BRAIN AND LANGUAGE 43, 519–527 (1992)

Not Everything Is the Same: Some Things Are Worse Than Others

A Response to Tesak

ARGYE E. HILLIS AND ALFONSO CARAMAZZA

Johns Hopkins University

In his paper "Everything is the same: A note on Caramazza and Hillis (1989) The disruption of sentence production: Some dissociations," Tesak (this volume) has raised a number of issues concerning the use of data to test models of cognitive processing. He takes issue with the methodology we have chosen for informing models of the cognitive processes underlying specific language tasks. In particular, Tesak has found lacking in our studies a systematic method of determining which differences in levels of performance are important differences. He also questions the model of sentence processing we adopted in order to explain the pattern of performance in sentence production tasks by a patient, ML, in our 1989 paper. He goes on to offer an alternative hypothesis that he believes provides a more satisfactory account of the data we presented. In this response, we recount the crucial issues that Tesak has raised and explore the extent to which we have dealt with these issues appropriately in the studies he has cited. We also look briefly at the adequacy of Tesak's alternative hypothesis in providing a solution to the problems we have faced in this research.

The central point in Tesak's criticism is that in our work we have not specified a priori criteria for determining whether an observed variation in performance is important. This issue is obviously not specific to case studies of brain-damaged patients nor to neuropsychological research more generally; it is a problem common to virtually all research. This issue has, therefore, received a great deal of attention in the scientific literature,

The writing of this paper was supported in part by NIH Grant NS 22201. We thank Brenda Rapp for helpful comments on an earlier version of the paper. Address correspondence and reprint requests to Alfonso Caramazza, Department of Cognitive Science, Johns Hopkins University, Baltimore, MD 21218.

0093-934X/92 $5.00

TABLE 1
SPELLING ERRORS AS A FUNCTION OF LENGTH IN FIVE TASKS: M.L.

Letter length	Written naming	Writing to dictation		Delayed copy	Oral spelling
		Words	Nonwords		
3–4	4/17 (23.5)	34/75 (45.3)	7/12 (58.3)	8/15 (53.3)	1/7 (14.3)
5	12/22 (54.5)	95/123 (77.2)	4/7 (57.1)	26/41 (63.4)	7/19 (36.8)
6	11/12 (91.7)	69/86 (80.2)	6/9 (66.7)	50/59 (84.7)	18/28 (64.3)
7–8	—	40/42 (95.2)	4/6 (66.7)	11/11 (100)	8/8 (100)

and we do not propose to add anything to that discussion here. There is, of course, no simple solution; whether a difference is considered to be important depends on the theories we are willing to contemplate. Thus, we have no objective, "cookbook" approach for making such decisions. As an initial screening, we can use statistical tests to determine with what level of confidence we can assert that a variation is not just due to chance. But not every statistical difference is an important difference in testing a particular hypothesis, and not every difference that falls short of statistical significance is necessarily unimportant. We will elaborate on this relationship between data and theory by addressing it with respect to the work Tesak has criticized. Our discussion follows the order of Tesak's paper—we discuss in turn the data, the explanation, and the model (and the alternative)—keeping in mind, nevertheless, that these topics are not independent.

Tesak begins his critical analysis by showing that we take differences in accuracy levels between different types of stimuli or different tasks to be "very similar" even though smaller absolute differences between percentage levels are considered substantial in testing other models in other papers.[1] He illustrates this apparent inconsistency with data from patient ML, whose impaired spelling performance, characterized by phonologically implausible spelling errors that increased with increased length of words in letters in all spelling tasks, was interpreted as resulting from selective damage within the spelling system to the graphemic buffer (Hillis & Caramazza, 1989). We described ML's rate of spelling errors as "very similar across all spelling *tasks* (dictation, delayed copying, and naming) and . . . uniformly characterized by a substantial length effect" (new emphasis added). Indeed, we see in Table 1 (which we reproduce here

[1] This is an unusual criticism reflecting a lack of appreciation of the problem of measurement in the context of specific tasks and theories. Thus it is commonplace that a 10-msec difference may be an important difference in one specific context but not in another. Similarly there are contexts where a much larger difference, say 100 msec, may be unimportant.

from the original paper because we believe that the data omitted by Tesak are crucial to our conclusions) that in each task there is a substantial length effect. Although the conclusion he cited concerned this similarity between *tasks,* Tesak highlighted that there seems to be a substantial difference between different types of *stimuli* (words vs. nonwords). Thus, he claims that we have ignored the fact that there is only a 7% increase in errors in written spelling of dictated nonwords, compared to an 86% increase in errors in oral spelling (of words) across stimulus length. Did we ignore a difference that was pivotal to our conclusion? The answer is negative from two viewpoints.

