DeLuca and his colleagues (DeLuca and Cicerone, 1991; DeLuca, 1993; DeLuca and Diamond, 1995) have provided some of the most systematic observations of confabulating patients that we have available. Therefore his views about our current level of understanding of this fascinating neuropsychological phenomenon are of particular interest. In the interest of stimulating a conversation, I will focus on several areas where I think DeLuca's analysis is problematic, needs further clarification or development, or perhaps underestimates the progress that has been made in understanding the cognitive mechanisms of confabulation.

Is the Field Suffering from the Problem of a Lack of a Clear Definition of Confabulation?

Most definitions of confabulation have in common the idea that patients may make statements that are false or engage in behaviors that reflect false memories or beliefs, and that they do this without an intention to deceive (e.g., Whitlock, 1981; Moscovitch, 1989; Johnson, 1991). DeLuca adopts such a definition for what he calls "confabulation in the broad sense," but suggests that little progress has been made in the last hundred years in understanding confabulation because cognitive neuropsychologists have not come to a consensus about whether there is one type or two (or more) distinct forms of confabulation.

A more optimistic view is that researchers have pointed out that confabulations (even from the same patient) differ on a number of dimensions. Confabulations differ in general content—whether they are about current visual experience (Anton's syndrome); identity of persons (Capgras syndrome) or places (reduplicative paramnesia); experiences of body parts (denial of paralysis); or about autobiographical episodes, semantic knowledge, or beliefs. They also differ in bizarre nature, whether spontaneously offered or given in answer to questions, and whether they are acted upon. They may also differ in the degree to which patients are aware (at least initially, see below) that they are embellishing. Such observations provide the beginning of a systematic scheme for describing (i.e., coding) patients' behavior that is more nuanced (i.e., multidimensional) than the simple dichotomous schemes that DeLuca correctly suggests we should reject, but also more detailed than the one he outlines.

In his Figure 1 and the accompanying discussion, DeLuca proposes a classification scheme largely based on general content of confabulations, which has shown some relationship to lesion evidence. For example, confabulation about visual experience is associated with bilateral occipital brain damage; confabulation that denies paralysis is associated with

---

Dr. Johnson is Professor of Psychology, Department of Psychology, Yale University, New Haven, Connecticut.

Acknowledgments: Preparation of this paper was supported by NIA Grants AGO9253 and AG15793.
Confabulation

Parietal plus frontal lesions; confabulation about memories for past events is associated with basal forebrain plus frontal lesions. However, to assess just how specific the content of a given patient’s confabulations are it would be useful to have studies in which, for example, bilateral occipital lesion patients are tested for confabulation using the same procedures used for frontal/basal forebrain lesion patients, and vice versa. Such an approach would address issues similar to those that arise whenever category-specific brain regions are hypothesized; for example, in assessing whether or not the recognition deficits of prosopagnosics are limited to faces or extend to other sets of stimuli (e.g., Farah, Levinson, and Klein, 1995; Gauthier, Behrmann, and Tarr, 1999; de Gelder and Rouw, 2000).

Even within categories, especially the category of memory confabulators, there has been little standardization in assessing and reporting confabulation, making it difficult to compare across studies (Johnson, Hayes, D’Esposito, and Raye, 2000).

An important task is to develop a coding scheme that will provide enough specificity for us to begin to discover the relations between characteristics of confabulation and underlying brain damage. With better coding schemes for observed behavior, and with better techniques for localizing damage, we might be able to explicate differences among patients that clinicians have long felt are important, such as whether patients’ confabulations are bizarre or more mundane, whether confabulations seem spontaneous or in response to stimuli in the environment, whether patients act on them, and so forth.

