Double Dissociation: Understanding its Role in Cognitive Neuropsychology

MARTIN DAVIES

Abstract: The paper makes three points about the role of double dissociation in cognitive neuropsychology. First, arguments from double dissociation to separate modules work by inference to the best, not the only possible, explanation. Second, in the development of computational cognitive neuropsychology, the contribution of connectionist cognitive science has been to broaden the range of potential explanations of double dissociation. As a result, the competition between explanations, and the characteristic features of the assessment of theories against the criteria of probability and explanatory value, are more visible. Third, cognitive neuropsychology is a division of cognitive psychology but the practice of cognitive neuropsychology proceeds on assumptions that go beyond the subject matter of cognitive psychology. Given such assumptions, neuroscientific findings about lesion location may enhance the value of double dissociation in shifting the balance of support between cognitive theories.

Cognitive neuropsychology uses patterns of impairment and sparing in patients following brain injury in order to constrain theories of normal cognitive structures and processes. It emerged as a distinctive research programme in the 1960s and by the end of the 1980s had its own journal (volume 1, 1984) and its own textbook (Ellis and Young, 1988/1996). In the practice of cognitive neuropsychology, the evidential value of double dissociation as support for claims about separate modules in the normal cognitive system has been highlighted.

This highlighting is sometimes interpreted as the manifestation of a methodological assumption about a special *logic of cognitive neuropsychology*. For example, Karalyn Patterson and David Plaut say that 'the gold standard was always a double dissociation' (2009, p. 43) and describe 'traditional cognitive neuropsychology logic' as resting on an assumption that 'the functional organization of cognition can be unequivocally revealed by dissociation' (p. 44). I reject the claim that the practice of cognitive neuropsychology is based on assumptions about a special logic. Specifically, I reject the claim that cognitive neuropsychology relies on an assumption that the inference from a pattern of impairment and sparing to a structure of separate cognitive modules is underwritten by a special deductive rule of double dissociation inference.

I am grateful to Anne Aimola Davies, Max Coltheart, Kim Plunkett, Nick Shea, Michael Smithson, and two anonymous referees, for comments on an earlier version of this paper.

Max Coltheart was one of the founding editors of *Mind & Language* and has written with distinction and influence on the themes of this paper—the aims and assumptions of cognitive neuropsychology, double dissociation, computational modelling, and the relationship between cognitive psychology and neuroscience. My intellectual debts to him will be evident on every page and the paper is dedicated to him with gratitude.

Address for correspondence: Corpus Christi College, Merton Street, Oxford OX1 4JF, UK.
Email: martin.davies@philosophy.ox.ac.uk

Mind & Language, Vol. 25, No. 5 November 2010, pp. 500–540.

One reason to reject the claim is that it is difficult to locate texts in which cognitive neuropsychologists unambiguously claim that their approach has a distinctive logic with proprietary rules of inference.[1] A further reason is that it is a familiar point about science in general that there is no logically valid deductive inference from evidence to explanatory theory. To be explanatory, a theory must go beyond a summary of the evidence and so cannot be entailed by the evidence. Within the narrower domain of psychology, it is well understood that there is no logically valid deductive inference from data, such as reaction time data, to an explanatory theory about cognitive structures and processes. It is very unlikely that cognitive neuropsychologists regard their research programme as being different from all the rest of empirical science and take themselves to have access to evidence with magical properties.

There are good reasons to reject the claim that the practice of cognitive neuropsychology is based on assumptions about a special logic. Nevertheless, that claim seems to be part of the 'received wisdom' about cognitive neuropsychology, in the sense that many people believe—or many people believe that many people believe—that its practice does rest on such assumptions.[2] This hypothesis about received wisdom helps to render intelligible the existence of influential papers with titles including, 'What can we infer from double dissociations?' (Dunn and Kirsner, 2003), 'What do double dissociations prove?' (Van Orden, Pennington and Stone, 2001), and 'Why double dissociations don't mean much' (Juola and Plunkett, 2000). More precisely, the hypothesis helps to explain the existence of those papers, given the notable absence of papers called 'What can we infer from evidence?', 'What does data prove?', or 'Reaction times don't mean much'.

If the practice of cognitive neuropsychology really had relied on a special logic with a special rule of double dissociation inference, this would have had serious consequences for the research programme. Over the last twenty years or so, developments in computational modelling, and particularly in connectionist cognitive science, have revealed a broad range of potential explanations of double dissociation. It has been demonstrated that the special rule of inference is logically invalid. If cognitive neuropsychology had been relying on a special logic including the special rule then these developments might have led, if not to the death of cognitive neuropsychology as originally conceived, at least to its entry into a persistent vegetative state (PVS).[3]

---

[1] When Max Coltheart (1985, p. 17) speaks of 'remorselessly pursuing the logic of cognitive neuropsychology', he is not alluding to special inference rules. Rather, he is indicating the interest, for cognitive neuropsychology, of patients whose patterns of impairment and sparing would be readily explained in terms of particular proposals about separate modules in the normal cognitive system.

[2] I am indebted to David Chalmers for discussion about the notion of received wisdom.

[3] Here, I borrow the words of an anonymous referee. Following the publication of the *Handbook of Cognitive Neuropsychology* (Rapp, 2001), the journal *Cognitive Neuropsychology* published a review (Harley, 2004) under the challenging title, 'Does cognitive neuropsychology have a future?'.

Since cognitive neuropsychology does not really rely on the special rule of inference, a demonstration that the rule is logically invalid should not lead to the real death of cognitive neuropsychology, or to its real entry into a PVS. But the received wisdom remains unhelpful. In science as in life, being thought to be dead or in a PVS, or even being thought to be thought to be dead or in a PVS, is damaging to one's prospects. It is time to overturn the received wisdom about cognitive neuropsychology.

In this paper, I defend a very different account of cognitive neuropsychology and the role of double dissociation. First, cognitive neuropsychology was and is normal empirical science proceeding, not by special rules of deductive inference, but by inference to the best explanation. This first stage (Sections 1 to 3) involves close attention to the notion of double dissociation that was introduced by Hans-Lukas Teuber (1955) and the difference between that notion and the one that is employed in cognitive neuropsychology, to potential explanations of double dissociation in terms of damage to separate modules in the normal cognitive system, and to the problem of resource artefacts.

Second, the primary significance of computational modelling for cognitive neuropsychology is not that it demonstrates the logical invalidity of a special rule of inference. So we need an alternative account of the importance of the development of computational cognitive neuropsychology and the contribution of connectionist cognitive science. In this second stage (Sections 4 to 7), I focus on the domain of cognition in which computational modelling of normal and impaired performance is most developed, namely, reading aloud. Computational modelling broadens the range of potential explanations of double dissociation and, in the mature phase of cognitive neuropsychology, its abductive methodology is increasingly evident. Competing theories are assessed for their probability in the light of large bodies of available evidence (including, of course, double dissociation evidence) and also for their explanatory virtues. Here, and throughout the paper, I draw on Peter Lipton's (1991/2004) seminal work on inference to the best explanation.

Cognitive neuropsychology differs from its antecedents in 19th-century neurology in that it is not primarily concerned with the brain-behaviour nexus. It is conceived, instead, as a division of cognitive psychology. But the practice of cognitive neuropsychology proceeds on assumptions that go beyond the subject matter of cognitive psychology and range into neuroscience. This raises a large issue that requires a paper of its own; namely, whether cognitive neuropsychology—and cognitive psychology more generally—is autonomous from neuroscience. But the resources of the present paper allow us to take a first step towards addressing the autonomy question. Teuber's (1955) definition of double dissociation requires that the two patients have lesions in different locations, but this condition is absent from the notion of double dissociation that is employed in cognitive neuropsychology. In Section 8, I show how a neuroscientific finding about lesion location may enhance the value of double dissociation in shifting the balance of support between competing cognitive theories.

1. Double Dissociation and Localisation of Function

Teuber (1955) discussed localisation of function, and what he called 'specificity of function', beginning from the point that bilateral lesions of the temporal lobes result in impaired visual discrimination. The question whether this is a case of specificity of function can be developed in two ways. The *localisation question* is whether the temporal lobes are involved in visual discrimination *while other neural areas are not*. The *dedication question* is whether the temporal lobes are involved in visual discrimination *but not in other functions*. Teuber did not simply assume that each function is narrowly localised in a region of the brain that is, in turn, dedicated to it. Rather, he paid considerable attention to the sceptical views about localisation expressed by Karl Lashley (1930).

Two points from Lashley are particularly important. First, one function might be localised in a small region while a second function is localised in a larger region including the first; indeed, some functions are localised in the whole cortex. In Lashley's example, brightness discrimination is localised in the area striata but the rat's maze habit is localised in the whole cortex (1930, p. 13):

> The habit of brightness discrimination in the rat is abolished by injury to the area striata, and by injury to no other part of the cortex. Here is a clear case of specialization. But the maze habit is abolished by destruction of this same area or of any other of equal size.

Second, Lashley allowed for the possibility that two functions may be localised in the very same region and he provided an example of the way in which lesions might affect two such functions (1930, p. 15):

> [T]he habit of threading a complex maze is seriously disturbed by destruction of any part of the cortex, provided the lesion involves more than 15 per cent. The habit of a simpler maze is unaffected by lesions involving as much as 50 per cent of the cortex.

Referring to Lashley's work, Teuber said (1955, p. 280): 'Conceivably, any neocortical lesion large enough to produce symptoms has a double effect: a specific one, depending on locus, and a general one, perhaps depending on size.' So, given that temporal-lobe lesions impair visual discrimination, we should ask whether visual discrimination is narrowly localised in the temporal lobes (cf. brightness discrimination and the area striata) or is more broadly localised (cf. the maze habit and the whole cortex). Progress could be made on this localisation question by examining the consequences for visual discrimination of lesions outside the temporal lobes, and Teuber presented some evidence of this kind (1955, p. 282).

If evidence were to support the hypothesis that visual discrimination is localised in the temporal lobes, there would remain the question whether the temporal lobes are dedicated to visual discrimination. Is their function specifically visual or more

general? Evidence that temporal-lobe lesions do not impair functions other than visual discrimination would seem to support the hypothesis of dedication. But, the significance of such evidence could be challenged by appeal to Lashley's claim (1930, p. 16): 'a given area may function at different levels of complexity and lesions may limit the complex functions without disturbing the simpler ones.' As Teuber himself explained the dialectical situation (1955, p. 283):

> To demonstrate specificity of the deficit for visual discrimination we need to do more than show that discrimination in some other modality, e.g. somesthesis, is unimpaired. Such simple dissociation might indicate merely that visual discrimination is more vulnerable to temporal lesions than tactile discrimination. This would be a case of hierarchy of function rather than separate localization.

Teuber proposed that better evidence for specificity of function would be provided if temporal lesions were to produce the one-way dissociation—tactile discrimination is spared while visual discrimination is impaired—and *lesions of some other area were to produce the reverse pattern of dissociation*. He called this combined pattern of evidence *double dissociation*.

> **Teuber's definition of double dissociation** One brain-injured patient (A) shows unimpaired performance on Task I (e.g. tactile discrimination) but impaired performance on Task II (e.g. visual discrimination) while a second patient (B), with a different lesion site, shows the reverse pattern, unimpaired on Task II but impaired on Task I.

In outline, the argument from double dissociation to separate localisation of functions is as follows. The fact that patient A is impaired on Task II supports the hypothesis that Task II is localised in an area including the site of patient A's lesion. Patient B's pattern of impairment and sparing provides evidence for two further hypotheses. First, the fact that patient B is unimpaired on Task II supports the hypothesis that the area in which Task II is localised does not extend as far as the site of patient B's lesion. Second, the fact that patient B is impaired on Task I supports the hypothesis that patient A's pattern of impairment and sparing is not to be explained by appeal to different levels of complexity for the two tasks. Thus, the double dissociation provides evidence that Task II is localised in an area, including the site of patient A's lesion, that is not involved in Task I. Since a double dissociation is symmetrical, it also provides evidence that Task I is localised in an area, including the site of patient B's lesion, that is not involved in Task II.[4]

[4] Separate localisation may not be fully dedicated localisation. In Teuber's example, evidence supported the hypothesis that the temporal lobes are involved in visual, but not tactile, discrimination. But evidence on the involvement, or not, of the temporal lobes in *auditory* pattern discrimination was not available (1955, p. 284).

