Anthony I. Jack and Andreas Roepstorff

# *Why Trust the Subject?*

## Preliminary Remarks

It is a great pleasure to introduce this collection of papers on the use of introspective evidence in cognitive science. Our task as guest editors has been tremendously stimulating. We have received an outstanding number of contributions, in terms of quantity and quality, from academics across a wide disciplinary span, both from younger researchers and from the most experienced scholars in the field. We therefore had to redraw the plans for this project a number of times. It quickly became clear to us that the collection would expand beyond the scheduled double issue of the *Journal of Consciousness Studies*. A triple issue was then drafted, but the number of excellent contributions continued to grow. We therefore had to reconsider the publication plans again, and the decision was made to publish an extended collection of papers in discrete instalments. At present substantial progress has been made towards determining the content of a second double issue of the *Journal of Consciousness Studies*, due summer 2004. A third instalment now appears to be a real possibility. We welcome enquiries from authors interested in submitting to later instalments, especially those offering a novel perspective that is not otherwise represented. However, we do not intend continue this collection indefinitely. In putting together the first major interdisciplinary collection on this topic, we view our task as that of providing a starting point. Sufficient outlets exist to support ongoing debate.[1]

The idea for this collection first took shape when we proposed it to the managing editor of *JCS*, Anthony Freeman, at the 'Towards a Science of Consciousness' conference in Skovde, Sweden, August 2001. Since then, he has been involved in every stage of its development and construction. His editorial experience and his patient assistance have been invaluable to us and to the collection.

Correspondence:
Dr A. I. Jack, Washington University Campus Box 8225, 4525 Scott Avenue, St Louis, MO 63110, USA. *Email: ajack@npg.wustl.edu*
Dr A. Roepstorff, PET-Centre, Aarhus University, Nørrebrogade 44, DK-8000 Aarhus C, Denmark. *Email: andreas@pet.au.dk*

[1] In particular, authors may wish to submit to regular issues of the *Journal of Consciousness Studies*; to *Consciousness and Cognition*; and to the recently created journal *Phenomenology and the Cognitive Sciences*.

## Why 'Trusting the Subject?'

In the context of cognitive science, the title of this volume *Trusting the Subject?* carries a double meaning. As touched on by Tony Marcel (this volume), the establishment of scientific knowledge is inherently bound up with notions of trust, which carries a social history of its own. From the days of the gentleman scientists in the learned societies of the seventeenth century, via the Introspectionists at the turn of the twentieth century, and the Behaviourists that dominated psychology throughout the middle part of the twentieth century, to the cognitive neuroscientists who have recently begun to transform psychology, differing understandings of trust both in the experimenter and in the experimental subject have served as semi-stable foundations for the generation of facts and the establishment of knowledge about the human mind. In doing cognitive science and consciousness research, two different levels of trust are therefore at stake. On one level it refers to the interaction between two concrete persons, an experimenter and a volunteering experimental subject, and the extent to which the former uses the reports of the latter as some sort of evidence for scientific inquiry. On another, more general level, however, the notion of 'subject' refers to the actual scientific enterprise of inquiring into the mind. Can one trust the subject of cognitive research, the human mind? Scientists are used to relying on instruments that they themselves have manufactured — technologies whose mode of operation, and limitations, are usually well understood. The unique challenge facing a science of consciousness is that that the best instrument available for measuring experience depends on cognitive processes internal to the subject. So just how much faith can we place in the capacity of the mind to understand itself? In principle, the construction of a maximally robust methodology for introspective evidence would require a detailed understanding of the operation of introspective processes — the processes that mediate the acquisition of introspective knowledge and underlie the production of introspective reports (Jack & Shallice, 2001). Given that knowledge, absolute trust in introspective evidence could be warranted. The practical question is: What attitude should we take given our relative ignorance of introspective processes?

Most scientists do not have, or at least cannot coherently formulate, any principled objection to introspective reports; rather, they simply lack faith that introspective reports are reliable in practice. Some find support for this belief in the informal interviews they conduct at the end of experiments. Subjects, it appears, frequently vary in their take on the experiment. Yet experimenters should not be surprised that such undisciplined reports do not provide consistent results. Like all methodologies, introspective methods require a number of factors to be controlled: When is the subject attending to their experience? To what aspect of experience are they attending? What 'model' are they using to interpret and filter their experiences? In some cases accurate reports will require at least some minimal training, to provide subjects with concepts they can use to effectively communicate their experience.

Despite such widespread pessimism, there are numerous reasons to believe in the accuracy of introspective reports. First and foremost, could any normally socially functioning human seriously doubt that they have succeeded, at least on occasion, in accurately accessing information about their own internal states, for instance concerning their: emotions, state of concentration, thoughts, actions, cognitive strategies, confidence, imagery and focus of attention? Second, numerous experiments, notably in psychophysics, memory and problem solving, illustrate the reliability of reports when they are carefully collected. Third, an informal reliance on introspective evidence is *ubiquitous* in psychology and cognitive science. It generates many of the hypotheses that psychologists seek to test using objective sources of evidence, it underlies their understanding of cognitive tasks or 'task analysis', and it frequently informs the questions and objections they offer as referees. Introspective understanding even forms the basis of many of the categories used to describe branches of psychological research (e.g. 'attention', 'episodic memory', 'awareness'). If psychologists are reliant on introspection as a source of anecdotal evidence, then shouldn't scientific instinct suggest that a more formal, disciplined and systematic treatment of the evidence will prove more productive? At the very least, we should like to clearly understand what would limit this strategy.