The Distinction between Aware and Unaware Confabulation

DeLuca points to a distinction in the literature between patients who confabulate without realizing it (unaware confabulation) and those who confabulate intentionally to fill in gaps, for example, to cover embarrassment (aware confabulation or gap filling). DeLuca goes on to argue that there is little evidence that confabulating patients knowingly fill in gaps. For example, in discussing Tallard’s (1965) patient OJ, DeLuca says, “It was clear from his words, his affect and his behavior, that there were no ‘gap-filling’ responses” (p. 121). This may be true, but what are the criteria for such judgments? It is likely that the factors that influence our evaluations of the veridicality of what others say (including the intentionality of their errors) are similar to those that influence our monitoring of our own memories (e.g., see Johnson, Bush, and Mitchell [1998], on interpersonal reality monitoring). Such reality monitoring processes are imperfect (as discussed below). For example, a person who is embarrassed and trying to cover it could sound convincing to others, including researchers and clinicians. Equally important, people who knowingly embellish can later believe their own embellishments or ‘‘confabulations’’ (e.g., Johnson, Raye, Foley, and Foley, 1981; Ackil and Zaragoza, 1999). It would be interesting to assess which brain-damaged patients are especially likely later to experience such intentional embellishments as memories, and which are not. It seems quite plausible that if brain-damaged patients are induced to confabulate (e.g., encouraged to guess answers to questions), they may later be more likely than controls to believe these induced confabulations. Thus, in addition to accounting for why brain-damaged patients are less able than normals to sort out generated from perceived information, we may need to account for why some might also be more inclined than others to embellish in the first place. (Here, there might be a role for individual differences in motives, embarrassment about memory problems, etc., see p. 152).

The Distinction between Confabulation and Confusional States

DeLuca argues that the mechanisms that account for confabulations are different from those that account for false statements that patients make in confusional states—and that the former result from “disturbance involving the frontal lobes” (p. 12). It then seems inconsistent when he argues that the change in anterior communicating artery (ACoA) patients from spontaneous to provoked confabulations as their confusion clears is evidence that confabulations fall on a continuum rather than into distinct types because lesion location is constant. Also, acute confusional states may not necessarily be produced by the same lesions as those that produce longer-term confabulation (or only by those lesions), but confusional states are likely to be accompanied by disturbed frontal functioning (from swelling, disruption of neurotransmitters, etc.). Closer inspection of the similarities and differences between confusional and confabulatory experiences should be revealing, and should help provide a better understanding of the similarities and differences between the brain areas that are functionally disrupted during
confusional states and those where lesions produce longer-lasting confabulations.

The Distinction between Confabulations and Delusions

DeLuca correctly notes that in the literature, confabulation tends to be used to describe false memories about distinct episodes and delusions to describe false beliefs. Also, in practice, confabulation is used to describe the consequences of brain damage, and delusions to describe the consequences of psychopathology. Nevertheless, on theoretical grounds, false memories and false beliefs have much in common (Johnson, 1988; Johnson and Raye, 2000). That is, many of the same theoretical constructs can be applied to understanding confabulations (Johnson, 1991; Moscovitch, 1995; Johnson, Hayes, et al., 2000) and delusions (Johnson, 1988; Frith, 1992; David and Howard, 1994), as well as hallucinations (Bentall, Baker, and Havers, 1991), and dreams (Kahan, 1994; Solms, 1997).

The similarity between confabulations and delusions in cognitive mechanisms is not too surprising because memories are particular kinds of beliefs about what happened (Johnson and Raye, 1981). Also, while many confabulations described in the brain-damage literature seem to be about particular events (e.g., that the hospitalized patient attended a business meeting that afternoon), much of what is labeled confabulation looks more like statements about personal semantic knowledge or beliefs (that one has four adult children from a four-month-old marriage, that one’s arm can move when it cannot). It would be potentially informative for investigators to systematically explore and record whether false statements made by patients seem to refer to particular episodes or to more general beliefs (attitudes, knowledge, etc.) and, if possible, what the accompanying subjective experience of the patient is. Relating subjective experience (or experimenter ratings of qualities of patients’ memories, e.g., Johnson, O’Connor, and Cantor, 1997) to lesion location and other indicators of neural dysfunction (e.g., hypoperfusion, etc.) may help clarify the mechanisms underlying misattributed mental experiences that seem to be about events and those that seem to reflect more general knowledge, beliefs, or “episodic” memories (e.g., Neisser, 1981).

