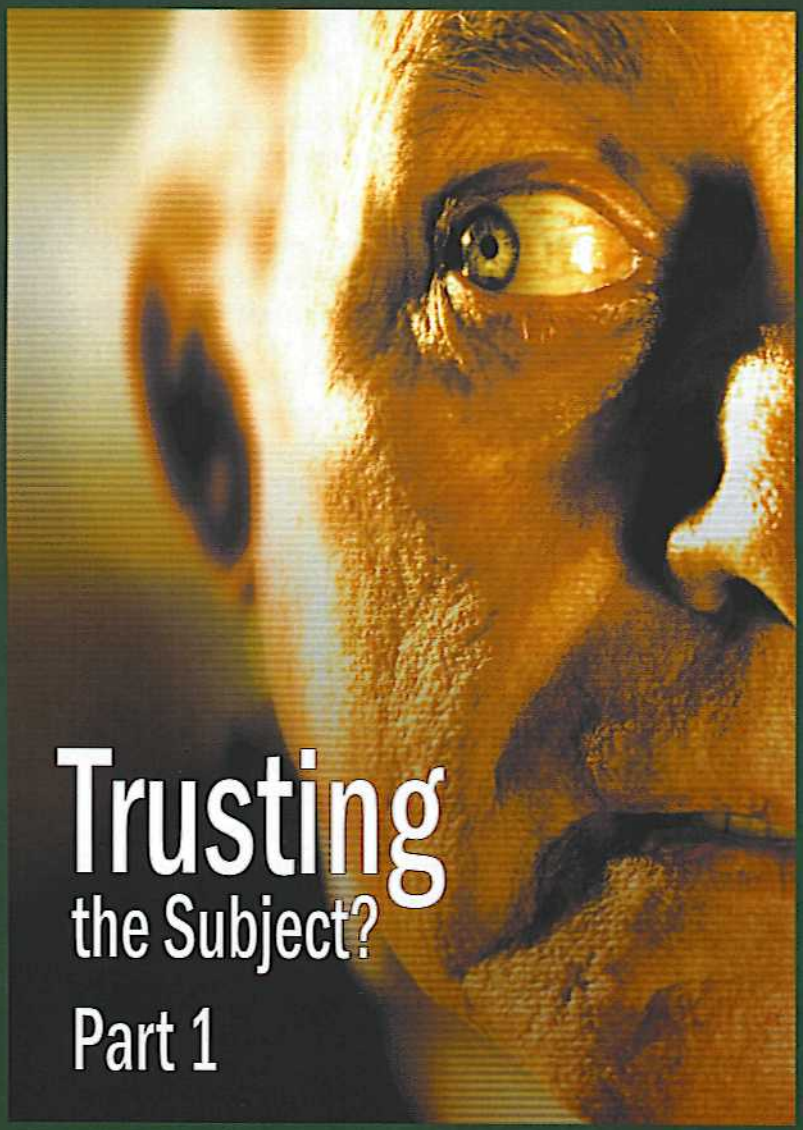
Offprint

Journal of Consciousness Studies

Volume 10, No.9-10 (2003)

controversies in science and the humanities



Trusting the Subject?

Part 1

K. Anders Ericsson

Valid and Non-Reactive Verbalization of Thoughts During Performance of Tasks

Towards a Solution to the Central Problems of Introspection as a Source of Scientific Data

Abstract: *Recent proposals for a return to introspective methods make it necessary to review the central problems that led psychologists to abandon those methods as sources of scientific data in the early twentieth century. These problems and other related challenges to verbal reports collected during the cognitive revolution during the 1960s and 1970s were discussed in Ericsson and Simon s (1980; 1993) proposal for a theoretically motivated procedure to elicit valid and non- reactive concurrent verbalization of thoughts while subjects were performing tasks. The same proposal explains why other verbal reports, such as introspections, detailed descriptions or explanations, require additional cognitive activity that often leads to reactivity and invalid reports. Finally, a new proposal is sketched for how the generation of introspective reports might be incorporated within a framework for non-reactive and valid verbalization of thoughts.*

The contributions to this special issue demonstrate rapidly growing interest in introspective methods for studying experience and cognitive phenomena. As long as the procedures for collecting evidence are consistent with existing cognitive theories, scrutiny of them is often limited. The real test of the validity of the data-collection procedures comes when the uncovered findings are in conflict with prevailing theories and investigators with opposing views attempt to replicate each other's findings. This paper argues that it is necessary to understand the fierce controversies over introspective evidence in order to allow the development of improved methods that will meet the standards for reproducible scientific data.

Correspondence: K. Anders Ericsson, Department of Psychology, Florida State University, Tallahassee, Florida 32306-1270, USA. *Email: ericsson@darwin.psy.fsu.edu*

Journal of Consciousness Studies, 10, No. 9-10, 2003, pp. 1-18

During the early twentieth century introspective data on 'imageless thought' was presented by Karl Buhler (1907) to refute the dominant theory of Wilhelm Wundt, one of the founders of laboratory research in psychology. During the subsequent fierce argument, introspective data failed to provide reproducible objective evidence that could resolve the argument. Following this and other related failures, the introspective method with its trust in observers' ability to analyse and report their conscious experience was impeached by John Watson (1913), who proposed a new methodological approach based on observable behaviour and performance. The criticism of introspective methods led eventually in laboratory psychology to the rejection of virtually all reports of thinking for several decades.

During the cognitive revolution in the 1950s and 1960s alternative types of verbal reports of thinking were used to gather information about cognitive structures and processes. Investigators soon discovered that some types of verbal reports were reactive, that is the performance of participants giving the reports differed from that of silent control subjects (Gagne and Smith, 1962). Other types of reports were directly inconsistent with the observed behaviour. In several reviews Herb Simon and I (Ericsson and Simon, 1980, 1993, 1998; Ericsson, 2002) showed that the detailed instructions and the methods to induce participants to give verbal reports influenced the validity and reactivity of collected verbal-report evidence. Most important, we proposed a new model for coordinating the processes responsible for generating 'think-aloud' reports on thinking with the ongoing task-directed processes to produce valid and non-reactive verbal reports. The same model explained how more complex verbalization processes induced changes in cognitive processes and led to reactive and invalid reports.

