Australian Journal of Psychology

*Australian Journal of Psychology* 2012; **64**: 19–28
doi:10.1111/j.1742-9536.2011.00047.x

# Strong inference in functional neuroimaging

**Greig de Zubicaray**

*School of Psychology, University of Queensland, Brisbane, Queensland, Australia*

**Abstract**

A recurring question for cognitive science is whether functional neuroimaging data can provide evidence for or against psychological theories. As posed, the question reflects an adherence to a popular scientific method known as 'strong inference'. The method entails constructing multiple hypotheses (*H*s) and designing experiments so that alternative possible outcomes will refute at least one (i.e., 'falsify' it). In this article, after first delineating some well-documented limitations of strong inference, I provide examples of functional neuroimaging data being used to test *H*s from rival modular information-processing models of spoken word production. 'Strong inference' for neuroimaging involves first establishing a systematic mapping of 'processes to processors' for a common modular architecture. Alternate *H*s are then constructed from psychological theories that attribute the outcome of manipulating an experimental factor to two or more distinct processing stages within this architecture. *H*s are then refutable by a finding of activity differentiated spatially *and* chronometrically by experimental condition. When employed in this manner, the data offered by functional neuroimaging may be more useful for adjudicating between accounts of processing loci than behavioural measures.

**Key words:** cognitive models, dissociations, fMRI, neuroimaging, speech production

## INTRODUCTION: BROADBENT'S HEIRS

A journal special issue entitled 'Neuroscience "versus" cognitive modelling' affords another opportunity to continue a debate that began in the late 1990s (e.g., Van Orden & Paap, 1997). Previous journal special issues have offered forums for exchanging ideas on this topic (e.g., *Cortex*, Volume 42, Issue 3, 2006), as have conference symposia (e.g., *14th Annual Meeting of the International Organisation for Human Brain Mapping*, 2008, Melbourne; *29th International Congress of Psychology*, 2008, Berlin), and more are apparently in the pipeline (e.g., a forthcoming issue of *Cognitive Neuropsychology*), attesting to its enduring interest among researchers in both fields. The early debates generally did not disappoint, even though they often involved an opening gambit invoking a distinction between information processing and brain mechanisms on behalf of the cognitive scientists (e.g., Bub, 2000; Coltheart, 2004; Fodor, 1999; Uttal, 2002). This appeal to functionalism represented little more than an attempt to define the terms of argument in favour of one of many philosophical responses to the mind–body problem. The premise having been assumed to be true, the subsequent argument denied the possibility of a materialist cognitive science that might choose to avail itself of neuroscience methods (de Zubicaray, 2006).

In addition to begging the question, a tactic often adopted by cognitive scientists in these debates was the appeal to authority. One such appeal involved citing David Marr's (1982) posthumously published monograph in which he proposed three different levels for machine implementations of information processing: The tri-level hypothesis. This hypothesis was often presented as if it were a psychological rule. David Marr's early demise means we do not know whether he would have adhered to the tri-level hypothesis through what would likely have been a prolific academic career, contemporaneous with the development of cognitive neuroimaging and connectionist modelling. Another appeal involved paraphrasing Donald Broadbent to justify a distinction between software and hardware, though not always with the intent to do his position justice.

The final chapter of Broadbent's (1958) famous monograph *Perception and Communication* entitled 'Retrospect and Prospect' includes this passage:

> The proper relation between the physiologist and the psychologist may be regarded as analogous to that between the automobile mechanic and the test driver. . . . And for many purposes a knowledge of the mechanism is not essential to the driver; no more so than a knowledge of the problems of driving is essential to the mechanic. (p. 306)

Correspondence: Greig de Zubicaray, PhD, School of Psychology, University of Queensland, Brisbane, Qld 4072, Australia. Email: greig.dezubicaray@uq.edu.au

Received 30 March 2011. Accepted for publication 1 November 2011.

This analogy and others of like type (e.g., computers, television, vacuum cleaners) frequently opened discussions about the relevance of neuroimaging to cognitive modelling (see Harley, 2004). The concluding sentence to Broadbent's passage above reads thus:

> Nevertheless, the driver and the mechanic are handling the same car and ought to speak a language which can be applied to the problems of either. (p. 306)

The shared language Broadbent refers to above now finds its expression in a materialist cognitive science. This involves an attempt to formulate information-processing models in terms of brain-like operations; hardware and software unified. Cognitive models antedating this language are less amenable to investigation with neuroscientific methods, if only because they tend to be underspecified, box-and-arrow models. Yet, these models do evolve, a prominent example being Baddeley's and Hitch's model of working memory function, now instantiated in neural networks (e.g., Burgess & Hitch, 2005). Following the 2006 *Cortex* special issue devoted to critiquing neuroimaging, the journal would host another entitled 'working memory in the brain' (Volume 43, Issue 1, 2007). The editors would introduce the special issue as directly addressing whether neuroimaging studies assess theories 'any more effectively than behavioural studies' (Logie & D'Esposito, 2007, p. 1).

