky et al. (1999) used the well-known "eyeball" criterion to evaluate the match between the observed and the simulated distribution of scores and concluded that the results passed this test. However, note that Fig. 3

casts doubts on claims that in Broca's aphasia, the speech production deficit may manifest without an accompanying comprehension problem. That is, cases in which a comprehension problem in passive and object relative and cleft seems absent, may be mere distributional artifacts: the patients may have performed at chance, yet their scores happened to be on the higher end of the distribution" (p. 143). This statement describes uncertainty, not probability. Although we can never be *certain* that a Broca's aphasic who scores 100% on passives on a comprehension test with 100 trials did not obtain that score by chance, we can be certain that such a chance event will occur approximately only once every 2¹⁰⁰ tests. For most of us this is certainty enough that the patient does not have a comprehension problem. And even if we reduce the number of trials to 20, getting all 20 correct happens by chance only once every 1,048,576 tests. Again, most of us would have serious reservations about dismissing such an outcome as a "mere distributional artifact." This is the essence of statistical testing.

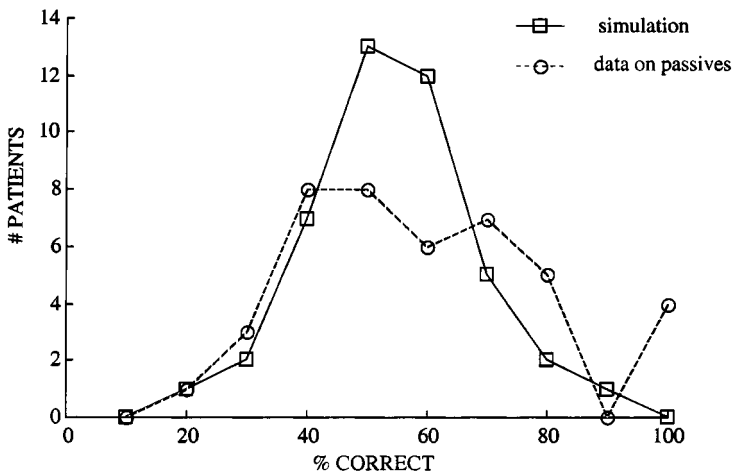


FIG. 3. Comparison between the passive comprehension data (dashed) and the simulation of Grodzinsky et al. (1999).

shows that in the range from 70 to 100% correct, the observed data include more subjects than the simulated ones and that in the central region real observations are underrepresented compared to the simulation. One needs to evaluate whether these are significant discrepancies by methods other than the eyeball criterion. But even if one were to carry out a more objective test of the observed and simulated distributions, it is not obvious that such a comparison would be meaningful. Given the properties of the binomial distribution, it is very likely that in the *simulated* distribution the tails mostly include “subjects” with small numbers of trials. However, we don’t know whether the tails of the *observed* distribution also include subjects with small numbers of trials. This crucial information is discarded by Grodzinsky et al. (1999) in their analysis of the two distributions. The contribution to the overall distribution of individual patient data based on various numbers of trials (and therefore more or less probably due to chance) does not play a role in their analysis.

The flaw in this approach to comparing distributions can be easily appreciated by considering two patients who obtain the same score on passive sentences: say 80% correct. Whether a score of 80% is consistent with Grodzinsky et al.’s (1999) hypothesis depends on the number of trials used in the test administered to the patients. If a test used only 10 sentences, we would not be able to confidently reject the null hypothesis that a score of 80% belongs to the chance distribution (cf. Fig. 1). However, the same score on a 40-item test allows us to confidently ($p < .0001$) reject the null hypothesis (cf. Fig. 2). Thus two patients with the same score but different numbers of trials would have drastically different implications for Grodzinsky et al.’s (1999) hypothesis. It should be clear from this example that without information about the number of test items associated with the comprehension score obtained by a patient, the analysis of the distribution of such scores is uninterpretable. This is because we don’t know whether a score in the distribution (say 90% correct) comes from a patient who was tested with only a few ($N = 6$) or with many ($N = 40$) sentences. In the former case, the score could not be used to reject the hypothesis of chance performance; in the latter case, we would be quite certain that the score does not come from a chance distribution. Thus the same score (or same point on the composite distribution) can have opposite meanings. The comparison of group distributions carried out by Grodzinsky et al. (1999) effectively masks the evidence that could be used to disconfirm their hypothesis; thus, it is not a valid test.

2.3 *The Limits of Group Mean Analysis*

As already noted, Grodzinsky et al. (1999) opted not to provide strong tests (patient-by-patient analyses) of their hypothesis but instead chose only to determine whether patients' performance as a group could be considered *consistent* with their claim.¹⁰ The problem with this approach is that the analyses are too weak to be useful, even had they been applied properly. The analysis of group means is not the correct test for Grodzinsky et al.'s (1999) hypothesis. Their hypothesis claims that agrammatic production is necessarily associated with agrammatic comprehension. This hypothesis can be tested only by assessing whether each patient who is categorized as an agrammatic Broca's aphasic also presents with agrammatic comprehension. We can better appreciate this point by considering two numerical examples.

2.3.1 The broad group mean approach. Consider first the following alternative hypothesis about the distribution of patterns of comprehension performance in Broca's aphasics: Comprehension performance in Broca's aphasia is more severely disrupted for passive than for active sentences.¹¹ The mechanisms responsible for this outcome could be the following. Let us suppose that damage to the brain regions that results in agrammatic production also *frequently and to varying extents* affects independent mechanisms that are involved in comprehension. Furthermore, suppose that comprehension of passives is generally harder than that of actives (even for neurologically intact subjects), perhaps because passives occur less frequently in the language.¹² Given these assumptions we would expect that the mean correct performance for passive sentences for a group of Broca's aphasics would be significantly worse than that for active sentences. However, performance for passive sentences need not be at chance level, although it could be. The exact level of performance would vary as a function of severity (and the particular strategies adopted by individual patients in the face of their deficit). The following numerical example illustrates one possible outcome.

To simplify, suppose that the 42 patients in the group analyzed by Grodzinsky et al. (1999) really consist of two subgroups, G1 and G2, with G1 being the more severely impaired subgroup. The nature of the damage in G1 is such that the mean performance of the subgroup is 40% on passive sentences and 80% on active sentences. However, individual scores are not exactly 40% and 80% correct, but are quasi-normally distributed around those means. The damage in G2 is such that the mean correct performance on passives and actives is 60 and 95%, respectively, with the same distributional constraints as the other subgroup. Table 1 presents the scores of 42 fictitious Broca's aphasics, and the distributions of these patients' scores for actives and passives are shown in Fig. 4. The mean correct performance for passives

¹⁰ In a more extreme statement that appears to further insulate their theory from falsification, Grodzinsky et al. (1999) assert that their analyses and conclusions ". . . weaken the diagnostic value of the active-passive comprehension contrast—it can be used just for a positive, yet not for a negative diagnosis of an individual as a Broca's aphasic, even though it is part and parcel of the overall behavior of the group. That is to say, if a patient performs at chance on passive and object relative clauses, and above chance on actives and subject relatives, we can use these scores for a positive diagnosis; yet the opposite is not true: the diagnosis is not ruled out by other results" (p. 143). Thus, if the results favor the theory, they are taken as support for the theory; if the results do not favor the theory, they can be ignored. This logic precludes the falsification, and therefore the fair testing, of the theory.

¹¹ Note that the examples developed here are intended only to illustrate why analyses of group means do not provide the correct test of Grodzinsky et al.'s (1999) hypothesis. They are not intended as substantive claims about the nature of the mechanisms responsible for comprehension performance in Broca's aphasics.

¹² Alternatively, it could be assumed that brain damage affects more severely the processing of grammatical morphemes, making it harder to understand passives.