First, although there was a large number of stimuli in each of the spelling tasks with words on which we drew our conclusion, there was a very small number of nonword stimuli, reported only for written spelling to dictation. Therefore, there was not a statistical difference between nonwords and words in any of the tasks. One can decide whether a trend revealed in a given task (or given set of stimuli) reflects the same trend observed in other tasks (or other sets of stimuli) by comparing it to the overall trend and using a χ^2 test for best fit to determine whether the difference exceeds that expected on the basis of random variation. In this example, the overall increase in spelling errors (for all tasks combined) across the word lengths was 43, 68, 79, and 94%. Therefore, if spelling nonwords to dictation followed this trend, we would expect about the same percentages of errors for each of the word lengths. For the number of stimuli presented, we would expect the following number of incorrect responses for each of the four nonword lengths in order: 5.2 (out of 12), 4.8 (out of 7), 7.1 (out of 9), and 5.6 (out of 6). The observed number of incorrect responses differed from the expected number by less than two responses for each nonword length (nonsignificant). Although we must admit that the small number of stimuli precludes appropriate use of a χ^2 test for best fit, certainly it is not possible to conclude that the difference between the trend observed for spelling nonwords to dictation is different from the trend observed for spelling words in any of the tasks. Thus, Tesak's suggestion that we ignored a substantial difference is unwarranted.[2]

But could there be an important difference between nonwords and words that simply did not reach statistical difference because the number of stimuli presented was too small? Of course, there could be a reliable difference between the two types of stimuli. However, even if such a difference were reliable it would not undermine the hypothesis that the patient has a deficit at the level of the Graphemic Buffer.[3] A possible

[2] Also, more generally, it is inappropriate when comparing trends across sets of data to select the two extremes for comparison, as Tesak did in his discussion of these data.

[3] Our model of the Graphemic Buffer does not predict that length effects should be identical for words and nonwords. What it does predict is that there should be a length

account for a difference in length effects for words and nonwords is that longer nonwords are spelled as though they are two (or more) short nonwords. So, for example, an eight-letter nonword like "mushrame" might be spelled by computing separate representations for "mush" (/maš/) then "rame" (/rem/). Thus, separate, shorter, representations would be held in the Graphemic Buffer in the course of spelling the eight-letter nonword; and we would expect (on the hypothesis that spelling errors increase as a function of the length in letters of representations held in the Graphemic Buffer) that the error rate on long nonwords would approximate that of spelling short (e.g., four-letter) words. We did not use a sufficient number of stimuli to adequately test this prediction, however, since we cannot control the "strategy" used for parsing nonwords into separate representations for spelling. Although the same parsing "strategy" could be used for spelling some longer words as well (e.g., mushroom), this strategy would not work for spelling many of the words presented as stimuli (e.g., surprise; schedule). Hence, the difference between long words and long nonwords probably is important in terms of the cognitive processes involved in spelling, but this difference was not crucial to our account of ML's pattern of performance (as a selective impairment within the spelling process to the Graphemic Buffer), nor to our proposed role of the Graphemic Buffer in the spelling process. What was crucial to our account and to our model was the similarity between tasks (oral and written spelling to dictation, written naming), since in our model the Graphemic Buffer is used to the same extent in all of these tasks for any given stimulus.

In his criticism of our work, Tesak points out that while we ignored the difference between length effects for words and nonwords spelled by ML, we described as "significant" a contrast in accuracy for high vs. low frequency words of 74% (109/146) vs. 49% (72/146), respectively (for a patient, DH, who we also proposed had selective damage to the Graphemic Buffer). Why was this difference considered significant? First, the difference in percentages with such a large number of stimuli cannot simply be due to chance ($\chi^2 = 19.9$; $df = 1$; $p < .00001$). Second, but of no less importance, this difference is not easily accommodated by our proposal of functional damage within the spelling process to the Graphemic Buffer. That is, this difference presented more of a problem to our explanation, since our model would predict that high and low frequency

effect for the representations held in the buffer. If the size of representations placed in the buffer were to differ for words and nonwords, then, we would not expect similar-sized effects for the two types of stimuli. Thus, it could be the case that the units of representations placed in the buffer for nonwords might correspond to syllables, whereas for words they might correspond to the whole word or morphemes (see Badecker, Hillis, & Caramazza, 1990).

words should be equally sensitive to damage to the Graphemic Buffer. Thus, this difference in accuracy for DH, unlike the difference in length effect between words and nonwords for ML, was statistically significant *and* relevant to the model we were testing. The point here is simple and straightforward: theoretical considerations dictate whether certain differences are deemed important.