## STRONG INFERENCE AND NEUROIMAGING

The central and enduring question concerning the ability of neuroimaging data to provide evidence for or against psychological theories has not proven easy for many cognitive neuroscientists to answer. As posed, the question assumes the superiority of a particular method of scientific hypothesis testing popularised by Platt in an influential article published in 1964 in the journal *Science*. Platt's (1964) article queried why some scientific fields made rapid progress compared to others. His conclusion was that the application of a particular scientific method was responsible; one that he designated 'strong inference'. According to Platt, strong inference entails steps of constructing multiple hypotheses (*H*s), designing and conducting experiments so that alternative possible outcomes will exclude or refute at least one of these *H*s (i.e., 'falsify' it), and repeating the procedure to exclude sub-*H*s or sequential *H*s. Platt (1964) was careful to acknowledge early influences for this 'falsificationist' method, including Bacon (1620/1939) and Chamberlin (1897/1965). Most cognitive psychologists use some kind of falsificationist criteria to evaluate the outcomes of experimental manipulations (see McDonald, 1992). Although his focus on hypothesis testing may have served as an inspiration, Platt was criticised for portraying some scientific fields as more exemplary than others, and for ignoring important issues in the application of scientific method (Davis, 2006). Both of these criticisms could be applied to the current debate concerning the relevance of neuroimaging to cognitive science.

A recent special issue of the journal *Perspectives in Psychological Sciences* (November 2010) hosted a series of articles on neuroimaging that managed to avoid discussing approaches for testing *H*s from competing psychological theories, although discussed methodology in detail. The issue contained articles with the by now familiar appeal to acquire more data and develop large-scale databasing efforts to build reliable neural network models of structure-to-function mappings. This has also surprisingly been marketed as a 'reboot' of the field, i.e., *Cognitive Neuroscience 2.0* (Yarkoni, Poldrack, Van Essen, & Wager, 2010). Somehow, the increased sophistication of the methods for establishing functional associations has not translated to increased testing of psychological theories. Instead, the development of new 'cognitive ontologies' has been proposed in an attempt to circumvent problems that neuroimaging has apparently encountered with those theories.

### Alternate and auxiliary hypotheses

If neuroimaging experiments are to be designed to test *H*s from rival psychological theories, then several well-documented limitations of Platt's (1964) 'strong inference' method need to be acknowledged. To begin with, devising alternate *H*s requires sufficient background information, yet this information may be incomplete particularly in nascent fields such as cognitive neuroscience. It is worth noting that even with background information, it is virtually guaranteed that selected *H*s will *not* be exhaustive (e.g., Davis, 2006; McDonald, 1992; O'Donohue & Buchanan, 2001). Moreover, the plausibility of *H*s depends on a range of factors that may be specific to a particular discipline or field (O'Donohue & Buchanan, 2001). For example, the assumption that there may be some systematic mapping from information processing to brain mechanisms was initially considered implausible by cognitive scientists who approached the issue from a functionalist perspective. There is now broad agreement concerning this assumption, despite the need to rely on converging evidence across studies to demonstrate such mappings (e.g., Coltheart, 2006; Page, 2006).

The corollary assumption of 'selectivity', or a systematic one-to-one mapping of structure and function ('strict localisation') is not entertained by most cognitive scientists or neuroimagers (see Henson, 2005, 2006a, 2006b; Poldrack, 2006; Shallice, 2003). Poldrack (2006) illustrates this problem in terms of a 'reverse inference' in which a neuroimager observes unpredicted activity in a particular brain region for a given task manipulation and refers to another

study's finding for this region with a different task manipulation to adduce a 'plausible' function. As Poldrack notes, this reflects the logical fallacy of affirming the consequent, and may be considered valid only to the extent that the prior probability of activity in this region conforms to the process of interest. In his conclusion, Poldrack (2006) proposed that reverse inferences about function-to-structure mappings could provide background information for deriving plausible *H*s for future experiments. O'Donohue and Buchanan (2001) note that proponents of strong inference tend to ignore this very common scientific scenario; 'situations where data are collected . . . and then hypotheses are formed that simply summarize the data' (p. 9). I will return to the issue of using background information to derive 'reverse inferences' for testing rival *H*s below.

In order to adhere to 'strong inference', neuroimaging experiments will need to be designed in such a way as to enable alternative possible outcomes, each of which excludes one or more *H*s. However, it is important to acknowledge this falsificationist approach represents an ideal that is rarely met in reality, a point Henson (2006a) made in his response to Coltheart (2006). Testing *H*s in psychological experiments entails a number of auxiliary *H*s that first need to be satisfied (e.g., representative sampling, measurement issues, etc). Even if the evidence falsifies one of the experimental *H*s, it could always be argued that this might be due to a weakness in one of the auxiliary *H*s (Davis, 2006; O'Donohue & Buchanan, 2001). An auxiliary *H* in functional magnetic resonance imaging (fMRI) experiments is that the blood oxygenation level-dependent (BOLD) response is a measure of brain activity related to information processing. Page (2006) criticised this auxiliary *H* on both physiological and theoretical grounds. According to his argument, although the BOLD signal can sometimes reflect neuronal output in terms of spiking rates, it is more likely to reflect local field potentials (LFPs) measuring both pre- and post-synaptic activity that might be modulated by processes such as inhibition, attention, or expectation (e.g., Logothetis, 2008; Sirotin & Das, 2009), and it therefore cannot be linked directly with a behaviour-generating process; and even if BOLD signals were a surrogate measure for neuronal spiking output, well-specified models explicitly relating cytoarchitectural mechanisms to information processing are yet to be devised.

