Parallel Distributed Processing Challenges the Strong Modularity Hypothesis, Not the Locality Assumption [Commentary on M. J. Farah, Neuropsychological inference with an interactive brain: A critique of the "locality" assumption]. *Behavioral and Brain Sciences, 17*, 77–78.

David C. Plaut
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, 15213–3890.
Electronic mail: *plaut@cmu.edu*

Farah presents convincing arguments that principles of Parallel Distributed Processing (PDP) provide more parsimonious explanations of a number of neuropsychological phenomena than do traditional, modular accounts. She ascribes this to the fact that PDP systems violate the "locality assumption" in that damage within an interactive network can have nonlocal effects. However, on closer inspection, Farah has misinterpreted the locality assumption as it has been formulated in the literature. Furthermore, Farah's version of the locality assumption is violated by modular systems as well, and thus does not provide a useful basis for distinguishing PDP and traditional accounts. Rather, this contrast is better understood at a more general level, in terms of a rejection of the strong modularity hypothesis (e.g., Fodor, 1983). In particular, the most fundamental contribution of PDP modeling to neuropsychology is that it enables a principled expression of nonmodular, interactive computation in which it is tractable to analyze the effects of damage.

The strong modularity hypothesis claims that the cognitive system is composed of informationally encapsulated components that receive input from only a few other components and produce all-or-none output in discrete stages. Information encapsulation entails that the knowledge required for a process is available only to the component dedicated to that process, and that any partial results of the process are unavailable to other components. Each of the central properties of PDP systems that Farah lists (section 1.4, paragraph 2, dpp. 8–9) contrasts in a specific way with these properties of modular systems: (1) the knowledge involved in a process is *distributed* in connection weights throughout the network rather than being localized to a particular component; (2) processing is *graded* and continuous rather than being staged and all-or-none, making partial results in one part of the network continually available to other parts; and (3) groups of units representing different types of information are highly *interactive*, instead of receiving inputs from only a few other components. Furthermore, while the strong modularity hypothesis says little about the nature of processing within each component, PDP systems employ a common set of computational principles both within and between groups of units: processing takes the form of graded interactions among distributed patterns of activity.

In arguing that PDP systems provide better accounts of neuropsychological data than do modular systems, Farah focuses on what she terms the "locality assumption." She interprets the locality assumption as implying that the effects of damage are local; that "nondamaged components will continue to function normally" (section 1.1, paragraph 1, dp. 2). This statement can be interpreted in two ways: (1) nondamaged components *behave* normally, in that their output to other components is unchanged, or (2) nondamaged components *compute* normally, in that their input/output function is unchanged. Farah clearly has interpretation (1) in mind. The central claim of Farah's target article is that the advantage of PDP accounts stems specifically from the occurrence of nonlocal effects of damage within these systems, in violation of the locality assumption. Yet clearly the nondamaged portions of the networks compute normally but behave abnormally in response to corrupted input from damaged portions.

However, the same is true of nondamaged components which receive input, either directly or indirectly, from a damaged component in a modular system. For instance, in De Haan, Bauer, and Greve's (1992) model of face processing (see Figure 10, section 2.3.2, paragraph 1), no one would claim that the "Response effector systems" would continue to function normally if, say, "Structural encoding" were impaired. In fact, Farah acknowledges this point, stating that the effects of damage in such a system are confined to "lesioned components *and the relatively small number of components downstream*" (section 1.3, paragraph 1, dp. 6, emphasis added). That damage alters the behavior of intact components downstream clearly violates Farah's interpretation of the locality assumption. But note that the claim that there are relatively few such components does not come from the locality assumption *per se*, but rather from more general assumptions about modular systems: processing tends to be feedforward (staged) and each component receives few inputs. In this regard, PDP systems differ from modular ones only in a quantitative way: portions of a PDP network typically receive a wider range of input than do components in a modular architecture. However, in both frameworks, nondamaged processes will be affected by corrupted input from damaged processes, and thus PDP and modular systems are on equal footing with respect to Farah's interpretation of the locality assumption.

In fact, in the neuropsychological literature, what corresponds most closely to Farah's interpretation of the locality assumption (1 above) is Caramazza's (1986) "transparency assumption," according to which it must be possible to relate the effects of damage on the *behavior* of the cognitive system to its normal operation in a principled way. This relationship had to be "transparent" in earlier formulations (Caramazza, 1986), but merely "tractable within the proposed theoretical frameworks" in later ones (Caramazza, 1992, p. 82). The locality assumption was originally formulated in the context of the transparency assumption, where it corresponds most closely to interpretation (2) above—that the damage itself must be local.

> My formulation of the transparency assumption implies that [neuropsychological evidence] E_i can only be related to [a cognitive model] M when the damage to the system is "local." This assumption may be too strong as an *in principle* claim—nonlocal, very general modifications of the system may still allow the possibility of relating E_i to M. However, *in practice*, given the tremendous complexity of the systems we are dealing with, it may *only* be possible to draw meaningful conclusions from impaired performance to normal cognitive systems under a restricted sort of condition. [Caramazza, 1986, p. 52, emphasis in the original]

Thus, the standard locality assumption is simply a way to ensure that the transparency assumption is tractable, and the transparency assumption is simply a way to ensure that the effects of damage are interpretable. In this light, PDP modeling in neuropsychology is important, not because it is "not constrained by the locality assumption" (section 1.3, paragraph 5, dp. 7) as Farah contends, but rather because it provides a rich, nonmodular theoretical framework in which it is nonetheless possible to relate normal and impaired behavior in a principled way.

