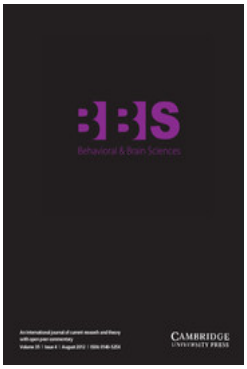

Précis of From neuropsychology to mental structure

Tim Shallice

Behavioral and Brain Sciences / Volume 14 / Issue 03 / September 1991, pp 429 - 438
DOI: 10.1017/S0140525X0007059X, Published online: 19 May 2011

BEHAVIORAL AND BRAIN SCIENCES (1991) 14, 429–469
Printed in the United States of America

Précis of *From neuropsychology to mental structure*[1]

Tim Shallice
Department of Psychology, University College London, London WC1E 6BT, England
Electronic mail: *ucjtsts@ucl.ac.uk*

Abstract: Neuropsychological results are increasingly cited in cognitive theories although their methodology has been severely criticised. The book argues for an eclectic approach but particularly stresses the use of single-case studies. A range of potential artifacts exists when inferences are made from such studies to the organisation of normal function – for example, resource differences among tasks, premorbid individual differences, and reorganisation of function. The use of "strong" and "classical" dissociations minimises potential artifacts. The theoretical convergence between findings from fields where cognitive neuropsychology is well developed and those from the normal literature strongly suggests that the potential artifacts are not critical. The fields examined in detail in this respect are short-term memory, reading, writing, the organisation of input and output speech systems, and visual perception. Functional dissociation data suggest that not only are input systems organised modularly, but so are central systems. This conclusion is supported by findings on impairment of knowledge, visual attention, supervisory functions, memory, and consciousness.

Keywords: amnesia; aphasia; attention; case study; cognition; consciousness; dyslexia; memory; modularity; neuropsychology; planning; reading

From neuropsychology to mental structure (hereafter *Neuropsychology*) is concerned with what we can learn from investigations of the behavior of brain damaged patients about the organisation of the normal cognitive system. The status of such neuropsychological evidence has changed dramatically over the last 20 years. The hallmark of classical neurology was the description of a great variety of highly specific and surprising impairments resulting from brain damage. In the 1970s, such observations were considered fascinating in their own right as well as clinically important but they were not often taken to be particularly relevant to an understanding of normal functions. So Fodor et al. (1974) could write "remarkably little has been learned about the psychology of language processes in normals from over a hundred years of aphasia study" (p. xiv), and Postman (1975) could write regarding memory that "The existing data do not impress us as unequivocal; more important extrapolations from pathological data to the study of normal function are of uncertain validity" (p. 308). By the 1980s these views seemed strange. Thus, neuropsychological dissociations became central pillars for Fodor's (1983) own modular model of mind and in normal memory research on the explicit/implicit contrast, which stemmed directly from discoveries on amnesic patients (see Roediger, 1990, for review) became very fashionable.

1. Neuropsychological findings: Their attraction and their problems

There are at least three reasons why interest in neuropsychological findings should have increased. Empirical phenomena in the corresponding study of normal processes – human experimental psychology – are very slippery things. Many factors affect any experimental procedure. Make a slight change in one aspect – rate of presentation, stimulus material, recall delay, amount of practice, and so on – and the effect disappears or reappears, although according to theory, it should not. Therefore, even if a phenomenon is narrowly robust, the experimental result provides only a most insecure platform for theoretical inferences. The first attraction of neuropsychological evidence is that all these factors shrink in significance by comparison with the size and specificity of observed deficits.

The weakness of the empirical methods available in "normal" human experimental psychology has a second consequence; we have little idea of the vastness of the problems that need to be tackled. A pessimist can view them as producing islands of detailed empirical knowledge surrounded by a sea of ignorance, whose size we conceal from ourselves by vague theorising (see Newell 1973). Neuropsychology can help in this second respect, too. Advanced clinical practice contains the distilled "craft" knowledge of more than 100 years of observation of patients by neurologists and neuropsychologists. As neurological disease can affect just about every part of the brain the disorders that have been described will probably encompass damage to nearly all the cognitive mechanisms. "Inverting" the set of disorders that exist might enable us to map the subcomponents of mind.

The third reason for the increase in interest in neuropsychological findings is that the information processing approach to modelling cognitive processes that was developing in the 1960s and 1970s provided a conceptual

 0140-525X/91 $5.00+.00

system which lent itself to the evaluation of neuropsychological observations. Observed syndromes appeared to fit with how a system with damaged subsystems would operate.

The increasing respectability of neuropsychological evidence in discussions of normal function, however, hid a growing rift among those working within neuropsychology. Very different methods continue to be used. In particular, those with a neuroscience or a clinical background tend to favour standard group study procedures, but a growing minority of neuropsychologists with a cognitive science background strongly favour single-case studies. Nor is this just a matter of a preference between two accepted approaches having different types of payoff. Many practitioners have been dubious about the scientific value of approaches different from their own, and their doubts have recently become public. Thus Badecker and Caramazza (1985) argued that the approach in which the performance of groups of Broca's and Wernicke's aphasics are contrasted is inherently flawed. Caramazza (1986) went on to reject group studies as inappropriate in principle for making extrapolations to normal function because the performance pattern of individuals in a group may differ qualitatively from the mean. The complementary criticism, that single-case studies are unscientific, has been voiced less frequently. Even relatively favourable critics like Zurif et al. (1989) point out, however, that single-case studies seem especially prone to problems from the selection of premorbidly atypical subjects and from the adoption of idiosyncratic strategies by individual patients. In addition, to extrapolate from either group or single-case studies one must assume in practice that no critical reorganisation of function takes place after the lesion; there is little or no evidence for this. Moreover, as novel tasks are often used, neuropsychological studies are prone to standard problems of experimental psychology methodology such as differences in task difficulty and variations by subject strategy.