Logothetis (2008) illustrated the problems of implicitly assuming that BOLD signal changes reflect neuronal spiking by noting early fMRI studies had concluded erroneously that their results matched those of the neurophysiological literature on motion and direction sensitive neurons in cortical area medial temporal (MT). Later fMRI research instead supported an interpretation of the BOLD activity in terms of the attention-capturing properties of motion stimuli—a processing explanation. Page's (2006) argument involves an implicit assumption that neuronal spiking is strongly related to behavioural responses. This assumption is questionable, as correlations between neuronal spiking and behavioural responses are typically weak, and even then, reflect sources of correlated noise (see Nienborg & Cumming, 2010). Logothetis' (2008) example of area MT is apposite, as the spiking rates of motion and direction sensitive neurons correlate only weakly with behavioural response times (RTs; e.g., Cohen & Newsome, 2009; Masse & Cook, 2008). This finding is perhaps not surprising given RT distributions are modulated by factors such as attention, expectation, and working memory capacity that serve to mask a decision process (Balota & Yap, 2011; Luce, 1991), notwithstanding the range of factors that complicate the measurement of neuronal spiking rates (Cohen & Kohn, 2011). Consequently, while the relationship between the BOLD signal response and neuronal spiking may be characterised as ambiguous or 'noisy', that relationship cannot be used to characterise the relationship between BOLD signals and behavioural responses, as neuronal spiking is only weakly correlated with both (cf. Page, 2006).

Whether BOLD responses are involved in a behaviour-generating process can be investigated by calculating correlations directly with behavioural measures. Strong correlations have been observed with RT for some time (e.g., Menon, Lucknowsky, & Gati, 1998). In fact, the *very* strong correlations often reported between BOLD signal responses and behavioural measures across various tasks led to some debate among neuroimagers, with the consensus now being that they are plausible despite concerns about the methods used to calculate them in some early studies (see Poldrack & Mumford, 2009 and the entire May 2009 issue of *Perspectives in Psychological Science*). These findings support the use of the BOLD signal response as an additional DV in terms of measuring task-related processing, one that has the advantage of providing evidence about a potential modular organisation of the nervous system in relation to cognition. This level of explanation, rather than the cytoarchitectural, is also consistent with recent evidence indicating that haemodynamic responses might reflect predominantly local processing within a particular functional circuit (e.g., Harris, Jones, Zheng, & Berwick, 2010).

Precisely how the parameters of spatially localisable BOLD responses relate to cognitive processes across various tasks is a matter for empirical research (e.g., Grinband, Wager, Lindquist, Ferrera, & Hirsch, 2008; Yarkoni, Barch, Gray, Conturo, & Braver, 2009). This is analogous to the interpretation of RT distributions, where caution has been advised in ascribing specific cognitive processes to various parameters (e.g., Balota & Yap, 2011; Matzke & Wagenmakers, 2009). For example, Balota and Yap (2011) emphasise the need to rely on converging evidence across studies, in conjunction with modelling, to inform interpretations of how variables

influence RT distributions: So too with BOLD responses. A reliance on converging evidence across studies contrasts with the emphasis placed on a single experimental outcome as is often the case with strong inference (O'Donohue & Buchanan, 2001).

### Forward, reverse, and strong inference

Discussions of how neuroimaging might be employed to test *H*s from rival psychological theories typically invoke the logic of dissociation for establishing evidence for modular information processes/processors (e.g., Henson, 2005; Shallice, 2003). Henson (2006b) introduced the term 'forward inference' (in contradistinction to Poldrack's (2006) 'reverse inference') to denote a pattern of a statistically significant 'reversed association' (see Dunn & Kirsner, 1988) across three levels of an experimental factor and two brain regions to provide critical support for the *H* of more than one underlying process/processor. Such an approach may be adopted to refute *H*s from single-process theories in favour of less parsimonious options (Keren & Schul, 2009). Henson (2006b) illustrated this approach with a neuroimaging 'test' of single- versus dual-process models of recognition memory (see also Shallice, 2003). Poldrack and Foerde (2008) have also used neuroimaging data similarly to support a dual-process theory of category learning. These examples exploit the spatial dimensions of neuroimaging data to demonstrate dissociations. However, there is also a temporal dimension to neuroimaging data that can be relevant to information processing models.

Elsewhere, I have argued that neuroimaging data may serve as a basis for constructing hypotheses to obtain evidence for or against different modular information processing models of speech production, and offered some examples from fMRI studies (de Zubicaray, 2006). I therefore disagree with Coltheart's (2010) recent assertion that the reason why we have learned so little from cognitive neuroimaging is its 'almost universal failure to use contemporary modular information-processing models of cognition' (p. 265). Below, I outlined a modification of strong inference for neuroimaging experiments to test alternate *H*s from psychological theories. This perspective involves the use of background or summary information to first delineate reliable structure-to-function mappings for a common modular architecture comprising representational stages through which processing flows in terms of spreading activation. Alternate *H*s are then constructed from psychological theories that attribute the effect of manipulating the same experimental factor to distinct processing stages within this common architecture. These *H*s are then refutable by a finding of activity differentiated spatially *and* chronometrically by experimental condition. Here, brain activity is being employed as a DV that converges with behavioural observations (e.g., RTs), yet has the advantage of providing information to identify the processing locus/processor responsible for the cognitive/neural operations underlying a given effect. In addition, by explicitly assuming a common architecture through which activation spreads across experimental conditions, the fallacy of interpreting a significantly activated region as the sole region involved in task performance is avoided. Two examples are given as illustrations of the method applied to test contemporary theoretical accounts of experimental effects in speech production.