What of the difference between various sentence production tasks reported for ML in our 1989 paper? Clearly, reading, repeating, writing, and conversing are very different tasks with quite different demands on word retrieval, memory load, and so on. Tesak delineates a number of points of divergence in these tasks, which may well account for variation in the absolute rates of errors across tasks. The variations in pattern of performance can be addressed by asking two questions: Are the differences between tasks in ML's performance *explained by* our proposal of damage to the positional level of sentence production in Garrett's model? Are the differences *consistent with* this proposal? The answer to the first question is certainly "no." In the 1989 paper we emphasized the inadequacy of Garrett's model in accounting for ML's pattern of performance, as reflected in the following excerpts: "Although it is possible to plausibly argue that M.L.'s pattern of impairment is consistent with the hypothesis that she has a functional lesion to the positional level of representation, it must be pointed out that this claim fails to account for some important features of our patient's performance. Furthermore, the theoretical framework we have adopted as a basis for M.L.'s performance is unsatisfactory in a number of respects" (p. 640–641); "Two sorts of problems vitiate the possibility for strong conclusions about the nature of the deficit responsible for our patient's performance, and therefore, undermine the possibility for strong claims about the processing structure of the language production system. One problem concerns the relatively undeveloped nature of the theoretical model guiding our interpretation of the data" (p. 643).[4]

Tesak echoes our observation that "the model is inadequate." He goes on to claim that the inadequacy of Garrett's model of sentence production to account for the variation between tasks serves as a basis for challenging our interpretation of ML's performance. However, he does not show that any aspect of ML's performance is inconsistent with the model or our proposal of a functional lesion within this model. Because the model does not allow predictions about variations in performance as a function of memory load, word retrieval, or other dimensions on which the different sentence production tasks vary, we would argue that her pattern of performance, while not entirely explained by the model, is nevertheless con-

[4] It is unclear why Tesak failed to mention in his critique our reservations about the adequacy of Garrett's model to fully account for ML's performance.

sistent with the model. That is, similarity between tasks in terms of absolute rates of errors or distribution of different types of morpheme errors is not relevant to the hypothesis we were testing. Rather, the crucial evidence for our hypothesis of a selective deficit to the positional level of sentence processing consists of a characteristic impairment (described below) in all tasks that involve positional level of processing in sentence production together with sparing of tasks that do not involve this level of processing. The pattern of performance that is predicted by assuming functional damage to the positional level of processing in Garrett's model is characterized by disruption of: (1) the phrasal geometry of the sentence to be produced and (2) the specification of grammatical morphemes to be inserted in specific sites within the sentence frame. A disruption in these two aspects of sentence form are clearly reflected in ML's performance in reading, repeating, and writing sentences and in producing sentences in story retelling and spontaneous speech. And, consistent with damage specific to this level of processing, ML showed spared comprehension of sentences. Thus, ML's performance provides evidence for proposing that reading, writing, repetition, and spontaneous speech all require a level of processing in which a sentence frame is specified, along with the grammatical morphemes in specific sites of this frame—the "positional" level in Garrett's model of sentence processing.

Tesak argues that Garrett's model, because it is based on speech error data from normal subjects, is only a model of "sentence generation in free conversation," and that we misuse the model by applying it to tasks of repetition, writing, and so on. He seems to be saying that the type of data used to formulate a model dictate its explanatory scope, i.e., that a model of language processing can only explain the type of data on which it was originally based. On this reasoning, cognitive neuropsychology would lay claim to models of producing paraphasias, models of lexical decision, and models of priming, but few, if any, models of normal language processing! We hope that patterns of impaired performance do not reflect novel cognitive mechanisms that result from brain damage, but instead reflect specific deformations of normal cognitive processing. And we surely do not want to suppose that the representations and transformations involved in producing sentences in story retelling are entirely different from those representations and transformations involved in producing sentences in free conversation. Furthermore, although it is conceivable that there are distinct processes of speaking, reading, repeating, and writing sentences, each with an independent mechanism for specifying sentence frames and the location of morphemes within the sentence frames for that specific task, it is our hypothesis that there is a single mechanism for this function in all tasks that require it. This hypothesis could be wrong, but it is not logically excluded. Thus, its validity is an empirical matter that cannot be decided by mere assertion as done by Tesak. Also,

ML's performance, which showed deficits in this function across all tasks of sentence production, provides evidence favoring the hypothesis of a single mechanism for the generation of positional frames.