A common, and historically motivated, misconception of introspective methodology views it as in competition with 'objective' (behavioural and neural) methods. In contrast to this, we have argued that the interpersonal perspective involved in the communication of experience is already an integral part of standard methodology in cognitive science (Jack & Roepstorff, 2002; Roepstorff & Frith, in press). It is reflected both by the experimenter's attempts to offer the subject a model for how they should carry out the experimental task (the task instructions or 'script'); and again when the subjects attempt to communicate their actual experience of the task, typically elicited in the informal post-hoc interview that is considered good experimental practice. Both the experimenter's model of what the task involves, and the reports elicited from subjects, frequently serve to inform the interpretation of cognitive experiments. Our aim is therefore to expand and improve upon current practice, through the explicit and formal recognition of the larger framework of 'script-report' that encompasses the standard formalisation of 'stimulus-response' in behavioural methods. The advantages of acknowledging this larger framework, and formalising new methods for capitalising on it, cannot be accomplished unless we also maintain attention to the behavioural factors that allow for tight experimental control and inference to underlying mechanism. In our view, 'stand-alone' introspective methods and 'armchair' introspection are not likely to carry us very far. Hence our emphasis on 'triangulation' — the use of introspective, behavioural, and physiological evidence in concert (Jack & Roepstorff, 2002) — and the specific emphasis of this collection on the role of introspective evidence in cognitive science.

## The Validity of Introspective Evidence

It is important to realize that no principled problem stands in the way of the scientific assessment of various types of introspective evidence. The testing of the reliability, consistency and validity of various types of introspective report measures lies well within the orbit of currently available methods.

A measure[2] may be called 'reliable'[3] if it yields the same results when tested in multiple sessions over time ('test–retest reliability') and across individuals (a cousin of 'inter-rater' and 'inter-observer' reliability). Of course, subjects' reports may differ, and so appear to be unreliable, simply because their internal mental processes and states vary. Thus it is critical to establish well controlled experimental conditions for eliciting reports. The considerable advances in behavioural science since the time of the Introspectionists offers experimenters considerable advantages in this regard (see Ericsson, this volume). Not only do these advances make it much more probable that experimenters can establish conditions under which introspective measures can be shown to be reliable, they also provide much greater insight into the behavioural and neural correlates of experiential phenomena.

A measure may be called 'consistent' when it can be shown that the results are not due to specific features of the measurement technique. Tests of consistency provide a means of checking that the observed effect is not due to a methodological artefact. Thus we might test the consistency of introspective evidence by comparing immediate forced-choice button-press reports with retrospective and open-ended verbal reports. In this way we might establish, for instance: that the results of forced-choice button-press reports have not been influenced by variations in the criterion for response or by automatisation of response such that they no longer constitute true introspective reports; and that retrospective reports have not been distorted by forgetting or memory interference effects.

'Validity' is the most important factor to establish, yet it is also the most theoretically complex, and a particularly vexed issue in cognitive science. A measure is validated when it can be shown to accurately reflect the phenomenon it purports to measure. Validity is complex because scientific measures are often simultaneously interpreted as providing evidence for phenomena at a number of different levels. A rough characterisation of three major sources of evidence in cognitive science might read as follows:

- Data from functional Magnetic Resonance Imaging (fMRI) serves most directly as evidence of cerebral blood flow (which has been validated), less directly as evidence for neural activity (which is in the process of being properly validated), and least directly as a means of identifying and localising specific cognitive functions (far from well validated).

[2] The term 'measure' is used in a very general sense here, and is not meant to imply the ability to map the underlying process in any specific manner (e.g. using a continuous scale). Thus a 'measure' might just be an experimental method for identifying the presence or absence of a particular internal state.

[3] Epistemologists often use the term 'reliability' to refer to the accuracy and/or validity of a particular knowledge source. It has a slightly different meaning in scientific contexts.

- Behavioural measures (e.g. the averaging of reaction time measures over multiple trials) serve most directly as evidence for stable patterns of behaviour, less directly as a means of assessing information processing, and least directly as means of establishing the existence and operation of specific cognitive functions.
- Introspective reports serve most directly as evidence about the beliefs that subjects have about their own experience, less directly as evidence concerning the existence of experiential phenomena, and least directly as evidence concerning the operation of specific cognitive functions.