The Role of Premorbid Personality Factors

DeLuca argues that there is little evidence that premorbid personality traits affect which brain-damaged patients will confabulate or what they will confabulate about. Although I would agree that the evidence is sparse (e.g., see Weinstein [1996] for of this issue), I would expect that if investigators look for it they will find such evidence. Researchers have reported correlations between individual difference factors and susceptibility to false memories in neurologically intact individuals. For example, there is some evidence that people high in imagery ability (Johnson, Raye, Wang, and Taylor, 1979; Markham and Hynes, 1993), high in dissociative tendencies (Hyman and Billings, 1998; Winograd, Peluso, and Glover, 1998), and high in anxiety (Johnson, 1999) are more likely to misattribute the origin of memories. Self-focus on one’s own emotional reactions can lead to increases in false memories (Johnson, Nolde, and DeLeonardis, 1996), as can stress (Payne et al., submitted). Based on studies conducted by cognitive and social psychologists, we would also expect available schemas, motives, and goals to influence the development of false memories (e.g., Johnson and Sherman, 1990; Santisioso, Kunda, and Fong, 1990; Mather, Johnson, and DeLeonardis, 1999; Sherman and Bessenoff, 1999).

There is no particular reason to believe that such variables would become inoperative under conditions of brain damage. An interesting possibility is that some of these influences might be exaggerated with brain damage. For example, we have found that older adults are more likely than younger adults to misattribute a statement made by one speaker to another based on activated schemas for the speakers (e.g., misattributing a statement such as “I’m pro-choice” to a Democrat rather than a Republican). In addition, older adults’ source accuracy for such schema-inconsistent statements was positively correlated with their scores on a battery of neuropsychological tests often used to assess frontal function (Mather et al., 1999). It is not much of a leap to expect that people who hold strong stereotypes, or have recurring preoccupations, may reveal them even more so under conditions of brain damage.

Identifying Brain Regions Whose Disruption Produces Confabulation

There is general agreement that confabulating patients tend to show deficits on memory tests and tests of
executive function. This behavioral pattern (confabulation + memory deficits + poor executive function) has especially been observed in ACoA patients, who tend to have basal forebrain and frontal damage, as DeLuca describes. Nevertheless, confabulation has been reported with other lesions or combinations of lesions (for a review of cases, see Johnson, Hayes, et al., 2000). Because of the likely number of brain regions in various circuits relevant to the cognitive processes underlying memory, and the difficulty of finding clinical cases of damage confined to a single identified structure (or confined to discrete combinations of two or more structures), the particular lesion or combinations of lesions that account for different characteristics of confabulation remain to be specified.

For example, DeLuca notes that “Despite decades of ACoA research, few have specifically implicated the anterior cingulate alone [emphasis added] as critical for confabulation” (p. 22). However, if confabulation results from lesions in more than one structure, even pure cases of lesions in a single structure will not implicate that structure in confabulation. In fact, DeLuca’s own observations of confabulating patients have convinced him that disruption of more than one structure typically is necessary. Many processes distributed among interacting circuits go into producing both true and false memories in neurologically intact individuals. Thus it seems likely that there is more than one way to disrupt the cognitive system in such a way that some kind of confabulation occurs (for additional discussion see Johnson, Hayes, et al. [2000]; Johnson and Raye [2000]).

The suggestion I am making here is parallel to the argument against identifying any single structure (e.g., the hippocampus) as the memory structure (e.g., Morton, 1985; Johnson, 1990, p. 126). Furthermore, memory and executive function are intertwined—there is unlikely to be any episodic memory without some executive processing (e.g., Johnson, Hashtroudi, and Lindsay, 1993). That is, it is increasingly recognized that frontal cortex is important for establishing, retrieving, and evaluating event memories, and that other regions (including posterior cortical areas) play important roles as well. Hence, a number of areas are implicated in normal memory functioning. Disruption of medial temporal structures is likely to produce more profound amnesia (“classical” anterograde amnesia) than disruption of frontal structures, but this does not mean that all critical memory processes reside in the medial temporal area. Just as there is more than one way to disrupt cognition that produces errors of omission, there is likely to be more than one way to disrupt cognition that produces errors of commission.