## EXAMPLES

### Speech production models and neuroimaging

In 2004, in the journal *Cognition*, Indefrey and Levelt published an update of their meta-analysis linking neuroimaging data to a modular architecture for speech production that immediately attracted the commentary of cognitive scientists (Indefrey & Levelt, 2000; e.g., Coltheart, 2004; Harley, 2004; Page, 2006). Virtually, all contemporary speech production models assume common processing stages (see Dell & Sullivan, 2004) and Indefrey and Levelt's (2004) meta-analysis identified brain regions associated reliably with those stages across neuroimaging (fMRI, PET, EEG, and MEG) experiments: a systematic mapping of processes-to-processors. Specifically, the meta-analysis identified roles for the left mid-temporal cortex in lexical-semantic processing and the posterior superior and middle temporal cortex (Wernicke's area) in phonological word form retrieval, respectively. In terms of picture naming, the time course of activation in these two regions occurred reliably between 150 and 400 ms following object perception, followed by post-lexical syllabification and articulatory processes in the left inferior prefrontal cortex and pre-motor cortical areas (400–600 ms), respectively. This broad localisation of processing stages (both spatial and temporal) has been supported by patient lesion data and virtual lesioning via transcranial magnetic stimulation (TMS), indicating it is unlikely to be epiphenomenal (Laganaro, Morand, & Schnider, 2009; Schuhmann, Schiller, Goebel, & Sack, 2009, in press). More recent neuroimaging work has provided finer-grained information about sub-processes within these broad areas (Indefrey, 2011).

While not disputing this mapping, it was not clear to some cognitive scientists how the results of the meta-analysis might be used to design neuroimaging experiments to test alternate *H*s from models of speech production; having simply established 'where' a stage of processing was broadly localised in the brain was not deemed especially informative (e.g., Harley, 2004; Page, 2006). As I detailed below, this mapping of *what happens where and when* can be used to test alternate *H*s concerning the different stages of processing at

which various experimental effects in spoken word production are proposed to occur. The method was originally applied in de Zubicaray, Wilson, McMahon, and Muthiah (2001) using information from the initial meta-analysis of Indefrey and Levelt (2000; and see the initial outline of the method in de Zubicaray, 2006).

### Loci of interference effects in picture naming

The speech production literature is replete with experiments demonstrating target picture naming latencies being affected by the co-presentation of picture or word distractor stimuli. Since its introduction by Rosinski, Golinkoff, and Kukish (1975), the picture–word interference (PWI) paradigm has been used extensively for investigating the chronometric characteristics of normal speech production. In PWI, participants are required to name the same target pictures while ignoring accompanying distractor words that are varied according to experimental factors. When a categorical relationship between the distractor, and the target picture is manipulated (e.g., picture DOG, word *jackal* versus the unrelated word *trumpet*), naming responses are slowed, an effect referred to as semantic interference (SI). Accounts of the SI effect have formed the bases of modern theories of language production. In many of these theories, competition is assumed to occur naturally during lexical access (see Goldrick, 2006; Levelt, 1999). These lexical selection by competition (LSC) accounts propose that presentation of a related distractor increases the activation level of a competitor at the lexical–semantic stage of processing, making it more potent, thus increasing the time taken to resolve the process for selecting the appropriate target name.

#### *Testing rival accounts of the distractor frequency effect in picture naming*

Experimental manipulations of distractor frequency have proven problematic for LSC models. As the LSC account proposes the time taken to select the target name depends on the activation levels of lexical competitors, high-frequency (HF) words should be recognised more slowly and produce greater interference in PWI because recognition models typically assume low-frequency (LF) words have lower resting activation levels (e.g., McClelland & Rumelhart, 1981). Instead, LF distractors have been shown to slow picture naming more than HF distractors (Catling, Dent, Johnston, & Balding, 2010; Dhooge & Hartsuiker, 2010; Miozzo & Caramazza, 2003). In order to explain these results within LSC architecture, existing models have had to adopt additional mechanisms (e.g., Roelofs, 2005).

An alternative to the LSC account locates distractor effects in PWI at the level of post-lexical mechanisms leading to the assembly and execution of articulatory programmes (e.g., Mahon, Costa, Peterson, Vargas, & Caramazza, 2007; Miozzo & Caramazza, 2003). Word distractors are assumed to have a privileged relationship with the articulators, and enter an output buffer as phonologically well-formed responses. In addition, the speed at which a response enters the output buffer is assumed to be related to frequency and to influence the speed at which the distractor can be excluded according to task-relevant criteria by the operation of a control or decision mechanism. According to this account, by entering the buffer faster, HF distractors will be excluded earlier by the control mechanism, leading to shorter picture naming latencies relative to LF distractors.

Attempts to confirm or refute the alternate loci for the distractor frequency effect have proven difficult due to the tendency of cognitive scientists to modify their models when faced with conflicting data. For example, Dhooge and Hartsuiker (2010) tested the viability of the post-lexical account by masking distractor words, reasoning that this would preclude the formation of a phonologically well-formed response in the output buffer and eliminate the distractor frequency effect. Masking did eliminate the effect, consistent with the post-lexical account. However, Roelofs, Piai, and Schriefers (2011) subsequently demonstrated their LSC model was able to account for Dhooge and Hartsuiker's (2010) result by incorporating additional assumptions.

The above example serves to illustrate the utility of employing a DV with the capacity to independently identify a processing locus/processor. Roelofs et al.'s (2011) model was designed to preserve the assumption of a lexical-level locus for the distractor frequency effect when the LSC account was challenged by behavioural data apparently consistent with a post-lexical locus. New behavioural data can expect to be treated in a similar manner—if so, the model becomes unfalsifiable *unless* additional information from another level of explanation is brought to bear on the problem. Recall that lexical and post-lexical stages of processing in spoken word production are chronometrically and spatially differentiated in terms of brain activity according to the Indefrey and Levelt (2004) meta-analysis. The former stage occurs reliably in left middle-posterior temporal cortex up to 400 ms post-picture presentation, while the latter stage occurs reliably in left pre-motor cortex 400 to 600 ms post-picture presentation. This leads (via reverse inference) to relatively straightforward alternate a priori *H*s for lexical and post-lexical accounts of the distractor frequency effect; *H*s that we tested in a 'strong inference' neuroimaging experiment.