The issue of validity is particularly vexed in cognitive science because there has been a long history of theoretical and philosophical disagreements about the nature of the mental — about what psychological evidence is ultimately serving as evidence for. The early psychologists regarded consciousness as the mark of the mental (Wilkes, 1988). Thus it is often said that scientific psychology began with 'psychophysics' — a project initially conceived by Fechner and Weber as an attempt to find law-like relationships between the physical properties of the stimulus and the experiential properties of the percept. Specifically the Weber-Fechner law was put forward to describe the relationship between physical intensity and 'felt' intensity. In stark contrast, the Behaviourists rejected any reference to internal mental states, and defined the purpose of psychology as that of identifying stable patterns of behaviour. This eventually gave way to a growing sense that behaviourist science (e.g. the description of 'processes' such as habituation, classical conditioning, overshadowing, etc.) served primarily as a means of re-describing the data. Information processing accounts provided a well grounded way of making inferences from this data to internal processes and states. Yet the dominance of the information processing model, and in particular its strong emphasis on behavioural performance, has sometimes made it difficult for other sources of evidence to find a purchase. Information processing accounts are primarily concerned with what subjects are capable of doing with the information in the stimulus, as indicated by the appropriateness of their behaviour for achieving a specific goal (whether that be a sub-personal goal, such as making an accurate visual saccade, or a personal goal, such as achieving good performance in a logical reasoning task). The observation that particular parts of the brain are preferentially involved in different tasks appeared, at least initially to many psychologists, to have little direct relevance to understanding cognition. Similarly, introspective reports do not, at least at first, appear to provide data relevant to information processing accounts. The goal of introspective report is to provide an accurate description of experience. Since the experimenter cannot directly observe subjects' experiences, there is no easy way to assess the accuracy of their performance. Putting the point another way, without knowing what information subjects have internal access to, psychologists can't use introspective reports in the same way they use objective behavioural measures to aid in the construction of information processing models. Thus the publicly inaccessible nature of experience can seem to militate against the validity of introspective evidence.

Introspective reports cannot be treated in the same way as other behavioural measures, yet this does not preclude their use to inform information processing accounts. The expanding view afforded by the increasing influence of neurophysiological evidence provides greater opportunity for introspective evidence to find other points of purchase. A consequence of this is that results previously thought to indicate the unreliable nature of introspective reports may now be seen in a different light. A good example comes from the historical account provided by Anders Ericsson (this volume), as follows:

> [A] large body of research has attempted to relate the level of accurate recall of a presented picture to the reported vividness of the memory (McKelvie, 1995; Richardson, 1988). To everyone's surprise, no clear relation between the amount of accurately recalled information and reported vividness has been found. Participants who reported recalling a presented stimulus as vividly and clearly as if it remained visible did not recall more accurate information than those who reported diffuse memory images. These and other puzzling findings, such as the reported persistence of visual eidetic images (Haber, 1979), confirmed the opinions of many experimental psychologists that introspective judgments about experience were frequently misleading and inconsistent with measures of performance (p. 6).

Ericsson is describing the sort of thinking that has led psychologists to conclude that introspective reports are invalid. Let us carefully consider the logic of this conclusion. What the research shows is that the intuitively appealing idea that memory accuracy and reported vividness should correlate turns out to be wrong. Yet the conclusion that the reports are invalid depends on how you interpret them. If the reports are interpreted as being reports about memory accuracy, then clearly they are invalid. Experimenters should not trust reported vividness as a guide to memory accuracy.

However, we might approach these reports in another way. Instead of attempting a direct translation of these reports into information processing terms (i.e. as describing the efficiency and thus accuracy of the processes underlying recall) we might more literally construe them as 'introspective judgments about experience'. According to this strategy, we should remain agnostic, at least for the time being, about the correspondence between experience and information processing. This gives us a three way relationship. We have memory accuracy, experienced vividness, and reported vividness. Given this framework, we can see that at least one of the two relationships must break down. Either memory accuracy does not correspond to experienced vividness, or experienced vividness does not correspond to reported vividness. Further, we can see that two separate suppositions support these different relationships. The first relationship is supported by a folk-psychological belief, the belief that perceptually vivid memories should be more accurate than diffuse non-vivid memories. The second relationship is supported by the view that the reports in question are accurate and valid. Given this framework we can see that one strong possibility is that the reports are accurate but that the folk-psychological belief is false. Furthermore, we can seek evidence that would provide some support for this view. For instance, it is reasonable to hypothesize that activity in visual cortical areas will correlate with the

experienced vividness of memories (Wheeler *et al*., 2000). Thus, evidence of a correlation between reported vividness and visual activity would support the view that the reports are valid whilst the folk-psychological belief is false. We have demonstrated a closely related result when subjects are asked to provide immediate reports of difficult to perceive (masked) visual stimuli. Summerfield *et al.* (2002) used EEG to show that gamma band activity over occipital (visual) cortex correlates with reported vividness, even for stimuli that were *incorrectly* identified.

Ericsson's example, and other similar cases,[4] illustrate that part of the reluctance of psychologists to ascribe validity to introspective report measures derives from a tendency that might be called 'the rush to operationalize'. For historical reasons, deriving from the positivism of the behaviourists, experimental psychologists are highly reluctant to adopt the strategy of interpreting introspective reports in the most straightforward and direct manner, as telling us about experience. Instead they seek what are called 'operational definitions' — they seek to define internal states and processes in terms of their behavioural effects. This emphasis on operational definitions ensures that the claims that psychologists make are concrete, specific and falsifiable. Yet the problem with adopting this strategy when interpreting introspective reports is obvious: despite the prevalence of folk-psychological beliefs, the true relationship between experience and behaviour is often difficult to ascertain. When scientists are forced to make a choice between appearing somewhat vague on the one hand and relying on an untested and intuitive assumption on the other, scientific progress is often better served by temporarily maintaining a degree of vagueness.