There are even major differences between the methods used by different practitioners within both the group and the single-case approaches. In about 1980, researchers using the single-case approach tended to study individual patients as exemplars of particular syndromes, sets of symptoms found to co-occur frequently (e.g., deep dyslexia, Coltheart 1980a). Some argued that the most appropriate syndromes to study were those arising from "fractionation," namely, the analysis of increasingly selective impairments on the grounds that they would be increasingly likely to result from damage to individual subsystems (Beauvois & Derouesne 1979; Shallice 1979). By the mid-1980s, however, the most common belief among those who advocated the single-case approach was that "a syndrome thought at time t to be due to damage to a single unitary module is bound to have fractionated by time $t + 2$ years into a host of awkward subtypes" (Ellis 1987). This would favor studying unique cases: Generalising across patients pretheoretically is likely to be functionally misleading; *only* the performance of individual patients should be related to theory (Caramazza 1986; Coltheart 1985; Morton & Patterson 1980). The classification of patients as exemplars of particular syndromes is also rejected (e.g., Caramazza & Badecker 1989; Ellis 1987). Although this approach may be seem logically satisfactory, it is difficult to see how a science could operate effectively as a social system if cross-patient replication and classification are abandoned. The database would become extremely cumbersome.

Such basic disagreements about method in neuropsychology make it important to examine empirical and inferential methodology, especially as the negative aspects of both positions could be correct, in which case the dismissive views of the 1970s about neuropsychology such as those of Fodor et al. (1974) and Postman (1975) referred to earlier may have been valid. In general, for a particular set of observations on the performance of a patient, even if there existed a complete theory of how the relevant tasks were performed there would be no direct way to know which putative component(s) were damaged. Moreover, if one took each subcomponent in turn, to know what behavior the damaged system could produce would require a complete specification of how damage might affect the operation of each subcomponent. Inferring the underlying undamaged system from an observed impairment requires rather strong assumptions about how the normal cognitive system might operate and what the consequences of damage might be.

The large number of highly specific, qualitatively diverse disorders in clinical neuropsychology suggests some form of modularity. There are also many other arguments in favour of modularity – from linguistics (e.g., Chomsky 1980), neurophysiology (e.g., Cowey 1985), system design (Marr 1982; Simon 1969) as well as from information-processing psychology, especially mental chronometry (e.g., Fodor 1983; McLeod et al. 1985; Posner 1978; Sternberg 1969). *Neuropsychology* accordingly examines the implications of modular assumptions for neuropsychological research.

Modularity provides the central theoretical assumption for much of the first half of *Neuropsychology* (Chapters 2–10), which asks what neuropsychological research can reveal about a system organised in this way. Chapter 11 addresses the more abstract question of whether any other forms of cognitive architecture – in particular connectionist ones – are as compatible with the types of selective impairments observed. Chapters 12 to 16 ask what other types of systems are required to enable a set of qualitatively distinct processing systems to function effectively, and in particular, what information neuropsychology can provide on this point.

2. Inferences to the organisation of modular systems

The approach advocated in *Neuropsychology* for drawing inferences from neuropsychological findings about the nature of modular systems is an old one. In 1885, Lichtheim drew a distinction between a "pure case" in which only a single subsystem is impaired and a "mixed case" in which more than one subsystem is affected and favoured the observation of pure cases. Pretheoretically, however, one cannot know whether any observed disorder is pure or mixed. It is argued that fractionation is generally an appropriate procedure for obtaining "pure cases," and that a strong emphasis should be given to dissociations as those aspects of a mixed case will also be found in a pure one. The approach advocated leads to a bias toward

single-case studies but not to the exclusion of group studies.

As the criticisms of both feuding camps indicate neuropsychological inferences are subject to a variety of potential artifacts. *Any* method for relating neuropsychological findings to normal function therefore requires justification. A procedure for obtaining data in science can be validated either "internally," by showing how its assumptions are valid, or "externally," by demonstrating that on known terrain its conclusions agree with those of more established procedures. *Neuropsychology* adopts both internal and external validation. In this Précis I concentrate on "internal" validation (*Neuropsychology* Chapters 2, 9, 10), because "external" validation – Section 2 (Chapters 3–8) – depends on a level of detail that cannot be provided in such a synopsis.

What follows is a set of methodological assumptions for inferring normal function from neuropsychological findings. The aim is either to test existing theories or to stimulate new theories by producing counterintuitive findings. One can accomplish the latter, however, only if the relation between damage to a subcomponent and impaired task performance is fairly transparent; Gregory (1961) has pointed out that this is often not the case in machines.

The methodological assumptions fall into three groups:

The type of models to be considered.

1. The cognitive system being investigated contains a large set of isolable processing subsystems (in the sense of Posner 1978) or modules (in the sense of Marr 1982).
2. The modularity operates on a number of levels. As far as neuropsychology is concerned, however, there is a limit to the fineness of the grain of the modularity
3. Following Marr (1982), isolable processing subsystems may be viewed as having functions carried out by algorithms implemented by particular mechanisms.

From the model to performance.

4. Cognitive systems are qualitatively similar across individuals for tasks that are routinely performed in a culture.
5. Task performance requires the use of a "procedure" – a temporary activating or inhibiting of sets of intersubsystem transmission routes, which leaves a particular route or set of "routes" through the network of subsystems active. The concept of a "procedure" is intended to be a realisation of the idea of a "control process" (Atkinson & Shiffrin 1968).
6. Tasks may at times be carried out by more than one procedure – that is, more that one combination of subsystems. If the procedure for carrying out each task is specified, however, the overall pattern of performance – namely, the gross pattern of associations and dissociations shown by the patient – depends on how much is needed of the available resources in each subsystem; task performance is monotonically related to that amount. (The concept "resource" is used in the sense of Norman & Bobrow 1975.)