TABLE 1
 Scores of 42 Fictitious Broca's Aphasics, 21
 Belonging to Group 1 (Being the More Severely
 Impaired Subgroup) and 21 to Group 2

| Patient no. | Group 1 | | Group 2 | |
|----------------|---------|----------|---------|----------|
| | Actives | Passives | Actives | Passives |
| 1 | 80 | 30 | 90 | 90 |
| 2 | 100 | 40 | 90 | 40 |
| 3 | 60 | 50 | 100 | 50 |
| 4 | 80 | 60 | 80 | 80 |
| 5 | 100 | 50 | 100 | 70 |
| 6 | 50 | 20 | 100 | 30 |
| 7 | 70 | 40 | 100 | 60 |
| 8 | 70 | 40 | 90 | 70 |
| 9 | 90 | 30 | 90 | 40 |
| 10 | 100 | 30 | 100 | 80 |
| 11 | 80 | 20 | 100 | 60 |
| 12 | 70 | 20 | 100 | 60 |
| 13 | 60 | 50 | 100 | 50 |
| 14 | 80 | 40 | 100 | 50 |
| 15 | 80 | 50 | 90 | 70 |
| 16 | 70 | 60 | 100 | 50 |
| 17 | 70 | 40 | 90 | 40 |
| 18 | 90 | 30 | 90 | 80 |
| 19 | 90 | 30 | 100 | 60 |
| 20 | 90 | 40 | 100 | 60 |
| 21 | 80 | 70 | 80 | 70 |
| <i>Mean</i> | 79 | 40 | 95 | 60 |

for the whole group is 52.4 (median = 50, mode = 40); the mean correct performance for actives is 86.9 (median = 90, mode = 100). The distributions of scores on passives and actives are therefore similar to those obtained by Grodzinsky et al. (1999). As is immediately apparent, we have exactly replicated the *group* results of Grodzinsky et al. (1999) with a drastically different hypothesis about the nature of the underlying deficit in Broca's aphasia. In other words, we have two very different hypotheses about the nature of the underlying deficit(s) in Broca's aphasia—Grodzinsky et al.'s (1999) and the "severity hypothesis"—and they are equally consistent with the results reviewed by Grodzinsky et al. (1999)¹³ *when only group mean analyses are performed on those results*. This example illustrates that the analysis of group means does not provide the appropriate test of Grodzinsky et al.'s (1999) hypothesis.

We did not need to resort to a contrived numerical example to illustrate the inadequacy of group mean analyses to test Grodzinsky et al.'s (1999) hypothesis (although the merit of that example is that it shows clearly the locus of the weakness in that analysis). We can use the actual data reported by Grodzinsky et al. (1999) to make much the same point, albeit with the important caveat that the interpretation of the distribution of patients' scores is not meaningful without information about the number of test trials used with each patient. Therefore, the analysis presented here is offered only as a general illustration of the limitations of the group mean approach

¹³ An important caveat is in order here. The "severity" hypothesis predicts that there should be a correlation between active and passive scores. The point here is not to promote the "severity" hypothesis (which is false in any case) but simply to illustrate the inadequacy of the group mean analyses carried out by Grodzinsky et al. (1999).

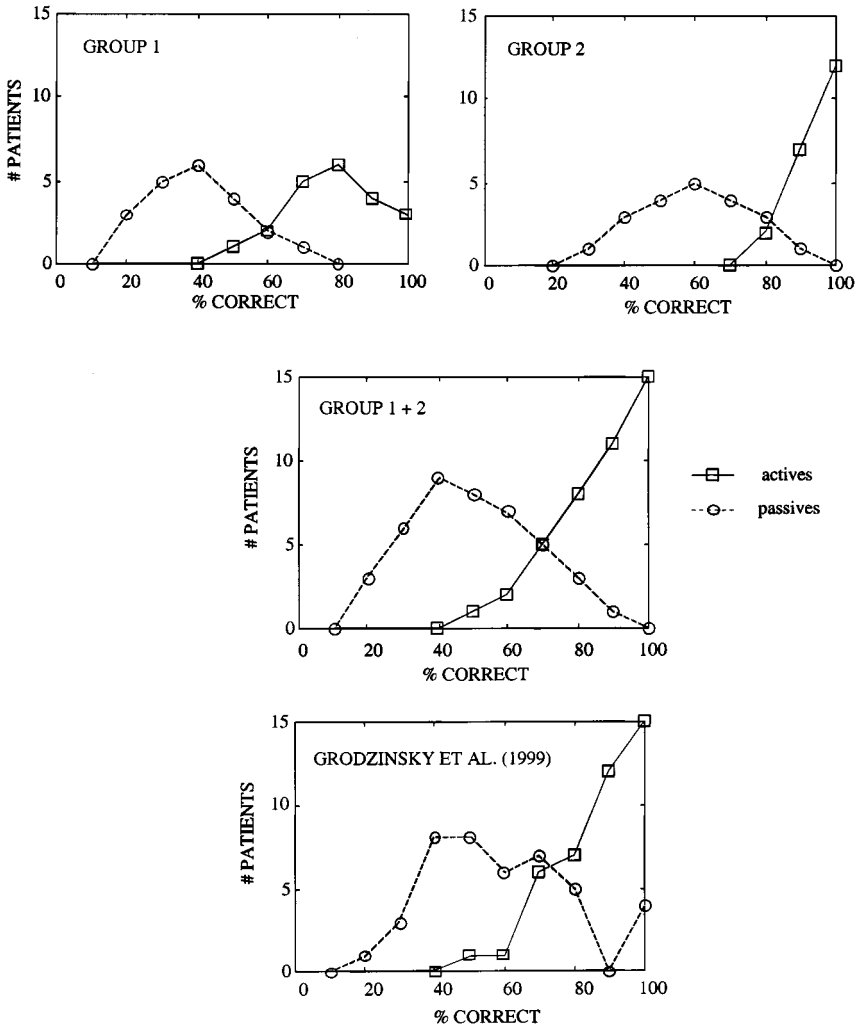


FIG. 4. Number of patients vs. performance level in actives (full line) and passives (dashed); top left panel shows the distributions for group 1; top right panel shows the distributions for group 2; middle panel shows the distributions for groups 1 + 2; bottom panel shows the distributions of Grodzinsky et al. (1999).

to test Grodzinsky et al.'s (1999) hypothesis, even if the appropriate information had been available.

Table 2 presents a reorganization of the 42 Broca's aphasics from Grodzinsky et al.'s (1999) review into two subgroups. One subgroup has a mean correct comprehension performance of 74 and 48% on actives and passives, respectively; the other subgroup's performance on these two sentence types is 96 and 63%, respectively. The distributions of these patients' scores on active and passive sentences are shown in Fig. 5. As is immediately apparent by comparing the obtained distributions with those shown in Fig. 4, they are very similar. In other words, the data reported by Grodzinsky et al. (1999) are consistent with the "severity hypothesis" of agrammatic comprehension *when only group mean analyses are performed on those results*. This example illustrates that the analysis of group means does not provide an unambiguous test of Grodzinsky et al.'s (1999) hypothesis.