The reluctance of psychologists to interpret introspective reports as telling us about experience is also shared by the philosopher Daniel Dennett (this volume). We regard Dennett's work as important for a number of reasons: First because he has long been at the forefront of a movement to discuss experiential phenomena in cognitive science and encourage debate concerning their interpretation. Second because he has formulated an explicit position concerning the scientific use of introspective evidence, which he calls 'heterophenomenology'. Third, because in our view, his position provides the best representation of the underlying philosophy that guides current practice in cognitive science. Dennett (this volume) argues that scientists should only go so far as to make claims about the beliefs that subjects have about their experiences. Scientists should stop short or 'reserve judgment' about the truth of these claims. We agree that this approach, which Dennett calls the 'bracketing' of experience, must play a key role in the establishment of a methodology for introspective evidence. Only by remaining

[4] Jack & Shallice (2001) and Jack (2001) discusses a similar mistake made by researchers in the field of perception without awareness. Until recently, forced-choice discrimination performance was seen as the gold-standard 'objective' measure of awareness, and the lack of correspondence with introspective reports was interpreted as illustrating their invalidity. After 50 years, researchers recently realized that the lack of correspondence was not due to problems with the introspective reports, but to contamination of the 'objective measure of awareness' by non-conscious information. Currently favored behavioral measures of awareness, such as those based on the Jacoby Process Dissociation Procedure (Jacoby, 1991), correspond much more closely with introspective reports of awareness.

neutral about the accuracy of any particular introspective report, can we critically assess the evidence concerning the reliability, consistency and validity of different types of report. Dennett's cautious approach thus helps to avoid the trap of over-interpreting introspective evidence — the problem of undue trust. Dennett's view is closely related to that espoused by many (but by no means all) psychologists: they are willing to acknowledge the role of introspective evidence in generating hypotheses and in influencing preliminary interpretations of results, yet they view the final arbiter, and the real business of science, as lying in the collection of objective evidence.

We believe that cognitive science can do better than this, and we are sceptical of certain aspects of Dennett's position. Specifically, Dennett appears to insist that we must *always* reserve judgment about the veracity of subjects' beliefs about their experiences, pending verification of their claims using objective evidence. Our view differs in two ways. First, we do not believe that it is possible to use objective evidence to directly test or 'verify' the accuracy of subjects' reports. Second, we do not find motivation for the claim that it is *always* necessary to reserve judgement about the accuracy of introspective reports. Instead we take the view that we should place a degree of trust in introspective reports, proportional to the evidence of their validity.

Again, both these issues come down to how the introspective evidence is interpreted. When introspective evidence is interpreted as evidence about the operation of cognitive processes, such as a claim about information processing accuracy, then it is clear that objective evidence can be used to directly test its validity. Thus Ericsson's example illustrates how reported vividness was found to be an invalid measure of memory accuracy. Alternatively, in their perceptual masking experiment, Summerfield *et al.* (2002) found that reported vividness consistently correlated well with discrimination accuracy across a range of masking times, except when masking time was very short, at which point the correlation between reported vividness and discrimination accuracy broke down.

When introspective evidence is interpreted more directly, as evidence about the subject's experience, objective measures can no longer serve to directly test their validity. In this case, as the discussion of Ericsson's example above illustrates, it may be possible to find convergent evidence that lends support to the view that the reports are valid. Yet objective evidence cannot be used to *directly* verify or falsify the accuracy of introspective reports about experience. This does not mean that reports are not falsifiable. It means falsification can only be achieved indirectly, by means of inference to the best explanation. For example, it may be that subjects who are suddenly placed in highly exasperating situations have an initial tendency to deny that they are angry. This might happen because the onset of anger causes attention to focus exclusively on the perceived source of irritation, so diverting attention from inner states and preventing accurate self-ascription. In this case, the falsity of the subjects' reports might be established by a convergence of evidence: the inconsistency of concurrent reports with later retrospective reports, the presence of behavioural indicators of anger,

and other evidence that supports the proposed hypothesis about the effects of anger on attention.

Dennett's heterophenomenological perspective differs from ours because he does not recognize the same gap between experience and behaviour. His philosophical position identifies mental states with patterns of behaviour — just as many psychologists are apt to do in practice. In our view patterns of behaviour, neural processes and experience exist as distinct facets of the mental. Thus we maintain that objective measures can only provide tangential evidence about experience, by means of an underlying theory of mental processes. Introspective evidence provides the most direct view of the experiential facet of mental processes. Just as establishing the reliability and consistency of behavioural measures assures a reasonable degree of trust in their validity as measures of information processing; so establishing the reliability and consistency of introspective reports should assure a reasonable degree of trust in their validity as measures of experience. We take it to be obvious that introspective evidence, and only introspective evidence, has 'face validity' in the measurement of experience. No doubt introspective reports will sometimes be mistaken, and this may be established by convergent evidence, yet the balancing of equivocal evidence should always be weighted in favour of introspective reports.