Evidence from neuroimaging normal subjects as they engage in monitoring the origin of memories (Raye, Johnson, Mitchell, Nolde, and D’Esposito, 2000) and a recent neuroimaging study of a confabulating patient’s changes in patterns of activation over the course of improvement (Benson et al., 1996) suggest that anterior cingulate may play a role in memory monitoring. Further neuroimaging work with normal subjects is needed to clarify which processes are subserved by which circuits and, in combination with increasingly better evidence about lesion sites from brain-damaged patients, we should be able to develop a richer characterization of the impact of dysfunction in various structures on confabulation.

Explicating the Cognitive Mechanisms of Confabulation and Relating These to Neural Mechanisms

DeLuca notes that poor “self-monitoring” is a theoretical idea that has often been invoked to explain confabulation and quotes Pak and Schneider (1999) that “poor self-monitoring may serve as a descriptive explanation . . . [but] does not disclose the specific mechanism of confabulations.” He cites Moscovitch and Mello (1997) and three papers from my lab (Johnson, 1991; Johnson, O’Connor, and Cantor, 1997; Johnson, Hayes, et al., 2000) as examples of work that goes beyond simply positing poor self-monitoring and that sheds some light on potential cognitive mechanisms of self-monitoring. For example, both Moscovitch and I point to disrupted strategic retrieval processes that are a normal part of evaluating memories. I appreciate this acknowledgment; however, I was also struck by the fact that the references to our work were to papers that explicitly made an effort to connect to the neuropsychology and neuroscience literatures. There were no references to strictly cognitive—behavioral papers (from our lab or other labs) in DeLuca’s discussion of confabulation. This is not unusual in cognitive neuropsychology and cognitive neuroscience papers, just as there are many cognitive—behavioral papers that do not draw on relevant neuropsychological evidence. This insularity is both common and understandable and highlights the importance of bridging efforts such as this journal.

In this spirit, I want to call attention to the fact that there is an extensive tradition in cognitive psychology of investigating the mechanisms of memory
distortion. (Similarly, in his commentary, Mark Solms highlights the psychoanalytic approach to memory distortion.) In particular, there are theoretical characterizations of processes that account for true and false memories that may help explicate the various ways in which confabulation could come about. Such ideas about process can provide hypotheses about some of the ways in which confabulations might differ, hypotheses about the different types of neural dysfunction that might produce them, and, at least, an interpretive framework for case studies and other findings from cognitive neuroscience.

For example, the source monitoring framework (SMF; Johnson and Raye, 1981, 1998, 2000; Johnson, 1988; Johnson, Hashtroudi, and Lindsay, 1993; Mitchell and Johnson, 2000) outlines various factors that can contribute to distortions of memories and belief. This theoretical approach highlights the fact that features of mental experience, for example, familiarity, perceptual, spatial and emotional detail, and metacognitive beliefs about memory, have an impact on what is considered by an individual to be a memory. It also highlights that features of mental experiences include information about current cognitive operations or about prior cognitive operations (e.g., intentional imagery processes). Furthermore, the mental experience that is being evaluated is not only a product of what actually happened, but it likely includes information generated by associative processes, schemas, and scripts—that is, by knowledge and beliefs—and is colored by needs and desires (many of which originate from social-cultural circumstances). The more similar memories are from various sources (e.g., perception and imagination or, more generally, in source feature A and source feature B), the more likely there will be source misattributions. But similarity alone does not create errors. Many factors can contribute to the distortion of memories and beliefs; failure to encode or access a particular type of information (e.g., visual detail, cognitive operations); failure in encoding/consolidation of combinations of features (e.g., feature binding); failure to consider multiple sources of evidence during remembering; failure to retrieve additional evidence beyond what is initially activated; adopting a low criterion for what constitutes evidence; overweighting nondiagnostic evidence, and so forth. That is, remembering involves relatively automatic activation of information by cues, but it also involves reflective (executive) functions that help select for, retrieve, and weigh appropriate evidence, and perhaps select against (or inhibit or reduce the weights of) inappropriate evidence.