TABLE 2
 Reorganization into Two Separate Groups of the
 Patients Included in Grodzinsky et al.'s Review

| Patient no. | Group 1 | | Group 2 | |
|----------------|-------------|-------------|-------------|-------------|
| | Actives | Passives | Actives | Passives |
| 1 | 50 | 54 | 86 | 64 |
| 2 | 52.6 | 56 | 86.8 | 59.75 |
| 3 | 67 | 33 | 88 | 63 |
| 4 | 67 | 33 | 90 | 40 |
| 5 | 67 | 35.5 | 90 | 45 |
| 6 | 67 | 54 | 90 | 50 |
| 7 | 67 | 71 | 92 | 95 |
| 8 | 70 | 70 | 96 | 42 |
| 9 | 71 | 29 | 96.5 | 76 |
| 10 | 71 | 29 | 100 | 29 |
| 11 | 79 | 33 | 100 | 40 |
| 12 | 79 | 42 | 100 | 50 |
| 13 | 79 | 71 | 100 | 50 |
| 14 | 80 | 40 | 100 | 55 |
| 15 | 80 | 50 | 100 | 66 |
| 16 | 83 | 17 | 100 | 67 |
| 17 | 83 | 50 | 100 | 67 |
| 18 | 83 | 57.5 | 100 | 72.38 |
| 19 | 83 | 67 | 100 | 90.5 |
| 20 | 85 | 35 | 100 | 100 |
| 21 | 85 | 73.25 | 100 | 100 |
| <i>Mean</i> | <i>73.7</i> | <i>47.6</i> | <i>96.0</i> | <i>62.9</i> |

Note. Group 1 has the more severely impaired patients.

2.3.2 Patients "hidden in the mean." The other example we will consider reveals even more serious problems with the use of group averages to test Grodzinsky et al.'s (1999) hypothesis. Let us suppose for the sake of argument that when a Broca's aphasic is impaired in sentence comprehension the form of the impairment will be exactly as claimed by Grodzinsky et al. (1999)—that is, chance on passives and normal on actives. But let us also further assume, contrary to Grodzinsky et al.'s (1999) hypothesis, that not all Broca's aphasics show the impairment in comprehension. This situation might arise, for example, if the brain region implicated in Broca's aphasia is immediately adjacent to a region of the brain that is necessary for normal sentence comprehension. Given the vagaries of brain damage, these two regions would tend to be damaged together very frequently—say about 90% of the time. Note that this hypothesis is drastically different from Grodzinsky et al.'s (1999). The claim here is that agrammatic production and comprehension tend to co-occur simply because of the neural proximity of the mechanisms whose damage is responsible for the two forms of impairment. Note, however, that the two impairments are *functionally* independent. On this hypothesis, the presence of agrammatic production in a patient can serve as an excellent clue as to whether the patient will show agrammatic comprehension, but it does not determine its presence. We cannot use mean patient-group performance to distinguish the latter hypothesis from Grodzinsky et al.'s (1999) hypothesis of a necessary co-occurrence of the two symptoms. A numerical example will further illustrate this point.

Suppose that 4 of the 42 patients tested by Grodzinsky et al. (1999) do not show

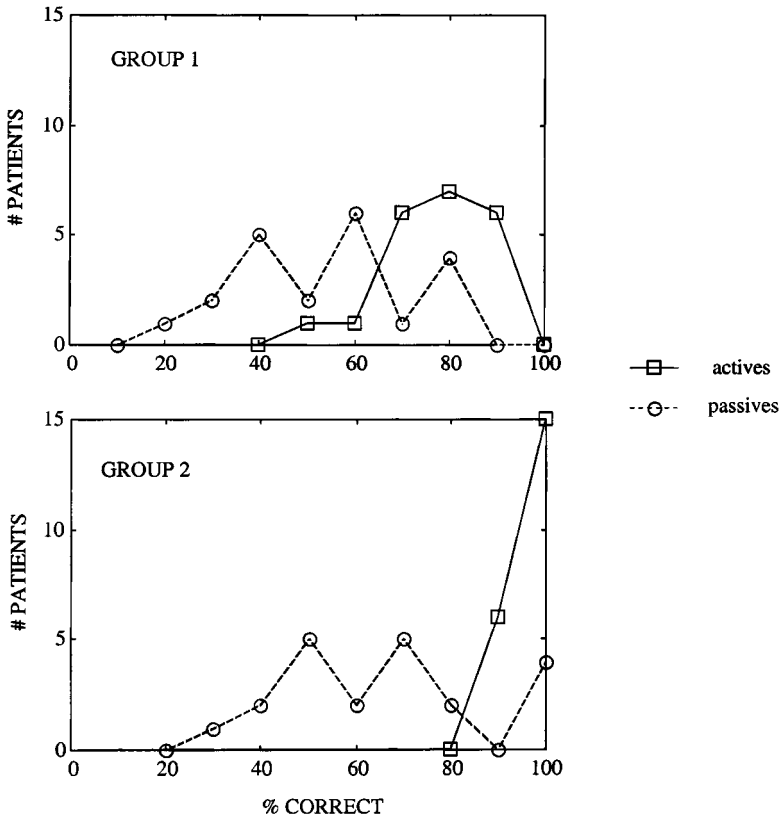


FIG. 5. Number of patients vs. performance level in actives (full line) and passives (dashed) for two groups of patients (groups 1 and 2, respectively, top and bottom) from Grodzinsky et al.'s (1999) set of patients.

the predicted association between agrammatic production and comprehension. Suppose also that the remaining 38 patients perform exactly as predicted by Grodzinsky et al. (1999). Table 3 shows the performance of this hypothetical set of patients. The mean correct performance of this group on passives is 54.8 and 100 on actives. Clearly, the group performance is again consistent with Grodzinsky et al.'s (1999) claim, even while the performance of 4 patients is clearly inconsistent with the claim. The present extreme numerical example demonstrates that by considering the group mean or the group distribution, the performance of patients that might be crucial for our understanding of human language/brain relationships are just "lost in the mean." There were such patients in Grodzinsky et al.'s (1999) sample and they are detectable by inspection of the right tail of the distribution of passive scores (see their Fig. 1), but they were simply ignored. Finally, the alternative hypothesis we have developed is not distinguishable from Grodzinsky et al.'s (1999) by means of the tests they used. That is, the group mean tests are fully compatible with the hypothesis that agrammatic production is not necessarily associated with agrammatic comprehension, and therefore the results of these tests do not constitute support for Grodzinsky et al.'s (1999) theory.

2.4 Summing Up

On the basis of their patient-group analyses, Grodzinsky et al. (1999) drew strong conclusions about Broca's aphasia and about methodology in cognitive neuropsych-

TABLE 3

A Hypothetical Set of 42 Patients, 38 of Them Performing at Chance (50%) on Passives and Perfectly on Actives (100%) and 4 Performing Perfectly on Both Passives and Actives (i.e., Patients 39 to 42)

| Patient no. | Passives | Actives |
|-------------|-------------|------------|
| 1 | 50 | 100 |
| 2 | 50 | 100 |
| — | 50 | 100 |
| — | 50 | 100 |
| 38 | 50 | 100 |
| 39 | 100 | 100 |
| 40 | 100 | 100 |
| 41 | 100 | 100 |
| 42 | 100 | 100 |
| <i>Mean</i> | <i>54.8</i> | <i>100</i> |

chology. We have seen that the group analyses carried out by Grodzinsky et al. (1999) are either flawed (the distribution comparison) or flawed and inappropriate for testing their theory (the group mean tests). Therefore these analyses cannot be used to support either their substantive claims about language processing and the brain or their methodological prescriptions for cognitive neuropsychology. What, then, is the status of Grodzinsky's hypothesis about the comprehension performance of patients clinically classified as Broca's aphasics? Do the available data, when analyzed appropriately, lend support to the hypothesis? We turn next to this issue. To anticipate our conclusion, we show that published results are clearly inconsistent with the hypothesis and that this conclusion could have been reached by Grodzinsky et al. (1999) had they analyzed their results correctly.

3. A FURTHER TEST OF THE HYPOTHESES

Grodzinsky et al. (1999) formulate two clear and testable research hypotheses about Broca's aphasia: (i) Broca's aphasics have no difficulty understanding active sentences and (ii) Broca's aphasics perform at chance level with passive constructions. Although Grodzinsky et al. (1999) focused primarily on the latter hypothesis, here we will consider both. Hypothesis (ii) can be tested by determining whether the percentage correct on passive comprehension in a forced choice paradigm can be described by a binomial distribution with a mean corresponding to chance level performance. This is the statistical hypothesis that Grodzinsky et al. (1999) attempted to test in their study. In this section, we test both this hypothesis and several other statistical hypotheses that can be derived from Grodzinsky et al.'s substantive claims about comprehension performance in Broca's aphasia.