The effect of lesions.

7. Lesions vary greatly in the subsystems they affect, with respect both to their number and their identity; in any particular case, the identity and the number of impaired subsystems is not ascertainable independently of the behavior of the patient.
8. The effect of a lesion on task performance is determined by (a) the pattern of quantitative loss of resources across the normal set of subsystems, with (b) the procedure adopted by the subject. (Which procedure is being used should be determined by empirical investigation. However, with no further information it is assumed that the procedure adopted is the one that optimises performance given the impairment by allowing the task to be carried out using less damaged subsystems *or* it is the one normal subjects typically use to carry out the task.)
9. Individual differences among normal subjects in the extent of the resources available are small compared to the destructive effects of neurological disease on resources.

To make predictions in any particular case, one needs additional assumptions about specific models. Most of the assumptions are fairly standard, being similar to those advocated by Caramazza (1986, although I draw different conclusions from them). Two concepts require explanation, however. "Resources" and "procedures" are introduced because methodological problems arising from differences in task difficulty and subjects' strategies are at least as pervasive in neuropsychology as in standard experimental psychology. If the foregoing assumptions are accepted certain consequences follow:

1. The theoretical relevance of neuropsychological research will depend on what can be learned about the operation of a system from how it behaves when it is damaged. Consider in particular, following assumption 3, a system characterisable as a set of isolable subsystems, each having a function for the system as a whole, and realised by an algorithm implemented by a specific mechanism. In his discussion of levels of explanation, Marr (1982) argues that to understand the normal functioning of such a system, one must proceed from the highest level down; he argues against attempting a mechanistic explanation on the level of actual or hypothetical hardware in the absence of functional or algorithmic specifications. In particular cases, this may prove to be too sweeping a position because the particular hardware available (e.g., slow parallel circuitry) may make some types of (mathematical) functions much easier to compute than others. Marr's argument may apply, however, to attempts to infer how hypothetical subsystems might work from behavioural impairments following lesions. Trying to determine the particular mechanisms employed by the subsystem from this type of evidence would be like trying to deduce how computer hardware works by examining the malfunctioning of a program caused by a machine fault when one does not know the program's structure. This would almost certainly be a hopeless task; inferences to the algorithmic level would be almost as difficult.

If one tries instead to ascertain overall modular organisation – i.e., what functionally isolable subsystems exist and what each one does – the picture is more promising. If a single subsystem were severely damaged, the ability of the organism to perform a particular microfunction would be grossly impaired. In many cases, specific consequences would be expected. If in such cases one looks at the whole range of tasks the organism can perform, those on which performance is impaired are likely to be relatively insensitive to the specific nature of

the damage in a subsystem (or transmission route), provided that the impairment is severe. The set of tasks on which impaired performance occurs will be relatively unchanged when there is a change in the nature of the damage within the subsystem. This is expressed by Assumptions 6 and 8.

The most promising level for interaction between neuropsychology and the theory of normal processing is not that of detailed computational models, but more global functional architecture. This corresponds to information-processing theory as represented by the work of Morton (1970), Posner (1978), and Bruce and Young (1986). Neuropsychological evidence may also be useful for testing theories at the algorithmic or implementational levels, but this is still an open question.

2. Nearly all theories of normal function that are derived from neuropsychological findings have assumed that the relevant patients exhibited a "pure syndrome," that only a single subsystem was damaged. Newcombe and Marshall (1980), for example, in the context of a 2-route model of reading, argued that reading using the so-called "semantic route" alone is inherently unstable; they assumed that the semantic errors in *deep dyslexia* arise because the syndrome corresponds to normal reading lacking only spelling-to-sound translation.

Patients who exhibit a set of dissociations – they perform normally on a group of tasks except for one where they are severely impaired – provide the best opportunity for testing theories. The most selective dissociations also occur in pure cases, making them especially important. Assumptions 1 and 7 suggest, however, that for typical subsystems pure patients will be rare. Hence there will usually not be data from a group large enough for effective averaging. Thus as finer and finer aspects of the cognitive architecture are investigated in attempts to infer normal function, neuropsychology will be forced to resort more and more to single-case studies. By Assumptions 4 and 9, however, valid inferences to normal function should be feasible from findings on single patients.

3. The problem exists for concentrating on patients with pure syndromes, that there is no theory-independent way to determine whether a set of symptoms exhibited by one patient arises from a single-component or a multiple-component disorder.

Ellis (1987) has made a further objection to focusing on patients with pure syndromes: "The cognitive neuropsychologist will pass over 999 patients to find the one thousandth who comes close to being a pure case of 'word meaning deafness' or whatever" (p. 402). There are two answers to this kind of objection. First, the concept of a pure case – like that of an "ideal gas" in physics – may be useful even if it is not realised in any real patient. Any dissociation observed in a multicomponent syndrome also occurs in a pure syndrome. Consider a prototypic dissociation in which normal performance is obtained on one task (I) and grossly impaired performance is obtained on another task (II) of roughly comparable difficulty. Assume that subsystems operate in an all-or-none fashion – when damaged they will not support any procedure (see Assumption 5) that requires them. Then the existence of the dissociation implies that there are one or more subsystems involved in the carrying out of task II but not of task I. The same dissociation is produced if just one of these critical subsystems is damaged (a pure case) as more than one, together possibly with other subsystems unrelated to tasks I and II (a mixed case). Therefore, in a dissociation, even when observed in a mixed syndrome, both the intact and the impaired performance will coexist in at least one pure syndrome. An analogous argument applies to the more important case of a set of dissociations or a group of tasks in which only one is impaired. (These arguments apply properly only to so-called "pure tasks," see *Neuropsychology*, Chapter 10.)