Individual differences would be expected to affect the likelihood of memory distortion insofar as they are likely to reflect differences in one or more of these factors. For example, individuals may differ in the likelihood that they generate candidates for potential misattribution (e.g., fantasy-prone personalities may embellish reality more; experts in a domain may make inferences without realizing it). Individuals may differ in the qualities of what is being evaluated (e.g., people high in imagery ability may generate information that is more difficult to distinguish from perceived information). Or they may differ in the quality of the evidence required (e.g., individuals high in suggestibility, dissociative abilities, or impulsivity may make confident memory attributions on the basis of little evidence or inappropriate evidence). Furthermore, individuals may differ in level of distraction, anxiety, or stress (any of which may, e.g., disrupt feature binding or strategic retrieval processes), and so forth.

Such cognitive mechanisms have been actively investigated in many labs (some factors have received more attention than others, of course), and individual differences are beginning to be explored as well. Although there is much work to be done, it is probably reasonable to say that the basic cognitive mechanisms of memory distortion are not a great mystery. It also seems reasonable to suggest that clinically significant symptoms such as confabulations and delusions represent various types of dysfunction of these normally imperfect processes. However, the relations between cognitive and neural mechanisms largely remain to be specified. In this regard, rapidly developing neuro-

Within the SMF framework, temporal information (when an event occurred) is a type of source "feature." Others include spatial information, perceptual details such as colors and sizes and sounds, emotional qualities, semantic information, information about cognitive operations (e.g., imagistative processes), modality (e.g., pictures vs. words), etc. As DeLuca notes, a number of investigators have emphasized the role of temporal confusions in confabulations. On the one hand, this is bound to be the case because confusion between any features that did not occur simultaneously could be described as a temporal confusion (e.g., misattributing advice heard on a television talk show to one's doctor involves temporal confusion as well as person confusion, location confusion, etc.). On the other hand, a more interesting possibility is that temporal attributions might suffer more under conditions of brain damage than some other source information because temporal information often is not a direct feature of individual events but must be inferred from relational information (e.g., Tzeng and Cotton, 1980; Friedman, 1993). Thus, some types of temporal information may require more complex reflective processes to access than, say, size or color. That is, temporal judgments may be the result of reflective retrieval of related information (e.g., Johnson, 1983, p. 113). From the limited available evidence, source monitoring on the basis of a number of different features, including temporal judgments, is associated with left prefrontal cortex (PFC) activity (Raye et al., 2000).
imaging techniques are particularly exciting. This active area of investigation is beginning to yield evidence (in addition to that available from brain-damaged patients) about particular brain regions engaged in episodic memory tasks (see Cabeza and Nyberg [2000] for an extensive recent review). Of particular relevance here are findings showing activation associated with feature binding (e.g., anterior hippocampus and medial PFC [BA 10], Mitchell, Johnson, Raye, and D’Esposito [2000]; Prabhakaran, Narayanan, Zhao, and Gabrieli, [2000]), and retrieval and evaluation of memories (e.g., right BA 10, Tulving, Kapur, Craik, Moscovitch, and Houle [1994]; Lepage, Ghaffar, Nyberg, and Tulving [2000]), and left BA 9/10/46 and anterior cingulate (e.g., Raye et al. [2000]). Furthermore, we see increased bilateral activity when episodic tasks require more complex processing (Raye et al., 2000). A hypothesis that my lab is currently exploring is that source monitoring often involves active interhemispheric cooperation (e.g., Banich, 1998) as mental experiences are taken as objects of reflection (e.g., Johnson and Reeder, 1997; Johnson and Raye, 1998, 2000; Nolde, Johnson, and Raye, 1998). If so, disruptions in interhemispheric processes, such as may be likely with medial damage, especially to areas that supply neurotransmitters to broad regions of frontal cortex, should disrupt source monitoring, resulting in confabulations.