We take this point to be important to establish because a degree of trust is actually essential to scientific progress. Although some degree of scepticism is always advisable, undue scepticism prevents the bold hypotheses that push science forward. The scientist who never dares to trust her methods will, of course, never allow herself to discover a thing. Ericsson's psychologists, who were so quick to interpret the reports of vividness as invalid, would never find the motivation to look for other correlates of those reports. So long as cognitive science continues to doubt the face validity of introspective reports, it will never conduct the investigations necessary to provide full validation of those measures; nor can it ever hope to provide scientific accounts of experience. We shall need to take the time to explore and understand experience, before we can hope to generate strong hypotheses about its behavioural and neural correlates.

If this view is correct, then it will have profound implications for methodology in cognitive science. At present the experimental assessment of awareness, in fields such as implicit learning, memory and perception without awareness, is largely achieved by means of objective performance measures. Our view certainly does not preclude the use of objective measures of awareness; however it does turn the current line of thinking about validation, expressed in Dennett's philosophy, on its head. Where experiential phenomena are concerned, it is *objective* measures that must seek validation by establishing their correspondence with *introspective* measures, and not vice versa. Furthermore, psychologists should be willing to accept the value of investigations that focus primarily on data from introspective reports, provided the cognitive tasks employed are also well controlled. At present few mainstream psychological journals would accept such investigations for publication. Instead, at present, they accept experiments that purport to use objective evidence to substantiate claims about experience.

## The Significance of Introspective Evidence

Why should we care about introspective evidence? What can it tell us, and how can it benefit cognitive science? We will discuss three advantages of introspective evidence.

### *(1) Understanding mechanism*

Introspective evidence may assist in the normal business of cognitive science, as an additional source of evidence that can inform and guide mechanistic accounts of mental function. To motivate this view, we need only make a few minimal and plausible assumptions about introspection. First, we assume that introspective processes have access to some limited subset of the functional properties of mental states (Jack & Shallice, 2001). Second, we assume that introspective processes are capable of performing some basic information processing operations on this information, which may be understood by analogy to perception:

(i) We have a capacity to learn to recognize internal states that have occurred on a number of previous occasions, such that we are able to recognise further recurrences of those states (Siegler & Stern, 1998), given that we are attending internally.
(ii) We have a capacity both to encode and to recall information about previous internal states, so allowing us to make comparisons between states.
(iii) We are able to attend selectively to specific features of internal states, so allowing us to compare states along a number of dimensions. For example 'This headache is sharper than the one I had yesterday, it was duller. The headache yesterday was throbbing, the one today is continuous.'

Finally, we assume that we evolved our capacities to recognize and distinguish between our own internal states, perhaps by the extension of existing perceptual and mnemonic processes (Jack, 2001), and that these capacities serve a useful function. In order to support an evolutionary advantage, we suppose that introspective processes are at least reasonably successful at recognizing and discriminating between internal states.

If these assumptions are correct, then there are straightforward ways in which we can use introspective reports to provide clues about functional differences between mental states. A strong example of this comes from work on synaesthesia (Cytowic, 1997). Synaesthetes report that particular sorts of experiences (e.g. hearing a word) are similar along a certain dimension to other, normally quite different sorts of experience (e.g. heard words have the colour properties of visual experience). This has led to hypotheses about the functional similarities of the two states, which have been borne out by neural tests (Paulesu *et al.*, 1995).

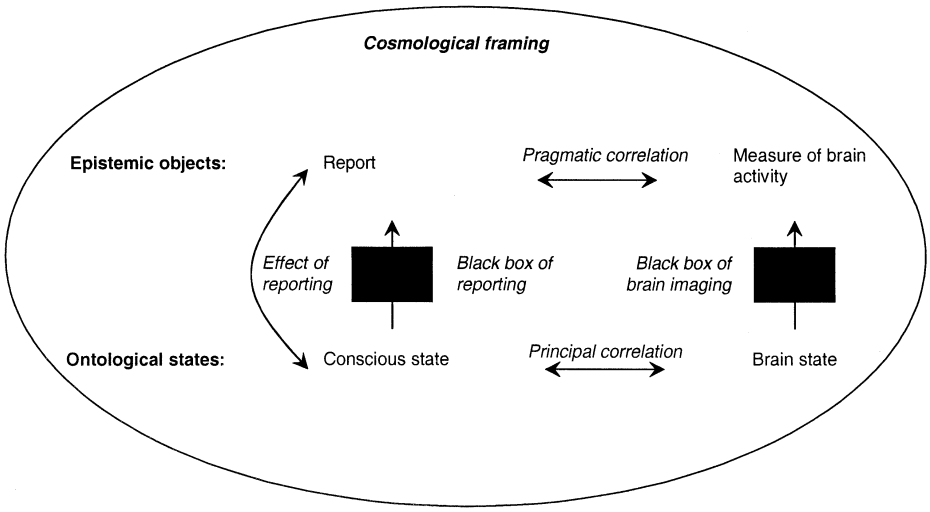
Another more general way in which introspective evidence may prove useful to cognitive science concerns the strategy that psychologists use to identify putative cognitive processes pending further investigation. It is common practice for psychologists to identify processes by reference to a particular behavioural