3.1 Data Samples

Two partially overlapping samples of data were selected for analysis. For the reasons set forth in the sections above, any statistical test of Grodzinsky et al.'s (1999) hypothesis concerning the Broca population must incorporate information about the

number of trials (N) given to each patient. We intended to analyze Grodzinsky et al.'s (1999) data from the 42 patients included in their analysis, but it was not possible to determine the N of trials administered for 9 patients included in their sample, because the original study was unpublished, the patients were not individually identified in the original study, or the number of trials was not included in the original study. We also excluded data from 7 patients tested by Hagiwara (1993), where the number of trials (six) was too small for meaningful analysis. Data from the remaining 26 patients included in Grodzinsky et al.'s (1999) analysis constitute our Sample 1 (S1).

A second sample of patient data (S2) was analyzed to increase the number of patients included and to assure that the sample was entirely unbiased with regard to the hypothesis at issue. As noted elsewhere (Berndt & Caramazza, 1999), 22 of the patients in Grodzinsky et al.'s (1999) sample were chosen for study in the original papers because they showed the expected pattern of active/passive performance.¹⁴ Clearly, these patients were not selected independently and should be excluded from any analysis that attempts to test, rather than simply to reiterate, the theory's predictions. When these data are excluded (some of which had already been excluded from S1 for the reasons stated above), there were 16 patients from Grodzinsky et al. (1999) that could be included in the unbiased S2. To this group we added data from 16 patients that were included in Berndt et al.'s (1996) analysis. These data were excluded from Grodzinsky et al.'s (1999) study presumably because they were not "true" Broca's aphasics. We scrutinized the patient descriptions from the studies included in Berndt et al. and concluded that 3 patients from that original sample demonstrated a comprehension profile for single words that may have excluded them from the Broca category. The remaining 16 Berndt et al. patients excluded from Grodzinsky et al.'s (1999) analysis are described in enough detail that it is quite clear that they conform to the clinical profile of Broca's aphasia. Thus, data from these 16 patients were included in S2. In addition, data from 4 new patients described by Berndt, Mitchum, and Wayland (1997) and 13 patients studied by Benedet, Christiansen, and Goodglass (1998) were included in S2, producing a total of 49 Broca's aphasics for that sample. The patients in this sample were selected for study on the basis of their sentence production performance and single word comprehension, i.e., not on the basis of their sentence comprehension.

¹⁴ We base this assertion on the following statements from the "Subjects" sections of the cited papers: "[patients] had to show intact comprehension of SVO active sentences (above-chance performance) and impaired comprehension of SV verbal passive sentences (at-chance performance)" (Beretta, Harford, Patterson, & Pinango, 1996, p. 739); "Among 70 patients who were tested, 10 patients of Broca's type were selected on the basis of: . . . (b) demonstrating above-chance performance on actives and at-chance level performance on direct passives" (Hagiwara, 1993, p. 327); "All [patients] demonstrated agrammatism, both in production and comprehension" (Grodzinsky, 1995, p. 489); "in order to satisfy the precondition for inclusion in the experiment, patients had to have a testing history of performing well (i.e., above chance) on the tasks involving the interpretation of active sentences and performing poorly (i.e., at chance) on verbal passives" (Grodzinsky et al., 1991, p. 437). Hickok and Avrutin (1995) do not explicitly state that their two patients were preselected for study; however, these patients were tested in a series of earlier studies of "agrammatic" comprehension in which the comprehension pattern at issue was clearly determined prior to inclusion for experimental testing (e.g., patient RD: "On a pretest administered in 1991 he performed at 100% in the comprehension of semantically reversible active voice sentences and at chance in the comprehension of semantically reversible passive voice sentences (Hickok, Zurif, & Canseco-Gonzalez, 1993, p. 376; see also Zurif, Swinney, Prather, Solomon, & Bushell, 1993, p. 454); Patient FC: "On our comprehension test he scored in the fashion of most Broca's aphasics—better on actives (95%) than on passives (70%) (Zurif et al., 1993, p. 455; see also Grodzinsky, 1989, p. 493)). Our point is not to question the need for preselecting patients for the specific study carried out, but only to note that these data cannot be used to evaluate the prevalence of the pattern in the general population of Broca's aphasics because the patients were not chosen on independent grounds.

3.2 Statistical Methods

We conducted a number of analyses using these two data samples. The analyses were designed to test various aspects of Grodzinsky et al.'s (1999) hypothesis. The first two analyses (a and b) evaluate the performance of individual Broca's aphasics. We argued above that the analysis of individual patient performance provides the most appropriate test of Grodzinsky et al.'s (1999) hypothesis about comprehension in Broca's aphasia. However, since Grodzinsky et al. (1999) have objected to this type of analysis, we have also carried out two analyses (c and d) that consider separately the group performance of the two samples of patients. Thus, four types of analyses were carried out:

(a) First we calculated the confidence limits of the percentage correct of active and passive sentences for each subject in the two samples (26 from S1 + 49 from S2 - 16 (overlap) = 59). If Grodzinsky et al.'s (1999) hypothesis is correct, there should be no patient with a percentage correct on passives significantly greater than 50% (or there should be no more than 5% of such cases if the risk of a Type I error is set at .05). Moreover, all subjects should produce 100% correct performance on active sentences. It is important to note that if the probability correct is 100%, no variability due to chance can be found. The confidence limits were calculated using the method of Blyth and Still (1983), later refined by Casella (1987). This method is available in the StatXact4 package (Mehta & Patel, 1998). The Type I error risk was set separately for actives and passives at the .05 level for each subject. The importance of Type I error risk protection is discussed below.

(b) Next we compared the active and passive sentence performance of each subject. Here Grodzinsky et al.'s (1999) hypothesis holds that for all patients the difference should be significant. The comparison was carried out by means of Fisher's exact test, with a unidirectional hypothesis at the .05 level. In order to better interpret negative results, we also calculated the power of the Fisher comparison (unidirectional), assuming that the expected difference in the population between actives and passives is 50% (100% - 50% = 50%). The power depends on the number of trials tested for each patient. This analysis was also carried out using StatXact4 (Mehta & Patel, 1998).

(c) Next we analyzed the group distribution of the percentage correct of actives and passives. According to Grodzinsky et al. (1999), in the case of actives all subjects should be at ceiling, whereas in the case of passives the distribution should have a mean = .50 (i.e., 50%). For the reasons discussed above, the variability in the N of trials across patients does not allow a simple comparison of the distributions with a definite binomial function. To overcome this problem, we normalized the scores based on the number of trials so that the data from each patient could be compared directly. We proceeded as follows. In the study of passives, we first subtracted .50 from the observed percentage correct for each subject, and then we divided this difference by the standard deviation appropriate for the number of trials (N) given to each subject, i.e., the square root of $(pq)/N$, with $p = q = .50$. The resulting points are approximately distributed as the standardized normal curve (McNemar, 1962). In the z approximation for the subjects who had been given a small number of trials ($N < 20$), a correction for continuity was applied. Using this method, the observed and the expected frequencies can be directly compared to check the goodness of fit. The number of cells was set so that there were at least two expected observations in each cell (Armitage, 1971). Each test of the goodness of fit was then based on five frequencies. The bottom rows in Tables 7a and 7b show the frequencies expected for each patient sample on the basis of the corresponding areas of the normal curves, multiplied by the sample size. In the inter-

pretation of the resulting χ^2 values, we assumed a value of $5 - 1 = 4$ *df* for each table.¹⁵

(d) Finally we analyzed the relationship between correct actives and correct passives. According to Grodzinsky et al. (1999), because one of these variables (actives correct) should always be at ceiling and the other (passives correct) should be at chance, no correlation between active and passive performance should be observed.