Teuber did not say merely that double dissociation provides better evidence for specificity of function than simple one-way dissociation does. He said that double dissociation is 'what is needed *for conclusive proof*' (1955, p. 283; emphasis added). This may suggest that double dissociation provides a *logical guarantee* of functional specificity; but it has long been recognised that any such suggestion would be incorrect. In an early commentary, Lawrence Weiskrantz speaks of 'the logical insufficiency of double dissociation for drawing an inference' (1968, p. 419) and he later reaffirms (1989, p. 105): 'The conclusion does not follow with logical certainty; it is a pragmatic argument, but no less valuable for that.'

2. Cognitive Neuropsychology

Neuropsychology as conceived by Teuber made common purpose with behavioural neurology because it was concerned with the brain-behaviour nexus. Indeed, an Editorial in the first issue of the journal *Neuropsychologia* described neuropsychology as 'a particular area of neurology'. Neuropsychology as behavioural neurology went beyond the limits imposed by a strictly behaviourist psychology but it was still very different from cognitive neuropsychology, which is a division of—a research programme within—cognitive psychology. As Max Coltheart puts the point (2001, p. 4): 'Cognitive neuropsychology is not a kind of neuropsychology . . . because, to put the matter in a nutshell, cognitive neuropsychology is about the mind, while neuropsychology is about the brain.'

2.1 Aims and Assumptions of Cognitive Neuropsychology

Research in cognitive neuropsychology has two complementary aims, set out by Coltheart in the first issue of the journal *Cognitive Neuropsychology* (1984, pp. 1–2):

> The first is to evaluate models of normal cognition by exploring their success in explaining the precise patterns of performance exhibited by people suffering from disorders of cognition. The second aim is to offer theoretically-motivated explanations of precisely what has gone wrong, and what remains intact, in the multicomponent mental system responsible for the relevant mental activity.

As an approach to the investigation of normal cognitive structures and processes, cognitive neuropsychology relies on a number of assumptions. The first is that the mind is modular. The generic notion of *modularity* is familiar from everyday life, where furniture and stereo systems are described as 'modular'. They are built from components, each of which makes a somewhat independent contribution to the functionality or the performance of the system as a whole. In the context of cognitive neuropsychology, the modularity assumption is that the normal cognitive system is made up of somewhat independent cognitive components. It is a further question whether these cognitive components are Fodorian modules and it is

important to observe that this further question is not simply equivalent to the question whether the components have *all* the marks of modularity listed by Jerry Fodor (1983). The Fodorian notion of a module is associated with a cluster of properties that tend to go together and it may apply to a cognitive component to a greater or lesser degree (see Coltheart, 1999, for discussion).

The second assumption is that, when one cognitive component is damaged, this does not bring about massive reorganisation of the prior modular structure. Consequently, the undamaged components continue to operate as before, so far as this is compatible with the impaired operation of the damaged component. Alfonso Caramazza (1986, p. 52) calls this the assumption of *transparency*. Coltheart calls it the *subtractivity* assumption (2001, p. 10): 'brain damage can impair or delete existing boxes or arrows [components] in the system, but cannot introduce new ones: that is, it can subtract from the system, but cannot add to it'.

Following brain injury that impairs the normal method of performing a task, a patient may still be able to perform the task or may learn anew how to perform the task—and may achieve normal levels of performance—by using a compensatory strategy or workaround. It is important to notice that this possibility does not, by itself, call into question the subtractivity assumption. A compensatory strategy may not require new cognitive components but may just make novel use of the operation of intact components of the normal cognitive system.[5]

The third assumption is that the modular structure or *functional architecture* of the mind as a whole, and of the cognitive systems responsible for the performance of particular tasks, is the same for all normal (neurologically intact) subjects. This assumption of *universality* (Caramazza, 1986, p. 49) does not, of course, exclude individual differences (within the normal range) in the performance of a cognitive component or in the efficiency of a pathway from one component to another. But the total absence of a component, or of a pathway, will be regarded as an abnormality.

When we study neurologically healthy individuals, the assumption of universality licenses the averaging of data across groups of subjects in order to assess hypotheses about the normal information-processing system. When we study brain-injured individuals, however, we cannot antecedently assume that the information-processing systems of different patients have been damaged in identical ways—even if the patients have been given the same clinical diagnosis or are described as exhibiting the same syndrome. Thus, Caramazza argues (1986, pp. 54–5):

> We would be justified in averaging the performance of a group of subjects *only* if we could assume that the nature of damage . . . to a particular cognitive system in each patient is identical in *all theoretically relevant respects*. . . . Without the assumption of identical [damage] . . . the grouping of patients' performance results in meaningless entities.

[5] The possibility of rehabilitation and re-learning does call for care in interpreting the results of neuropsychological assessment of patients following brain injury.

Instead of using averaged data from groups of patients, cognitive neuropsychologists invoke hypotheses about damage to components of the normal cognitive system as potential explanations of specific patterns of impaired performance shown by individual patients. Thus, cognitive neuropsychology often proceeds by the study of single cases. A *series* of single-case studies yields multiple constraints on the functional architecture of the normal cognitive system that these patients shared, according to the assumption of universality, before their injuries. A theory about the normal cognitive system is supported to the extent that, for the pattern of impaired performance shown by each patient in the series, the theory allows an explanatory hypothesis that is consistent with the subtractivity assumption.

2.2 The Beginnings of Cognitive Neuropsychology

Studies of patients with impaired memory provide an early example of the cognitive neuropsychology approach. Consider a model of normal memory in which the route to laying down traces in long-term memory goes through short-term memory, so that long-term memory requires all that short-term memory requires and more. A hierarchical model of this kind allows an explanation of the pattern of impaired memory exhibited by patient HM, studied by Brenda Milner and her colleagues (Scoville and Milner, 1957; Milner, Corkin and Teuber, 1968). Following surgery to remove parts of the medial temporal lobe on both sides of his brain, HM could not commit new events to long-term memory, although his short-term memory was intact—for example, his scores on the Digit Span subtest of the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955) were six forward and five backward. An explanatory hypothesis for patient HM's pattern of impairment can appeal to subtraction of the 'more' that long-term memory requires, while the remainder of the memory system continues to operate as before.

The hierarchical model of memory that was dominant in the 1960s predicts that, if short-term memory is impaired, then long-term memory will be impaired as well. But patient KF (Warrington and Shallice, 1969; Shallice and Warrington, 1970) showed impaired short-term memory with spared long-term memory.[6] Following a severe head injury that caused left parietal damage, patient KF performed very poorly on Digit Span and, when presented with two digits, could recall only the first; his performance was also poor when letters or words were used instead of digits. But his ability to commit new material to long-term memory—assessed, for example, by the Paired-Associate Learning subtest of the Wechsler Memory Scale (WMS; Wechsler, 1945)—was normal. This finding shifted the balance of support

[6] Tim Shallice (1988) reports that, when he first heard from Elizabeth Warrington the suggestion that patient KF had impaired short-term memory but intact long-term memory, he told her that 'this was theoretically impossible' (pp. 41–2): 'Waugh and Norman's (1965) impressive paper had just appeared, and it seemed to settle the relationship between short- and long-term memory, with the short-term store (or primary memory) the vehicle for the laying down of traces in long-term storage.'

towards heterarchical models of memory in which there is a route to the long-term memory store that does not go via the auditory-verbal short-term memory store that underpins performance on Digit Span and similar span tasks (see Shallice, 1988, chapter 3).

In a similar way, the reading performance of patient GR (Marshall and Newcombe, 1966, 1973) challenged then-dominant hierarchical model of reading comprehension as dependent on phonological mediation. According to this model, access to lexical semantics from orthography requires conversion of orthography to phonology and then, as for spoken words, access to lexical semantics from phonology. Following a severe missile injury to the left temporo-parietal region, patient GR would often produce semantic errors when asked to read single words aloud. For example, he read 'daughter' as 'sister' and 'guilty' as 'hangman'. These errors had two features. First, GR was unable to achieve the pre-lexical conversion of orthography to phonology; when the word 'guilty' was read as 'hangman', not a single phoneme was correct. But, second, patient GR achieved some access to (perhaps degraded) lexical semantics. The pronounced word was semantically related to, and apparently semantically constrained by, the orthographically presented word. In short, patient GR was able to achieve access to lexical semantics from orthography without pre-lexical conversion of orthography to phonology—contrary to the dominant model at that time.[7] This finding shifted the balance of support towards heterarchical models of reading comprehension in which there is a direct route from orthography to lexical semantics as well as an indirect route via phonology.

In these two examples of early work in cognitive neuropsychology, patterns of impairment exhibited by patients, KF and GR, shifted the balance of support away from dominant hierarchical or single-route models of memory and of reading. The balance of support was shifted towards a heterarchical model of memory with two routes to long-term memory and towards a heterarchical model of reading with two routes to reading comprehension.

2.3 Double Dissociation in Cognitive Neuropsychology

The notion of double dissociation used in cognitive neuropsychology is usually defined in terms of complementary dissociations of impairments in a pair of patients:

> One patient (A) shows unimpaired performance on Task I but impaired performance on Task II while a second patient (B) shows the reverse pattern, unimpaired on Task II but impaired on Task I.

[7] Andy Young (2001) reports John Marshall's reaction when he first heard from Freda Newcombe about the case of patient GR: '[He] reacted with disbelief when Newcombe told him that one of the ex-servicemen misread words for others of similar meaning (for example, reading "canary" as "parrot"). Knowing that theories of reading all stipulated that print must be converted to sound before its meaning could be understood, Marshall pointed out that a reading error based on common meaning without a shared similarity in sound was "impossible", and that she "must be mistaken".'

This differs from Teuber's (1955) definition in that it imposes no requirement on the sites of the patients' lesions. (We shall return to Teuber's definition in Section 8.) In cognitive neuropsychology, double dissociation evidence is used to support claims about separate cognitive components or modules rather than about separate localisation. Separate cognitive components are susceptible, in principle, of independent damage. If two separate components are distinctively implicated in two tasks then independent damage to each of the components affords an explanation of a double dissociation of impairments on those tasks.

Patients HM and KF instantiate a double dissociation of memory impairments. Patient HM showed intact short-term memory as assessed by Digit Span but impaired ability to commit new information to long-term memory while patient KF showed the reverse pattern. A heterarchical model of normal memory in which there is a route into long-term memory that is relatively independent of the system for auditory-verbal short-term memory can account for this double dissociation. In the case of patient HM's pattern of impairment, an explanatory hypothesis appeals to damage to cognitive components that are implicated in long-term memory but not in auditory-verbal short-term memory. In the case of patient KF's reverse pattern, an explanatory hypothesis appeals to damage to components that are implicated in auditory-verbal short-term memory but not in long-term memory. The balance of support is shifted towards a heterarchical or two-route model of normal memory because it allows a better explanation of the double dissociation of impairments than the previously dominant hierarchical model does.

2.4 Association, One-Way Dissociation and Double Dissociation

Arguments from double dissociation of impairments to separate modules in the normal cognitive system play an important role in cognitive neuropsychology. Association of impairments and simple one-way dissociation are standardly regarded as having less evidential value. First, it is said that association of impairments is apt to be misleading if it is taken as evidence that two tasks draw on the same component, rather than separate components, of the normal cognitive system. The reason is that, while separate components are in principle susceptible of independent damage, the details of their neural localisation may make it overwhelmingly likely that a brain injury that damages one of them will also damage the other. An association of impairments that arose from contiguous or overlapping localisation of separate cognitive components would be 'uninteresting to cognitive psychology' (Coltheart, 1985, p. 9).

Second, it is said that one-way dissociation of impairments is apt to be misleading if it is taken as evidence of separate cognitive components. The reason is that, even if two tasks draw on the same component of the normal cognitive system, damage to that component may result in performance of the more difficult task being impaired while performance of the easier task remains intact.

Arguments from association of impairments to shared modules in the normal cognitive system, or from one-way dissociation of impairments to separate modules,

must address the possibility of alternative explanations of the evidence—explanations that appeal to contiguous localisation or to differences in task difficulty. In the case of double dissociation of impairments, these particular kinds of alternative explanation are not clearly available. In this sense, double dissociation is cognitive neuropsychology's 'single most powerful tool, which was and is currently used to reveal the functional independence of discrete components of the cognitive system' (Vallar, 2004, p. 45).