It could be argued that the post-lexical account makes no particular claims about brain regions or processors as it was formulated in purely information processing terms, hence any results from neuroimaging could be claimed to be irrelevant—this is essentially an approach that Page (2006) has adopted. Yet, the post-lexical account shares lexical and

articulatory processing stages with other modular speech production models, and the Indefrey and Levelt (2004) meta-analysis linked these stages with spatially and temporally dissociable processors—a mapping that Page (2006) acknowledged. Elsewhere, the authors of the post-lexical account have argued for a systematic mapping of object representations to brain mechanisms, indicating support for the use of neuroimaging to map processes to processors according to modular architectures (Mahon & Caramazza, 2011). However, as Dhooge and Hartsuiker (2010) noted, the post-lexical account does not specify the nature of the control mechanism hypothesised to operate on the articulatory buffer. They proposed that the verbal self-monitoring system might be responsible for this function, documenting a range of experimental findings in support. In their meta-analysis, Indefrey and Levelt (2004) ascribed the monitoring of both internal (i.e., the output of the phonological word form stage of processing) and external speech to bilateral posterior superior temporal gyri (pSTG) thus providing another *H* for testing.

The results of our fMRI experiment supported a post-lexical account: significantly slower naming latencies for LF compared to HF distractors were accompanied by significantly different activity in pre-motor and primary motor cortices between conditions, and in posterior STG (de Zubicaray, Miozzo, Johnson, Schiller, & McMahon, in press, Exp. 1). Note that based upon the common architecture assumption, middle temporal cortex activation is explicitly assumed to have occurred in both HF and LF distractor conditions, however, the null *H* of 'no differential activity' in this region was not refuted, nor was it refuted in other processing stages/processors.

Strong inference can comprise an additional step of conducting experiments to exclude sequential or sub-*H*s. We adopted this approach to examine whether other distractor manipulations conform to the pattern of differential activity observed for lexical frequency. If so, this would broaden (generalise) support for the post-lexical account of PWI effects. The impetus for this additional experiment was the finding that the distractor frequency effect does not interact with SI, implying the two do not share a common processing locus (Miozzo & Caramazza, 2003). We therefore opted to investigate the distractor age-of-acquisition (AoA) effect in PWI (Catling et al., 2010), in which late AoA words slow naming compared to early AoA distractors, as AoA had been shown to interact with SI in picture naming and a lexical level (LSC) explanation for its effects proposed (Belke, Brysbaert, Meyer, & Ghyselinck, 2005).

The alternate *H*s for testing in the second fMRI experiment were the same as the first. In addition, to account for a common articulatory buffer we refined the reverse inference supporting the *H* for the post-lexical account to include the pre-motor and motor cortical regions identified for the distractor frequency effect in the first experiment. The results of the second experiment differed from the first, this time supporting the lexical level (LSC) explanation: The significant AoA effect (late > early) in the naming latencies was paralleled by significant differential activation encompassing left posterior-to-middle temporal cortex (de Zubicaray et al., in press, Exp. 2). As per the first experiment, the common architecture assumes pre-motor cortex activation occurs in both AoA distractor conditions concomitant with articulation, however, the null *H* of 'no differential activity' in this region was not refuted, nor was it refuted for other processes/processors involved in spoken word production. We also determined that the posterior-to-middle temporal cortex activation in Experiment 2 did not overlap significantly with the posterior STG activation observed in Experiment 1 that was attributed to engagement of the verbal self-monitor, indicating the two experiments were unlikely to involve activation of the same processor/processing stage.

The results of the two fMRI experiments indicate that PWI effects may have different loci according to the type of distractor employed, a scenario not envisaged by either account tested. The results of Experiment 1 may also be interpreted as indicating that the modifications made to the LSC account to preserve an assumption of a lexical locus for the distractor frequency effect in PWI are not plausible (cf. Roelofs et al., 2011). Each experiment involved a falsification of a specific *H* as exemplified in strong inference; however, when viewed together the results cannot be interpreted as providing unambiguous support for either account of PWI effects. Proponents of both accounts are therefore not likely to view these results as definitive. In fact, it is hard to identify examples in the history of science when the results of experiments conducted per strong inference have ever been deemed to be definitive (see O'Donohue & Buchanan, 2001). This is why converging evidence is needed across studies, in conjunction with modelling, to inform interpretations of PWI effects. Of note, the results of Experiment 2 converge with the results of an earlier fMRI study of the SI effect in PWI that was interpreted as supporting a lexical processing account (de Zubicaray et al., 2001; see Dell & Sullivan, 2004).

### *The locus of semantic interference in post-cue naming*

Picture distractors have also been used to demonstrate interference effects in naming tasks. An example is the post-cue naming paradigm introduced by Humphreys, Lloyd-Jones, and Fias (1995). In this paradigm, the target picture to be named is cued by a colour name subsequent to an overlapping presentation of the target and a distractor picture; target and distractor having been presented in different colours. As in PWI, an SI effect is observed with categorically related versus unrelated distractors. Humphreys et al. (1995)

proposed the SI effect in this task occurred because of a competitive lexical selection mechanism, essentially the same lexical locus proposed to account for SI effects in PWI tasks.