The Ericsson-Simon (1993) Model of Verbalization of Thinking has been accepted as a useful foundation for discussing introspection (see the entry on 'Psychology of Introspection' in the Routledge Encyclopedia of Philosophy by Von Eckardt (1998). This framework for collecting verbal reports on cognitive constructs, such as memory and rules, has even met the standards of evidence for behaviourists (Austin and Delaney, 1998). As a further recognition of its validity, protocol analysis now plays a central role in applied settings, such as in the design of surveys and interviews (Sudman, Bradburn and Schwarz, 1996) and user testing of computer tools and applications (Henderson *et al*, 1995). It is fair to say that there has been a dramatic advance in the development of rigorous methods for collecting evidence on mediating cognitive processes and structures. At the same time, some investigators (Jack and Shallice, 2001) have viewed the Ericsson-Simon model as too constraining to study consciousness. Jack and Roepstorff (2002) propose a revision of the type of introspective methods embraced by pioneers such as William James (1890, p. 185), who recommended that we should 'first and foremost and always' be 'looking into our minds'.

I will argue here that a reintroduction of introspective methods must be based on a deep understanding of the past controversies over introspective reports. Now, a century later, we have a better understanding of some fundamental issues with evidence collected using the introspective method. In a brief historical

overview I will describe how philosophers and pioneering psychologists discovered the problems associated with trusting observers' reports from their analysis of complex and dynamic mental states. I will also describe how they changed their research methods to avoid these problems and designed tasks to elicit reproducible performance. In the main section of this paper, I will describe how Herb Simon and I tackled the problems of verbal reports of thinking, specifically how we proposed theoretical accounts of the act of verbalizing thoughts with minimal interference and how we designed reporting procedures that yielded non-reactive overt expression of task-directed thinking. I will then review evidence on validity where this type of verbalization of thoughts is shown to be consistent with task analyses of how the task can be performed, with other types of process data, such as latencies and eye-fixations, and with experimental tests of proposed cognitive mechanisms. Herb Simon's and my goal was to design a highly constrained measurement situation where trust is not an issue. In performing these tasks not even devious participants would be able to produce verbalizations of thoughts that are inconsistent with their task-directed processes that mediate performance. At least, they would not be able to do so without violating our checks for validity.

Turning to my central argument about introspection I will show that verbal-report methods that attempt to collect more detailed observations on the mental states than can be unobtrusively verbalized are associated with reactive effects on performance. In my conclusions I will argue that such reporting methods can integrated into Herb Simon's and my theoretical framework by proposing detailed accounts of the additional cognitive processes that mediate the supplementary observations. Finally, I will propose how phenomena discovered by introspective methods can be captured through the design of new tasks that explicitly target those phenomena and where mediating processes can be studied with valid and non-reactive reports of task-directed thinking.

Varieties of Introspection and the Central Problems Encountered at the Beginning of the Twentieth Century

The controversy over 'imageless thought' should be seen as a culmination of efforts to refine introspective methods for studying thinking. The method of introspection changed from an informal method to study one's own experiences in the seventeenth and eighteenth centuries to a scientific method applied to specific mental phenomena in the laboratory in the late nineteenth century. To convey my view of the central problems of the introspective method, I will first describe my understanding of its historical development, its demise and its subsequent transformation into alternative types of verbal reports.

Historic sketch of the development of introspection

The observation of one's own spontaneous thinking by 'looking within' has a long history that can be traced back at least to the Greeks. Aristotle is generally given credit for the first systematic attempt to record and analyse the structure of

thinking. He focused particularly on recalling a specific piece of information from memory. On one occasion he even reported on a specific sequence of thoughts corresponding to the recall of 'autumn': 'from milk, to white, from white to air, and this to fluid, from which one remembers autumn, the season one is seeking' (Aristotle translated by Sorabji, 1972, p. 56). More generally, Aristotle argued that thinking corresponds to a sequence of thoughts (see the boxes in Figure 1), where the brief transition periods between consecutive thoughts (see the arrows in Figure 1) do not contain any reportable information. Hence the processes that determined how one thought triggered the next thought could not be directly observed, but had to be inferred by retrospective reflection of the relations between consecutive thoughts. By examining his memory of which thoughts tended to follow one another, Aristotle inferred that previously experienced associations were the primary determining factor.

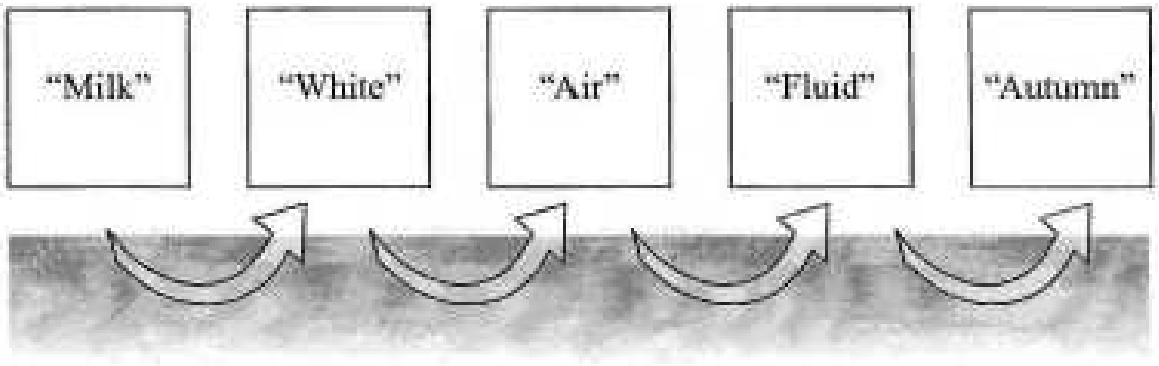


Figure 1. An illustration of Aristotle's view of thinking as a sequence of states with reportable thoughts, while the processing responsible for transitions between consecutive thoughts remains inaccessible.

Aristotle's account of thinking as a reportable sequence of thoughts has never been seriously challenged. This account also by implication recognizes limits for information that can be gained directly by introspection on thinking. However, philosophers in the seventeenth, eighteenth and nineteenth centuries (Ericsson and Crutcher, 1991) raised deeper questions about the nature of thoughts that could not be answered by such a simple description of thinking. For example, could all thoughts be described as mixtures of sensory images derived from past experiences? In order to evaluate these claims empirically, philosophers would typically relax and daydream, allowing their thoughts to wander. Once a thought emerged, it would be inspected carefully and studied to assess its sensory modality and components. The British philosopher David Hume even argued that these thoughts were as detailed as original perceptions (e.g. we cannot imagine a printed page of text without imaging every letter on the page at the same time) (James, 1884).