Neuroimaging techniques provide us with new opportunities to try to tie the processes postulated from cognitive studies to neural circuits, through manipulating conditions while studying neurologically intact individuals as they encode, consolidate, retrieve, and evaluate information. These cognitive-behavioral and neuroimaging studies should, in turn, provide neuropsychologists with a more specific understanding of the nature of normal memory processing from which to generate interpretations about disrupted processing in brain-damaged individuals. In a reciprocal fashion, neuropsychologists have already provided information about structures that should be clear targets for special attention in future neuroimaging studies (see studies reviewed in Johnson, Hayes, et al. [2000]). Of particular interest are structures of the basal forebrain and ventromedial frontal region disrupted in ACoA patients, as highlighted by the work of DeLuca and colleagues. Finally, improved methods for identifying lesions and metabolic dysfunction in brain-damaged patients (e.g., Benson et al., 1996), along with more systematic coding of confabulations and theoretically motivated cognitive testing of such patients, should provide invaluable converging evidence regarding hypothesized cognitive processes and help specify brain regions whose disruption produces confabulation.

Potential Connections between Cognitive Approaches to Reality Monitoring and the Psychoanalytic Perspective

In the multiple-entry, modular memory system (MEM) that provides a description of the types of component processes recruited during the encoding of events and during reality–source monitoring, executive functions arise from the operation of various component processes of reflection (e.g., Johnson, 1997; Johnson and Reeder, 1997). These are the processes, for example, by which two or more representations are compared (noting), by which related information is strategically activated through the self-generation of cues (retrieving), and by which mental experiences are kept active (refreshed, rehearsed) and taken as objects to be evaluated (as in reality monitoring). Solms (this issue) discusses a key idea advanced by Freud that much can be learned from the operation of associative processes unconstrained by attentional–executive processes. Potential consequences of the disruption of reflection in MEM include the characteristics described by Solms: insensitivity to mutual contradiction (because of a failure to keep multiple representations coactive or to compare them), failures in temporal identification (see footnote 1), and absence of reality testing/monitoring (including the various ways monitoring can fail as described in the previous section).

Furthermore, what behaviors we see when reflection is suspended are perhaps driven primarily by needs, attitudes, and desires interacting with the vicissitudes of external stimuli and the associative pathways linking concepts, schemas, and habits (e.g., Shallice, 1988). Such statements and actions can indeed seem impulsive, uncritical, and sometimes bizarre.

One possibility discussed by Solms is that reflection (the "secondary processes") involves an active inhibitory process by which instinctive actions are inhibited (or delayed), giving thought (other reflective processes) a chance to have its influence. Cognitive psychologists, cognitive neuropsychologists, and cognitive neuroscientists have proposed at least two ways that instinctive, or habitual, or prepotent responses can be circumvented: (1) by active inhibitory processes, or (2) by active goals or agendas that add activation ("bias") to appropriate representations or action plans that then have an activation advantage over pre-
potent but inappropriate representations or action plans. It is possible that these are two sides of the same coin or two aspects of the same process and cannot, strictly speaking, be separated. It is also plausible that there are two distinguishable reflective functions, one corresponding to active inhibitory processes that introduce a delay in responding, giving slower reflective processes a chance, and one corresponding to the active maintenance of goals or agendas. One possibility is that ventromedial or orbitofrontal regions are associated with the first function and dorsolateral prefrontal regions with the second (e.g., Cummings, 1993; Johnson, Hayes, et al., 2000). Although Freud's idea of "binding" that Solms discusses sounds more like the inhibitory function, it is likely that both inhibitory and agenda-maintaining processes are necessary to keep thought and action in line with reality (that is, to provide the opportunity for normal reality monitoring processes to be engaged). Furthermore, some of the differences in the qualitative characteristics of confabulations (e.g., how bizarre, how impulsive, or spontaneous they seem) might be related to variations in the differential disruption of these (and other) reflective functions.

Cognitive psychology, cognitive neuropsychology, cognitive neuroscience, and psychoanalysis, all recognize the key role that reflecting on reflection—as in reality or source monitoring of memories and beliefs—plays in normal cognition. The combined insights from these approaches provide a relatively clear roadmap for future investigations that might further clarify this critical cognitive function.

References


Confabulation