The argument does not apply for an association in which the performance on two tasks is impaired with no indication of qualitative or quantitative equivalence in the way they are impaired. If the impairment on Task I arises from damage to subsystems that are not involved in Task II, and vice versa, the association of deficits in the two may be observed only in the mixed syndrome; it may not occur in any pure syndrome. Hence sets of dissociations are heuristically important. They are safer than associated deficits in that they necessarily mimic what happens in a pure syndrome. Moreover, they occur reasonably frequently in neuropsychological practice. So they make the pure-syndrome approach a viable one.

The second response to Ellis's (1987) criticisms relates to his fear that "recognised syndromes will inevitably be prone to multiply and change at an alarming rate" (p. 410). If we are to be alarmed at this prospect, the set of syndromes that are multiplying and changing are presumably doing so in a chaotic fashion, so they do not provide any solid clues about the organisation of the underlying structure. A fractionation approach (selecting more selective impairments for study) however, will lead to purer and hence more informative syndromes. As a heuristic procedure, though, it should not be rigidly applied.

4. If Assumptions 6 and 9 are valid and the various procedures drawing upon a subsystem rely on it equally, the overall pattern of impairment from damage to a particular subsystem in different patients will be qualitatively similar and the same set of dissociations will be found. This suggests a useful classification system in which patients are categorised theoretically in terms of the hypothetical subsystems impaired. When the organisation of subsystems remains theoretically unclear the empirical "units" of the classification scheme correspond to the most selective impairments observed, although from the logic of the fractionation procedure, such an empirical classification scheme can only be provisional.

This classification scheme provides: (i) a set of patients with whom more details theory about the hypothesised component can be tested. This also allows for replication or failures thereof; (ii) the possibility of finding anatomical correlates of the hypothesised subsystem; and (iii) a principled basis for forming functionally derived groups if enough patients are available. Group studies of amnesia defined as a set of dissociations including impaired performance on (episodic) memory tasks (see Chapter 15) provide one example of this classification scheme.

5. Although it is the most important neuropsychological source of information about the functional architecture investigating patients with orderly sets of dissociations is not the only possibility. Another is the critical variable method, where the performance of a certain type of task is affected in one patient (A) by a change in variable

X but not by a change in variable *Y*, and the complementary effect is observed in another patient (B). Patterson (1981) used this approach to differentiate a number of different forms of acquired dyslexia and interpreted them as arising from damage to different components in a three-route model of reading.

Error patterns can also be revealing. The semantic errors that occur in deep dyslexia (Coltheart 1980a; Marshall & Newcombe 1973) suggest that such patients cannot use spelling-to-sound translation and so they must be using a second (semantic) route to access semantics without phonological mediation, one that normal subjects presumably have available, too. In addition, because different procedures can draw upon the same subsystem, certain patterns of association between deficits can be informative too, although general problems with making inferences from associations make them appropriate only if the impairments on the two tasks can be equated quantitatively. For example, Caramazza et al. (1987) described an agraphic patient who had equivalent error rates across letter positions when writing both words and nonwords. The patient had no difficulty in writing letters, so the authors inferred that there must exist a stage prior to motor output that is common to both writing words and nonwords.

6. Group studies have recently been strongly criticised because of potential averaging artifacts (e.g., Caramazza 1986; Caramazza & Badecker 1989). These authors argued that only single-case studies are relevant for inferring normal function. Dissociations and sets of dissociations can be obtained from group studies, too, however, with group assignment based on syndrome classification, lesion location or disease process (e.g., Parkinsonian). Such findings can be used to infer normal function because performance dissociations between two groups will standardly also manifest themselves between pairs of individuals, one from each group. The mean contrast is hence representative of the behavior of certain individuals. Thus, one can normally make the same inferences from *dissociations* whether they are produced by individuals or groups. Although group studies are in many areas less informative than single-case studies, they are just as legitimate. Moreover, they suffer less from certain other potential artifacts.

It is noteworthy that critics of group studies have concentrated their attacks on the studies in which groups are defined in terms of complex mixed syndromes – for example, the classical aphasias (Badecker & Caramazza 1985; Caramazza & McCloskey 1988). To my knowledge there has been no specific criticism of group studies that compare control subjects with amnesics (e.g., Graf et al. 1984; Warrington & Weiskrantz 1970; see *Neuropsychology* Chapter 15), patients with perceptual disorders (see Warrington 1982a; *Neuropsychology* Chapter 8) or patients with supervisory or executive disorders (see Milner 1982; *Neuropsychology* Chapter 14). In all three of these areas group selection procedures are more straightforward than in aphasia research. The amnesia studies alone have been very influential in recent memory research on normal subjects (see Roediger 1990); hence the critics of group studies would need very strong and concrete arguments to show that these three lines of research have no relevance for understanding normal function. In my view such arguments have not been produced.

3. More specific problems for the approach

A methodology based on a set of assumptions requires us to consider whether the assumptions are valid in particular cases. Four seem particularly problematic – those concerning the patient's strategy (Assumption 8b of section 2), task difficulty (Assumption 6), individual differences (Assumption 9), and reorganisation of function (Assumption 8a). A selective impairment found in a particular task in some patient could just reflect: the patient's idiosyncratic strategy, the greater difficulty of that task compared with the others, a premorbid lacuna in that patient, or the way a reorganised system but not the original normal system operates. How can one guard against these possibilities?