Towards the end of the nineteenth century, some philosophers proposed 'self experiments' to assess the accuracy and validity of such memory images. For example, Hamilton (1859) proposed that one could throw a number of pebbles behind one's back and then rapidly turn around and take apeek before closing

one's eyes. When a memory image of such a glance was compared to the actual perceptual scene by opening one's eyes again, it was clear that the memory image contained accurate information about only a small number of specific pebbles. More generally, it was becoming clear that thoughts and mental images differed fundamentally from the perceptions and sensations corresponding to external objects.

External objects can be inspected and details about them can be noticed without changing their content and structure. In contrast, when thoughts and images are inspected the associated mental image also changes. For example, when one generates a visual image of rose and then focuses on one of its components, such as the petals, other aspects of the image of the rose are reported to change as well. The most influential laboratory psychologists of the late nineteenth century, Wilhelm Wundt (1897), argued that introspective analysis of experience was possible only when simple physical stimuli, such as points of light and brief sounds, were presented repeatedly to observers under controlled conditions. In fact, he argued forcefully that introspective analysis of complex thoughts as they emerge during spontaneous thinking was not possible, because such analysis would change the corresponding mental states and thus disrupt the course of thinking.

Around the beginning of the twentieth century investigators at the University of Wiirzburg generated considerable controversy with their attempts to study spontaneous thinking with introspective methods. Trained introspective observers were invited to their laboratory and asked prepared questions, such as 'Do you understand the following proverb, "We depreciate everything that can be explained"?' (Buhler, 1907). The observers were asked to give their answers to the questions as quickly as possible and then after each answered question they gave detailed retrospective reports about their thoughts that mediated the question-answering process. The retrospective reports were extensive and detailed. Most reported thoughts consisted of visual and auditory images, but some observers claimed to have experienced thoughts without any corresponding imagery (imageless thoughts).

The proposed existence of imageless thoughts had far-reaching theoretical implications for Wundt's theory that all 'experience' corresponds to neural activity. The original paper by Karl Buhler (1907) led to a heated exchange between him and Wilhelm Wundt (1907) who argued that these reports were artifacts of the reporting methods and the poor training of the observers. Ultimately, the most devastating aspect of this controversy was that it revealed that the issue of imageless thoughts could not be resolved empirically, which in turn impeached introspection as a scientific method by which to study thinking.

The central problems of introspective analysis of thoughts

The central problem of introspection was not so much 'looking within', but rather the problem of inducing the same mental states in many observers where the states were sufficiently stable to allow consistent judgments across observers. In fact, the introspective analysis of sensory stimuli in psychophysical

studies of perception was never rejected by John Watson (1913), and this work continued uninterrupted for the whole twentieth century. In psychophysical studies experimenters could repeatedly present each of a collection of simple sensory stimuli, and the ability to discriminate these stimuli could be evaluated objectively. But no comparable method was available to reproduce the same thought or complex mental state for many different participants. Therefore, it was necessary to trust the observers' reports of their idiosyncratic thoughts, because it was not possible to assess the observers' thoughts or mental states independently in order to allow objective verification of the accuracy of their reports.

(a) From trust to reproducible performance. At the beginning of the twentieth century researchers were naive about problems of trust. For example, in studies of memory the participants were presented with stimuli and later the same stimuli were presented to assess memory and provide introspective reports. The subjects performing the introspections knew that the same stimuli were shown both at presentation and at the recognition test, so one would have to trust their reports that they remembered the stimuli. New procedures for objective tests of memory were later developed; the participants were given recognition tests where they had to identify the presented stimuli among a mixture of old and not-previously-seen stimuli. This measurement procedure relieved the need to trust the participants' reports of memory, because highly accurate recognition (reliably higher than guessing) provides proof that the participants remembered the presented stimuli. Even more compelling is the case when a participant could recall most of the presented stimuli. The probability is small that a participant can correctly guess which word was presented from items sampled from a pool of several thousand words.

Subsequent research relied on these objective tests of memory to evaluate the validity of introspective judgments of participants. For example, 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.

More generally, experimental psychologists developed standardized tests with stimuli and instructions where the same pattern of performance could be replicated in the same laboratory as well as different ones. Furthermore, psychologists redirected their research away from complex mental processes, such as thinking, and towards processes that were unaffected by prior experience and knowledge. For example, subjects were given well-defined simple tasks, such as memorization of lists of nonsense syllables, e.g. XOK, ZUT. In these tasks it is easy to measure objective performance and thus there is no need to trust participants'

honesty or their willingness to cooperate. In addition, the formation of simple basic associations in memory led experimenters to predict no cognitive mediation and thus no reports of mediating thoughts, and the issue of trusting subjects' verbal reports appeared to become essentially irrelevant.

(b) Studying complex thought with non-introspective verbal reports. The interest in studying complex processes of thinking did not completely stop. Interestingly, the first investigator to come up with a new method for studying thinking was John Watson (1920). He instructed a friend to solve a problem and asked him to 'think aloud' while working on it. According to Watson, thinking was accompanied by covert neural activity of the speech apparatus that is 'inner speech'. Hence thinking aloud did not require observations by any hypothetical introspective capacity, and thus all that was necessary to think aloud was merely to give overt expression to these subvocal verbalizations. Many other investigators proposed similar types of instructions to give concurrent verbal expression of one's thoughts (see Ericsson and Simon, 1993, for a brief historical review).

Renewed Interest in Verbal Reports of Thinking During the Cognitive Revolution

The cognitive revolution in the 1950s and 1960s brought renewed interest in higher-level cognitive processes. Investigators started to explore how problem solving, concept formation and decision making could be explained by mediating thought processes. Cognitive theories were proposed in which strategies, concepts and rules were central to the account of human learning and problem solving (Miller, Galanter and Pribram, 1960). Information-processing theories (Newell and Simon, 1972) sought computational models that could regenerate human performance on well-defined tasks by the application of explicit procedures. Much of the evidence for these complex mechanisms was derived from self-observation, informal interviews and systematic questioning of participants. Almost immediately some investigators raised concerns about the validity of these data (Ericsson and Simon, 1993). For example, Robert Gagne and his colleagues (Gagne and Smith, 1962) demonstrated that requiring participants to verbalize reasons for each move in the Tower of Hanoi reduced the number of moves in the solutions and improved transfer to more difficult problems when compared to a silent control condition. Other investigators criticized the validity and accuracy of the retrospective verbal reports. For example, Nisbett and Wilson (1977) showed that participants frequently gave explanations that were inconsistent with their observed behaviour. Consequently, it might appear at first glance that all of these types of verbal reporting would have to be rejected on the same grounds that led to the rejection of introspection. Herb Simon and I (Ericsson and Simon, 1980; 1993), however, were able to show that the methodology had advanced and that it was possible to identify conditions where participants are able to produce consistently valid non-reactive reports of their thinking.