paradigm, good performance on which is hypothesized to require the process in question. One problem with this strategy of defining processes in terms of behavioural tasks is that psychologists are then apt to generalize on the basis of the task. They assume that similar tasks should evoke similar processes, when this is often not the case. For example, psychologists frequently talk about 'recognition memory' and 'working memory' in a manner that suggests they are referring to an actual cognitive process. Yet there is now a wealth of evidence that both working memory tasks and recognition memory tasks involve a range of different cognitive processes (which have been shown to be engaged to differing extents depending on manipulations of the task, e.g. Chein & Fiez, 2001; Mandler, 1980; Rowe & Passingham, 2001). As a result, the only presently salvageable notion of 'working memory' as a cognitive process is so vague as to have almost no explanatory value. In contrast 'episodic memory', a construct that is almost unique in psychology in that it was explicitly defined by Endel Tulving (1972) in phenomenological terms, appears to be maintaining its explanatory value very well. Given the clear phenomenology associated with different aspects of working memory tasks (e.g. imagery, sub-vocal rehearsal, 'refreshing' of information, 'chunking', 'holding information in mind', etc.) one cannot help but wonder whether research in this area would have fared better if its constructs had been defined at the outset in phenomenological terms. In general, it seems that higher cognitive processes such as those involved in working memory tasks, tests of executive function and problem solving, are particularly amenable to analysis using introspective reports and verbal protocols (Jack & Roepstorff, 2002). Thus Tulving's strategy may prove particularly productive for investigations of these processes.

### *(2) Understanding consciousness*

The pursuit of scientific theories of consciousness has clearly become a hot topic in cognitive science. Yet, surprisingly, few researchers have explicitly acknowledged the central importance of understanding introspection to this enterprise. When we have pushed consciousness researchers on the need to produce some sort of account of the mechanisms that allow us to acquire knowledge about experience, many have replied that such higher-order processes are not their primary interest. They claim to be interested in the nature of experience itself ('first-order awareness' or 'phenomenal consciousness'), not the processes that allow us to make judgments and reports concerning our experiences ('second- order awareness' or 'reflexive consciousness'). These researchers have simply missed the point. From an epistemological point of view, introspection is the *sine qua non* of consciousness. Without introspection, we simply wouldn't know about the existence of experience. And without a good theory of introspection, we have no way of establishing what sorts of claims about experience are justified.

Good scientific theories succeed because they make sense of the data. A theory of consciousness must not only make sense of the neural and behavioural data — it must also make sense of experience. Yet it is important to realize that




making sense of the data on experience will not always mean following the most intuitive and obvious interpretation of that data. A strong theory isn't likely to fit comfortably with all our intuitions about experience. Part of the promise of a strong theory of consciousness is that it should cause us to view our experiences in a new light.

Consider the logic behind attempts to find the neural correlates of consciousness (NCC). Studies of the NCC seek to establish a correspondence between the subject's experience and their brain state. To provide evidence for this, the researcher demonstrates a correspondence between the neural measure used (here we consider brain imaging) and the introspective reports elicited by subjects. We can understand these measures as *epistemic objects*. That is, these measures derive from the application of a set of methods and criteria that ensure some kind of validity. Within science studies, it has become customary to describe this process as 'black boxing'. This means that when the process is running smoothly 'one need focus only on its inputs and outputs, and not on its internal complexity' (Latour, 1999). Yet, as has been demonstrated by a whole range of science studies, the relationship between the resulting epistemic objects and the underlying states is in no way trivial. The 'black boxes' involve complex transformations that do not always succeed in achieving a smooth translation. The commonly applied counter strategy to this problem is to 'open the black box' in order to follow in minute details the actual transformations, reductions and amplifications involved in settling the epistemic objects.

Although it may not be generally known outside the brain imaging community, it is relatively uncontroversial that the colourful pictures of brain activity obtained by PET, fMRI, MEG or EEG are very far from realistic photographs of the brain. They are rather to be seen as complicated graphs, the outcome of a set of mathematical procedures and transformations, that could have been done differently (Roepstorff & Gjedde, 2003).To complicate things even further, there are serious discussions in the brain imaging field about what the relationship is

between the largely metabolic and circulatory measures obtained and the actual behaviour of neurons. These discussions occur at two levels. They are a matter of settling the link between, for instance, fMRI measurements of the BOLD signal or PET measurements of blood flow or oxygen consumption and the underlying neuronal activity. More fundamentally, however, there is no agreement as to what should count as a proper description of brain states — should they, for instance, be identified by synaptic processing or by the firing of individual neurons? This means that our understanding of the link between the measure of brain activity and the putative brain state is constantly evolving.

A science of consciousness will also require us to address the link between reports and conscious states. We must recognize that introspective reports do not represent a transparent reflection of inner experience, but are instead the products of a complex 'black box' set of processes. There will be times when we shall need to 'open the black box' of introspective processes, before we can hope to generate a stable view of the nature of experience. One line of evidence for instabilities in our present view of experience is demonstrated by an interesting set of experiments due to Tony Marcel (1993). They show that even in a very simple psychophysical setting, the measurement of experience depends on the actual method of reporting, be that button pressing, verbal account or eye blink (see also Marcel, this volume). The relationship between the reports on the one hand, and the putative underlying conscious states on the other, is not trivial. As with the relation between brain images and brain states, it is the result of the application of particular epistemic technologies, and only through a careful interplay between black-boxing and opening the box will it become possible to elucidate this relation.