Subsequent cognitive research suggested alternative accounts of the SI effect in post-cue naming. Dean, Bub, and Masson (2001) showed that the effect also occurred when participants named the object colour rather than the object, and was modulated by the use of distinctive object attributes. They instead proposed the effect occurred because of greater demands on integrating related target and post-cue attributes (form/feature/colour) in visual short-term memory. That is, although the post-cue task involves a naming response, Dean et al. (2001) account ascribes the effect to the cognitive operations of a processing stage (object recognition) characterised by Indefrey and Levelt (2004) as a 'lead-in' to the core word production system (i.e., a *pre-lexical* locus), after which processing in the word production system is assumed to progress through its stages in a comparable manner for both related and unrelated distractor conditions.[1]

The next account of the post-cue naming SI effect would also depart from the LSC mechanism proposed by Humphreys et al. (1995) by emphasising a bottleneck between production ready distractor and target naming responses in an articulatory response buffer—essentially the same post-lexical account as that proposed for PWI effects (see Mahon et al., 2007). According to this account, both target and distractor occupy the articulatory buffer until the presentation of the post-cue. Because the categorically related distractor satisfies some response-relevant criteria compared to unrelated distractors (e.g., a TARGET-*distractor* pairing of WOLF-*jackal* entails a distractor that is an animal name like the target), a decision mechanism takes longer to clear the distractor from the buffer on presentation of the cue.

Our *H*s concerning areas of differential brain activity expected for Humphreys et al. (1995) lexical locus for the SI effect in post-cue naming followed our previous fMRI studies of SI in PWI (e.g., de Zubicaray et al., 2001), namely the left posterior-to-middle temporal cortex regions identified by Indefrey and Levelt (2004). The *H* for the post-lexical account (Mahon et al., 2007) again corresponded to the articulatory regions identified in Indefrey and Levelt's (2004) meta-analysis, including the pre-motor cortex. An a priori *H* for Dean et al. (2001) pre-lexical account was based on Indefrey and Levelt's (2004) characterisation of visual and conceptual 'lead-in processes' as involving the ventral surface of temporal-occipital cortex within 175 ms of picture onset. Within temporal-occipital cortex, Martin's (2007) review of neuroimaging and lesion studies identified two reliable candidate processes/processors for colour and form/feature representations corresponding to Dean et al. (2001) locus for the SI effect. These were the lingual and fusiform gyri (see also Mahon & Caramazza, 2011). Consequently, if the SI effect is due to relatively greater demands on integrating object/cue attributes in the categorically related condition, then differential activity should be observable in these structures.

Again, it could be argued that none of these theories of the SI effect in post-cue naming make explicit predictions about brain regions as they were formulated in purely information processing terms. However, the use of a systematic mapping of processing stages to processors enables *H*s to be developed at this level of explanation consistent with the common architecture underlying all theories. The fMRI data revealed significant differential activity in the fusiform and lingual gyri for categorically related versus unrelated conditions, concomitant with a significant slowing of naming latencies for the former condition (the SI effect). We interpreted this result as being consistent with Dean et al.'s (2001) pre-lexical account as the null *H* of 'no differential activity' was not refuted in the other regions defined a priori for the alternate *H*s (see Hocking, McMahon, & de Zubicaray, 2010).

Here, it can be argued that fMRI data did not assist in the adjudication between rival theories of speech production. Rather than support or refute competitive lexical versus post-lexical accounts, the fMRI data instead indicated the SI effect in the post-cue naming task might not be a phenomenon relevant to arguments concerning the architecture of the speech production system. This contribution is nonetheless informative.

## SUMMARY AND FUTURE DIRECTIONS

This article has outlined a 'strong inference' method for testing alternate *H*s from psychological theories with neuroimaging data, representing a more formal description of the approach introduced in de Zubicaray et al. (2001; see also de Zubicaray, 2006). Limitations of *H* falsification, and of Platt's (1964) method in particular (e.g., McDonald, 1992; O'Donohue & Buchanan, 2001) were noted. Strong inference applied to neuroimaging data starts with *H*s about stages of processing/processors within a common modular architecture (using 'reverse inference'; Poldrack, 2006) and manipulates a factor hypothesised to influence one processor selectively. The critical finding of activity differentiated chronometrically and spatially according to a specific processing stage/processor is then used to refute alternate *H*s from theories that attribute the outcome of the manipulation to the operations of another stage/processor.

Like the method of forward inference (Henson, 2006b), strong inference is clearly theory dependent. It also shares the assumption that different processors do not support the same cognitive process during the different conditions involved in manipulating an experimental factor. This assumption is further constrained by the associations

between processes and processors first established through reverse inference (Poldrack, 2006). However, it differs from forward inference by assuming a common architecture/system responsible for processing across theories. Forward inferences are therefore necessary to test alternate *H*s framed at the systems level or between different classes of theory (single vs multiple processes, e.g., Keren & Schul, 2009).

The examples presented in Sections 3.2.1 and 3.2.2 benefit from over 40 years of research supporting well-specified modular information processing models of spoken word production. Over this period, the field of psycholinguistics has relied on converging evidence from experimentation, lesion patients, and modelling. The use of neuroimaging represents a logical next step. The addition of the Indefrey and Levelt (2004) meta-analysis represents a relatively coarse mapping of processes-to-processors within a commonly assumed modular architecture. However, as the examples I presented earlier demonstrate, even a coarse mapping may be used to construct alternate *H*s concerning the nature of the processing stage engaged by manipulation of an experimental factor. This information may then be used to further inform and constrain the architecture (e.g., Indefrey, 2011).

Given the above, I am of the view that the future of cognitive neuroimaging should not be defined in terms of large-scale databasing efforts in which functional associations are made in terms of new 'cognitive ontologies' generated without reference to existing modular information processing models (cf. Yarkoni et al., 2009). Coltheart (2010) helpfully lists a range of cognitive domains that have been characterised in terms of their specific information-processing modules; this would seem a good place to start. Both sides of the current debate should exploit the apparent consensus concerning the possibility of a systematic mapping of processes-to-processors in these modular architectures. This will enable strong inference to be applied more widely.