Critically, all of the advantages of Farah's PDP accounts can be most naturally understood as arising from ways in which the nature of computation in these systems violates various aspects of the strong modularity hypothesis. In both the simulation of semantic memory impairments (Farah & McClelland, 1991), and the simulation of impaired attentional allocation (Cohen, Romero, Servan-Schreiber, & Farah, 1994), the ability of the networks to account for the data arises out of graded cooperative and competitive interactions among portions of a network that are not meaningfully interpretable as informationally encapsulated components. Furthermore, Farah acknowledges that the success of the simulation of impaired face processing (Farah, O'Reilly, & Vecera, 1993) less clearly stems from violating her interpretation the locality assumption. On the other hand, the fact that residual knowledge after partial damage can support performance on implicit tasks stems directly out of violating a central aspect of the strong modularity hypothesis: knowledge is not encapsulated in separate components but rather is distributed throughout the network (see Hinton & Shallice, 1991; Plaut & Shallice, 1993b; 1993a, for similar results).

Viewing PDP systems as challenging the strong modularity hypothesis rather than Farah's locality assumption also provides a better understanding of the finer-grained analyses that she mentions. In Hinton and Shallice's (1991) simulation of deep dyslexia, visual and semantic errors co-occur because the knowledge of how visual representations relate to semantic representations is distributed throughout the network, rather than being confined to a "visual" component and a "semantic" component, respectively (see Plaut & Shallice, 1993a, for further results and discussion). In Patterson, Seidenberg, and McClelland's (1989) simulation of surface dyslexia, poor performance on low-frequency exception words and the occurrence of regularization errors occur because the knowledge of all spelling-sound correspondences is embedded in the same set of weights, rather than being split into "lexical" and "nonlexical" components, and the robustness of a given correspondence in the face of damage depends on its frequency of occurrence (but see Behrmann & Bub, 1992, for criticism of the adequacy of the account). In Mozer and Behrmann's (1990) simulation of neglect dyslexia, the lexicality effects after visual damage arise simply from the fact that lexical knowledge can reconstruct corrupted word input but not corrupted nonword input. The only sense in which damage has nonlocal effects in any of these simulations is the same sense that applies to modular systems: intact components downstream from a lesion are affected by corruption of their input.

In summary, I strongly agree with Farah that PDP principles provide a way of characterizing cognitive processes that is fundamentally different from more traditional, modular frameworks, and that systems which embody these principles can generate more satisfactory accounts of a wide range of psychological and neuropsychological phenomena. However, I disagree with her in the specific aspects of PDP systems which distinguish them from modular systems. Both PDP and modular systems exhibit nonlocal effects of damage, and thus violate Farah's interpretation of the locality assumption. However, the distributed, interactive, graded processing within PDP systems contrasts sharply with the encapsulated, staged, all-or-none processing within systems adhering to the strong modularity hypothesis. It is exactly these distinctions, and not a violation of Farah's locality assumption, that are fundamental to the strengths of the PDP approach in cognitive neuropsychology.

Acknowledgments

I'd like to thank Marlene Behrmann, Jay McClelland, and Tim Shallice for helpful comments on an earlier draft of this commentary.

References

Behrmann, M., & Bub, D. (1992). Surface dyslexia and dysgraphia: Dual routes, a single lexicon. *Cognitive Neuropsychology*, *9*, 209–258.

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. (1992). Is cognitive neuropsychology possible? *Journal of Cognitive Neuroscience*, *4*, 80–95.

Cohen, J. D., Romero, R. D., Servan-Schreiber, D., & Farah, M. J. (1994). Mechanisms of spatial attention: The relation of macrostructure to microstructure in parietal neglect. *Journal of Cognitive Neuroscience*, *6*, 377–387.

De Haan, E. H. F., Bauer, R. M., & Greve, K. W. (1992). Behavioral and physiological evidence for covert recognition in a prosopagnosic patient. *Cortex*, *28*, 77–95.

Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *Journal of Experimental Psychology: General*, *120*, 339–357.

Farah, M. J., O'Reilly, R. C., & Vecera, S. P. (1993). Dissociated overt and covert recognition as an emergent property of a lesioned neural network. *Psychological Review*, *100*, 571–588.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*, 74–95.

Mozer, M. C., & Behrmann, M. (1990). On the interaction of selective attention and lexical knowledge: A connectionist account of neglect dyslexia. *Journal of Cognitive Neuroscience*, *2*, 96–123.

Patterson, K., Seidenberg, M. S., & McClelland, J. L. (1989). Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neuroscience* (pp. 131–181). London: Oxford University Press.

Plaut, D. C., & Shallice, T. (1993a). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*, 377–500.

Plaut, D. C., & Shallice, T. (1993b). Perseverative and semantic influences on visual object naming errors in optic aphasia: A connectionist account. *Journal of Cognitive Neuroscience*, *5*, 89–117.