A first complication is that patients can adapt to their impairments by using strategies that are hardly ever found in normal subjects. Many pure alexics read "letter-by-letter." A less obvious possibility is that the neologisms found in some jargon aphasics are a strategy, filling the pauses that their naming difficulty causes (Butterworth 1979). Kolk and van Grunsven (1984) have argued that the metalinguistic judgements of aphasics can only be understood if one considers the strategy the aphasic might adopt to carry out an unnatural task.

A number of methodological heuristics can make the procedure used by the patient more transparent to the investigator – testing patients with adequate performance on baseline tests, using ecologically valid procedures, using converging operations, using strategy-control tests for the critical comparisons (see Bub et al. 1985), and training the patient in the appropriate procedure (e.g., Beauvois & Derouesne 1982). The more of these safeguards one uses the more likely that Assumption 8b will be satisfied.

A second problem is that tasks are often of different levels of difficulty. For example, well-learned information is easier to retrieve, so a difference between the ability to retrieve long-term knowledge and recently learned information as in amnesia might merely reflect a difference in difficulty. The classical neuropsychological solution was to seek a complementary dissociation and thereby demonstrate a double dissociation (e.g., Teuber 1955). For instance, the idea that the good performance of amnesic patients on short-term memory or semantic memory tasks merely arises because these tasks are easier or better learned is countered by the existence of patients with selective deficits on these tasks who have preserved episodic long-term memory (see *Neuropsychology* Chapters 3 and 15). The application of the resource Assumption (6) shows that with some minor modifications this remains a valid method. Indeed, it works in this particular case. A patient has been described who has grossly impaired knowledge but in appropriate circumstance is normal at retrieving recently learned information (see Coughlan & Warrington 1981).

The third complication is that a dissociation could arise from a selection artifact because the patient was weak at certain tasks before illness. If the dissociation is a *classical* one in which all tasks are performed at normal levels and none are resource-limited (except for one or more that are significantly worse and below the normal range) then such an individual-differences account is implausible. In many theoretically relevant cases, however, the perfor-

mance of the patients on the tasks they carry out better is still below the normal range even though the contrast between the better and worse performed tasks is striking – a *strong* dissociation. It is hence difficult to estimate the statistical significance of a difference in scores. In specific situations a counterargument can be made against this type of artifactual explanation but this is one type of problem for which group data provide a more solid response.

Perhaps the most difficult of these four inference problems is the last one: Could the dissociation merely reflect the operation of a reorganised system that is qualitatively different from the normal one? For example, according to the right hemisphere theory of deep dyslexia (e.g., Coltheart 1980b; Saffran et al. 1980), any observations of deep dyslexic reading might have no bearing on our understanding of the normal reading system. The nature of the dissociations is again relevant here. Classical dissociations are less prone to this problem as there is no reason a reorganised system should produce the same quantitative level of performance as the original one. This is also true if a strong complementary or classical dissociation exists; it is implausible that two reorganised systems with opposing characteristics should substitute for the original one in different patients.

Considering the four potential artifacts besetting inferences from neuropsychological evidence, the overall danger of misleading conclusions is not that severe if selective impairments are studied and if patients with the complementary dissociation exist. It is necessary, however, to assess the approach in practice. *Neuropsychology* Chapters 3 to 8 are concerned with this.

4. Converging inferences from neuropsychological and normal evidence

To assess how well inferences from neuropsychological findings converge with those from normal experimentation, it is most appropriate to consider areas where both approaches are well developed. *Neuropsychology* considers auditory-verbal short-term memory, reading, writing, and the relation between input and output speech processes in most detail, but it also touches on visual object perception, visual attention, and long-term memory.

Consider first, selective disorders of auditory-verbal short-term memory in the presence of relatively intact language, intelligence, and auditory word perception (*Neuropsychology* Chapter 3). Eight patients (see, e.g., Warrington & Shallice 1969) with these characteristics are discussed in *Neuropsychology* (and six more patients have since been described; see Shallice & Vallar 1990). That these dissociations can be plausibly attributed to a specific impairment in short-term retention can be seen in the nine patients in whom it has been investigated by so-called Brown-Peterson recall (decline is exceptionally rapid) or the "recency" effects in recall, which is very restricted in the number of serial positions affected.

What short-term retention system might be implicated and what would its function be? The patients show a set of dissociations between impaired auditory-verbal span and (relatively) intact visual-verbal short-term retention, intact auditory nonverbal short-term retention, intact verbal long-term memory, and intact speech production. This suggests that a store exists that is specific to the retention of speech input at the phonological level. This is supported by complementary findings in the normal literature (e.g., separate visual short-term store (STS) – Broadbent et al. 1978; Margrain 1967; separate auditory nonverbal STS – Rowe 1974; separate verbal-LTS – Craik & Watkins 1973; Geiselman et al. 1982; separate output-speech STS – Salame & Baddeley 1982). In addition, for certain of the inferences, complementary support also comes from other neuropsychological syndromes (e.g., separate visual STS – Caramazza et al. 1983; separate verbal LTS – Baddeley & Warrington 1970; Drachman & Arbit 1966; separate output-speech STS – Damasio & Damasio 1980). Thus the idea that the retention of speech input – presumably phonological representations – is separable from a number of other related cognitive operations is strongly supported. What specific functions within the speech comprehension process the traces facilitate remains the subject of much debate (see Vallar & Shallice 1990, Chapters 7, 8, 14–18).

Reading disorders present a more complex picture (*Neuropsychology* Chapters 4 and 5). First, there are a considerable number of qualitatively distinct disorders to consider, not just one. Second, certain observed impairments cannot be understood merely by subtracting one or more components from the normal system. Thus, letter-by-letter reading, the way pure alexic patients frequently read, is clearly a compensatory procedure; one school holds that deep dyslexic reading is likewise realised by a system distinct from the normal one. In this domain the possible artifacts to which neuropsychological inference is subject are more than merely theoretical.