None of this should be allowed to suggest that double dissociation rules out the possibility of any kind of alternative explanation. There is no special rule of inference taking us from double dissociation to separate modules. Double dissociation arguments in cognitive neuropsychology, like arguments from evidence to theory throughout normal empirical science, are not deductive but abductive; they work by inference to the best—not the only possible—explanation (Coltheart, 2001, pp. 15–16; Lipton, 2004).

3. Double Dissociation and Resource Artefacts

According to Teuber, a simple one-way dissociation of impairments on two tasks is not sufficient by itself to demonstrate separate localisation. In his example, visual and tactile discrimination may both be temporal-lobe functions, with visual discrimination more vulnerable than tactile discrimination to temporal-lobe damage. A similar reservation is standard in cognitive neuropsychology and Tim Shallice offers a simple example (1988, p. 232): '[C]onsider a patient who could read frequent words, but not infrequent ones. One would not want to argue that different subsystems were involved in the tasks, but merely that the latter task was the more demanding.' A one-way dissociation of impairments might be explained without appeal to separate modules. It might be, in Shallice's terminology (*ibid.*), a *resource artefact*.

Suppose that, for each cognitive component or module, there is an associated notion of resource meeting two conditions: first, damage to the component reduces the amount of resource that is available; second, performance of a task declines, or at least does not improve, as the available resource is reduced. Then, if performance of a task is plotted against available resource, the performance/resource curve is non-decreasing; equivalently, the performance/damage curve is non-increasing. Subject to that restriction, a component system may have different performance/resource curves for different tasks. We have a resource artefact when a pattern of performance can be explained in terms of different performance/resource curves for different tasks, on the assumption that the tasks are performed by the same system. A one-way dissociation of impairments on two tasks may have a natural explanation along these lines. As the available resource is reduced by damage to the system, performance of the more difficult task may fall away more steeply than performance of the easier task.

3.1 Defining Double Dissociation

The standard view in cognitive neuropsychology, like Teuber's view, is that double dissociation avoids the problem of resource artefacts. But Shallice shows that the problem may persist unless care is taken over the definition of double dissociation.

We have defined double dissociation as a pair of complementary *within-patient* dissociations. Each patient's performance should be unimpaired or within the normal range on one task but impaired or below the normal range on the other task. However, this definition faces the problem that, whatever criterion is set for the normal range (for example, performance within two standard deviations of the mean), the patient's performance on both tasks could be arbitrarily close to the boundary between the normal and the impaired. In order to deal with this problem, John Crawford and colleagues propose that the requirement for a one-way dissociation should include 'a statistically significant difference between the patient's scores on the two tasks' (Crawford, Garthwaite and Gray, 2003, p. 361).[8]

Adopting Crawford and colleagues' proposal, we could define double dissociation, once again, as a pattern of complementary within-patient dissociations:

> **DD1** One patient (A) performs Task I significantly better than Task II while a second patient (B) shows the reverse pattern, performing Task II significantly better than Task I.

But Shallice shows that pattern DD1 by itself does not avoid the problem of resource artefacts. We might account for this pattern of performance in terms of two levels of damage to a single system with different non-decreasing performance/resource curves (equivalently, different non-increasing performance/damage curves) for the two tasks (Shallice, 1988, p. 234, Figure 10.5). To see this, suppose that performance of Task I is resilient to modest degrees of damage to the system but falls away dramatically with more severe damage, while performance of Task II declines linearly with damage. Suppose, too, that patient A has suffered modest damage to the system and patient B has suffered severe damage. Then it is clear that patients A and B may exhibit pattern DD1.

Shallice (1988, p. 235) contrasts pattern DD1 with a pattern of complementary *between-patient* comparisons that does avoid the problem of resource artefacts:

> **DD2** On Task I, patient A performs significantly better than patient B while on Task II the situation is reversed, with patient B performing significantly better than patient A.

[8] Crawford, Howell and Garthwaite (1998) provide a method to compare the difference between a single patient's scores on two tests with the corresponding inter-test differences in the control sample. Crawford and Garthwaite (2005) provide an improved method for making this comparison, the Revised Standardized Difference Test.

We cannot account for pattern DD2 in terms of a single system with non-decreasing performance/resource curves for both tasks. The combined pattern DD1 with DD2 is *double dissociation with (significant) cross-over*.

One problem with pattern DD1 by itself was that it does not require that the two between-patient differences should be complementary. For example, it does not exclude the possibility that patient B should perform worse than patient A on both tasks. An apparent way to avoid this problem is to require that each patient's superior performance (Task I for patient A, Task II for patient B) should be within the normal range, while performance on the other task should be impaired. This is what Teuber required for double dissociation and what we required in the previous section (see also Coltheart, 1985, p. 10). When this requirement is met, each patient's dissociation of impairments is described as *classical*, as is the double dissociation (Shallice, 1988, p. 227).

It is important to observe, however, that this requirement may leave a residual problem. Even a double dissociation that exhibits pattern DD1 *and is classical* might not exhibit pattern DD2. Both dissociations might be classical without both between-patient differences being significant. Consequently, there are hypothetical cases in which a classical double dissociation that exhibits pattern DD1 is accounted for in terms of damage to a single system with different non-decreasing performance/resource curves for the two tasks. Pattern DD2 is crucial if the problem of resource artefacts is to be avoided, while the requirement that the double dissociation should be classical turns out to be less important.

3.2 Double Dissociation and Discovery

Some commentators may have interpreted Shallice's (1988) careful examination of the problem of resource artefacts as supporting the idea of a special inference rule. They may have read it as proposing that there is a logically valid inference from double dissociation, *properly defined*, to the claim that there are separate modules distinctively implicated in the performance of the two tasks. However, no such interpretation of Shallice's text would be warranted. He himself says (1988, p. 247; emphasis added): [W]ith the ease of explaining a dissociation in terms of isolable subsystems ... such explanations began to be treated *illicitly* as almost having the force of a logical inference.' Shallice shows how to define double dissociation in a way that excludes explanation in terms of different levels of damage to a single system with a single notion of resource and different non-decreasing performance/resource curves for the two tasks. But it certainly does not follow that there is only one possible explanation of double dissociation so defined.

Given a double dissociation of impairments on Task I and Task II, the simplest and boldest explanatory hypothesis would be that, in the normal cognitive system, the two tasks are subserved by two completely separate component systems. In accordance with the subtractivity assumption, if the Task I system were damaged

then Task I would be impaired while Task II was spared; if the Task II system were damaged then the pattern of impairment and sparing would be reversed. But this is not the only possible explanation of double dissociation. It might be that, in the normal cognitive system, the systems subserving Task I and Task II are not completely separate but overlap, with some modules in common. It may still be that some of the modules implicated in Task I are not implicated in Task II while some of the modules implicated in Task II are not implicated in Task I. In such a case, the double dissociation could still be explained in terms of *damage to separate modules that are distinctively implicated in the two tasks*.

Suppose now that one system subserves both tasks and that each component module of the system is implicated in both tasks. For simplicity, assume that there are just two modules, M1 and M2. Since both modules are implicated in both tasks, double dissociation of impairments cannot be explained in terms of damage to separate modules that are *distinctively* implicated in the two tasks. But this functional architecture allows an explanation of double dissociation if Task I is more demanding of module M1 resources than Task II is, while the situation is reversed for module M2. Thus, suppose that, with partial damage to module M1, performance on Task I falls away steeply but performance on Task II is resilient; and suppose that partial damage to module M2 has the opposite result. Given these complementary resource artefacts, double dissociation of impairments could be explained in terms of partial damage to each of the two modules.

Shallice notes that we can move much further away from explanations of double dissociation in terms of damage to separate modules that are distinctively implicated in the two tasks. In some cases, double dissociation can be explained in terms of damage to different regions of a continuous processing space with no relevant modular structure (1988, p. 249):

> It is obvious that two patients with scotomas (visual-field deficits) in different parts of the visual field can be conceived of as forming a double dissociation. One might be able to see perfectly at 9° eccentricity but not at 15°, and the other could show the inverse pattern. Yet it does not make much sense to say that the simple cells analysing the input at 9° eccentricity are part of a different isolable subsystem from those analysing it at 15° eccentricity. They do not lie within different discrete functional units.

The availability, in principle, of these different kinds of explanation of double dissociation has been argued to have the consequence that data about patterns of impairment and sparing in patients may not be sufficient for discovery of the correct theory about the functional architecture of the normal cognitive system (Glymour, 1994). But if normal empirical science works by inference to the best explanation then we need not ask of cognitive neuropsychology that there should be a method for discovering the correct theory of the normal cognitive system from patterns of

impairment and sparing. We only ask that data from patients should be capable of shifting the balance of support between competing theories.[9]

4. Box-and-Arrow Diagrams and Computational Models

More than twenty years ago, Mark Seidenberg raised a concern that, in cognitive neuropsychology, models of the normal cognitive system were typically presented as box-and-arrow diagrams, with very little detail about either representational format or processing algorithm (1988, p. 405):

> There seems to be a basic characteristic of this research that limits its interest, and that is the commitment to explanations framed in terms of the 'functional architecture' of the processing system. One of the main characteristics of the cognitive neuropsychological approach as it has evolved over the past few years . . . is that very little attention is devoted to specifying the kinds of knowledge representations and processing mechanisms involved.

Box-and-arrow models of this kind are not sufficiently explicit to be implemented as computer programmes (1988, p. 417): 'There is no way of knowing if the proposed mechanism could actually yield the observed pattern of results because it isn't specific enough to support detailed predictions.'

A box-and-arrow diagram is not an explanation of any aspect of cognitive performance; it is at most an outline within which an explanation might be developed. Given some pattern of normal or impaired performance and a pair of competing box-and-arrow diagrams, one outline might promise a good explanation of the performance and it might seem less plausible that a good explanation could be developed within the other outline. But, in Seidenberg's words, this 'does not represent much of an advance over intelligent intuition' (1988, p. 407).

Seidenberg focused his concern on box-and-arrow models that figured in the cognitive neuropsychology of language and, particularly, on the dual-route model of reading aloud. Since reading aloud is undoubtedly the domain in which computational modelling of normal and impaired cognition has subsequently

[9] According to Shallice, cognitive neuropsychologists at the end of the 1980s regarded their discipline as providing 'the royal road to discovery of the structure of the mind' (2004, p. 41). But even a 'royal road to discovery' might not amount to a discovery process in the sense that concerns Glymour and, in any case, Andy Ellis and Andy Young say, in their textbook (1988, pp. 5–6): '[I]t would be unwise to regard the search for double dissociations as some sort of royal road to understanding the structure of the mind. Having unearthed a double dissociation, there is a lot of work to be done in determining just what processes mediate aspects of tasks 1 and 2 independently, and what processes, if any, the two tasks share in common.' Jennifer Gurd and John Marshall add (2003, p. 192): 'to the best of our recall, no-one ever believed that double dissociations are ''some sort of royal road to understanding the structure of the mind'''.

become most developed, I shall adopt the same focus. But my aim is not to adjudicate between competing models of reading aloud. It is, rather, to understand the impact of computational modelling on the practice of cognitive neuropsychology, given that its impact has *not* been to bring about the early demise of a research programme that was based on a special inference rule.

4.1 The Dual-Route Model of Reading Aloud

An a priori analysis of the task of reading words aloud suggests that one way of carrying it out would involve, for each orthographic input representation, a direct mapping to a phonological output representation, drawing on *lexical* information about the orthography and phonology of words. On the face of it, this first way of carrying out the task would not extend to generating a pronunciation for a letter string that is not a word. A second way of carrying out the task would involve, for each orthographic input representation, the assembly of a phonological output representation, drawing on *non-lexical* information about letter-sound (more accurately, grapheme-phoneme) correspondences.[10] On the face of it, this second way would not generate the correct pronunciation for an exception (irregular) word.

The *dual-route* model of the processes involved in mature reading aloud starts from the idea that there is both a lexical route and a non-lexical route from print to speech. In the case of a regular word like 'MINT', both routes would deliver the same correct pronunciation. In the case of an exception word like 'PINT', the lexical route would be essential for a correct pronunciation, while in the case of a pronounceable nonword letter string like 'SLINT', only the non-lexical route would deliver a pronunciation (Coltheart, 1985).