Of course, the hypothesis that the cognitive processes underlying reading, writing, repeating, and speaking share a common mechanism for specifying phrasal geometry and grammatical morphemes does not deny that there are points at which the tasks diverge. As Tesak says, "speaking and writing are different." For one thing, speaking requires articulation, whereas writing requires upper extremity control. Tesak correctly points to several other differences (e.g., speed, memory load, etc.) that may well account for differences between absolute rates of errors in various sentence production tasks (including the possibility differential damage to mechanisms that are specific to particular tasks). However, he suggests that these differences, while unexplained by the hypothesis of damage to the positional level of processing, are explained by proposing a "reduced capacity for linguistic computations." He claims that ML's performance gets worse in the order predicted by the hypothesis of reduced computational capacity: reading, repeating, writing, and speaking. But even if we assume that the variation in error rates across tasks reflects differences in the computational demands of the tasks (as do variations in error rates of normal subjects), we are left with the question of why ML omits closed class morphemes selectively in all tasks that involve producing sentences.

Tesak gives two reasons for selective difficulty with closed class items. First, "since the positional level follows the functional level, the later level is more affected, if resources for representation are restricted." This explanation is curious, since he previously claimed that Garrett's model is irrelevant to reading, repeating, and so on. Also, it is not clear why later levels should be more affected. Is it because the subject has "used up" all of her computational resources on previous levels? Then, does this claim boil down to her difficulty being only at the positional level of processing? Or shouldn't she have even more difficulty at even later levels, say in articulatory or motor output, on this reasoning? Tesak's second explanation for her selective difficulty with closed class items is that "closed class items are retrieved differently." He claims that they are retrieved more slowly and less automatically (by aphasic patients?) in comparison with normals. But this observation does not explain anything. Why are they retrieved differently? What is it about these words that results in slow processing by some individuals and results in omission by ML in reading, repeating, and writing sentences, although she has no difficulty producing them individually in reading, repeating, and writing?

Tesak also claims that the reduced capacity hypothesis accounts for ML's strikingly reduced phrase length because "only short structures fit into a reduced computational system." But other so-called agrammatic patients who have been reported (e.g., by Miceli, Mazzuchi, Menn, &

Goodglass, 1983; Berndt, 1987; and Nespoulous, Dordain, Perron, Ska, Bub, Caplan, Mehler, & Lecours, 1988) do not show this reduced phrase length. Persumably, Tesak would respond that such patients have a computational capacity that is not as reduced as that of ML. Yet, they show similar patterns with respect to the proportion of closed class words omitted.

What patterns of performance would be inconsistent with the "reduced computational capacity" hypothesis? One might have guessed that selective impairment in production with spared comprehension would be inconsistent with such a hypothesis, but Tesak argues that "reduced computational capacity for production" explains patterns of performance like that of ML. If we found a patient who was impaired only in producing grammatical sentences in writing, would he account for it by proposing "reduced computational processing for written production?" It seems to us that the "reduced capacity hypothesis" is not wrong, but is merely too powerful to be tested. Like the proposal of a "noisy system" it can be invoked to explain just about any pattern of performance: an unspecified mechanism accounts in an unspecified way for an unspecified range of facts. Thus, Tesak has not really offered an *alternative* hypothesis to one we proposed since his "reduced capacity" account can "explain" everything and, hence, nothing.

In short, we remain dismayed by our inability to provide a satisfactory account of *all* aspects of impaired sentence production by our patient, ML; but we maintain that the data we presented can be best accounted for by proposing damage to the positional level of processing in Garrett's model of sentence production. We also maintain that this study illustrates appropriate use of patterns of performance, including trends observed across language tasks that share a particular component of processing and dissociations observed between those tasks that do and those that do not share a given component of processing, to provide evidence favoring a particular model of normal language processing.

REFERENCES

Badecker, W., Hillis, A. E., & Caramazza, A. 1990. Lexical morphology and its role in the writing process: Evidence from a case of acquired dysgraphia. *Cognition,* **35,** 205–243.

Berndt, R. S. 1987. Symptom co-occurrence and dissociation in the interpretation of agrammatism. In M. Coltheart, G. Sartori, & R. Job (Eds.), *The cognitive neuropsychology of language.* London: Erlbaum.

Caramazza, A., & Hillis, A. E. 1989. The disruption of sentence production: Some dissociations. *Brain and Language,* **35,** 625–650.

Hillis, A., & Caramazza, A. 1989. The Graphemic Buffer and attentional mechanisms. *Brain and Language,* **36,** 208–235.

Miceli, G., Mazzucchi, A., Menn, L., & Goodglass, H. 1983. Contrasting cases of Italian agrammatic aphasia without comprehension disorder. *Brain and Language,* **33,** 273–295.

Nespoulous, J-L., Dordain, M., Peron, C., Ska, B., Bub, D., Caplan, D., Mehler, J., & Lecours, A. R. 1988. Agrammatism in sentence production without comprehension deficits: Reduced availability of syntactic structures and/or grammatical morphemes? *Brain and Language,* **33,** 273–295.

Tesak, J. 1992. Everything is the same: A note on Caramazza and Hillis (1989) "The disruption of sentence production: Some dissociations." *Brain and Language,* **43,** 512–518.