Despite these difficulties, if one makes appropriate allowance for the possibility of nontransparent syndromes, complementary inferences both across syndromes and with normal experimental findings again occur. Principal amongst these is the separation of phonological and semantic reading processes that first came to attention on the basis of neuropsychological research (contrast Marshall & Newcombe, 1973, with Rubinstein et al. 1971). Thus four phonological readers have been described (e.g., Schwartz et al. 1980) who read aloud fluently but have little or no ability to comprehend the words they read. In a complementary way, not only deep dyslexics, of whom many have been described (e.g., Marshall & Newcombe 1966), but also three patients with little preserved phonology (e.g., Levine et al. 1982) can read to meaning although they have no useful access to phonological information. In normal subjects dual task experiments also support the possibility of accessing semantic representations independently of phonological ones (e.g., Kleiman 1975). More subtle inferences are that spelling-to-sound units exist that are intermediate in size between grapheme-to-phoneme and lexical (from phonological readers, Shallice et al. 1983; from phonological alexia, Derouesne & Beauvois 1985), that morphemic spelling-to-sound correspondences exist independent of the semantic system (from certain phonological readers, e.g., Schwartz et al. 1980), and that reading-to-meaning is not idiographic (Saffran 1980). The first and third of the inferences also received support from studies of normal subjects (e.g., McClelland 1977; Patterson & Morton, 1985). Overall, the inferences from differ-

ent acquired dyslexic syndromes and from findings on normal subjects support each other.

The situation with respect to writing is similar (*Neuropsychology* Chapter 6). We know much less about the writing system in normal subjects than about the reading system and indeed models of the normal writing process are largely based on a priori analysis of the sort of system that would be required to produce correct written forms in, say, a language as irregular in its sound-to-spelling correspondences as English (e.g., Morton 1980) or on an analysis of spontaneously occurring errors (e.g., Ellis 1979). At least for the central parts of the writing process, however, the neuropsychological evidence is more clear-cut than for reading, as three pure syndromes exist. These are lexical agraphia, preserved nonword writing with impairment of word writing (e.g., Beauvois & Derouesne 1981) (7 cases), the complementary syndrome, phonological agraphia (e.g., Shallice, 1981, 2 cases) and graphemic buffer disorder, in which writing of both words and nonwords is similarly affected, yet motor execution is rapid and fluent (Caramazza et al. 1987). All three syndromes fit well with impairments to components on Morton's model, which was derived from a theoretical analysis of the normal writing process.

It might be argued that reading and writing represent especially productive areas for dissociation analysis because they are skills that cross between evolutionarily more basic domains, although their learned nature would conflict with Fodor's (1983) claims that modules are necessarily innate. In other areas, however, comparable mappings of dissociations between normal and neuropsychological findings also exist. Thus the separability of speech input and output processes at the level of phonology (*Neuropsychology* Chapter 7) is supported both by the syndrome of the auditory parallel to deep dyslexia (e.g., Michel & Andreewsky, 1983, 2 cases) and by dual-task experiments on normal subjects (e.g., Shallice et al. 1985).

Better known are the dissociations in the domain of memory (*Neuropsychology* Chapter 15) between recall/recognition and completion/cueing in amnesic patients (e.g., Warrington & Weiskrantz 1970) and normal subjects (e.g., Jacoby & Dallas 1981) and between episodic and semantic processes in amnesics (Kinsbourne & Wood 1975) and semantic-memory patients (Coughlan & Warrington 1981) and complementarily in normal subjects (Herrmann & Harwood 1980; Tulving 1972; but see also McKoon et al. 1986; Tulving 1986). In this domain, though, it remains an open issue whether the dissociations are appropriately explained in terms of separable systems (see Roediger 1990; Schacter & Tulving 1990).

The amnesia literature is particularly interesting methodologically. The dissociations that the amnesics show in both short and long-term memory (e.g., Baddeley & Warrington 1970) and completion/cueing (Warrington & Weiskrantz 1970) were both established in about 1970 (and incidentally, on the basis of group studies). The effects were rejected as artifactual later in the 1970s, however, for two of the reasons discussed earlier in section 3, problems 1 and 2 (e.g., Butters & Cermak 1974; Woods & Piercy 1974). Later again, they were accepted as valid (e.g., Cermak 1976; Graf et al. 1984). The potential artifacts proved less serious than originally feared. Indeed, I know of no area in which inferences from neuropsychological dissociations were at one time in conflict with those from the normal literature and were later shown to be artifactual. The types of artifact considered earlier have not proved too dangerous in practice.

5. Alternative interpretations of dissociations

What gives neuropsychological findings their force is the specific impairments observed. Dissociations have often been treated as evidence of isolable subsystems (e.g., Fodor 1983; Shallice 1979), a simple explanation of the findings. The presence of dissociations in no way entails the existence of isolable subsystems, however, even if the artifacts, discussed in section 2, which complicate the subtraction approach are not relevant. Dissociations can arise from damage to other types of architecture. For example, *Neuropsychology* Chapter 11 considers as alternatives continuous processing spaces, overlapping processing regions and coupled systems. Dissociations can also occur from impairments to different levels or modes of operation of the same subsystem. Is it possible to distinguish between these alternatives? At least three lines of inquiry seem possible in principle: How do different neuropsychological syndromes relate? How "strong" are individual dissociations? How do the neuropsychological observations map onto results from experimental paradigms using normal subjects? To my knowledge no attempt has been made to distinguish between rival architectures along these lines. The first and third, have been attempted on rare occasions to seek evidence of isolable subsystems. The first was used to predict the existence of further central dyslexic syndromes before they were discovered (e.g., Shallice & Warrington 1980).