The dual-route model of reading aloud involves two processing systems that are, in principle, susceptible of independent damage. If the lexical route were damaged while the non-lexical route continued to operate normally (in accordance with the subtractivity assumption) then the predicted pattern of performance would be preserved reading of regular words and nonwords but regularisation errors on exception words (for example, 'PINT' pronounced to rhyme with 'MINT'). If the non-lexical route were damaged while the lexical route continued to operate normally then the predicted pattern of performance would be preserved reading of both regular and exception words but impaired pronunciation of nonwords.

Each of these patterns of performance is found in patients with acquired disorders of reading: the first is surface dyslexia; the second is phonological dyslexia. The dual-route model promises explanations of these disorders in terms of selective damage to some components of the normal reading system while other components continue to operate as before. If these are the best explanations of the patients'

[10] A grapheme is a letter or combination of letters that corresponds to a single phoneme in the pronunciation of a word. Thus, in the word 'THE', the combination of letters 'TH' is a grapheme.

patterns of impairment and sparing then the balance of support is shifted towards the dual-route model of reading aloud and away from competing models.

4.2 Computational Models of Reading Aloud

Seidenberg calls the idea from which dual-route models begin 'the central dogma of dual-route models' and formulates it as follows (1988, p. 412; emphasis added):

> There *must be* (at least) two ways to pronounce letter strings, because *no single mechanism could yield* the correct pronunciations of both exception words (such as HAVE and DEAF) and regular nonwords (such as BANT and RAFE).[11]

He then calls this dogma into question (p. 413):

> [W]hat if the central dogma is false? That is, if one actually worked out a mechanism capable of generating the pronunciations of words and nonwords, would it necessarily involve separate routes for exception words and nonwords?

Seidenberg's challenge is based on a connectionist model of reading aloud developed with Jay McClelland (Seidenberg and McClelland, 1989; Seidenberg, 1989). After a single large network is trained on 2,897 monosyllabic words, the weights on the connections are responsible for producing pronunciations of regular words, exception words, and pronounceable nonwords. In this network, there are not two separate processing systems corresponding to the lexical route and non-lexical route (1988, p. 414): 'All types of words and nonwords are pronounced by this single mechanism.' The network's performance in generating pronunciations of words in its training corpus is impressive (more than 97% correct) and many experimental effects in reading aloud by healthy adults are simulated. Seidenberg and McClelland's (1989) connectionist model (SM89) is not, however, as good as healthy adult readers at pronouncing nonwords (Besner, Twilley, McCann and Seergobin, 1990).

The poor nonword reading of the SM89 model is attributable, at least in part, to the highly distributed coding schemes used for orthographic inputs and phonological outputs and performance is substantially improved in subsequent connectionist models. David Plaut and colleagues (Plaut, McClelland, Seidenberg and Patterson, 1996) used a network with localist input representations of graphemes and localist output representations of phonemes. After training on 2998 monosyllabic words

[11] The idea from which dual-route models begin can be interpreted in a less dogmatic and less essentialist-sounding way, along the following lines. *If* reading exception words aloud depends on accessing lexical information then some other process must be implicated in reading nonwords; and *if* reading nonwords aloud depends on grapheme-phoneme correspondence rules then some other process must be implicated in reading exception words (Coltheart, 1985, pp. 8–9).

(including the 2,897 words used in the earlier simulation) the network generated correct pronunciations for all the words in the training set and performed similarly to healthy adult readers on nonwords.

It might seem that this is a straightforward case of competing models of a cognitive process. But Seidenberg rejects the suggestion that the connectionist models and the dual-route model of reading aloud are symmetrically related (1988, p. 414):

> We have one [implemented connectionist] model that specifies a plausible computational mechanism that accounts for quantitative aspects of performance. We have other models that say that performance could derive from other mechanisms but don't actually show how. It is possible that the mechanisms given in [a box-and-arrow diagram of the dual-route model] would work if fully specified, but we simply do not know.

Here, Seidenberg compares box-and-arrow diagrams unfavourably with the SM89 connectionist model. But the virtues of explicitness and implementation need not be associated exclusively with connectionist cognitive science. Coltheart and his colleagues restored symmetry to the debate by developing an implemented computational version of the dual-route model of reading aloud, the dual-route cascaded (DRC) model (Coltheart, Curtis, Atkins and Haller, 1993; Coltheart and Rastle, 1994; Coltheart, Rastle, Perry, Langdon and Ziegler, 2001; Rastle and Coltheart, 2006; Coltheart, Saunders and Tree, 2010; see Figure 1).

5. Computational Cognitive Neuropsychology of Reading Aloud

In computational cognitive neuropsychology, computational models of normal adult performance are 'lesioned' and the performance of the models following damage is compared with findings from patients with acquired disorders. The leading question for the computational cognitive neuropsychology of reading aloud has been whether computational models can be lesioned to produce the patterns of impairment and sparing that are characteristic of surface dyslexia and phonological dyslexia.

5.1 Simulating Surface Dyslexia and Phonological Dyslexia in the DRC Model

In the case of the DRC model, the answer to this question is straightforward. By setting parameters of the model to different values, it is possible to simulate actual patients with surface dyslexia and actual patients with phonological dyslexia (Coltheart *et al.*, 2001, pp. 241–4). It is also possible—indeed, 'almost too simple' (Coltheart, 2006a, p. 102) and 'trivial' (Coltheart *et al.*, 2001, p. 242)—to have the model produce extreme versions of these disorders that are never seen in actual patients: extreme surface dyslexia in which every exception word is regularised while

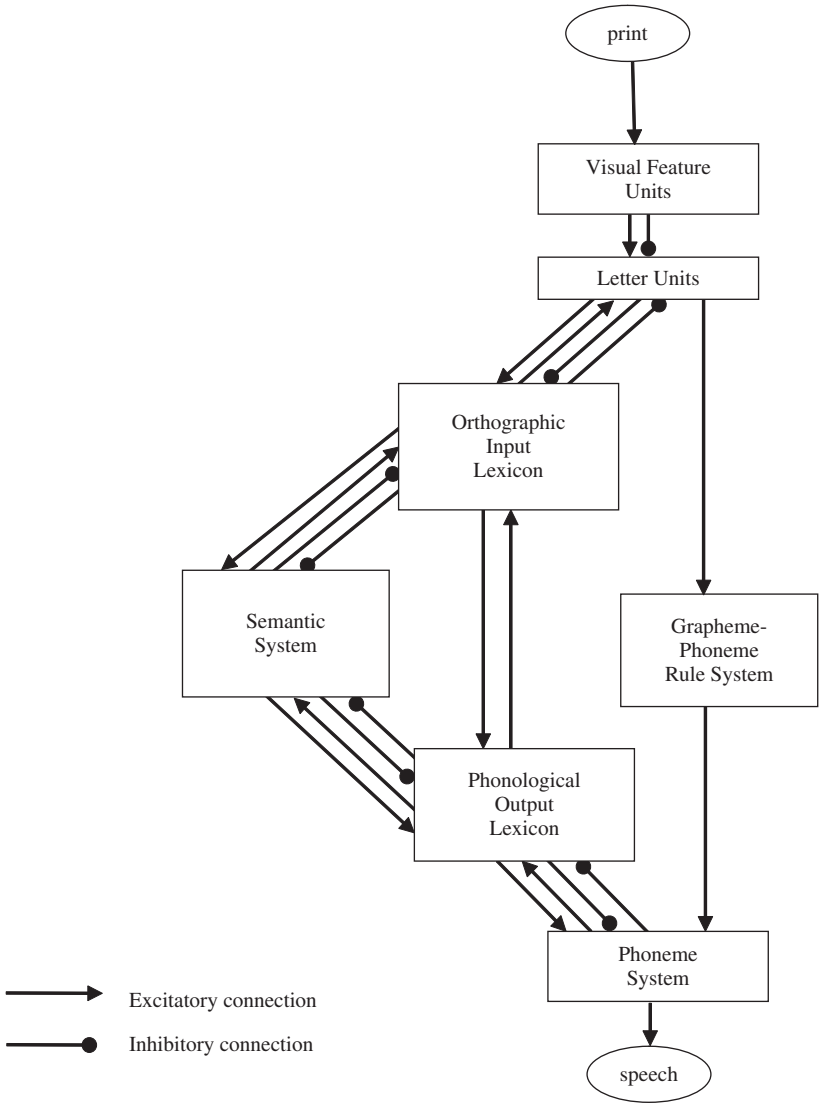

Figure 1 *The dual-route cascaded (DRC) model of reading aloud. The route from the orthographic input lexicon to the phonological output lexicon via the semantic system has not been implemented.*

reading of regular words and nonwords remains perfect; and extreme phonological dyslexia with zero correct nonword reading while reading of regular and exception words remains perfect.

For each of these extreme patterns of impairment and sparing, the dual-route model of the normal reading system allows an explanatory hypothesis that is

consistent with the subtractivity assumption. But the dual-route model does not predict that these extreme patterns will actually occur and it is no objection to the model if they do not occur (Coltheart, 2006a, p. 103):

> Whether one route of the dual-route reading system can be completely abolished by brain damage without any effect at all on the other route is not a question to do with the model, but a question to do with how the various components of the model are represented in the brain.

The actual occurrence of extreme versions of surface dyslexia and phonological dyslexia depends on matters about which the dual-route model, as a contribution to cognitive psychology, is silent.

5.2 The Connectionist Triangle Model

The connectionist model of reading aloud that Seidenberg and McClelland implemented was only part of the theoretical model that they envisaged (1989, p. 536): 'The larger framework assumes that reading words involves the computation of three types of codes: orthographic, phonological, and semantic.' The theoretical model that provides the framework for implemented connectionist models of reading aloud is a triangle with orthography, phonology, and semantics at its vertices. It includes, not only the direct route from orthography to phonology that was implemented by Seidenberg and McClelland (1989) and by Plaut and colleagues (1996), but also a second route connecting orthography and phonology via semantics (Harm and Seidenberg, 2001, 2004; see Figure 2). The question about lesioning models to simulate surface dyslexia and phonological dyslexia can then be answered in terms of implementations of this triangle model.

5.2.1 Simulating Surface Dyslexia. The connectionist triangle model, like the DRC model, is a two-route model. So, it might seem that the triangle model can explain surface dyslexia in terms of damage to the orthography-semantics-phonology route, just as the DRC model explains it in terms of damage to the lexical route. However, this is not quite right. In Plaut and colleagues' (1996) implementation of the direct orthography-phonology route of the triangle model, the trained network produces nonword reading that is similar to that of healthy adults and generates correct pronunciations for regular words *and exception words*. So it is not clear how damage to a second route via semantics would lead to the regularisation errors on exception words that are characteristic of surface dyslexia.

In order to address this issue, Plaut and colleagues propose that, in the triangle model, there would be a *premorbid division of labour* between the two routes. The triangle model would learn to read regular words, exception words, and nonwords but 'if the semantic pathway contributes significantly to the pronunciation of words, then the phonological pathway need not master all of the words by itself'

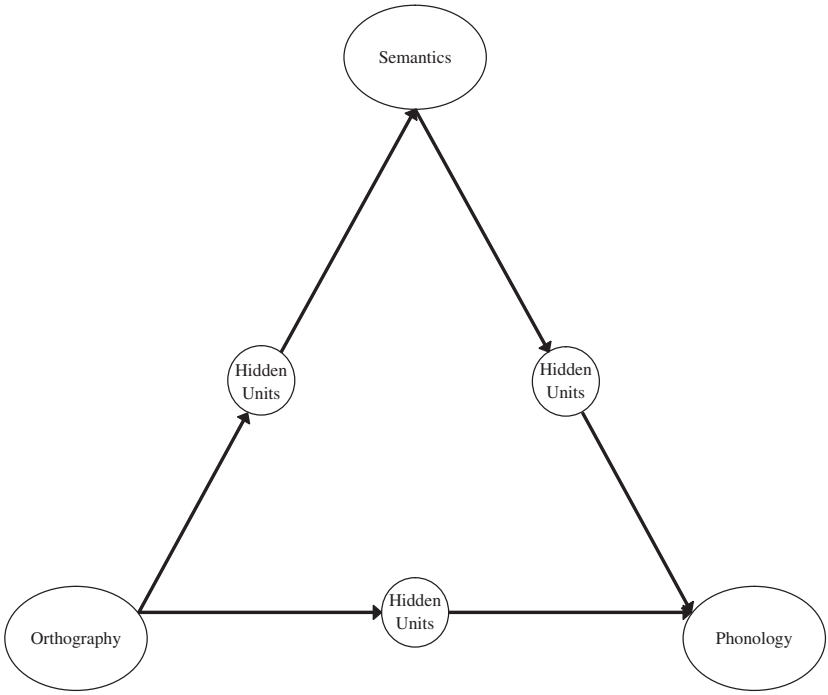

Figure 2 *The connectionist triangle model of reading aloud. Implemented models vary in (a) the components of the triangle that are included, (b) the representations that are used for orthography, phonology and semantics, (c) whether recurrent connections are used to create basins of attraction for the patterns of activation over the units at the vertices of the triangle, and (d) whether there are direct connections as well as connections via hidden units on the sides of the triangle.*

(1996, p. 91).[12] Following learning, 'brain damage that impaired or eliminated the semantic pathway would lay bare the latent inadequacies of the phonological pathway' (p. 92).