3.3 Results

Tables 4 and 5 show the data for the subjects of the two samples considered in this study, S1 and S2, along with the results of the statistical analysis for each patient.

3.3.1 Confidence Limits of Active and Passive Correct for Each Subject

Sample 1, N = 26 (subjects from Grodzinsky et al. 1999): Actives. Only 5/26 patients (19.2%) performed at ceiling with actives: by definition, these are the only patients for whom the confidence limits can include 1.00. For 8 patients (36.8%), the lower confidence limits for actives include chance level (.50). However, it could be argued that the confidence limits might be too wide when the number of trials (*N*) is small. It is reasonable, therefore, to consider only the confidence limits based on *N* greater than or equal to 16 (in fact, 16 is the *N* value for which sufficient power is provided). Even with the latter restriction, there are 6/26 patients (23.1%) who are not significantly different from chance on actives. On the whole, this outcome is strong evidence against Grodzinsky et al.'s (1999) hypothesis that p (actives) = 1.00.

Passives. For 4 subjects (15.4%) the confidence limits of passives correct did not include .50, and 1 of them was at ceiling. However, it is possible that some number of subjects could have confidence limits that do not include .50 by chance. The expected rate of false rejections is 5%, but higher values might be observed, although less probably. To safeguard against this possibility, we can set a rejection level of the chance hypothesis at the individual subject level that would survive the higher protection required to control for Type I error at the level of the whole sample. A Bonferroni correction requires that we set the significance level to prevent Type I error at the level of .05/26 or even at .05/42 (as 42 was the number of subjects originally considered by Grodzinsky et al. (1999)). In the latter case, evaluating significance with $\alpha = .001$ for each patients allows a .05 level protection for the whole set of 42 patients (.05/42 = .0012). Let us work with an even greater protection, fixing $\alpha = .000001$ (i.e., one per million) for each subject. This significance level means that, on a Bonferroni basis, the probability of a single false rejection of the null hypothesis in a set of 42 patients will be smaller than $.000001 \times 42 = .000005$, i.e., a Type I error is expected about once over 20,000 experiments. Inspection of Table 6 shows that even with the strongest protection, 3 patients have confidence limits still not inclusive of the chance level of .50. We conclude that in the population of subjects from which Grodzinsky et al.'s (1999) cases were sampled there is strong

¹⁵ In tests of goodness of fit where an observed distribution is compared to the normal distribution, it is customary to base the significance assessment on $N-3$ *df*. This is because in the standardization of the observed frequencies, the mean, the standard deviation, and the *N* of subjects are constants. In our case, we evaluated χ^2 points in a more conservative way, considering as a constant only the sample size. This was possible because, after the transformation explained above, our points can already be viewed as *z* scores without the use of constants derived from the group distribution. In fact, our approach is conservative, and the significance levels observed with 4 *df* would be even greater if evaluated with 2 *df*.

TABLE 4
Data and Statistical Analyses for the Subjects of Sample S1

| Patient | A | P | Confidence limit | | Fisher | Power |
|-----------------------------------|--------|---------|-------------------------------|---------------------------------|-------------------|-------|
| | | | Pa | Pp | | |
| AB | 8/10 | 4/10 | .800 (.444–.963) | .400 (.150–.733) | .085 | .71 |
| AK ^a | 19/24 | 10/24 | .792 (.594–.914) | .417 (.234–.634) | .009 | .99 |
| AT ^a | 49/72 | 27/72 | .681 (.566–.784) | .375 (.264–.489) | <.001 | 1.00 |
| B | 14/14 | 4/14 | 1.000 (.770–1.000) | .286 (.104–.581) | <.001 | .87 |
| BL ^a , ED ^a | 17/24 | 7/24 | .708 (.500–.874) ^b | .292 (.126–.500) | .004 | .99 |
| D | 12/14 | 9/14 | .857 (.581–.974) | .643 (.371–.847) | .192 ^b | .87 |
| EB ^a | 16/24 | 13/24 | .667 (.447–.831) ^b | .542 (.339–.745) | .278 ^b | .99 |
| EG ^a | 19/24 | 8/24 | .792 (.594–.914) | .333 (.169–.553) | .002 | .99 |
| ER | 9/10 | 4/10 | .900 (.556–.995) | .400 (.150–.733) | .029 | .71 |
| ES | 17/20 | 7/20 | .850 (.639–.958) | .350 (.154–.589) | .002 | .97 |
| FM ^a | 93/186 | 102/182 | .500 (.426–.574) ^b | .560 (.485–.634) | .145 ^b | 1.00 |
| GV | 9/10 | 5/10 | .900 (.555–.998) | .500 (.187–.813) | .070 | .71 |
| HR ^a | 23/24 | 10/24 | .958 (.808–.998) | .417 (.234–.634) | <.001 | .99 |
| HT ^a | 12/24 | 13/24 | .500 (.308–.692) ^b | .542 (.339–.745) | .500 ^b | .99 |
| JG | 10/10 | 4/10 | 1.000 (.733–1.000) | .400 (.150–.733) | .005 | .71 |
| JR ^a | 16/24 | 17/24 | .667 (.447–.831) ^b | .708 (.500–.874) | .500 ^b | .99 |
| LS ^a | 52/72 | 40/72 | .722 (.607–.815) | .556 (.434–.673) | .028 | 1.00 |
| ME ^a | 45/48 | 45/48 | .938 (.831–.989) | .938 (.831–.989) ^b | .661 ^b | 1.00 |
| NF | 20/20 | 12/20 | 1.000 (.846–1.000) | .600 (.361–.791) | .002 | .97 |
| PJ ^a | 71/72 | 67/72 | .986 (.931–.999) | .931 (.851–.972) ^b | .104 ^b | 1.00 |
| POE ^a | 20/20 | 20/20 | 1.000 (.846–1.000) | 1.000 (.846–1.000) ^b | n.p. ^b | .97 |
| ROO+ ^a | 18/20 | 9/20 | .900 (.685–.982) | .450 (.231–.685) | .003 | .97 |
| VS ^a | 95/108 | 65/108 | .880 (.804–.933) | .602 (.506–.693) ^b | <.001 | 1.00 |
| YM | 8/10 | 5/10 | .800 (.444–.963) | .500 (.222–.777) | .175 | .71 |
| YY | 20/20 | 13/20 | 1.000 (.846–1.000) | .650 (.411–.846) | .004 | .97 |

Note. “A” and “P,” number of correct responses to active and passive sentences, respectively, of the total number of trials. “Pa” and “Pp”: confidence limits of the percentage of correct actives and passives, respectively. “Power,” power of Fisher’s exact test. “n.p.”: not possible because the four cells contain equal frequencies.

^a Subjects are also included in S2 (see text).

^b Results are inconsistent with Grodzinsky et al.’s (1999) hypothesis; when this corresponds to a failed rejection of the null hypothesis, only cases with power greater than .90 are marked. For other statistical details see the text.

evidence against the hypothesis that all Broca’s aphasics’ performance with passives is governed by chance.

Sample 2, N = 49 (new sample): Actives. For the patients in this sample (see Tables 4 and 5), and considering only cases where the statistical power is sufficient, the confidence limits for actives correct included the chance level for 7/49 patients (14.3%). This is a lower rate than was found with S1, probably because the patients in S2 were somewhat less severely impaired as a group.

Passives. For passives, 21/49 subjects (42.9%) performed significantly better than chance. We will not repeat here the calculations and the considerations discussed for S1; however, it is clear that the hypothesis that passive performance is at chance level is not tenable for a substantial number of patients in S2. In fact, for several of these cases the probability that the passive score is very good (or even at ceiling) simply by chance is negligible.