### *(3) Types of psychological explanation*

In our view, by far the most significant role that introspective evidence can play is that of elucidating the links between different types of psychological explanation. At present, psychology is a highly fractured discipline that lacks any sound and over-arching theoretical framework. On the one hand much of psychology, in particular those areas that now come under the umbrella of 'cognitive science', is concerned with providing mechanistic accounts of mental processes. On the other hand, many branches of psychology, in particular social psychology and the therapeutic branches, are concerned with giving accounts that work at the personal, experiential, level of explanation. Thus the challenge of relating introspective evidence to objective evidence directly reflects a key challenge for psychology — that of resolving the tension between different types of psychological explanation, so finally unifying the discipline.

Timothy Wilson (this volume), whose earlier work set the standard for a generation of work on verbal reports (Nisbett & Wilson, 1977), notes that despite the frequent attacks on introspective methods, they are used successfully in many areas of psychology. Nonetheless, it is clear that introspective methods have played a far greater role in areas of psychology that might be considered branches of 'social science' by hard-nosed cognitive scientists. It seems that

many cognitive scientists fear their scientific credibility would be threatened by introspective reports.

How should we understand the difference between these two types of psychological explanation? In our experience, most people have little trouble understanding what it means to provide mechanistic accounts. This is the dominant model of explanation in psychology, borrowed from the hard sciences. Yet many people fail to recognize the critical differences between these mechanistic accounts and accounts that work at the personal level. Personal-level accounts are accounts that we can make sense of on our own terms, rather than from a removed third-person perspective. Personal-level accounts help us to make sense of our experience, they inform our conscious strategies, they alter our interpersonal perceptions, and they help us to understand the implications of mechanistic accounts for our everyday lives.

A simple example of a scientific finding that can be understood at the personal level comes from the work on vividness and memory recall discussed in the previous section. The finding that the perceptual vividness of recalled information does not serve as a good guide to memory accuracy is something that we can make sense of a personal level. More than that, it is a finding that we can actively make use of at a personal level. We can use this information to inform the cognitive strategies that we employ to check the veracity of our own memories, by altering the criteria we use to ascribe confidence. Meta-cognitive strategies of this sort have been shown to influence performance in the lab, specifically on free report tasks (Koriat & Goldsmith, 1996). Increasingly, cognitive psychologists are coming to realize that a great deal of the variation observed both in long-term and working memory performance can be accounted for by the use of more or less sophisticated meta-cognitive strategies — personal first-person knowledge about how best to encode, retrieve and assess the accuracy of memories (Ericsson, 2003).

Our experiences, and the ways we think about them, are far from epiphenomenal. They make a major contribution to performance on all but the most simplified cognitive tasks, they influence our life decisions, and they directly affect our sense of well-being. By better characterizing the information we have internally available to us, and the ways in which we categorise and process that information, we may greatly improve our understanding of meta-cognitive and self-regulatory processes, and so find ways to improve them.

If many cognitive scientists find it hard to recognise personal level psychological accounts, then even more overlook their central significance for cognitive science. Scientific enquiry serves two basic purposes:

(i) As a means of understanding — the pursuit of knowledge for its own sake.
(ii) As a means of intervention — the generation of technologies and methods that allow us to influence the world.

First, let us consider science as a means of understanding. The question we need to ask ourselves is this: What is involved in understanding the mind? If we were to possess a complete mechanistic account of brain function, would we

have a full and complete understanding of the mind? We find this claim implausible. Surely in order to claim that we understand the mind, we must be able to understand our own minds. For many philosophers, this just is the 'problem of consciousness'. According to this perspective, consciousness does not represent a specific and tractable issue for scientific investigation. Rather it represents the diagnosis of a serious failure in the whole discipline. The basic argument of these philosophers is clear: scientific accounts leave something out — they leave out experience. No wonder that this charge should sting. How can we expect to smuggle experience in through the back door, when we are so reluctant to collect systematic data on it? Cognitive scientists spend a great deal of time discussing the interpretation of objective evidence, yet we have never read nor heard a cognitive scientist suggest that further work is needed to understand what it is like to carry out an experimental task.

Understanding personal-level explanations does not merely constitute the greatest intellectual challenge facing cognitive science. There are also eminently practical reasons for wanting to develop better personal-level accounts and seeking to understand their points of contact with mechanistic accounts. It is the personal level that we primarily care about. With chronically ill patients, it is the level of pain they experience that concerns us. Only by extension do we concern ourselves with their galvanic skin response, their cortisol levels, or their neural activity.

Second, let us consider science as a means of intervention. How might the science of the mind intervene to improve our lives? One of the most interesting features of accounts that work at the personal level is that they can serve to directly alter mental function. For instance, cognitive therapy is a method that works simply by encouraging subjects to observe their own experiences and to think about them in different ways. It is a highly effective method. It has long been a treatment of choice for anxiety-related disorders, and it has recently been shown to be highly effective in preventing relapses in depression (Teasdale *et al*., 2002).

More generally it is clear that the population at large has a powerful hunger for interventions that work at the personal level. Walk into any major bookshop, and you will find that popular psychology makes up one of the largest sections. Personal-level explanations and training protocols represent both the most humane and the most publicly acceptable means of intervention that any science of the mental can hope to offer. So shouldn't a major goal for cognitive science be the use of modern scientific methods to better inform our understanding of the processes that mediate the influence of personal-level explanations?