The proposal that surface dyslexia results from damage to the semantic route is challenged by the existence of patients with semantic impairment but intact reading aloud of exception words.[13] Plaut and colleagues explain this pattern of impairment and sparing by appeal to *individual differences* in the premorbid division of labour

[12] There is an important general point here. Two components or pathways in a cognitive system may be independent in their *operation* but interdependent in their *learning*. This interdependence may help to explain associations between developmental impairments. It does not cast doubt on the proposal that developmental impairments are to be defined and understood in terms of models of the normal adult cognitive system (Castles, Bates and Coltheart, 2006).

[13] The reverse dissociation—patients with surface dyslexia but no semantic impairment—need not be a problem for the proposal since damage to the orthography-semantics-phonology route might leave semantic representations themselves intact (Plaut, 1997, p. 776).

between the semantic and non-semantic routes. In a simulation, Plaut (1997) manipulated two parameters, semantic strength and weight decay, to illustrate how the division of labour could vary. High values for these parameters produce less pressure or less capacity for the non-semantic route to learn. This results in a less competent non-semantic route and so the prospect of greater impairment of exception word reading if the semantic route is subsequently damaged. Low values for the parameters produce greater pressure and capacity to learn and thus result in a more competent non-semantic route—in the limit, a non-semantic route that produces correct pronunciations for all exception words.

Thus, the overall proposal by Plaut and colleagues (Plaut *et al*., 1996; Plaut, 1997; Woollams, Lambon Ralph, Plaut and Patterson, 2007, 2010) is that surface dyslexia is produced by damage to a route from orthography to phonology via semantics and that the severity of surface dyslexia depends, not only on the degree of semantic impairment, but also on the extent of the premorbid division of labour between the semantic and non-semantic routes to reading aloud.

Plaut and colleagues (1996) also consider a second explanation of surface dyslexia as arising from 'partial impairment of the phonological pathway in addition to severe impairment of the semantic pathway' (p. 92). In the most dramatic case, total removal of the semantic route would eliminate the redundancy between the two routes but leave a non-semantic route able to read nonwords, regular words and exception words. Partial damage to the non-semantic route would then result in impaired exception word reading with spared regular word and nonword reading, as the result of a resource artefact.

5.2.2 Simulating Phonological Dyslexia. In the connectionist triangle model, one possible explanation of phonological dyslexia is that it results from impaired phonological representations. This proposal arises first in the context of developmental phonological dyslexia (Plaut *et al*., 1996, p. 104; Harm and Seidenberg, 1999) but Harm and Seidenberg (2001) report a simulation of acquired phonological dyslexia and present several considerations suggesting that 'phonological dyslexia is caused by an impairment in the representation of phonological information rather than [as the dual-route model claims] grapheme-phoneme conversion' (2001, p. 72).[14]

The strong proposal that *all* cases of phonological dyslexia result from damage to phonological representations is challenged by the existence of patients with phonological dyslexia but no phonological impairment (Coltheart, 2006a).[15]

[14] According to this proposal, phonological dyslexia is the result of a resource artefact (Harm and Seidenberg, 2001, p. 72): 'the advantage of word reading over nonword reading derives from nonwords having a less stable phonological representation than words; therefore phonological impairment yields more errors on nonword reading than word reading.'

[15] It might seem that the reverse dissociation—patients with phonological impairment but intact reading aloud of nonwords—would present a problem for the strong proposal (and for the more inclusive proposal to follow). But such patients are not found and no theory predicts that they should be found.

But this pattern of impairment and sparing can be accounted for by a more inclusive proposal that allows an alternative explanation of phonological dyslexia; namely, that it can also result from disruption of the direct route from orthography to phonology in the presence of an intact second route from orthography to phonology via semantics. This would be similar to the DRC model's explanation of phonological dyslexia in terms of damage to the non-lexical route.

6. Connectionist Explanations of Double Dissociation

When Seidenberg (1988) challenged the dogma that normal reading aloud requires separate routes for exception words and nonwords, he used a single-route connectionist model, corresponding to the direct orthography-phonology route of the two-route triangle model. In contrast, the explanations of surface dyslexia and phonological dyslexia that are offered by Seidenberg, Plaut and their colleagues appeal to both routes of the triangle model. Thus, in computational cognitive neuropsychology of reading aloud, there is a measure of convergence between the research programmes using the DRC model and the connectionist triangle model (Woollams *et al*., 2007; Coltheart, in press).

The DRC model's explanation of surface dyslexia in terms of damage to the lexical route is similar to the triangle model's explanation in terms of damage to the indirect orthography-semantics-phonology route. The DRC model's explanation of phonological dyslexia in terms of damage to the non-lexical route is similar to the triangle model's explanation in terms of damage to the direct orthography-phonology route. But these similarities must not be allowed to obscure significant differences between the two models' explanations of the double dissociation of impairments on the tasks of exception word reading and nonword reading.

6.1 Double Dissociation in the Triangle Model

In the DRC model, the lexical route is implicated in exception word reading but not in nonword reading, while the non-lexical route is implicated in nonword reading but not in exception word reading. Consequently, the double dissociation is explained in terms of *damage to separate modules that are distinctively implicated in the two tasks*. The connectionist triangle model's explanation of the double dissociation is bound to be different. While the indirect semantic route is implicated in exception word reading but not in nonword reading, the direct non-semantic route is implicated in exception word reading as well as nonword reading.

Because of this redundancy for exception word reading between the two routes of the triangle model, the semantic route is not essential for reading exception words and even total removal of the semantic route is not guaranteed to impair exception word reading. As we have seen, the triangle model's explanation of

surface dyslexia appeals to an idea that does not figure in the DRC model's account; namely, individual differences in a premorbid division of labour between the two routes. An alternative explanation of surface dyslexia appeals to damage to *both* routes—severe damage to the semantic route and also damage to the non-semantic route.

For nonword reading, there is no such redundancy between the two routes of the triangle model. The non-semantic route is essential for reading nonwords and damage to this route provides one possible explanation of phonological dyslexia. Another possible explanation appeals to damage to a phonological component that is shared by *both* routes.

In summary, whichever way the triangle model's possible explanations of surface dyslexia and phonological dyslexia are combined, the double dissociation is *not* explained in terms of damage to separate modules that are distinctively implicated in the two tasks of exception word reading and nonword reading. This illustrates the general point that connectionist cognitive science has revealed a broader range of potential explanations of double dissociation. Further illustrations abound.

6.2 Pathways and Parameters in Connectionist Models

Plaut (1995; Plaut and Shallice, 1993) used an implemented connectionist model of the route to reading via semantics in order to simulate the double dissociation of impairments on reading aloud concrete versus abstract words.[16] In this model, there are no separate modules that are specific to, or essential for, the task of reading concrete words aloud or the task of reading abstract words aloud. All the connections are involved in processing both categories of words. In particular (Plaut, 1995, p. 317): 'the direct [feedforward] pathway generates an initial approximation of the semantics of the stimulus word which are gradually refined by the clean-up [recurrent] pathway into the exact semantics of the word'. This computational difference between feedforward and recurrent pathways is typical of *attractor networks*.

Lesioning the model by removing twenty percent of the feedforward connections (from orthography units to hidden units) severely impaired abstract word reading with relative sparing of concrete word reading. Removing seventy percent of the recurrent connections (from semantic units to clean-up units) impaired concrete word reading more than abstract word reading. The explanation of the double dissociation is that the computational difference between initial approximation and subsequent refinement interacts with a difference in semantic representation between concrete words and abstract words. Concrete words have more (and more intercorrelated) semantic features than abstract words and so, during training, 'the network develops stronger semantic attractors for concrete words than for abstract

[16] Impaired reading aloud of abstract words with relatively spared reading of concrete words is typical of patients with deep dyslexia. Elizabeth Warrington (1981) describes a patient with the reverse dissociation.

words' (Plaut, 1995, p. 317). Consequently, the feedforward and recurrent pathways exhibit *complementary resource artefacts*. Abstract word reading is more sensitive than concrete word reading to damage to the feedforward pathway while the situation is reversed for the recurrent pathway.

Christopher Kello and colleagues (Sibley and Kello, 2005; Kello, Sibley and Plaut, 2005) simulated a double dissociation of impairments by varying one parameter, *input gain* for the hidden units, between low and high values.[17] The network was trained (with input gain for the hidden units set to 1) until it produced the correct output for each item in the training set. Input gain for the hidden units was then varied to lower and higher values. Low levels of input gain resulted in more regularity-based or componential processing, with performance on novel items (analogous to nonwords) better than performance on exception items. High levels of input gain resulted in more item-based or conjunctive processing, with performance on exception items better than performance on novel items.[18]

6.3 Random Damage to a Single-Route Connectionist Model

Kim Plunkett and colleagues (Juola and Plunkett, 2000; Plunkett and Bandelow, 2006) propose an even more radical departure from explanations of double dissociation in terms of damage to separate modules that are distinctively implicated in the two tasks. They demonstrate that double dissociation of impairments may arise from different *random* patterns of damage to a *single-route* network.

In order to investigate the effects of random damage, Plunkett and Bandelow (2006) trained a network to effect transitions from input phonological representations of 2280 monosyllabic nouns (of which twenty-six were irregular) and 946 monosyllabic verbs (122 irregular) to output representations of their plural or past tense forms. The network achieved a very high level of performance on all four inflectional categories: regular nouns 99.91%; irregular nouns 96.15%; regular verbs 99.64%; irregular verbs 98.36%.

Copies of the trained network were then subjected to 13,720 separate lesions, each involving removal of a randomly selected 580 (one percent) of the connections. For each lesioned network, performance on each of the four inflectional categories was evaluated, relative to the performance of the undamaged trained network. Plunkett and Bandelow (2006, p. 203, Table 5) show that, on three different definitions

[17] Without the input gain parameter, the net input to a unit, u_j, is $\Sigma_i\ a_i w_{ij}$, the sum, over all the units, u_i, connected to u_j, of the product of the activation of the unit and the weight on the connection from u_i to u_j. Input gain is a multiplier, γ, on this sum, so that the activation of u_j is a non-linear function of $\gamma \Sigma_i\ a_i w_{ij}$.

[18] Reducing or increasing input gain for hidden units is equivalent to making weights on connections to hidden units smaller or larger after training. The results of the simulations are thus consistent with the idea that keeping weights small improves generalisation to novel items while exception items, like exclusive-or, require large weights on connections to non-linear hidden units.

of dissociation, these randomly lesioned networks reveal double dissociation of impairments between regular verbs and irregular verbs.[19]

6.4 Connectionism and a Special Rule of Inference

Connectionist cognitive science has contributed to cognitive neuropsychology by revealing potential explanations of double dissociation that depart in more or less radical ways from explanation in terms of damage to separate modules that are distinctively implicated in the two tasks.

These developments demonstrate that the following rule of *double dissociation inference* is logically invalid:

> From a double dissociation of impairments infer that the normal cognitive system includes separate modules that are distinctively implicated in the two tasks.

If—contrary to fact—the practice of cognitive neuropsychology had been based on methodological assumptions about a special logic including this special rule of inference then the demonstration of invalidity might well have led to the early demise of the research programme. But that is not why the contribution of connectionist cognitive science is important.