Towards Valid and Non-Reactive Verbal Reports of Thinking

The most significant advances since the earlier controversy over 'imageless thought' concerned the methodology for inducing thinking by presenting well-defined tasks as well as the formal theoretical analysis of possible mediating processes (task analysis).

Task analysis

The introspectionists in Wurzburg presented their observers with tasks to induce spontaneous thinking. Their tasks, such as whether the observers had understood a presented proverb, didn't even have a correct answer. A devious participant could conceivably decide to answer 'yes' or 'no' even before he heard the proverb, and the experimenter would not be able to tell. Consequently, cognitive researchers developed collections of tasks that would induce some specific mental activity and where participants could emit consistently accurate responses only after seeing the stimulus and after some cognitive processing. For example, it would be impossible to guess the correct answers before seeing a maths problem, such as '258 + 893 = ?' or '24 x 36 = ?'.

Completing a series of tasks correctly requires relevant knowledge. In fact it is possible for the investigator to identify various procedures that people could possibly use, in light of their prior knowledge and skills, to generate correct answers. This type of analysis (task analysis) provides a set of possible thought sequences for the successful performance of a task, where the application of each alternative procedure is associated with a different sequence of thoughts (intermediate steps). Let me illustrate how this general type of analysis can be applied to the mentally demanding task of multiplying two 2-digit numbers 'in one's head'. Typically, many adults have acquired basic mathematical knowledge: they know their multiplication table and only the standard 'pencil and paper' procedure taught in school for solving multiplication problems. Accordingly, one can predict that most adults will solve a problem such as 36 x 24 'in their head' by first calculating 4x3 6 then adding 20 x 36. But this specific problem can be solved using alternative methods that are more efficient for adults who know some of the squares of 2-digit numbers. People with more advanced knowledge of mathematics may recognize that 24 x 36 is equivalent to (30 - 6) x (30 + 6) and use the formula $(a + b) \times (a - b) = a^2 - b^2$, thus calculating 36 x 24 as $30^2 - 6^2 = 900 - 36 = 864$. Other subjects may recognize other shortcuts, such as $36 \times 24 = (3 \times 12) \times (2 \times 12) = 6 \times 12^2 = 6 \times 144 = 864$.

Observations on processes that mediate successful task performance

Whereas introspection requires additional observation and analysis by the participant, cognitive researchers obtain information on concurrent processes by simply observing the participants completing the task. Although a covert sequence of thoughts generated during the performance of a task (illustrated in the centre of Figure 2) is never directly observable, it is associated with several types of

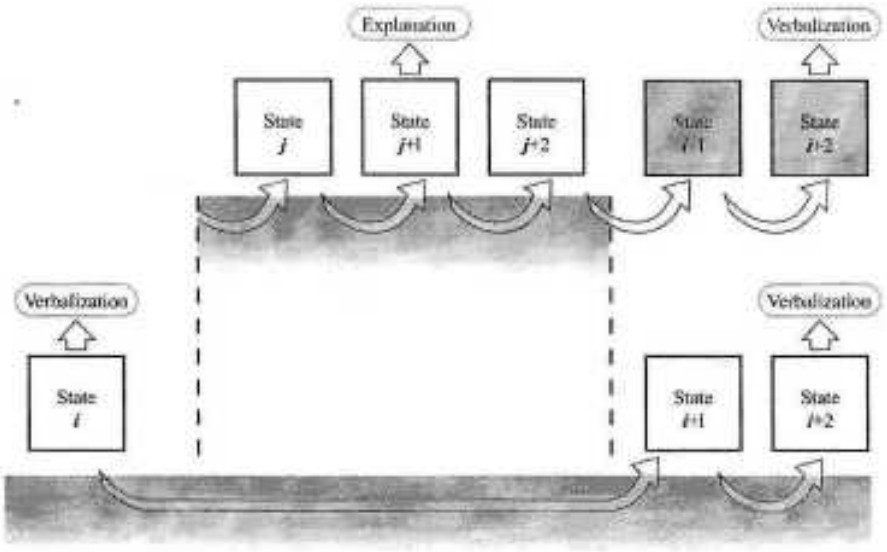


Figure 2. An illustration of a covert sequence of thoughts (centre) generated in response to a presented task along with its associated observable non-verbal indicators.

observable indicators. At the top of Figure 2, the total time required by subj ects to generate their answer (response latency) can be viewed as the sum of the component times for generation of each thought, such as intermediate products and results, when the corresponding thought processes are not overlapping. From the task analysis it is often possible to infer the number of intermediate steps that would be required for solving different problems with alternative methods. From an analysis of average latencies across a set of different problems it will be possible to determine which procedures can best account for the subjects' observed latency pattern. In addition, most tasks are presented to subjects visually and it is necessary for subjects to direct their visual gaze sequentially to process the essential presented information, while an eye-tracking device can register their sequences of eye-fixations. Given that different hypothesized procedures typically predict different sequences of fixating displayed information, it is possible to identify the procedure that shows the highest agreement with the observed eye-fixations.

Other types of observations on thought processes can be collected by instructing participants to provide observable evidence on their thought processes. First, the subjects can be instructed to verbalize their thinking concurrently with performing the task (see Figure 2). Alternatively, the subjects can be asked to give a retrospective report of their thought sequence immediately after the completion of the task. Finally, the subject can give a general description of her strategies (post-session reports) once she has completed numerous tasks during the testing session. All of these types of verbal reports share the problem of introspection however, their generation can be explained only by proposing some additional

cognitive activity because participants do not spontaneously generate these reports when they perform tasks.

Models of the generation of verbal reports

On the basis of their theoretical analysis, Ericsson and Simon (1993) argued that the closest connection between thinking and verbal reports is found when subjects verbalize thoughts generated during task completion (see Figure 2). When subjects are asked to think aloud, some of their verbalizations correspond to merely vocalizing 'inner speech', which would otherwise have remained inaudible. Non-verbal thoughts can also be given verbal expression by brief expressions, labels and referents. For example, when one subject was asked to think aloud while mentally multiplying 36 x 24 on two test occasions one week apart the following protocols were obtained:

1. OK, 36 times 24, um, 4 times 6 is 24, 4, carry the 2, 4 times 3 is 12, 14, 144, 0,
   2 times 6 is 12, 2, carry the 1, 2 times 3 is 6, 7, 720, 720, 144 plus 720, so it would be 4, 6, 864.
2. 36 times 24, 4, carry the — no wait, 4, carry the 2, 14, 144, 0, 36 times 2 is, 12, 6, 72, 720 plus 144, 4, uh, uh, 6, 8, uh, 864.