3.3.2 Comparison of Performance on Actives and Passives

According to Grodzinsky et al.’s (1999) hypothesis, the difference between the percentages of active and passive correct responses should nearly always be signifi-

TABLE 5

Data and Statistical Analysis for the Subjects of the Sample S2 (Patients marked with ^a in S1 should also Be Considered Part of This Sample)

| Patient | A | P | Confidence limit | | Fisher | Power |
|-------------|-------|-------|-------------------------------|---------------------------------|---------------------|-------|
| | | | Pa | Pp | | |
| BA (K&V) | 17/20 | 12/20 | .850 (.639-.958) | .600 (.360-.791) | .078 | .97 |
| HE (K&V) | 18/20 | 15/20 | .900 (.685-.982) | .750 (.533-.896) ^a | .204 ^a | .97 |
| KOE (K&V) | 20/20 | 19/20 | 1.000 (.846-1.000) | .950 (.769-.997) ^a | .500 ^a | .97 |
| LA (K&V) | 19/20 | 17/20 | .950 (.769-.997) | .850 (.639-.958) ^a | .302 ^a | .97 |
| OO (K&V) | 19/20 | 19/20 | .950 (.769-.997) | .950 (.769-.997) ^a | .756 ^a | .97 |
| ZO (K&V) | 18/20 | 17/20 | .900 (.685-.982) | .850 (.639-.958) ^a | .500 ^a | .97 |
| AB (D&M) | 16/16 | 14/16 | 1.000 (.802-1.000) | .875 (.646-.977) ^a | .242 ^a | .91 |
| CD (D&M) | 16/16 | 14/16 | 1.000 (.802-1.000) | .875 (.646-.977) ^a | .242 ^a | .91 |
| ED (D&M) | 16/16 | 8/16 | 1.000 (.802-1.000) | .500 (.272-.728) | .001 | .91 |
| GH (D&M) | 13/16 | 6/16 | .812 (.571-.947) | .375 (.178-.646) | .014 | .91 |
| AK (Martin) | 4/8 | 5/8 | .500 (.193-.807) | .625 (.249-.889) | .500 | .55 |
| GL (Martin) | 4/8 | 5/8 | .500 (.193-.807) | .625 (.249-.889) | .500 | .55 |
| JS (Martin) | 41/48 | 30/52 | .854 (.724-.933) | .578 (.437-.704) | .002 | 1.00 |
| AP (Martin) | 25/40 | 20/44 | .625 (.458-.773) ^a | .454 (.312-.610) | .089 | 1.00 |
| NB (Martin) | 37/40 | 40/44 | .925 (.806-.979) | .909 (.795-.968) ^a | .554 ^a | 1.00 |
| RW (Martin) | 38/40 | 30/44 | .950 (.834-.991) | .682 (.532-.814) ^a | .002 | 1.00 |
| HH (Berndt) | 24/24 | 23/24 | 1.000 (.874-1.000) | .958 (.808-.998) ^a | .500 ^a | .99 |
| JD (Berndt) | 24/24 | 24/24 | 1.000 (.874-1.000) | 1.000 (.874-1.000) ^a | n.p. ^a | .99 |
| DH (Berndt) | 24/24 | 24/24 | 1.000 (.874-1.000) | 1.000 (.874-1.000) ^a | n.p. ^a | .99 |
| NC (Berndt) | 22/24 | 20/24 | .917 (.744-.985) | .833 (.634-.941) ^a | .333 ^a | .99 |
| 1 (Benedet) | 2/10 | 8/10 | .200 (.037-.556) | .800 (.444-.963) | .012 ^{a,b} | .71 |
| 2 (Benedet) | 6/10 | 3/10 | .600 (.267-.850) | .300 (.087-.619) | .185 | .71 |
| 3 (Benedet) | 9/10 | 4/10 | .900 (.556-.995) | .400 (.150-.733) | .028 | .71 |
| 4 (Benedet) | 7/10 | 6/10 | .700 (.347-.933) | .600 (.267-.850) | .500 | .71 |
| 5 (Benedet) | 5/10 | 6/10 | .500 (.222-.777) | .600 (.267-.850) | .500 | .71 |
| 6 (Benedet) | 9/10 | 4/10 | .900 (.556-.995) | .400 (.150-.733) | .029 | .71 |
| A (Benedet) | 10/10 | 8/10 | 1.000 (.733-1.000) | .800 (.444-.963) | .237 | .71 |
| B (Benedet) | 10/10 | 9/10 | 1.000 (.733-1.000) | .900 (.556-.995) ^a | .500 | .71 |
| C (Benedet) | 10/10 | 5/10 | 1.000 (.733-1.000) | .500 (.222-.777) | .016 | .71 |
| D (Benedet) | 10/10 | 10/10 | 1.000 (.733-1.000) | 1.000 (.733-1.000) ^a | n.p. | .71 |
| E (Benedet) | 10/10 | 10/10 | 1.000 (.733-1.000) | 1.000 (.733-1.000) ^a | n.p. | .71 |
| F (Benedet) | 8/10 | 4/10 | .800 (.444-.963) | .400 (.150-.733) | .085 | .71 |
| G (Benedet) | 10/10 | 10/10 | 1.000 (.733-1.000) | 1.000 (.733-1.000) ^a | n.p. | .71 |

Note. K&V, Kolk and VanGrusven, 1985; D&M, Druks and Marshall, 1991; Martin, Martin et al., 1989, and Martin, 1987; Berndt et al., 1997; Benedet, Benedet et al., 1998; 1-6 Spanish patients A-G English speaking patients. n.p.: not possible because the four cells contain equal frequencies.

^a Results inconsistent with Grodzinsky et al.'s (1999) hypothesis.

^b In this case, Fisher's test is significant, but with a paradoxically greater success rate on passives.

cant. Since failure to reach significance can derive from insufficient power, we have calculated the power of Fisher's exact test in every case. This power index (i.e., the probability that the test will detect significant differences) depends on (1) the number of items each patient was given, (2) the magnitude of the expected difference (1.00 - .50 = .50, according to Grodzinsky et al., 1999), and (3) the acceptable probability of Type I error (which was set at $\alpha = .05$, unidirectional). In light of these considerations, failure to reject the null hypothesis (that actives do not differ from passives) will be assumed only if the power of the analysis was at least .90, corresponding to a 0.10 Type II error risk.

Subjects from Grodzinsky et al.'s (1999) sample (S1). Results are shown in Table 4. For 8/26 subjects (30.7%) actives were not significantly better than passives. Seven of these patients are also included in the S2 sample.

TABLE 6

Study of the Confidence Limits of the Observed Correct Passive Performance (Hits) of Three Patients from the Original Sample of Grodzinsky et al. (1999)

| Case | Passive hits (percentage) | Confidence limits | | |
|------|------------------------------|--------------------|---------------------|------------------------|
| | | ($\alpha = .05$) | ($\alpha = .001$) | ($\alpha = .000001$) |
| POE | 20/20 (1.000) | .846–1.000 | .649–1.000 | .613–1.000 |
| ME | 45/48 (.937) | .830–.983 | .749–.999 | .668–.999 |
| PJ | 67/72 (.931) | .851–.972 | .784–.989 | .723–.996 |

Note. We adopt different protection levels, setting Type I error risk, respectively, at (i) .05 for individual patients, (ii) .001 for individual patients, and .05 for the whole group of 42 patients in the original sample size of Grodzinsky et al. (1999), (iii) .000001 for individual patients, and .00005 for the whole group of 42 patients. For further statistical details see text.

Subjects from the extended sample (S2). For the extended sample, we cannot reject the null hypothesis in 20 cases of 49 (40.8%). Thus, even with the very stringent measures taken to safeguard against inappropriate acceptance of the null hypothesis, for a substantial proportion of both samples there was not a significant difference between active and passive sentence performance.