One type of architecture deserves special attention because of widespread current interest in it – connectionist (or PDP) architectures (e.g., McClelland & Rumelhart 1986). Wood (1978) suggested that lesions in the input or output layers of a simple two-layer distributed-memory system could produce double dissociations. This critical example is not interestingly generalisable, however, as it depends on the precise input and output vectors used. Some multilayered connectionist models can be viewed as detailed realisations of isolable subsystem. The degree of interaction between the different layers, however, means that the relation between the nature of observed impairments and the global operation of a damaged subsystem is much more opaque than for simple models, say, where individual subsystems compute unambiguous representations. The theory-stimulation function of neuropsychological evidence would be weaker. A model of this sort, which is to my knowledge the first to provide a mechanistic account of detailed characteristics of a syndrome, is that of Mozer and Behrmann (1990), who apply the visual attention model MORSEL to the properties of neglect dyslexia and attentional dyslexia (syndromes discussed in *Neuropsychology* Chapter 13).

Connectionist models may have a more complex methodological effect in cognitive neuropsychology. Hinton and Sejnowski (1986) lesioned single units in the middle layer of a simple three-layer connectionist network that mapped graphemic strings to semantic features. When it was wrong, the system tended to produce explicit errors,

not merely arbitrary collections of semantic features. These errors were, on average, more similar both semantically and visually to the target word than to an average member of the word set. More recently, Hinton and Shallice (1991) lesioned a related but more complex network in a systematic fashion. Wherever the lesion was made, the same qualitative error pattern occurred, a mixture of semantic errors, visual errors, and mixed visual and/or semantic errors, an error pattern characteristic of deep dyslexia (ignoring derivational errors that were outside the domain of the model). Methodologically, the rejection by recent cognitive neuropsychologists (e.g., Caramazza 1984; Ellis 1987) of symptom complexes based primarily on error types may need revision, at least when the theories being considered are connectionist ones with a strong "attractor" structure. If so, this would reinforce the assumption made earlier that the appropriate methodology for cognitive neuropsychology research will depend on the general type of model of the cognitive system being considered.

6. Above the modules (or nets)

Whether the systems that carry out the processing required in complex but routine cognitive operations like face recognition or phonological analysis of an utterance are best characterised as modules or networks, one may pose the question whether the higher level processes that must also exist – particularly those concerned with the allocation of mental resources – operate according to similar principles or different ones. The last section of *Neuropsychology* (Chapters 12–16) is concerned with the neuropsychological evidence on such complex processes – knowledge, visual attention, supervisory operations, memory, and consciousness. On the whole, the methodological scruples of the earlier sections are abandoned, and a dissociation is assumed to represent the operation of a damaged subsystem.

The best known answer in cognitive science to the general question posed in the preceding paragraph is Fodor's (1983). He argues that, in addition to modular input systems, there exist equipotential central systems. About these he says we can learn very little, claiming that nothing is known about the neuropsychology of thought, and that "there is good reason why nothing is known about it – namely that there is nothing to know about it . . . in the case of central processes you get an approximation to universal connectivity, hence no stable neural architecture" to describe (Fodor 1983, p. 119). [See also multiple book review of Fodor's *Modularity of Mind*, *BBS* 8(1) 1985.]

Closer examination of the neuropsychological evidence suggests that Fodor's answer is inadequate. Domains of knowledge can be selectively impaired (see *Neuropsychology* Chapter 12) – for example, such routine thought operations as elementary arithmetic, as in acalculia (see Warrington 1982b). Yet the carrying out of elementary arithmetic operations must depend on systems that would not be modular according to a number of Fodor's (1983) criteria for modules. Yet, if one provisionally adopts his assumption about input systems that an observed selective impairment is evidence for a module – and for Warrington's acalculia (1982b) patient the dissociation was both classical and deep – then elementary arithmetic operations are not the product of a global central equipotential system.

At the very least it would seem that Fodor's thesis requires some revision. One possibility might be to move the upper boundary of what counts as the input system and to separate such routine thought operations from the central system. The processes involved in the so-called frontal syndrome must be ineluctably central, however. These have been characterised by Luria (1966) as being concerned with the programming, regulation, and verification of activity. Can disorders of these processes be related to more standard cognitive scientific ideas, and could they be characterised as impairments of Fodor's central system?

A model developed by Norman and Shallice (1986) can be viewed as an attempt to anchor the overall theory Luria applied to "frontal functions" within a cognitive scientific framework. Our model was based on two main premises. The first, introduced at the beginning of Chapter 13, is that the routine selection of routine operations is decentralised. It was suggested that the basic units underlying action or thought are a very large but finite set of discrete programs – thought or action schemas – that either place a particular pattern of demands on the mosaic of functionally specific subsystems directly or when certain specific circumstances arise. A schema is selected if its level of activation exceeds a given threshold; once selected it remains active even if its level of activation falls, unless it attains its goal or is actively inhibited by a competitor or by any higher level controlling schema. Schemas are independently activated by triggers and are in mutually inhibitory competition.

This process of selection between routine action or thought operations is termed "contention scheduling." In certain artificial intelligence work on problem solving, however, it has been found necessary to add to a system that routinely executes whatever solution procedure is in operation a planning component that operates differently and learns from its mistakes (see Boden 1977; Charniak & McDermott 1985, for review). We therefore argued that there is an additional system – the Supervisory System – which has access to representations of the environment and of the organism's intentions and cognitive capacities. This system operates not by directly controlling behaviour, but by modulating the lower level contention-scheduling system by activating or inhibiting particular schemas. It would be involved in the genesis of willed actions and required in situations where the routine selection of actions is unsatisfactory – for example, in coping with novelty, in decision making, in overcoming temptation, or in dealing with danger.