In these two examples, the reported thoughts are not introspectively analysed into their perceptual or imagery components, but merely verbally expressed and referenced, such as 'carry the 1', '36' and '144 plus 720'. Similarly, subjects are not asked to describe or explain how they solve these problems. Instead, they are instructed to remain focused on solving the problem and merely to give verbal expression to those thoughts that emerge in attention while generating the solution under normal (silent) conditions.

If the act of verbalizing subjects' thought processes doesn't change the sequence of thoughts while completing well-defined tasks, then subjects' task performance should not change as a result of thinking aloud. In a comprehensive review of more than forty studies, Ericsson and Simon (1993) found no evidence that merely giving concurrent verbal expression to one's thoughts (cf Ericsson and Simon's procedure for inducing 'think aloud') altered accuracy of performance compared to that of subjects who completed the same tasks silently under otherwise similar conditions. But some studies showed that participants would take somewhat longer to complete the tasks while thinking aloud — presumably because of the additional time required for completing the overt vocalization of the verbal expression of the thoughts.

The same theoretical framework can also explain how other types of concurrent verbal report procedures would change cognitive processes. For example, when subjects are instructed to explain or carefully describe their thoughts, in contrast to merely verbalizing each thought as it emerges, they are not able to remain completely focused on the task. To be able to verbalize the required explanations and descriptions the participants need to change their thought processes to generate the corresponding thoughts, as is illustrated in Figure 3. This

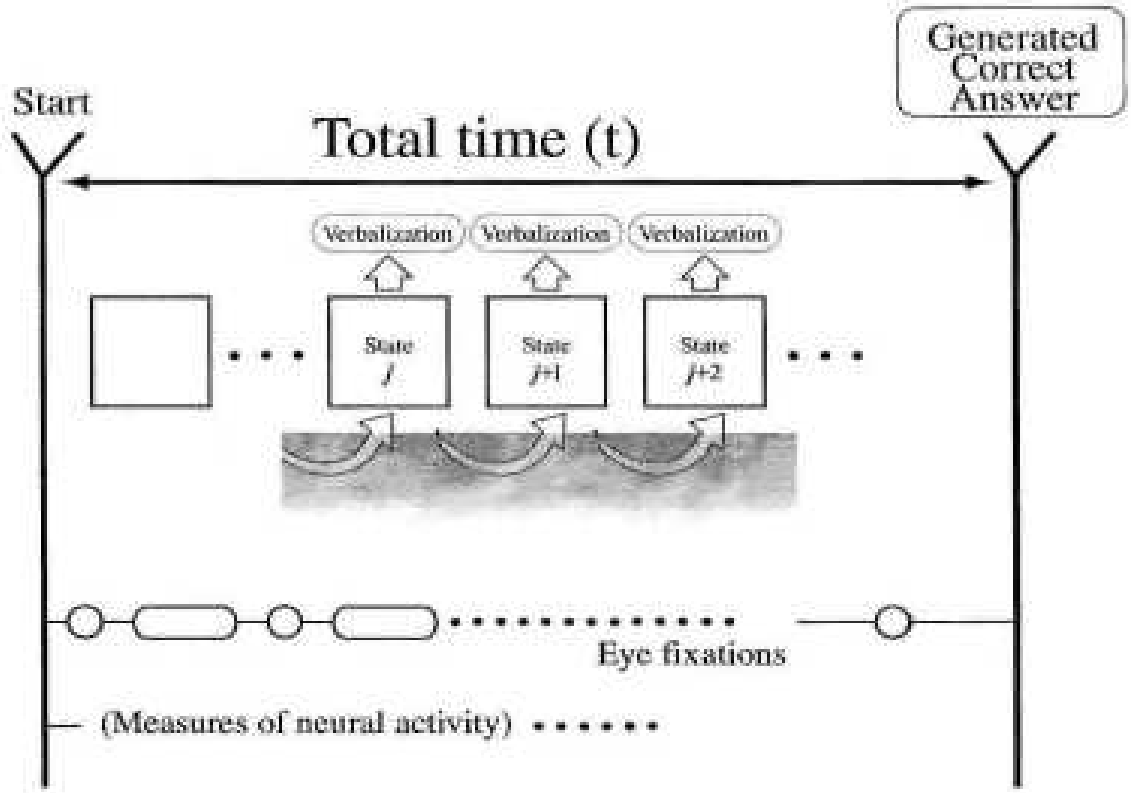


Figure 3. An illustration of how the interspersed cognitive activity of explaining and describing can change the sequence of thoughts as thinking is resumed.

additional cognitive activity must necessarily change the sequence of mediating thoughts. Therefore when participants resume their task-directed processes after the explanation, the sequence of thoughts will differ and thus lead to changes in the accuracy of performance. Our reviews (Ericsson and Simon, 1993, 1998; Ericsson, 2002) have found that instructions to explain one's thinking or to give extended detailed descriptions of presented stimuli (cf. 'verbal overshadowing' reviewed by Schooler, Fiore and Brandimonte, 1997) are associated with reliable changes in memory and task performance.

In sum, after brief instruction and familiarization in giving verbal reports, subjects can think aloud without any systematic changes to the sequential structure of their thought processes (see Ericsson and Simon 1993, for detailed instructions and associated warm-up tasks recommended for laboratory research). The fact that subjects must already possess the necessary skills for efficient verbalization of thoughts is consistent with the extensive evidence on the acquisition of self-regulatory private speech during childhood (Diaz and Berk, 1992) and on the spontaneous vocalization of inner speech by adults, especially in noisy environments (Ericsson and Simon, 1993).

Reproducibility of cognitive phenomena, performance and associated process data

In sharp contrast to the tasks used in traditional research on introspection, cognitive psychologists carefully design well-structured experimental tasks where the successful generation of accurate responses is mediated by the cognitive activity under investigation. For example, it is possible to design tasks that predictably induce one or more of the conscious processes (Type-C processes, such as noticing, planning and monitoring) that Jack and Shallice (2001) propose mediate introspective judgments.