3.3.3 Distribution of the Individual Percentages of Correct Passive Responses

The hypothesis that comprehension of passive sentences in Broca's aphasia is systematically at chance can also be tested at the "group" level (but see section 2). As discussed above, the approach followed by Grodzinsky et al. (1999) is not appropriate, both because of the composite nature of the group distribution and because Grodzinsky et al. (1999) did not compare the distributions statistically. We used the approach to comparing distributions set forth above, in which z scores can be directly compared with the normal distribution tables. Figure 6 reports the distribution of these z scores and the corresponding normal distributions, and Table 7 shows the statistical comparison between the observed distributions and the normal distribution corresponding to a binomial with mean = .50 and the appropriate sample size (26 for S1 and 49 for S2). For both samples we can confidently reject the null hypothesis of a chance distribution about the expected mean of .50.

3.3.4 Relationship Between Actives and Passives

Grodzinsky et al.'s (1999) research hypothesis implies that the proportion of correct responses to active and passive sentences should be uncorrelated, and this can be easily tested statistically. This is the only case where the original data reported by Grodzinsky et al. (1999) could be used to test this hypothesis, since the number of trials is not required for this analysis.

Subjects from Grodzinsky et al.'s (1999) full sample (N = 42). The Pearson correlation between active and passive hits was .339. The regression analysis is reported in Table 8a. The percentage of correct actives is significantly predictive of the percentage of correct passives, and the square of the actives' percentage is a better predictor than the linear term. This means that a parabola better interprets the relation between passives and actives. Accordingly, we calculated the minimum of this parabola, and the value of actives for which passives are minimal is 69.1. In other words,

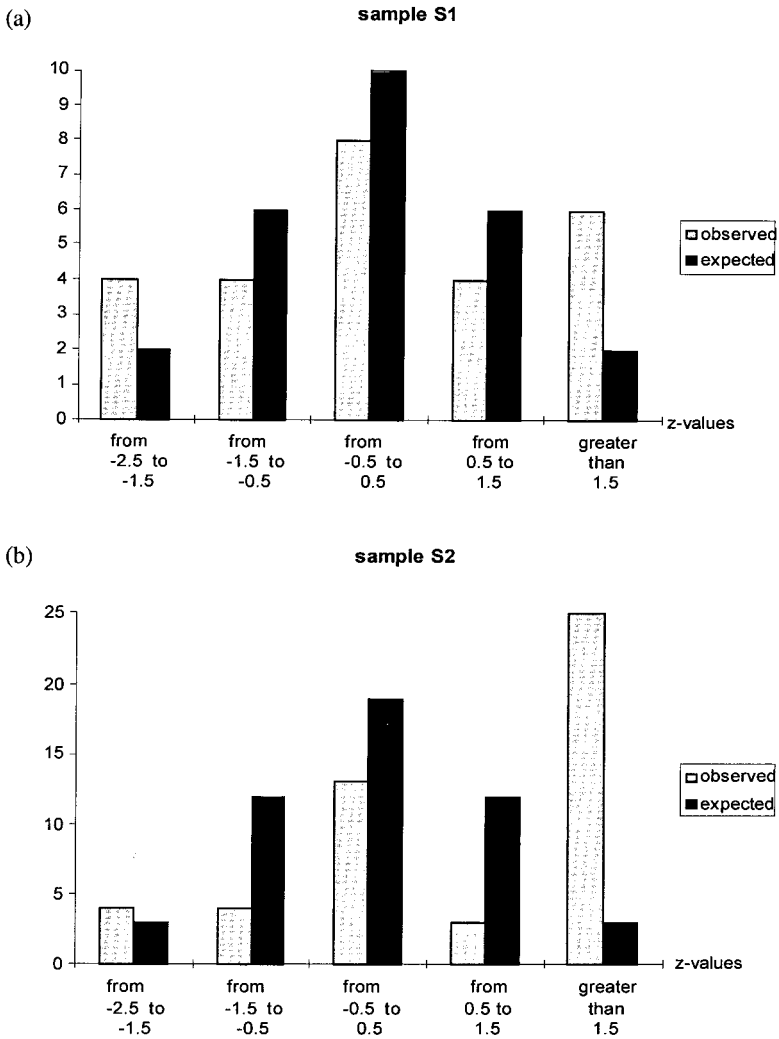


FIG. 6. Distribution of the observed deviations from the frequencies expected by chance. For further details see statistical methods and Results. (a) Sample S1; (b) sample S2.

TABLE 7

Comparison between the Frequencies Observed in Samples S1 and S2 and the Frequencies Observed in the *z* Approximation to the Binomial Distribution with Mean .50 (See Statistical Methods)

| Z Interval | -2.5/-1.5 | -1.5/-0.5 | -0.5/+0.5 | +0.5/+1.5 | >1.5 |
|--|-----------|-----------|-----------|-----------|------|
| (a) Goodness of fit for sample S1 (26 subjects) ($\chi^2 = 11.374, df = 4, p = .023$) | | | | | |
| Observed | 4 | 4 | 8 | 4 | 6 |
| Expected | 2 | 6 | 10 | 6 | 2 |
| (b) Goodness of fit for sample S2 (49 subjects) ($\chi^2 = 175.644, df = 4, p < .001$) | | | | | |
| Observed | 4 | 4 | 13 | 3 | 25 |
| Expected | 3 | 12 | 19 | 12 | 3 |

The number of cells has been established in order to have expected frequency values of at least 2.

TABLE 8
Regression Analysis of the Relationship between
the Percentage of Correct Passives (Dependent
Variable) and of Correct Actives (Predictor)

| Source of variability | <i>F</i> (<i>df</i> 1, 40) | <i>p</i> |
|---|-----------------------------|----------|
| (a) Original Grodzinsky et al.'s (1999) sample (42 subjects) (Regression equation: passives = 149.661 - 3.002 actives + .02171 actives ²) | | |
| Linear | 5.186 | .028 |
| Quadratic | 5.916 | .020 |
| (b) Sample S2 (49 subjects) (Regression equation: passives = 158.575 - 3.6845 actives + .02962 actives ²) | | |
| Linear | 12.565 | .0009 |
| Quadratic | 19.425 | .00006 |

subjects with a moderate–severe deficit for actives are worse on passive comprehension than are those with a very severe deficit for actives. The basis for the latter effect is not obvious. It may be that patients with more severe deficits on active voice were truly guessing for both constructions, while the patients with somewhat better performance on actives are attempting to employ a strategy for the passive sentences that systematically resulted in incorrect responses. However, this possibility would have to be examined carefully in individual patients' performances.

Subjects from sample S2. The same outcome is observed in this sample (see Table 8b). In this case the correlation between actives and passives is stronger (.459). The significance of the quadratic term is higher than that of the linear term, and the minimum of the parabola (i.e., of the passives) corresponds to a percentage of correct actives of 62.2. The similarity of the latter result between the two samples is striking, given that the overlap between Grodzinsky et al.'s (1999) sample ($N = 42$) and the S2 sample ($N = 49$) was only 16 subjects. We can conclude that Grodzinsky et al.'s (1999) implicit hypothesis of independence between active and passive performance is not supported by the results either for their sample of patients or for the extended sample.

4. DISCUSSION AND CONCLUSION

The analyses presented in section 3 appear to falsify the hypothesis that patients classified as Broca's aphasics exhibit a single, systematic pattern of performance. Statistical tests of the performance of individual cases and of the distribution of group scores fail to support the hypothesis. It is clear from these analyses that a substantial number of patients fail to show the required performance pattern. This is true even when we limited the analysis to patients identified by Grodzinsky et al. (1999) as "true" Broca's aphasics.