7. Normal Empirical Science with Longer Shortlists

According to the account that I am defending, cognitive neuropsychology was and is normal empirical science proceeding, not by special rules of deductive inference, but by inference to the best explanation. Peter Lipton draws attention to 'two senses in which something may be the best of competing potential explanations' (2004, p. 59). A potential explanation might be the most probable given the available evidence—the *likeliest*—or it might be 'the one which would, if correct, be the most explanatory or provide the most understanding' (*ibid.*)—the *loveliest*. Likeliness

[19] Plunkett and Bandelow (2006, p. 203) offer three definitions of dissociation: Stringent—intact performance is 95% of baseline performance or better, impaired performance is 50% of baseline performance or worse; Relaxed—intact performance is 90% of baseline performance or better, impaired performance is 70% of baseline performance or worse; Variance—intact performance is at least two standard deviations above the mean of the 13,720 lesioned networks, impaired performance is at least two standard deviations below the mean. On the stringent definition, 126 of the lesioned networks showed intact performance for regular verbs and impaired performance for irregular verbs, while one of the lesioned networks showed the reverse pattern. On the same definition, there was also a double dissociation of impairments between irregular verbs and irregular nouns. For comparison with the case of reading aloud, we note that there was a double dissociation of impairments between irregular and novel verbs on the relaxed definition of dissociation, and between irregular and novel nouns on the variance definition.

and loveliness are distinct properties of potential explanations but Lipton argues persuasively that loveliness may be a guide to likeliness and may sometimes function as a proxy for likeliness in our actual inferential practices.

Lipton (2004, pp. 148–51) proposes a two-stage structure for inference to the best explanation, *hypothesis generation* then *hypothesis selection*. In the first stage, a limited list of 'live candidates' is drawn up. In the second stage, the best candidate is selected from this list. In both stages, explanatory considerations figure alongside considerations of probability.

7.1 From Box-and-Arrow Diagrams to Connectionist Cognitive Science

Before there were computational models, there were box-and-arrow diagrams. So let us consider first a case of inference to the best explanation in which double dissociation of impairments on two tasks is to be explained and the putatively explanatory hypotheses are cast in the form of box-and-arrow diagrams.

In the stage of hypothesis generation, one of the live candidates will surely be a diagram in which there are separate components (boxes or arrows) that are distinctively implicated in the two tasks. It is true that a box-and-arrow diagram is at most an outline for an explanation but this outline promises an explanation of double dissociation in terms of selective damage to separate components (Bullinaria and Chater, 1995, p. 231). Against this first box-and-arrow diagram, we might set a second diagram in which all the components that are implicated in one task are also implicated in the other. A good explanation of double dissociation might be developed within this outline but the second box-and-arrow diagram by itself is barely a live candidate.

In this situation, we might settle for a shortlist of one and omit the stage of hypothesis selection altogether. Even if we place both candidates on the shortlist, the assessment of the candidates in the selection stage will be perfunctory. Clearly, the first box-and-arrow diagram is the better candidate. It provides the outline of an explanation of double dissociation whereas the second diagram provides no explanation at all.

The contribution of connectionist cognitive science to cognitive neuropsychology is important because it allows longer shortlists, and serious competition between live candidates, in inference to the best explanation. It has long been acknowledged that there are, in principle, many more potential explanations of double dissociation than just those in terms of damage to separate modules that are distinctively implicated in the two tasks (Shallice, 1988, pp. 249–53; and Section 3.2 above). Connectionist models have helped to ensure that alternative explanations are sufficiently developed, plausible, and salient, to be regarded as live candidates. In inference to the best explanation, competing theories are assessed for their probability in the light of available evidence (including, but not restricted to, double dissociation evidence) and also for their explanatory virtues. When, as in the case of the two box-and-arrow diagrams, there are only two candidates on the shortlist and one is barely alive, this assessment is perfunctory. With longer shortlists,

the characteristic features of the assessment of competing potential explanations are more visible.

Competition has different consequences for likeliness and for loveliness. Competing potential explanations cannot both be highly likely but they may both exhibit a high degree of loveliness. Also, Bayes' theorem tells us how to update the probabilities of competing explanations in the light of evidence but there are no theorems to guide our assessments of explanatory virtue or vice.

7.2 Competing for Likeliness

Plunkett and Bandelow (2006) demonstrate that double dissociation of impairments on the two tasks of generating the past tense forms of regular and irregular verbs may result from random damage to a single-route connectionist model. This establishes that 'the occurrence of a double dissociation does not logically require the inference that separate underlying mechanisms are involved' (2006, p. 208). Double dissociation, like other evidence in psychology and throughout normal empirical science, does not mean as much as conclusive proof. But double dissociation evidence, DD, can certainly shift the balance of support between competing theories and, in particular, can shift the balance of probability towards a two-route theory, T_2, and away from a single-route theory, T_1.

According to Bayes' theorem:

$$[Pr(T_2|DD)/Pr(T_1|DD)] = [Pr(DD|T_2)/Pr(DD|T_1)].[Pr(T_2)/Pr(T_1)].$$

The ratio, $Pr(T_2|DD)/Pr(T_1|DD)$, gives a measure of the balance of probability between the two theories after any shift resulting from the evidence. If the ratio is greater than one then the balance of probability *favours* theory T_2 over theory T_1. The *likelihood ratio*, $Pr(DD|T_2)/Pr(DD|T_1)$, gives a measure of the shift itself. If the likelihood ratio is greater than one then the double dissociation evidence *shifts* the balance of probability *towards* theory T_2 and away from theory T_1, by comparison with the balance of the prior probabilities, $Pr(T_2)/Pr(T_1)$.

Plunkett and Bandelow's (2006) study allows an estimate for $Pr(DD|T_1)$ of about 1.3 in a million. A comparable estimate for $Pr(DD|T_2)$ might be provided by an investigation of the consequences of random damage (removal of a fixed proportion of the total number of connections) in a two-route connectionist network in which one route had been trained only on regular verbs and the other route only on irregular verbs. If the estimate for $Pr(DD|T_2)$ were to be greater than the estimate for $Pr(DD|T_1)$ then evidence of double dissociation following random damage to a network would shift the balance of probability towards the two-route theory, T_2, and away from the single-route theory, T_1.

In the competition for likeliness, theories are assessed in the light of *all* the available evidence and not just the double dissociation that is to be explained. Even if double dissociation evidence were to shift the balance of probability towards

theory T_2 and away from theory T_1, the two-route theory might not win the competition for overall likeliness. Double dissociation evidence might shift the balance towards the two-route theory but other available evidence—already taken into account in the prior probabilities—might strongly support the single-route theory. The likelihood ratio might be greater than one but the ratios of prior probabilities and of posterior probabilities might both be less than one. Double dissociation evidence might shift the prior balance without reversing it.[20]

7.3 Assessing Explanatory Virtue or Vice

The assessment of competing potential explanations against the criterion of loveliness is well illustrated in the literature on implemented computational models of reading aloud. The potential explanations that have emerged from computational cognitive neuropsychology are clearly superior to box-and-arrow diagrams in respect of explicitness and they aspire to the explanatory virtue of *scope*, offering simulations of an impressively wide variety of experimental findings.

Explanatory scope is particularly attractive when it is coupled with *unification* of phenomena that are superficially disparate. Anna Woollams and colleagues (Woollams, Lambon Ralph, Plaut and Patterson, 2007, 2010) claim that it is a virtue of the connectionist triangle model of reading aloud that it unifies the double dissociation between surface dyslexia and phonological dyslexia with the association between surface dyslexia and semantic impairment. The key to this unification is the model's 'assumption of a causal link between the integrity of semantic knowledge and accurate reading of low-frequency exception words' (Woollams *et al.*, 2007, p. 330).

The DRC model does not make this assumption. Consequently, 'on the DRC account of reading in semantic dementia, the association between a patient's semantic deficit and his or her reading accuracy is just that—an association, not a causal relationship' (Coltheart, Tree and Saunders, 2010, p. 261). The explanation of the association is not cognitive but neuroanatomical. Woollams and colleagues, in turn, regard the DRC model's account as 'distinctly unparsimonious' (2007, p. 333) and claim for the connectionist model of reading aloud, not only the virtue of *parsimony*, but also the virtue of making *novel predictions* beyond the domain of reading (Woollams *et al.*, 2010, p. 280; see also Seidenberg and Plaut, 2006).

Coltheart and colleagues give reasons to reject the assumption on which these explanatory virtues depend; that is, the assumption that surface dyslexia is caused by semantic impairment (Coltheart *et al.*, 2010, p. 258):

[20] Plunkett and Bandelow (2006) note that information about the relative frequencies of the two one-way dissociations is relevant to the probabilities of potential explanations of double dissociation. In their study, impaired regular verb morphology with intact irregular verb morphology is more than a hundred times rarer than the reverse pattern. A more even balance between the two dissociations might be expected following damage to a two-route network (p. 206): 'Modular accounts are not committed to one type of dissociation being more common than any other, whereas single-route accounts are.'

> [T]his claim would predict that all patients with semantic impairment will be surface dyslexic ... This is not so; ... numerous patients with semantic impairments but normal accuracy in reading aloud even for low-frequency exception words have been reported.

The triangle model explains the existence of patients with semantic impairment but intact reading in terms of individual differences in a premorbid division of labour. Coltheart describes this explanation as *circular* and *unfalsifiable* (Coltheart, 2006a, p. 100; see also Coltheart *et al*., 2001, p. 244; Rapp, Folk and Tainturier, 2001, p. 257).[21]

Finally in this rapid review, there is a suggestion from the side of connectionist cognitive science that the DRC model lacks *explanatory depth* (Seidenberg, Plaut, Petersen, McClelland and McRae, 1994; Seidenberg and Plaut, 2006; Patterson and Plaut, 2009). For example, Seidenberg and colleagues claim (1994, p. 1187): 'The dual-route approach is therefore much more in the spirit of fitting models to data rather than deriving models from more general explanatory principles.'

If I were trying to adjudicate between competing models of reading aloud then each of these judgements of explanatory virtue or vice would require careful and extended examination. But they are presented here as illustrative examples. In the assessment of loveliness, as in the competition for likeliness, the typical features of normal empirical science are more visible when shortlists are longer and there is serious competition between live candidates. As the result of developments in computational modelling, what was always true about the methodology of cognitive neuropsychology is made plain.

8. Double Dissociation and Lesion Location

The notion of double dissociation used in cognitive neuropsychology omits the requirement of different lesion locations that is present in Teuber's (1955) definition. In their discussion of the occurrence of double dissociation following random damage to a single-route connectionist network, Plunkett and Bandelow remark (2006, p. 208): 'When accompanied by evidence of non-overlapping lesions, the inference of separable mechanisms from behavioural double dissociations carries added weight.' If this is right then a neuroscientific finding—that two patients have lesions in different locations—may enhance the value of behavioural double dissociation in shifting the balance of support between competing cognitive theories.

[21] The problem here is that no independent evidence about the premorbid state of a patient's non-semantic route to reading may be available. See Lipton (2004, p. 24) for discussion of the benign circularity of 'self-evidencing' explanations.

In this final section, I present a schematic example showing that Plunkett and Bandelow are correct.[22]

8.1 Cognitive Neuropsychology and Neuroscience

The question of the evidential value of findings about lesion location for theories about the structure of the normal cognitive system is one aspect of a large issue. What is the relationship between, on the one hand, cognitive neuropsychology—and cognitive psychology more generally—and, on the other hand, neuroscience?

Shallice (1988, chapter 9) described a position that he called 'ultra-cognitive neuropsychology'. It is the position of cognitive neuropsychologists for whom (p. 203; emphasis added): 'the emphasis on the individual case, a rejection of the group study, and *a lack of concern with the neurological basis of behaviour* are becoming almost elements of a creed'. Shallice himself adopted a cautious stance towards neuroscience (p. 214): 'To hope for an advance in theories of the functional organisation of cognition by paying special attention to issues of localisation is not, at present, a promising strategy.' But he held out hope for the future (p. 215): 'With, say, advances in neurological measurement techniques, the situation might very well change.' And he firmly rejected the ultra-cognitive position.