The research methodology attempts to push participants to reach their best consistent performance for the collection of tasks by providing opportunities for 'warm-up' and familiarization with the task before collecting the experimental data on performance (Ericsson and Oliver, 1988). Under these constrained performance conditions it should not even be possible for individuals to exhibit a performance that would be superior to the one observed. Similarly it should not be possible for devious participants to 'fake' the observations on the processes mediating their task performance, that is to intentionally produce eye-movements or verbal reports consistent with one method while proficiently generating the answer to the tasks with a different method.

The processes mediating performance are hardly ever perfectly reproduced for the same task, even when the same subject is tested (cf. the 'think aloud' protocols above, collected a week apart). Even in the case of thinking aloud, where the connection between thoughts and reports is the closest, there is no perfect mapping. This lack of one-to-one correspondence is primarily due to the fact that thoughts that pass through attention are not always verbalized. Most important, verbal reports do not differ in kind from other observations of processes, such as latencies and eye movements, and verbal expression of thoughts is influenced by many uncontrollable factors and thus will vary to some degree from trial to trial.

Validity of concurrent verbal reports and other process data

When we restrict the analysis to those thoughts that are verbalized, the evidence for validity is consistently strong (Ericsson and Simon, 1993). First, when the participants complete the task correctly, their verbalized information is in agreement with one of the sequences of thoughts derived from the task analysis, which had been completed before the observation of the verbal reports. For example, the sample protocols on mental multiplication (reported above) are consistent with only one of the methods uncovered in the task analysis — namely the 'paper-and-pencil' method. Even if the verbalized information had been more sparse for a highly skilled subject and only contained '144' and '720', the reported information would still have been sufficient to reject each of the alternative multiplication methods, because neither of those methods involves generating both of the reported intermediate products. The general finding that a task analysis can identify, a priori, the specific intermediate products that are later verbalized by subjects during their problem solutions, provides compelling

evidence that the concurrent verbalizations reflect the processes that mediate the actual generation of the correct answer.

Second, the verbal reports are only one of several different kinds of observable indicators of the same thought process (see Figure 2). Given that each kind of empirical indicator can be separately recorded and analysed, it is possible to evaluate the agreement of different types of data. Ericsson and Simon (1993) found that the solution methods derived from an analysis of the verbal reports were consistent with those derived from analyses of response latencies and sequences of eye-fixations.

Third, the hypothesized mechanisms mediating thinking and the observed verbalizations can be examined with experimental methods (Ericsson and Simon, 1993). In particular, the reproducibly superior performance of expert and skilled performers offers a unique opportunity for studies of complex cognitive mechanisms using verbal reports. When the sequence of concurrently verbalized thoughts during the execution of representative tasks is reproducible for single experts, then it is possible to experimentally test the validity of the hypothesized cognitive mechanisms by traditional experimental methods. For example, verbal reports of people with exceptional memory have then been evaluated experimentally by presenting them with altered memory tasks where their performance has been predictably reduced in a decisive manner (Ericsson, 1988; Ericsson and Kintsch, 1995; Ericsson, Patel and Kintsch, 2000).

In sum, think-aloud reports provide the most informative data available on thinking during cognitive tasks. Those aspects of verbal reports that can be validated against other sources suggest a close correspondence between the reports and the cognitive processes that produce intermediate results in attention. On the other hand, the reports are not infallible; subjects may occasionally make speech errors and omit articulation of disambiguating information in their verbalizations of thoughts.

Other Types of Verbal Reports of Thinking and Cognition

The main goal for Herb Simon and me (Ericsson and Simon, 1980, 1993) was to identify the conditions and the verbal-report procedure that would elicit valid and verifiable verbalizations of thinking. It is, however, possible to use our model for these verbalization processes of thoughts during performance in discussing other types of verbal reports.

The type of retrospective verbal report that is most closely related to 'think aloud' verbalization involves asking the participants to report their thoughts immediately after the completion of a given trial. When the time to generate the response is brief (1s to 5s), it is likely that the participants can recall their sequence of thoughts reasonably accurately. Our review (Ericsson and Simon, 1993) showed that when participants are asked merely to recall their thoughts, the reported information is consistent with other observations of the same processes, such as latencies.

In contrast to this type of retrospective report, most other verbal reports of thinking require cognitive activities that go beyond immediate recall sequences of already generated thoughts. Ericsson and Simon (1980, 1993) identified two additional cognitive activities that contribute to the decreased validity of the verbally reported information. The first arises when the investigators try to obtain more information than the subjects' thought sequences can provide. For example, some investigators ask subjects *why* they responded in a certain way. Sometimes the subjects' recalled thoughts may provide sufficient information for an answer, but typically the subjects need to go beyond any retrievable memory of their processes to give an answer. As subjects can access the end-products of their cognitive processes only during perception and memory retrieval, they cannot report why only one of several logically possible thoughts entered their attention and thus must speculate to generate answers to such questions. In support of this argument Nisbett and Wilson (1977) found that subjects' responses to why-questions in many circumstances were as inaccurate as those given by other subjects that merely observed and explained another subject's behaviour. Since our review in 1993, Schooler, Fiore and Brandimonte (1997) have presented similar evidence for invalidity of retrospective reports when participants were instructed to give extended detailed descriptions of them (cf. 'verbal overshadowing'). In an attempt to replicate these effects, Meissner, Brigham and Kelley (2001) found no effect of merely recalling what the subjects could confidently remember (cf. the standard retrospective report). They were also able to pinpoint the source of invalidity to aspects of verbal-overshadowing instructions where participants were encouraged to keep producing verbal descriptions, even if they started to feel that they were guessing. In a review of a large number of related studies I (Ericsson, 2002) found no effects of merely giving verbal expression to memories, but consistent effects of attempting to report more than subjects could confidently recall.

Second, investigators often ask subjects to describe their methods for solving problems at the end of the experiment, when they have completed a long series of different tasks. If subjects generated and consistently applied a general strategy for solving all of the problems, they should be able to respond to such requests easily with a single memory retrieval. But subjects typically employ many methods and shortcuts and even change their strategies during the experiment, through learning. Under such circumstances subjects would have great difficulty describing a *single* strategy used consistently throughout the experiment, even in the unlikely event that they were motivated and able to recall most of the relevant thought sequences. It is therefore not surprising that subjects' descriptions are imperfectly related to their averaged performance during the entire experiment.