## ACKNOWLEDGEMENTS

The author thanks Mike Page, Rik Henson, and an anonymous reviewer for their helpful comments on an earlier draft of this article. This work was supported by the Australian Research Council (ARC) Discovery Project grant DP1092619. Greig de Zubicaray is supported by ARC Future Fellowship FT0991634.

## NOTE

1. There is the possibility that an interference effect at a pre-lexical locus could propagate resulting in differential effects at subsequent stages of the word production network. However, any propagation would likely be in terms of a relatively slower spread of activation through subsequent stages, rather than reflect differential demands on operations at each of these stages. Given the difference in naming latencies between conditions is usually on the order of several 10s of milliseconds, such propagation is unlikely to be detectable in the BOLD response. However, with a neuroimaging technique with greater temporal sensitivity, such as EEG or MEG, this might be possible. I thank Mike Page for bringing this issue to my attention.

## REFERENCES

Bacon, F. (1620/1939). Novum organum. In E. A. Burtt (Ed.), *The English philosophers from Bacon to Mill* (pp. 24–128). New York: Random House.

Balota, D. A., & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry: The power of response time distributional analyses. *Current Directions in Psychological Science*, *20*, 160–166.

Belke, E., Brysbaert, M., Meyer, A. S., & Ghyselinck, M. (2005). Age of acquisition effects in picture naming: Evidence for a lexical-semantic competition hypothesis. *Cognition*, *96*, B45–B54.

Broadbent, D. (1958). *Perception and Communication*. London: Pergamon Press.

Bub, D. N. (2000). Methodological issues confronting PET and fMRI studies of cognitive function. *Cognitive Neuropsychology*, *17*, 467–484.

Burgess, N., & Hitch, G. J. (2005). Computational models of working memory: Putting long-term memory into context. *Trends in Cognitive Sciences*, *9*, 535–541.

Catling, J. C., Dent, K., Johnston, R. A., & Balding, R. (2010). Age of acquisition, word frequency, and picture-word interference. *Quarterly Journal of Experimental Psychology*, *63*, 1304–1317.

Chamberlin, T. C. (1897/1965). The method of multiple working hypotheses. *Science*, *148*, 754–759.

Cohen, M. R., & Kohn, A. (2011). Measuring and interpreting neuronal correlations. *Nature Neuroscience*, *14*, 811–819.

Cohen, M. R., & Newsome, W. T. (2009). Estimates of the contribution of single neurons to perception depend on timescale and noise correlation. *Journal of Neuroscience*, *29*, 6635–6648.

Coltheart, M. (2004). Brain imaging, connectionism, and cognitive neuropsychology. *Cognitive Neuropsychology*, *21*, 21–25.

Coltheart, M. (2006). What has functional neuroimaging told us about the mind (so far)? *Cortex*, *42*, 323–331.

Coltheart, M. (2010). What is functional neuroimaging for? In S. J. Hanson & M. Bunzl (Eds.), *Foundational Issues of Human Brain Mapping* (pp. 263–272). Cambridge, MA: MIT Press.

Davis, R. H. (2006). Strong inference: Rationale or inspiration? *Perspectives in Biology and Medicine*, *49*, 238–249.

de Zubicaray, G. I. (2006). Cognitive neuroimaging: Cognitive science out of the armchair. *Brain and Cognition*, *60*, 272–281.

de Zubicaray, G. I., Miozzo, M., Johnson, K., Schiller, N. O., & McMahon, K. L. (in press). Independent distractor frequency and age-of-acquisition effects in picture-word interference: fMRI evidence for post lexical and lexical accounts according to distractor type. *Journal of Cognitive Neuroscience*. doi:10.1162/jocn_a_00141.

de Zubicaray, G. I., Wilson, S. J., McMahon, K. L., & Muthiah, S. (2001). The semantic interference effect in the picture-word paradigm: An event-related fMRI study employing overt responses. *Human Brain Mapping*, *14*, 218–227.

Dean, M. P., Bub, D. N., & Masson, M. E. (2001). Interference from related items in object identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 733–743.

Dell, G. S., & Sullivan, J. M. (2004). Speech errors and language production: Neuropsychological and connectionist perspectives. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 63–108). San Diego: Elsevier.

Dhooge, E., & Hartsuiker, R. J. (2010). The distractor frequency effect in picture-word interference: Evidence for response exclusion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 878–891.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, *95*, 91–101.

Fodor, J. A. (1999). Let your brain alone. *London Review of Books*, *21*, 68–69. Retrieved from http://www.lrb.co.uk/article.php?get=fodo01_2119

Goldrick, M. (2006). Limited interaction in speech production: Chronometric, speech error, and neuropsychological evidence. *Language and Cognitive Processes*, *21*, 817–855.

Grinband, J., Wager, T., Lindquist, M., Ferrera, V. P., & Hirsch, J. (2008). Detection of time-varying signals in event-related fMRI designs. *Neuroimage*, *43*, 509–520.

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

Harris, S., Jones, M., Zheng, Y., & Berwick, J. (2010). Does neural input or processing play a greater role in the magnitude of neuroimaging signals? *Frontiers in Neuroenergetics*, *2*, 15.

Henson, R. N. (2005). What can functional imaging tell the experimental psychologist? *Quarterly Journal of Experimental Psychology*, *58A*, 193–233.