In contrast, Coltheart proclaims, 'I'm still an ultra-cognitive-neuropsychologist after all these years' (2004, p. 21), and proposes that findings from neuroimaging have no evidential value for confirming or refuting cognitive psychological theories (p. 22):

> [N]o facts about the activity of the brain could be used to confirm or refute some information-processing model of cognition. This is why the ultra-cognitive-neuropsychologist's answer to the question 'Should there be any ''neuro'' in cognitive neuropsychology?' is 'Certainly not; what would be the point?'.

In a subsequent paper, Coltheart claims that 'no functional neuroimaging research to date has yielded data that can be used to distinguish between competing psychological theories' (2006b, p. 323; see also 2010a, b; and see McGeer, 2007; Roskies, 2009; and Hanson and Bunzl, 2010, for further discussion).

Suppose that one cognitive theory, T_1, says that a single cognitive system is implicated in both of two tasks while a competing theory, T_2, says that there are

[22] In their discussion of behavioural double dissociation, Plunkett and Bandelow say, 'the identification of such double dissociations, though *compatible* with a modular account . . . , is not *evidence* [conclusive proof] for such an account, since precisely the same effects can be observed in a single-route model' (2006, p. 206). By the same test, Teuber's double dissociation (with different lesion locations) is not conclusive proof for a modular account either. But Teuber's double dissociation may do more than behavioural double dissociation alone to shift the balance of probability between competing theories.

separate component systems that are distinctively implicated in each of the tasks.[23] Suppose, too, that functional neuroimaging reveals activation at different locations when the two tasks are performed. We might naturally think that the neuroimaging data supports the two-module theory over the single-module theory. But Coltheart argues that this neuroscientific finding does not support theory T_2 'because T_2 does not predict this result' (2006c, p. 423). As a cognitive theory, T_2 is completely silent on the topics of localisation of function and neural activation. If functional neuroimaging had instead revealed activation at all the same locations when the two tasks were performed, that would still have been consistent with the two-module theory, T_2.

8.2 The Silence of Cognitive Psychology

Theories in cognitive psychology speak, in more or less detail, of modularity, representation and algorithm, but they are silent on neuroscientific topics. It is important to acknowledge the consequences of this silence. A cognitive theory does not, by itself, predict or explain neuroscientific results, such as the results of functional neuroimaging studies. In order to bring a cognitive theory into contact with neuroscientific evidence, we need to appeal to auxiliary hypotheses that link mind and brain.

These consequences of the silence of cognitive psychology might seem to suggest that neuroscientific evidence cannot shift the balance of support between competing *purely cognitive* theories. But, for at least two reasons, this suggestion should be firmly resisted. First, cognitive theories are also silent on the topic of behaviour. The subject matter of cognitive psychology is the mind, conceived in terms of functional architecture and information processing. But, in the practice of cognitive psychology, behavioural evidence from healthy adults is certainly taken to shift the balance of support between competing purely cognitive theories.

Second, the leading idea of cognitive neuropsychology is that patterns of impairment and sparing in patients following brain injury constrain theories of normal cognitive structures and processes. But a theory about the normal cognitive system does not, by itself, predict or explain the pattern of patients' performance. In particular, a cognitive theory saying that separate modules are distinctively implicated in each of two tasks does not *predict* double dissociation of impairments on those tasks. For all that the cognitive theory says, it might be neurologically impossible for brain injury to result in double dissociation. So, either outcome—double dissociation or no double dissociation—would be consistent with the two-module theory. If two patients do show double dissociation of impairments on the two tasks then the two-module theory does not, by itself, *explain* the double dissociation. The

[23] Coltheart considers an example concerning verbal and spatial working memory (Smith and Jonides, 1997) but his argument generalises.

most that can be said is that the theory allows an explanation, consistent with the subtractivity assumption, in terms of independent damage to the separate modules.

No cognitive theory, by itself, predicts or explains double dissociation but it is essential to the practice of cognitive neuropsychology that double dissociation evidence can shift the balance of support between competing cognitive theories. How can this be so?

In inference to the best explanation, one cognitive theory may be supported because it allows a better explanation of double dissociation than the explanations that are allowed by competing theories. In the assessment of competing theories for probability in the light of available evidence, we need to calculate likelihood ratios and so we must estimate the probability of double dissociation following damage, given one or another cognitive theory. When we consider cognitive systems that are embodied in the brain rather than in a computational model, we need to make assumptions about the probability that separate modules will be independently damaged and, in particular, about the probability that separate modules are separately localised. A typical assumption would be that these probabilities are substantially greater than zero. Such assumptions go beyond the subject matter of cognitive psychology but the practice of cognitive neuropsychology relies on them. As Coltheart explains (2001, p. 10):

> [I]t is possible for there to be functional modularity but no anatomical modularity. If so, almost any form of brain damage must affect very many—even all—modules. In that case, cognitive neuropsychology would get nowhere because the functional modularity of cognition would not manifest itself in the performance of brain-damaged patients.

8.3 Double Dissociation and Lesion Location: Lashley and Teuber Revisited

If behavioural evidence from healthy adults and double dissociation evidence from patients can shift the balance of support between cognitive theories then so can neuroscientific evidence. A schematic example shows the evidential value of a neuroscientific finding about lesion location for theories about the structure of the normal cognitive system.

Consider, once again, two theories, T_1 and T_2, about a cognitive process such as generating the past tense forms of regular verbs (Task I) and irregular verbs (Task II). Theory T_1 says that a single-route system performs both tasks; theory T_2 says that there are separate routes for the two tasks. In each case, we consider the cognitive system as embodied, not in a computational model, but in a cartoon brain in which there are just two locations, L and R.

We assume that the single-route system is localised in the whole cortex and that lesions in different locations produce the same impairments—as Lashley (1930) said in the case of the rat's maze habit. For the two-route system, there are two localisation options: either both routes are localised Lashley-style in the whole

cortex (Lash) or the two routes are separately localised in the two locations (Teub). The two-route theory, T_2, is silent on the topic of localisation but, in line with the practice of cognitive neuropsychology, we assume that the probability of separate localisation of the two routes (Teub) is greater than zero. In fact, we assume that the two localisation options for the two-route theory, Lash and Teub, are equally probable; but the argument to follow does not depend on that simplifying assumption.

Two patients, A and B, are selected from a population recruited at a morphology clinic and we ask two questions:

> **Question 1** Does the finding (DD) that the two patients instantiate a behavioural double dissociation of impairments on the two tasks shift the balance of probability towards theory T_2 and away from theory T_1?
>
> **Question 2** Is the balance shifted further towards theory T_2 if we have the additional neuroscientific finding that the two patients have lesions in different locations (DL) rather than in the same location (SL)?

In order to answer these questions about *shifts* in the balance of probability, we need to estimate two likelihood ratios:

> **Question 1** $\Pr(DD|T_2)/\Pr(DD|T_1)$
>
> **Question 2** $\Pr(DD \,\&\, DL|T_2)/\Pr(DD \,\&\, DL|T_1)$

For Question 1, we suppose that we have no information about lesion location and we treat the two options, different locations and same location, as equally probable. The probability of double dissociation given theory T_1 is not zero, but it is very low, and it is independent of lesion location. For illustrative purposes, say that $\Pr(DD|T_1) = 0.001$ (see Table 1). To estimate the probability of double dissociation given theory T_2, we need to consider the two localisation options, Lash and Teub, separately. With the Lash option, we assume as a worst case that the probability of double dissociation is still only 0.001 (and independent of lesion location). With the Teub option of separate localisation of the two routes in L and R, the probability of double dissociation following lesions in *different locations* is fairly high; say, 0.1. In contrast, the probability of double dissociation following lesions in the *same location* is exceedingly low; zero for simplicity. Thus, $\Pr(E|T_2) = 0.0255$, and the likelihood ratio is 25.5. Since the likelihood ratio is greater than one, the finding of behavioural double dissociation does shift the balance of probability towards the two-route theory, T_2.

For Question 2, we simply eliminate the possibility that the two patients have lesions in the same location (SL). A suitable proxy for the likelihood ratio is

	Theory T_1	Theory T_2	
		Lash	Teub
DL	0.001	0.001	0.1
SL	0.001	0.001	0
$Pr(DD\|T_i)$	0.001	0.0255	
$Pr(DD\|T_i$ & DL)	0.001	0.0505	

Table 1 *The upper part of the table shows the conditional probability of double dissociation given assumptions about theory (T_1 or T_2), localisation (Lash or Teub), and lesion location (DL or SL). The lower part of the table shows the conditional probability of double dissociation given each of the theories, and given each of the theories and the additional finding of different lesion locations for the two patients.*

provided by the ratio $Pr(DD|T_2$ & DL)/$Pr(DD|T_1$ & DL).[24] Thus, the likelihood ratio is 50.5, confirming that the balance is shifted further towards the two-route theory if we have the additional neuroscientific finding that the two patients have lesions in different locations.

Although the real brain is immeasurably more complex than the cartoon brain in this schematic example, the trend of the results seems to be robust provided that the probability that separate modules are separately localised (Teub) is not close to zero. It is important to note, however, that we have assumed that localisation is consistent across the two patients. This assumption is crucial for the estimate of a fairly high probability of double dissociation given separate localisation of the two routes (Teub) and *different* lesion locations for the two patients (DL).

To see this, consider instead a situation in which the separate localisations of the two routes are swapped between the two patients so that the route localised in location L in patient A is localised in location R in patient B and *vice versa* (TeubSwap in contrast to the original TeubStay). In this situation, it is the probability of double dissociation following lesions in the *same location* that is fairly high, while the probability of double dissociation following lesions in *different locations* is exceedingly low (see Table 2). Thus, the support that is provided for the two-route theory by the additional finding of different lesion locations for the two patients depends on a further—though very plausible—neuroscientific assumption that TeubStay is more probable than TeubSwap.

[24] For Question 2, the likelihood ratio, $Pr(DD$ & $DL|T_2)/Pr(DD$ & $DL|T_1)$, is equal to the product of the ratio in the text, $Pr(DD|T_2$ & DL)/$Pr(DD|T_1$ & DL), and the ratio $Pr(DL|T_2)/Pr(DL|T_1)$. We assume that whether patients A and B have different lesion locations is probabilistically independent of which theory is correct, so that this latter ratio equals one.

		Theory T_2		
			Teub	
	Theory T_1	Lash	Stay	Swap
DL	0.001	0.001	0.1	0
SL	0.001	0.001	0	0.1
$Pr(DD\|T_i)$	0.001		0.0255	
$Pr(DD\|T_i$ & DL)	0.001		0.0255	

Table 2 *The conditional probabilities show that the additional support that is provided for the two-route theory, T_2, by the finding of different lesion locations (DL) depends on the assumption that it is more probable that separate localisation of separate modules is consistent across the two patients (TeubStay) than that it is reversed (TeubSwap).*

9. Conclusion

The received wisdom about the role of double dissociation in cognitive neuropsychology is inaccurate and unhelpful and it is time for it to be overturned. There is no special logic of cognitive neuropsychology and no special deductive rule of double dissociation inference. One-way dissociation of impairments can sometimes be explained as a resource artefact; that is, explained in terms of a difference of difficulty between two tasks that are performed by a single cognitive system. This kind of explanation is not available for double dissociation of impairments but it does not follow that double dissociation excludes all possible explanations except one. Cognitive neuropsychology was and is normal empirical science, proceeding by the method of inference to the best, not the only possible, explanation. The role of double dissociation, like other evidence, is to shift the balance of support between competing theories.

Double dissociation is often explained in terms of damage to separate modules that are distinctively implicated in the two tasks. The contribution of connectionist cognitive science to cognitive neuropsychology has been to reveal a range of potential explanations of double dissociation that depart from that pattern in more or less radical ways. The importance of the development of computational cognitive neuropsychology is not that it demonstrates the logical invalidity of a special rule of inference but that it allows longer shortlists in inference to the best explanation. With serious competition between live candidates, the characteristic features of the assessment of theories—assessment for probability in the light of all the available evidence and assessment for explanatory virtue or vice—are more visible. In the mature phase of cognitive neuropsychology, its abductive methodology is increasingly evident.

The notion of double dissociation that is employed in cognitive neuropsychology is different from Teuber's notion in that it imposes no requirement on the sites of the patients' lesions. This fits well with the conception of cognitive neuropsychology as a division of cognitive psychology. But a neuroscientific finding of different lesion

locations in two patients can enhance the value of behavioural double dissociation in shifting the balance of support between purely cognitive theories. It may provide additional support for a two-route theory over a single-route theory. It is not to be assumed, however, that evidence that supports a two-route theory over one competing theory would support the same theory over all other live candidates. Still less should it be assumed that the theory that is supported by one piece of evidence over one competitor is the likeliest in the light of all available evidence or the loveliest in explanatory virtue.