Conclusion

There is compelling evidence that some types of verbal-report procedures lead to reactive effects and elicit verbal reports that are inconsistent with the participants' observed performance (Ericsson and Simon, 1993). It is therefore not

reasonable to ask the scientific community to trust verbal reports in general. Grouping all types of verbal-report procedures together will bring us back to the unproductive controversies of the beginning of the twentieth century and to the renewed criticisms of verbal reports during the 1960s and 1970s. To avoid an indiscriminate rejection of verbal reports as data, it is essential that we distinguish different types of verbal reports. Herb Simon and I (Ericsson and Simon, 1980, 1993) identified a theoretically motivated procedure to elicit valid and non-reactive verbalization of thoughts. We proposed the existence of processes that could vocalize and verbalize thoughts passing through attention without disrupting the flow of thinking. Furthermore, we (Ericsson and Simon, 1980, 1993) showed in comprehensive reviews of a large number of empirical studies that our recommended procedures for thinking aloud and talking aloud provided valid verbalizations of thoughts where the generation did not influence the sequence of thoughts. In contrast, procedures demanding explanations and detailed descriptions led to reactive effects and thus would provide an inaccurate account of thought under typical (silent) task conditions.

In our book (Ericsson and Simon, 1993) we invited researchers to extend our theoretical analysis and methods of validation to other types of verbal-reporting procedures. Our recommended procedures for eliciting think aloud and talk aloud may well be unduly constraining, as Jack and Shallice (2001) propose. It is entirely possible that alternative and more introspective procedures would provide more information. The critical question is *how* the proposed additional introspective processes are coordinated with the task-directed processes to yield verbal reports of extra information beyond that included in think aloud and immediate retrospective reports. If the introspective processes are additional to the regular task-directed processes, wouldn't it be possible, or even likely, that these processes might be reactive and alter performance? In this paper I have reviewed empirical evidence for how requirements to go beyond the heeded sequences of thoughts, such as by generating explanations and extended detailed descriptions, results in reactivity and alters performance and memory. It would therefore be important for the proponents of new introspective procedures to account for these problems of validity and reactivity and describe if and how the new procedures can avoid those types of interference. If we follow an orderly approach for extending the repertoire of verbal-report procedures it would help us to steadily accumulate a body of valid evidence of verbal reports on thinking and, whenever empirically warranted, expand the repertoire of useful verbal-report procedures.

In our book (Ericsson and Simon, 1993), we argued that it is not necessary to dismiss reactive verbal report procedures. It is better to view the verbal reports as reports for a different task, where the demands for the verbalizations are an integral part. For example, most educators are very interested in understanding the reactive yet typically beneficial effects of instructing students to explain their performance. Recent research inducing students to generate self-explanations is a very interesting development in that direction (Chi *et al*, 1994; Renkl, 1997). The requirement to generate self-explanations has been shown to have a reliably different effect on problem-solving performance than does merely thinking

aloud (Neuman and Schwarz, 1998). Similarly, verbal-report instructions requiring explanations influence performance, whereas thinking aloud doesn't alter performance compared to a silent control condition (Berardi-Coletta *et al.*, 1995). For anyone interested in effective education there is no single correct procedure for inducing verbalizations during problem solving. In fact, a detailed analysis of the different verbalizations elicited during 'think aloud' and 'explain' instructions should provide investigators with an effective tool to identify those induced cognitive processes that are associated with desired changes (improvements) in the subjects' task performance, memory and understanding.

One could make a similar argument that asking individuals to introspect while performing a task is reactive but merely changes the task demands. One needs to consider that introspecting while performing the task differs from the traditional task in some important ways. In the traditional task the participants are helped to find a stable performance that allows repeated completion of a series of similar tasks in a very efficient and reproducible manner. In contrast, the introspecting subjects are instructed to engage in additional observation and noticing. Which aspects of the cognitive processes and for how long these aspects are observed are likely to be unpredictable to both the subjects and the experimenter. As a consequence, the traditional methods of validation, such as task analysis, agreement between reports and performance characteristics (e.g. latencies and eye-fixations) and designed experimental tests, will not be available as tools for the analysis of open-ended introspective reports. This type of introspective report can still provide valuable opinions and ideas that might lead to the generation of interesting and more targeted hypotheses. One of the pioneers of the study of the brain, Lashley (1923, p. 352), said: 'introspection may make the preliminary survey, but it must be followed by the chain and transit of objective measurement'. Many cognitive phenomena, such as afterimages or feelings of knowing, were originally discovered by introspective observation, presumably while the person was engaged in some other task-related activity. Once this type of experience has been recorded and proposed to reflect a general phenomenon it is possible for investigators to try to design tasks where this specific phenomenon can be reliably induced in many subjects. For example, subjects can be presented with bright colour stimuli and then asked to name the corresponding colour in the afterimage (cf. Comstock and Kittredge, 1922). To study the 'feeling of knowing', subjects have been asked to estimate the probability that they will be able to recognize the correct name of the capitol of a presented name of a country, where the estimates can be evaluated against performance on subsequent memory tests (Gruneberg and Monks, 1974). Once suitable tasks with well-defined criteria for performance have been developed, the thought processes of associated phenomena can then be studied with non-reactive reporting methods, such as think aloud and immediate retrospective reports.

In the current exciting quest to better understand consciousness it is hard to overestimate the importance of rigorous data-collection methods that produce independent scientific evidence. Crucial future advances in cognitive science and cognitive psychology will, by necessity, require that we keep rejecting

prevailing theories and fundamental intuitions. These types of advances will only be possible when we have methods that provide reproducible empirical evidence; only such evidence can compel scientists to change their beliefs and thus be capable of successfully resolving current and future theoretical controversies.

Acknowledgements

This article was prepared while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford. I am grateful for the financial support provided by the John D. and Catherine T. MacArthur Foundation, Grant #32005-0. I also want to thank Kathleen Much and Natalie Sachs-Ericsson for very helpful comments on earlier drafts of this article.

References

Austin, J. and Delaney, P.F. (1998), 'Protocol analysis as a tool for behavior analysis', *Analysis of Verbal Behavior*, 15, pp. 41-56.

Berardi-Coletta, B., Buyer, L.S., Dominowski, R.L. and Rellinger, E.R. (1995), 'Metacognition and problem solving: a process-oriented approach', *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 21, pp. 205-23.

Biihler, K. (1907), 'Tatsachen und Probleme zu einer Psychologie der Denkvorgange: I. Uber Gedanken [Facts and problems in a psychology of thinking: I. On thoughts]', *Archiv fur die gesamte Psychologie*, 9, pp. 297-365.