We have gone to some lengths to explain why the analyses set forth here are the correct tests of Grodzinsky et al.'s (1999) hypothesis and in what ways the analyses they performed are flawed. Here we will consider more closely the potential counterargument that, despite our care in selecting patient data, the nonconforming patients are not truly members of the class at issue, i.e., Broca's aphasics (see footnote 1). Aside from the invocation of uncertainty in determining chance performance (which has been discussed at length here), this has been the primary response to data describing nonconforming patients.

This type of argument—that nonconforming cases must not be "true" cases of

Broca's aphasia—is an example of reasoning that has been treated extensively in the philosophy of science as part of the topic of how one selects “the class of potential falsifiers.” Popper (1983) clearly described the problems encountered when this class is not explicitly set forth in the formulation of a hypothesis. The following example concerns the hypothesis “All swans are white”:

... This statement is falsifiable. Suppose, however, that there is someone who, when a non-white swan is shown to him, takes the position that it cannot be a swan, since it is ‘essential’ for a swan to be white. Such a position amounts to holding non-white swans as logically impossible structures (and thus also as unobservable). It excludes them from the class of potential falsifiers. Relative to this *altered* class of potential falsifiers the statement ‘All swans are white’ is of course unfalsifiable. In order to avoid such a move, we can demand that anyone who advocates the empirical-scientific character of a theory must be able to specify under what conditions he would be prepared to regard it as falsified. (p. xxi, emphasis in original)

Grodzinsky and colleagues have repeatedly altered the class of potential falsifiers for their hypothesis by essentially arguing that patients whose comprehension does not conform to their hypothesis must be excluded from the class of potential falsifiers (see footnote 9), thus rendering the thesis unfalsifiable. If the hypothesis is going to be regarded as part of the scientific endeavor to understand brain/language relationships, the next step is clearly to reformulate the hypothesis in such a way that it regains its falsifiability. In its current formulation, however, it has been falsified: agrammatic Broca's aphasia is not associated with a single pattern of comprehension performance.

REFERENCES

- Armitage, P. 1971. *Statistical methods in medical research*. Oxford: Blackwell.
- Benedet, M. J., Christiansen, J. A., & Goodglass, H. 1998. A cross-linguistic study of grammatical morphology in Spanish- and English-speaking agrammatic patients. *Cortex*, **34**, 309–336.
- Berndt, R. S. 1991. Sentence processing in aphasia. In M. T. Sarno (Ed.), *Acquired aphasia*. 2nd ed. New York: Academic Press. Pp. 112–270.
- Berndt, R. S., & Caramazza, A. 1999. How “regular” is sentence comprehension in Broca's aphasia? It depends on how you select the patients. *Brain and Language*, **67**, 242–247.
- Berndt, R. S., Mitchum, C. C., & Haendiges, A. N. 1996. Comprehension of reversible sentences in “agrammatism”: A meta-analysis. *Cognition*, **58**, 289–308.
- Berndt, R. S., Mitchum, C. C., & Wayland, S. 1997. Patterns of sentence comprehension in aphasia: A consideration of three hypotheses. *Brain and Language*, **60**, 197–221.
- Beretta, A., Harford, C., Patterson, J., & Piñango, M. 1996. The derivation of postverbal subjects: Evidence from agrammatic aphasia. *Natural Language and Linguistic Theory*, **14**, 725–748.
- Blyth, C., & Still, H. 1983. Binomial confidence intervals. *Journal of the American Statistical Association*, **78**, 108–116.
- 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., Berndt, R. S., Basili, A. G., & Koller, J. J. 1981. Syntactic processing deficits in aphasia. *Cortex*, **17**, 333–348.
- Caramazza, A., & Zurif, E. B. 1976. Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language*, **3**, 572–582.
- Casella, G. 1987. Refining binomial confidence intervals. *Canadian Journal of Statistics*, **14**, 113–129.
- Druks, J., & Marshall, J. C. 1991. Agrammatism: An analysis and critique, with new evidence from four Hebrew-speaking aphasic patients. *Cognitive Neuropsychology*, **8**, 415–433.
- Geschwind, N. 1965. Disconnexion syndromes in animals and man. *Brain*, **88**, 237–294, 585–644.
- Grodzinsky, Y. 1985. Trace deletion, theta roles, and cognitive strategies. *Brain and Language*, **51**, 469–497.
- Grodzinsky, Y. 1986. Language deficits and the theory of syntax. *Brain and Language*, **27**, 135–159.

- Grodzinsky, Y. 1989. Agrammatic comprehension of relative clauses. *Brain and Language*, **37**, 480–499.
- Grodzinsky, Y. 1990. *Theoretical perspectives on language deficits*. Cambridge, MA: MIT Press.
- Grodzinsky, Y. The neurology of syntax: Language use without Broca's area. *Behavioral and Brain Sciences*. [in press]
- Grodzinsky, Y., Pearce, A., & Marakovitz, S. 1991. Neuropsychological reasons for a transformational analysis of verbal passive. *Natural Language and Linguistic Theory*, **9**, 431–453.
- Grodzinsky, Y., Piñango, M. M., Zurif, E., & Drai, D. 1999. The critical role of group studies in neuropsychology: Comprehension regularities in Broca's aphasia. *Brain and Language*, **67**, 134–147.
- Grodzinsky, Y., Zurif, E., & Swinney, D. (1985). Agrammatism: Structural characteristics and processing antecedents. In M. L. Kean (Ed.), *Agrammatism*. New York: Wiley Publishing Co.
- Hagiwara, H. 1993. The breakdown of Japanese passives and theta-role assignment principle by Broca's aphasics. *Brain and Language*, **45**, 318–339.
- Heilman, K. M., & Scholes, R. J. 1976. The nature of comprehension errors in Broca's conduction and Wernicke's aphasics. *Cortex*, **12**, 258–265.
- Hickok, G., & Avrutin, S. 1995. Representation, referentiality, and processing in agrammatic comprehension: Two case studies. *Brain and Language*, **50**, 10–26.
- Hickok, G., Zurif, E., & Canseco-Gonzalez, E. 1993. Structural description of agrammatic comprehension. *Brain and Language*, **45**, 371–395.
- Kolk, H. H., & Van Grunsven, M. 1985. Agrammatism as a variable phenomenon. *Cognitive Neuropsychology*, **2**, 347–384.
- Kolk, H. H., Van Grunsven, M., & Keyser, A. 1985. On the parallelism between production and comprehension in agrammatism. In M. L. Kean (Ed.), *Agrammatism*. New York: Academic Press. Pp. 165–206.
- Martin, R. C. 1987. Articulatory and phonological deficits in short-term memory and their relation to syntactic processing. *Brain and Language*, **32**, 159–192.
- Martin, R. C., Wetzel, W. F., Blossom-Stach, C., & Feher, E. 1989. Syntactic loss versus processing deficit: An assessment of two theories of agrammatism and syntactic comprehension deficits. *Cognition*, **32**, 157–191.
- Mehta, C., & Patel, N. 1998. StatXact 4 for Windows. Cytel Software Corp., Cambridge, MA.
- McNemar, Q. 1962. *Psychological statistics*. New York: Wiley.
- Miceli, G., Mazzucchi, A., Menn, L., & Goodglass, H. 1983. Contrasting cases of Italian agrammatic aphasia without comprehension disorder. *Brain and Language*, **19**, 65–97.
- Popper, K. R. 1983. Realism and the aim of science. In W. W. Barley III (Ed.), *Postscript to the logic of scientific discovery*. London: Routledge.
- Schwartz, M. F., Saffran, E. M., & Martin, O. S. M. 1980. The word order problem in agrammatism: Comprehension. *Brain and Language*, **10**, 249–262.
- Zurif, E., Swinney, D., Prather, P., Solomon, J., & Bushell, C. (1993). An on-line analysis of syntactic processing in Broca's and Wernicke's aphasia. *Brain and Language*, **45**, 448–464.