It is unclear how much effort cognitive science is likely to put into understanding interventions at the personal level. However, it is abundantly clear that the mechanisms of science funding are in place to ensure a vast increase in the number of interventions available at the genetic, neural and pharmacological levels. Such progress will bring with it a clear and troubling concern: how will we be able to understand the effects of these interventions? Specifically, how can we hope to attain true informed consent from people undergoing such interventions, unless we are able to explain their effects at the personal level? It seems unlikely that explaining the neural and behavioural consequences of interventions in

brain function will prove adequate. If we are to pursue a truly ethical course, we shall surely need to do our best to explain to patients how these interventions may alter their concept of self — how they will alter the ways in which they experience their everyday life. If we continue to refuse to trust the subject, the subject will have no reason to trust us. Cognitive scientists should not fear that introspective evidence will impugn the scientific credibility of their work. They should fear the Frankenstein science they will create without it.

## References


Chein, J.M. & Fiez, J.A. (2001), 'Dissociation of verbal working memory system components using a delayed serial recall task', *Cereb Cortex*, **11** (11), pp. 1003–14.

Cytowic, R.E. (1997), 'Synaesthesia: phenomenology and neuropsychology', in *Synaethesia*, ed. S. Baron-Cohen & J.E. Harrison (Oxford: Blackwell).

Ericsson, K.A. (2003), 'Exceptional memorizers: Made, not born', *Trends in Cognitive Sciences*, **7** (6), pp. 233–5.

Jack, A.I. (2001), 'Paradigm Lost: A review of *Consciousness Lost and Found* by Lawrence Weiskrantz', *Mind & Language*, **16** (1), pp. 101–7.

Jack, A.I. & Roepstorff, A. (2002), 'Introspection and cognitive brain mapping: From stimulus-response to script-report', *Trends in Cognitive Sciences*, **6** (8), pp. 333–9.

Jack, A.I. & Shallice, T. (2001), 'Introspective physicalism as an approach to the science of consciousness', *Cognition*, **79** (1–2), pp. 161–96.

Jacoby, L.L. (1991), 'A process dissociation framework: Separating automatic from intentional uses of memory', *Journal of Memory and Language*, **30** (5), pp. 513–41.

Koriat, A. & Goldsmith, M. (1996), 'Monitoring and control processes in the strategic regulation of memory accuracy', *Psychological Review*, **103** (3), pp. 490–517.

Latour, B. (1999), *Pandora's Hope: Essays on the Reality of Science Studies* (Cambridge, MA: Harvard University Press).

Mandler, G. (1980), 'Recognising: The judgment of previous occurrence', *Psychological Review*, **87**, pp. 252–71.

Marcel, A. (1993), 'Slippage in the unity of consciousness', in *Experimental and Theoretical Studies of Consciousness*, ed. G.R. Block & J. Marsh (Chichester: Wiley).

Nisbett, R.E. & Wilson, T.D. (1977), 'Telling more than we can know: Verbal reports on mental processes', *Psychological Review*, **75**, pp. 522–36.

Paulesu, E., Harrison, J., Baron-Cohen, S., Watson, J., Goldstein, L., Heather, J., Frackowiac, R. & Frith, C. (1995), 'The physiology of coloured hearing', *Brain*, **118**, pp. 671–6.

Roepstorff, A. & Frith, C.D. (in press), 'What's at the top in the top-down control of action?', *Psychological Research*.

Roepstorff, A. & Gjedde, A. (2003), 'Subjectivity as a variable in brain imaging experiments' [in Danish], in *Subjektivitet og videnskab. Bevidsthedsforskning i det 21. århundrede*, ed. D. Zahavi & G. Christensen (Roskilde: Roskilde Universitetsforlag).

Rowe, J.B. & Passingham, R.E. (2001), 'Working memory for location and time: Activity in prefrontal area 46 relates to selection rather than maintenance in memory', *Neuroimage*, **14** (1 Pt 1), pp. 77–86.

Siegler, R.S. & Stern, E. (1998), 'Concious and unconscious strategy discoveries: A microgentic analysis', *Journal of Experimental Psychology: General*, **127** (4), pp. 377–97.

Summerfield, C., Jack, A.I. & Burgess, A.P. (2002), 'Induced gamma activity is associated with conscious awareness of pattern masked nouns', *Int J Psychophysiol*, **44** (2), pp. 93–100.

Teasdale, J.D., Moore, R.G., Hayhurst, H., Pope, M., Williams, S. & Segal, Z. V. (2002), 'Metacognitive awareness and prevention of relapse in depression: Empirical evidence', *J Consult Clin Psychol*, **70** (2), pp. 275–87.

Tulving, E. (1972), 'Episodic and semantic memory', in *Organization of Memory*, ed. E. Tulving & W. Donaldson (New York: Academic Press).

Wheeler, M.E., Petersen, S.E. & Buckner, R.L. (2000), 'Memory's echo: Vivid remembering reactivates sensory-specific cortex', *Proc Natl Acad Sci U S A*, **97** (20), pp. 11125–9.

Wilkes, K.V. (1988), '——, yishi, duh, um, and consciousness', in *Consciousness In Contemporary Science*, ed. A.J. Marcel & E. Bisiach (Oxford: Clarendon Press/Oxford University Press).