Chi, M.T.H., de Leeuw, N., Chiu, M.-H. and LaVancher, C. (1994), 'Eliciting self-explanations improves understanding', *Cognitive Science*, 18, pp. 439-77.

Comstock, C. and Kittredge, H. (1922), 'An Experimental Study of Children as Observers', *American Journal of Psychology*, 33, pp. 161-77.

Diaz, R.M. and Berk L.E. (1992), *Private Speech: From Social Interaction to Self-Regulation* (Hillsdale, NJ: Erlbaum).

Ericsson, K.A. (1988), 'Analysis of memory performance in terms of memory skill', in *Advances in the Psychology of Human Intelligence*, Vol. 4, ed. RJ. Sternberg (Hillsdale, NJ: Erlbaum), pp. 137-79.

Ericsson, K.A. (2002), 'Toward a procedure for eliciting verbal expression of nonverbal experience without reactivity: interpreting the verbal overshadowing effect within the theoretical framework for protocol analysis', *Applied Cognitive Psychology*, 16, pp. 981 -7.

Ericsson, K.A. and Crutcher, RJ. (1991), 'Introspection and verbal reports on cognitive processes — two approaches to the study of thought processes: a response to Howe', *New Ideas in Psychology*, 9, pp. 57-71.

Ericsson, K.A. and Kintsch, W. (1995), 'Long-term working memory[1], *Psychological Review*, 102, pp. 211-45.

Ericsson, K.A. and Oliver, W. (1988), 'Methodology for laboratory research on thinking: task selection, collection of observation and data analysis', in *The Psychology Of Human Thought*, ed. R.J. Sternberg and E.E. Smith (Cambridge: Cambridge University Press), pp. 392[A]t28.

Ericsson, K.A., Patel, V.L. and Kintsch, W. (2000), 'How experts' adaptations to representative task demands account for the expertise effect in memory recall: comment on Vicente and Wang (1998)', *Psychological Review*, 107, pp. 578-92.

Ericsson, K.A. and Simon, H.A. (1980), 'Verbal reports as data', *Psychological Review*, 87, pp. 215-51.

Ericsson, K.A. and Simon, H.A. (1993), *Protocol Analysis; Verbal Reports As Data* (revised edition; original edition published 1984) (Cambridge, MA: Bradford Books/MIT Press).

Ericsson, K.A. and Simon, H.A. (1998), 'How to study thinking in everyday life: contrasting think-aloud protocols with descriptions and explanations of thinking', *Mind, Culture, & Activity*, 5, pp. 178-86.

Gagne, R.H. and Smith, E.C. (1962), 'A study of the effects of verbalization on problem solving', *Journal of Experimental Psychology,* 63, pp. 12-18. Gruneberg, M.M. and Monks, J. (1974),' "Feeling of knowing" and cued recall', *Ada Psychologica,* 38, pp. 257-65.
Haber, R.N. (1979), 'Twenty years of haunting eidetic imagery: where's the ghost?", *Behavioral & Brain Sciences,* 2, pp. 583-629.
Hamilton, Sir W. (1859), *Lectures on Metaphysics and Logic,* Vol. I (Edinburgh: Blackwood).
Henderson, R.D., Smith, M.C., Podd, J. and Varela-Alvarez, H. (1995), A comparison of the four prominent user-based methods for evaluating the usability of computer software', *Ergonomics,* 39, pp. 2030-44.
Jack, A.I. and Roepstorff, A. (2002), 'Introspection and cognitive brain mapping: from stimulus-response to script-report', *Trends in Cognitive Sciences,* 6, pp. 333-9.
Jack, A.I. and Shallice, T. (2001), 'Introspective physicalism as an approach to the science of consciousness', *Cognition,* 79, pp. 161-96.
James, W. (1884), 'On some omissions of introspective psychology', *Mind,* 9, pp. 1-26.
James, W. (1890), *Principles of Psychology* (New York: Holt).
Lashley, K.S. (1923), 'The behavioristic interpretation of consciousness II', *Psychological Review,* 30, pp. 329-53.
McKelvie, S.J. (1995), 'The VVIQ and beyond: vividness and its measurement', *Journal of Mental Imagery,* 19, pp. 197-252.
Meissner, CA., Brigham, J.C. and Kelley, CM. (2001), 'The influence of retrieval processes in verbal overshadowing', *Memory and Cognition,* 29, pp. 176-86.
Miller, G.A., Galanter, E. and Pribram, K.H. (1960), *Plans and the Structure of Behavior* (New York: Holt, Rinehart, and Winston).
Neuman, Y. and Schwarz, B. (1998), 'Is self-explanation while solving problems helpful? The case of analogical problem-solving', *British Journal of Educational Psychology,* 68, pp. 15-24.
Newell, A. and Simon, H. A. (1972), *HumanProblem Solving* (Englewood Cliffs, NJ: Prentice-Hall).
Nisbett, R.E. and Wilson, T.D. (1977), 'Telling more than we can know: verbal reports on mental *processes', Psychological Review,* 84, pp. 231-59.
Renkl, A. (1997), 'Learning from worked-out examples: a study on individual differences', *Cognitive Science,* 21, pp. 1-29.
Richardson, J.T.E. (1988), 'Vividness and unvividness: reliability, consistency, and validity of subjective imagery ratings', *Journal of Mental Imagery,* 12, pp. 115-22.
Schooler, J. W, Fiore, S.M. and Brandimonte, M. A. (1997), At a loss for words: verbal overshadowing of perceptual memories', *The Psychology of Learning and Motivation,* 37, pp. 291-340.
Sorabji, R (1972), *Aristotle on Memory* (Providence, RI: Brown University Press).
Sudman, S., Bradburn, N.M. and Schwarz, N. (ed. 1996), *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology* (San Francisco, CA: Jossey-Bass).
Von Eckardt, B. (1998), 'Psychology of Introspection', in *Routledge Encyclopedia of Philosophy,* ed. E. Craig (London: Routledge), pp. 842-6.
Watson, J.B. (1913), 'Psychology as the behaviorist views it', *Psychological Review,* 20, pp. 158-77.
Watson, J.B. (1920), 'Is thinking merely the action of language mechanisms?', *British Journal of Psychology,* 11, pp. 87-104.
Wundt, W. (1897), *Outlines of Psychology,* trans. C.H. Judd (Leipzig: Wilhelm Engelmann).
Wundt, W. (1907), 'Uber Ausfrageexperimente und uber die Metoden zur Psychologie des Denkens [On interrogation experiments and on the methods of psychology of thinking]', *Philosophische Studien,* 3, pp. 301-60.