Characteristics of frontal patients such as their "stuck-in-set perseverations" (Sandson & Albert 1984) as exhibited in the Wisconsin card-sorting task (Milner 1963) and their apparently opposite tendency to respond to objects in their environment by using them even when they have no reason to do so – "utilisation behaviour" (Lhermitte 1983; Shallice et al. 1989) – can be explained as follows: Damage to the supervisory system releases the unmodulated operation of contention scheduling. In the case of

"stuck-in-set perseverations," it is assumed that a particular thought schema, which would control a set that is no longer appropriate, remains strongly activated by the stimulus situation through prior overlearning and cannot be effectively overridden without the supervisory system. In the latter case, the lack of any other strongly activated schema allows a schema activated only by a stimulus trigger to become selected. These behaviors are related to normal subjects' "action lapses" (see Reason 1984) when the supervisory system is occupied by some unrelated thought process; similar inappropriate data-driven activation by overlearned triggering stimuli would be involved in both cases.

The errors of frontal patients on more complex tasks are also well described as a loss of a planning or programming component, as Milner et al. (1985) point out. Yet does one need to postulate a presumably complex supervisory system to carry out this function and not just an ability, say, to inhibit inappropriate "central sets," as suggested by Rosvold and Mishkin (1961)? In fact, certain apparent difficulties of frontal patients seem difficult to explain by this simpler account. Knight (1984), for example, showed that frontal patients lack the special P300 response to novel stimuli exhibited by normal subjects – an absence of a positive response, not the lack of an inhibitory one. [See Donchin & Coles: "Is the P300 Component a Manifestation of Context Updating?" *BBS* 11(3) 1988.]

In addition, it would seem that a supervisory system modulating the operation of action and thought schemas would be far from internally equipotential as claimed by Fodor (1983) for his "central systems." Animal experiments show a considerable degree of specialisation in prefrontal cortex (see, e.g., Fuster 1980; Petrides 1987). In humans, fractionations of the frontal syndrome are beginning to be reported. Shallice and Burgess (1991) have described two patients who perform well on frontal lobe tasks except those that require the laying-down and realising of intentions.

A clearer example of a specific function with which the supervisory system should be concerned is dealing with what should happen if no existing routine thought or action schema were adequate to achieve a particular goal or, indeed, if none were strongly triggered by the present combination of goal and events. A person may be able to draw on a memory of what happened in an analogous situation, as Schank (1982) has pointed out. The whole process of remembering has been viewed by Norman and Bobrow (1979) as a series of cycles of specifying descriptions, matching with records, and verifying candidate memories retrieved. This approach fits well with the view that the primary function of episodic memory is to provide the supervisory system with an additional means of tackling nonroutine problems. According to the model, the articulation of descriptions and the process of verification would be controlled by the frontally located supervisory system.

Is there any evidence that supports this speculation? Aspects of the memory disorders of some frontal patients are hard to explain in terms of a general organisational problem in encoding material. Smith and Milner (see Milner et al. 1985) found that patients with frontal lobectomy had more difficulty on a frequency discrimination task than other subjects. The authors argued that the task calls for an orderly search through memory and that this may be the source of the difficulty in patients with frontal lesions. Damage to the description and verification stage of Norman & Bobrow's (1979) theory would produce just these types of difficulty.

More direct evidence can be obtained from the confabulations that occur in so called "frontal amnesia." An impairment at the verification stage in Norman & Bobrow's model seems to account well for the difficulty of one such patient (RW) studied by Delbecq-Derouesne et al. (1990). In a task like retention of a paired associate, where little strain is placed on the verification process, he performed well; in this respect, his disorder differed markedly from that of classical amnesics. By contrast, if stimuli were present that elicited an irrelevant association, RW was incapable of selecting which of the responses retrieved was the valid one: In a recall task confabulations were produced and in recognition tasks distracters tended to be confidently selected.

Frontal amnesia, then, appears to be an impairment of that part of the supervisory system concerned with formulating the description of any memories that might be required and of verifying that any candidate memories that have been retrieved are relevant. Classical amnesia, by contrast, would arise from an interruption of the flow of memory information from the processing systems to the supervisory system. On this theory, the contrast between semantic and episodic memory is replaced by a related one. The new contrast is between information accessible through the operation of some routine schema operating directly on the processing system (the semantic memory system) and information that requires the supervisory system to formulate a description and to verify any record retrieved. The damage, moreover, seems to be primarily to only part of the supervisory system because on most "frontal lobe" tasks RW was at most only mildly impaired.

Using the observation of a selective impairment as evidence for the existence of a specific processing system, as Fodor (1983) did for input systems, also supports modularity in Marr's sense for the so-called "central" systems. The systems that modulate the on-line processing systems may be almost as complex and variegated as the systems under their control.

Finally, in *Neuropsychology* Chapter 16 it is argued that the supervisory processes, together with those involved in contention scheduling, episodic memory, and the linking of language with other cognitive processes, are responsible for the existence of conscious experience. A functionalist account along these lines can at least provide an explanation for the counterintuitive neuropsychological findings related to consciousness that have been described such as blindsight, aspects of prosopagnosia, and the split-brain syndrome. [See also: Puccetti & Dykes: "Sensory Cortex and The Mind-Brain Problem" *BBS* 1(3) 1978; and Campion: "Is Blindsight an Effect of Scattered Light, Spared Cortex, and Near-Threshold Vision?" *BBS* 6(3) 1983.]

NOTE

1. For reprints, please contact Tim Shallice, Department of Psychology, University College London, Gower Street, London WC1E 6BT, England.