Henson, R. N. (2006a). What has (neuro)psychology told us about the mind (so far)? A reply to Coltheart (2006). *Cortex*, *42*, 387–392.

Henson, R. N. (2006b). Forward inference in functional neuroimaging: Dissociations vs associations. *Trends in Cognitive Sciences*, *10*, 64–69.

Hocking, J., McMahon, K., & de Zubicaray, G. (2010). Semantic interference in object naming: An fMRI study of the postcue naming paradigm. *Neuroimage*, *50*, 796–801.

Humphreys, G. W., Lloyd-Jones, T. J., & Fias, W. (1995). Semantic interference effects on naming using a postcue procedure: Tapping the links between semantics and phonology with pictures and words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 961–980.

Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology*, *2*, 255. doi: 10.3389/fpsyg.2011.00255.

Indefrey, P., & Levelt, W. J. M. (2000). The neural correlates of language production. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 845–865). Cambridge, MA: MIT Press.

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*, 101–144.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, *4*, 533–550.

Laganaro, M., Morand, S., & Schnider, A. (2009). Time course of evoked-potential changes in different forms of anomia in aphasia. *Journal of Cognitive Neuroscience*, *21*, 1499–1510.

Levelt, W. J. M. (1999). Models of word production. *Trends in Cognitive Sciences*, *3*, 223–232.

Logie, R. H., & D'Esposito, M. (2007). Working memory in the brain. *Cortex*, *43*, 1–4.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*, 869–878.

Luce, R. D. (1991). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Mahon, B. Z., & Caramazza, A. (2011). What drives the organization of object knowledge in the brain? *Trends in Cognitive Sciences*, *15*, 97–103.

Mahon, B. Z., Costa, A., Peterson, R., Vargas, K. A., & Caramazza, A. (2007). Lexical selection is not by competition: A reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 503–535.

Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, *58*, 25–45.

Masse, N. Y., & Cook, E. P. (2008). The effect of middle temporal spike phase on sensory encoding and correlates with behavior during a motion-detection task. *Journal of Neuroscience*, *28*, 1343–1355.

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*, 798–817.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception. *Psychological Review*, *88*, 375–407.

McDonald, J. (1992). Is strong inference really superior to simple inference? *Synthese*, *92*, 261–282.

Menon, R. S., Lucknowsky, D. C., & Gati, J. S. (1998). Mental chronometry using latency-resolved functional MRI. *Proceedings of the National Academy of Science of the United States of America*, *95*, 10902–10907.

Miozzo, M., & Caramazza, A. (2003). When more is less: A counterintuitive effect of distractor frequency in the picture-word interference paradigm. *Journal of Experimental Psychology: General*, *132*, 228–252.

Nienborg, H., & Cumming, B. (2010). Correlations between the activity of sensory neurons and behavior: How much do they tell us about a neuron's causality? *Current Opinion in Neurobiology*, *20*, 376–381.

O'Donohue, W., & Buchanan, J. A. (2001). The weakness of strong inference. *Behavior and Philosophy*, *29*, 1–10.

Page, M. P. A. (2006). What can't functional neuroimaging tell the cognitive psychologist? *Cortex*, *42*, 428–443.

Platt, J. R. (1964). Strong inference. *Science*, *146*, 347–353.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*, 59–63.

Poldrack, R. A., & Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience and Biobehavioral Reviews*, *32*, 197–205.

Poldrack, R. A., & Mumford, J. A. (2009). Independence in ROI analysis: Where is the voodoo? *Social Cognitive and Affective Neuroscience*, *4*, 208–213.

Roelofs, A. (2005). From Popper to Lakatos: A case for cumulative computational modeling. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 313–330). Hillsdale, NJ: LEA.

Roelofs, A., Piai, V., & Schriefers, H. (2011). Selective attention and distractor frequency in naming performance: Comment on Dhooge and Hartsuiker (2010). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1032–1038.

Rosinski, R. R., Golinkoff, R. M., & Kukish, K. S. (1975). Automatic semantic processing in a picture–word interference task. *Child Development*, *26*, 247–253.

Schuhmann, T., Schiller, N. O., Goebel, R., & Sack, A. T. (2009). The temporal characteristics of functional activation in Broca's area during overt picture naming. *Cortex*, *45*, 1111–1116.

Schuhmann, T., Schiller, N. O., Goebel, R., & Sack, A. T. (in press). Speaking of which: Dissecting the neurocognitive network of

language production in picture naming. *Cerebral Cortex*. doi: 10.1093/cercor/bhr155.

Shallice, T. (2003). Functional imaging and neuropsychology findings: How can they be linked? *Neuroimage*, *20*, S146–S154.

Sirotin, Y. B., & Das, A. (2009). Anticipatory haemodynamic signals in sensory cortex not predicted by local neuronal activity. *Nature*, *457*, 475–479.

Uttal, W. R. (2002). Précis of the new phrenology: The limits of localizing cognitive processes in the brain. *Brain and Mind*, *3*, 221–228.

Van Orden, G. C., & Paap, K. R. (1997). Functional neuroimages fail to discover pieces of mind in the parts of the brain. *Philosophy of Science*, *64*, 85–94.

Yarkoni, T., Barch, D. M., Gray, J. R., Conturo, T. E., & Braver, T. S. (2009). BOLD correlates of trial-by-trial reaction time variability in gray and white matter: A multi-study fMRI analysis. *PLoS ONE*, *4*, e4527.

Yarkoni, T., Poldrack, R. A., Van Essen, D. C., & Wager, T. D. (2010). Cognitive neuroscience 2.0: Building a cumulative science of human brain function. *Trends in Cognitive Sciences*, *14*, 489–496.