Faculty of Philosophy and Department of Experimental Psychology
University of Oxford

References

Besner, D., Twilley, L., McCann, R.S. and Seergobin, K. 1990: On the association between connectionism and data: are a few words necessary? *Psychological Review*, 97, 432–46.

Bullinaria, J.A. and Chater, N. 1995: Connectionist modelling: implications for cognitive neuropsychology. *Language and Cognitive Processes*, 10, 227–64.

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.

Castles, A., Bates, T.C. and Coltheart, M. 2006: John Marshall and the developmental dyslexias. *Aphasiology*, 20, 871–92.

Coltheart, M. 1984: Editorial. *Cognitive Neuropsychology*, 1, 1–8.

Coltheart, M. 1985: Cognitive neuropsychology and the study of reading. In M.I. Posner and O.S.M. Marin (eds), *Attention and Performance XI*. Hillsdale, NJ: Lawrence Erlbaum Associates, 3–37.

Coltheart, M. 1999: Modularity and cognition. *Trends in Cognitive Sciences*, 3, 115–20.

Coltheart, M. 2001: Assumptions and methods in cognitive neuropsychology. In B. Rapp (ed.), *The Handbook of Cognitive Neuropsychology: What Deficits Reveal About the Human Mind*. Hove, E. Sussex: Psychology Press, 3–21.

Coltheart, M. 2004: Brain imaging, connectionism, and cognitive neuropsychology. *Cognitive Neuropsychology*, 21, 21–5.

Coltheart, M. 2006a: Acquired dyslexias and the computational modelling of reading. In M. Coltheart and A. Caramazza (eds), *Cognitive Neuropsychology Twenty Years On*. Hove, E. Sussex: Psychology Press, 96–109.

Coltheart, M. 2006b: What has functional neuroimaging told us about the mind (so far)? *Cortex*, 42, 323–31.

Coltheart, M. 2006c: Perhaps functional neuroimaging has not told us anything about the mind (so far). *Cortex*, 42, 422–7.

Coltheart, M. 2010a: What is functional neuroimaging for? In S.J. Hanson and M. Bunzl (eds), *Foundational Issues in Human Brain Mapping*. Cambridge, MA: MIT Press, 263–72.

Coltheart, M. 2010b: Levels of explanation in cognitive science. In W. Christensen, E. Schier and J. Sutton (eds), *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Sydney: Macquarie Centre for Cognitive Science, 57–60.

Coltheart, M. in press: Dual-route theories of reading aloud. To appear in J.S. Adelman (ed.), *Visual Word Recognition*. Hove, E. Sussex: Psychology Press.

Coltheart, M., Curtis, B., Atkins, P. and Haller, M. 1993: Models of reading aloud: dual-route and parallel-distributed-processing approaches. *Psychological Review*, 100, 589–608.

Coltheart, M. and Rastle, K. 1994: Serial processing in reading aloud: evidence for dual-route models of reading. *Journal of Experimental Psychology: Human Perception and Performance*, 20, 1197–211.

Coltheart, M., Rastle, K., Perry, C., Langdon, R. and Ziegler, J. 2001: DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108, 204–56.

Coltheart, M., Saunders, S.J. and Tree, J.J. 2010: Computational modelling of the effects of semantic dementia on visual word recognition. *Cognitive Neuropsychology*, 27, 101–14.

Coltheart, M., Tree, J.J. and Saunders, S.J. 2010: Computational modeling of reading in semantic dementia: comment on Woollams, Lambon Ralph, Plaut, and Patterson (2007). *Psychological Review*, 117, 256–72.

Crawford, J.R. and Garthwaite, P.H. 2005: Testing for suspected impairments and dissociations in single-case studies in neuropsychology: evaluation of alternatives using Monte Carlo simulations and revised tests for dissociations. *Neuropsychology*, 19, 318–31.

Crawford, J.R., Garthwaite, P.H. and Gray, C.D. 2003: Wanted: fully operational definitions of dissociations in single-case studies. *Cortex*, 39, 357–70.

Crawford, J.R., Howell, D.C. and Garthwaite, P.H. 1998: Payne and Jones revisited: estimating the abnormality of test score differences using a modified paired samples *t* test. *Journal of Clinical and Experimental Neuropsychology*, 20, 898–905.

Dunn, J.C. and Kirsner, K. 2003: What can we infer from double dissociations? *Cortex*, 39, 1–7.

Ellis, A.W. and Young, A.W. 1988: *Human Cognitive Neuropsychology* (Enlarged Edition, 1996). Hove, E. Sussex: Lawrence Erlbaum Associates.

Fodor, J.A. 1983: *The Modularity of Mind*. Cambridge, MA: MIT Press.

Glymour, C. 1994: On the methods of cognitive neuropsychology. *British Journal for the Philosophy of Science*, 45, 815–35.

Gurd, J.M. and Marshall, J.C. 2003: Dissociations: double or quits? *Cortex*, 39, 192–5.

Hanson, S.J. and Bunzl, M. (eds) 2010: *Foundational Issues in Human Brain Mapping*. Cambridge, MA: MIT Press.

Harley, T.A. 2004: Does cognitive neuropsychology have a future? *Cognitive Neuropsychology*, 21, 3–16.

Harm, M.W. and Seidenberg, M.S. 1999: Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological Review*, 106, 491–528.

Harm, M.W. and Seidenberg, M.S. 2001: Are there orthographic impairments in phonological dyslexia? *Cognitive Neuropsychology*, 18, 71–92.

Harm, M.W. and Seidenberg, M.S. 2004: Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review*, 111, 662–720.

Juola, P. and Plunkett, K. 2000: Why double dissociations don't mean much. In G. Cohen, R.A. Johnston and K. Plunkett (eds), *Exploring Cognition: Damaged Brains and Neural Networks: Readings in Cognitive Neuropsychology and Connectionist Modelling*. Hove, E. Sussex: Psychology Press, 319–27.

Kello, C.T., Sibley, D.E. and Plaut, D.C. 2005: Dissociations in performance on novel versus irregular items: single-route demonstrations with input gain in localist and distributed models. *Cognitive Science*, 29, 627–54.

Lashley, K.S. 1930: Basic neural mechanisms in behavior. *Psychological Review*, 37, 1–24.

Lipton, P. 2004: *Inference to the Best Explanation* (2^{nd} edn.; 1^{st} edn. 1991). London: Routledge.

Marshall, J.C. and Newcombe, F. 1966: Syntactic and semantic errors in paralexia. *Neuropsychologia*, 4, 169–76.

Marshall, J.C. and Newcombe, F. 1973: Patterns of paralexia: a psycholinguistic approach. *Journal of Psycholinguistic Research*, 2, 175–99.

McGeer, V. 2007: Why neuroscience matters to cognitive neuropsychology. *Synthese*, 159, 347–71.

Milner, B., Corkin, S. and Teuber, H-L. 1968: Further analysis of the hippocampal amnesic syndrome: 14-year follow-up of H.M. *Neuropsychologia*, 6, 215–34.

Patterson, K. and Plaut, D.C. 2009: 'Shallow draughts intoxicate the brain': lessons from cognitive science for cognitive neuropsychology. *Topics in Cognitive Science*, 1, 39–58.

Plaut, D.C. 1995: Double dissociation without modularity: evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, 17, 291–321.

Plaut, D.C. 1997: Structure and function in the lexical system: insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes*, 12, 765–805.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S. and Patterson, K. 1996: Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review*, 103, 56–115.

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

Plunkett, K. and Bandelow, S. 2006: Stochastic approaches to understanding dissociations in inflectional morphology. *Brain and Language*, 98, 194–209.

Rapp, B. (ed.) 2001: *The Handbook of Cognitive Neuropsychology: What Deficits Reveal About the Human Mind*. Hove, E. Sussex: Psychology Press.

Rapp, B., Folk, J.R. and Tainturier, M.-J. 2001: Word reading. In B. Rapp (ed.), *The Handbook of Cognitive Neuropsychology: What Deficits Reveal About the Human Mind*. Hove, E. Sussex: Psychology Press, 233–62.

Rastle, K. and Coltheart, M. 2006: Is there serial processing in the reading system; and are there local representations? In S. Andrews (ed.), *From Inkmarks to Ideas: Current Issues in Lexical Processing*. Hove, E. Sussex: Psychology Press, 3–24.

Roskies, A.L. 2009: Brain-mind and structure-function relationships: a methodological response to Coltheart. *Philosophy of Science*, 76, 927–39.

Scoville, W.B. and Milner, B. 1957: Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery and Psychiatry*, 20, 11–21.

Seidenberg, M.S. 1988: Cognitive neuropsychology and language: the state of the art. *Cognitive Neuropsychology*, 5, 403–26.

Seidenberg, M.S. 1989: Visual word recognition and pronunciation: a computational model and its applications. In W. Marslen-Wilson (ed.), *Lexical Representation and Process*. Cambridge, MA: MIT Press, 25–74.

Seidenberg, M.S. and McClelland, J.L. 1989: A distributed, developmental model of word recognition and naming. *Psychological Review*, 96, 523–68.

Seidenberg, M.S. and Plaut, D.C. 2006: Progress in understanding word reading: data fitting versus theory building. In S. Andrews (ed.), *From Inkmarks to Ideas: Current Issues in Lexical Processing*. Hove, E. Sussex: Psychology Press, 25–49.

Seidenberg, M.S., Plaut, D.C., Petersen, A.S., McClelland, J.L. and McRae, K. 1994: Nonword pronunciation and models of word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 20, 1177–96.

Shallice, T. 1988: *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press.

Shallice, T. 2004: On Harley on Rapp. *Cognitive Neuropsychology*, 21, 41–3.

Shallice, T. and Warrington, E.K. 1970: Independent functioning of the verbal memory stores: a neuropsychological study. *Quarterly Journal of Experimental Psychology*, 22, 261–73.

Sibley, D.E. and Kello, C.T. 2005: A computational exploration of double dissociations: modes of processing instead of components of processing. *Cognitive Systems Research*, 6, 61–9.

Smith, E.E. and Jonides, J. 1997: Working memory: a view from neuroimaging. *Cognitive Psychology*, 33, 5–42.

Teuber, H-L. 1955: Physiological psychology. *Annual Review of Psychology*, 6, 267–96.

Vallar, G. 2004: The 2003 status of cognitive neuropsychology. *Cognitive Neuropsychology*, 21, 45–9.

Van Orden, G.C., Pennington, B.F. and Stone, G.O. 2001: What do double dissociations prove? *Cognitive Science*, 25, 111–72.

Warrington, E.K. 1981: Concrete word dyslexia. *British Journal of Psychology*, 72, 175–96.

Warrington, E.K. and Shallice, T. 1969: The selective impairment of auditory verbal short-term memory. *Brain*, 92, 885–96.

Waugh, N.C. and Norman, D.A. 1965: Primary memory. *Psychological Review*, 72, 89–104.

Wechsler, D. 1945: *The Wechsler Memory Scale*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. 1955: *The Wechsler Adult Intelligence Scale*. San Antonio, TX: The Psychological Corporation.

Weiskrantz, L. 1968: Some traps and pontifications. In L. Weiskrantz (ed.), *Analysis of Behavioral Change*. New York: Harper & Row, 415–29.

Weiskrantz, L. 1989: Remembering dissociations. In H.L. Roediger III and F.I.M. Craik (eds), *Varieties of Memory and Consciousness: Essays in Honour of Endel Tulving*. Hillsdale, NJ: Lawrence Erlbaum Associates, 101–20.

Woollams, A.M., Lambon Ralph, M.A., Plaut, D.C. and Patterson, K. 2007: SD-squared: on the association between semantic dementia and surface dyslexia. *Psychological Review*, 114, 316–39.

Woollams, A.M., Lambon Ralph, M.A., Plaut, D.C. and Patterson, K. 2010: SD-squared revisited: reply to Coltheart, Tree, and Saunders (2010). *Psychological Review*, 117, 273–83.

Young, A.W. 2001: Obituary for Freda Newcombe. *The Independent*, 13 April.