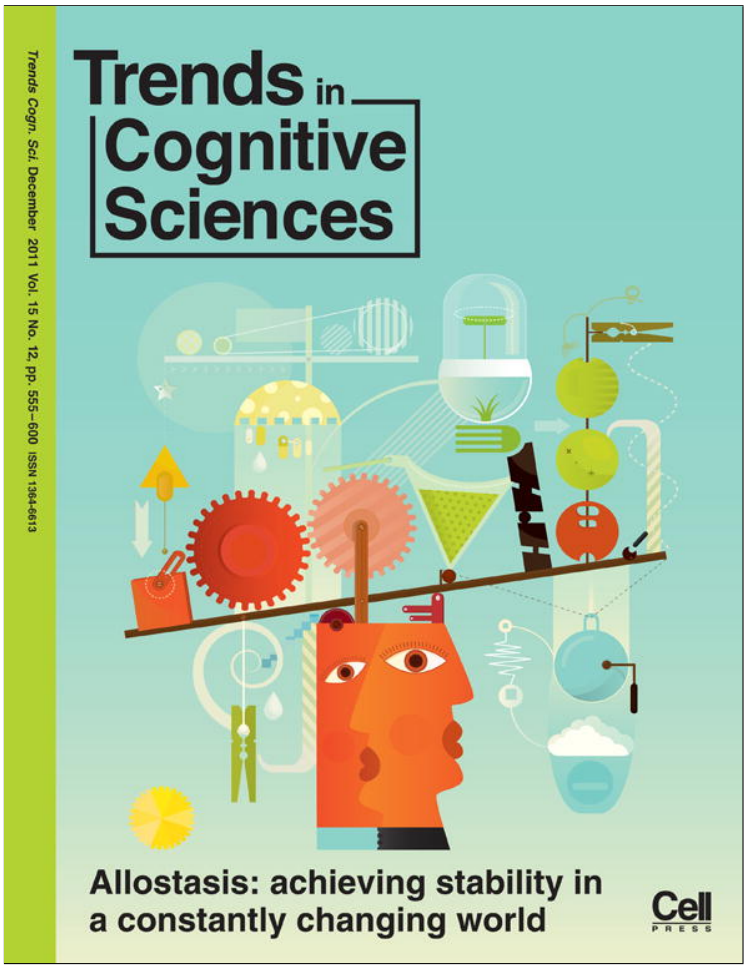

Opinion

Perceptual consciousness overflows cognitive access

Ned Block

Department of Philosophy, New York University, 5 Washington Place, New York, NY 10003, USA

One of the most important issues concerning the foundations of conscious perception centers on the question of whether perceptual consciousness is rich or sparse. The overflow argument uses a form of 'iconic memory' to argue that perceptual consciousness is richer (i.e., has a higher capacity) than cognitive access: when observing a complex scene we are conscious of more than we can report or think about. Recently, the overflow argument has been challenged both empirically and conceptually. This paper reviews the controversy, arguing that proponents of sparse perception are committed to the postulation of (i) a peculiar kind of generic conscious representation that has no independent rationale and (ii) an unmotivated form of unconscious representation that in some cases conflicts with what we know about unconscious representation.

The current status of the overflow controversy

The overflow argument appeals to visual iconic memory (see Glossary) to argue that a conscious perceptual system that has 'rich' contents 'overflows' – that is, has a higher capacity than – the 'sparse' system that cognitively accesses perception [1–6]. A key experimental paradigm that has provided support for the overflow argument was introduced in 1960 by George Sperling. Sperling [7] showed subjects an array of letters (for example, 3 rows of 4 letters as in Figure 1a) for a brief period. Although subjects thought they could see all or almost all of the letters, they could report only 3-4 of them from the whole matrix. However, they could also report 3-4 items from any row that was cued after stimulus offset, suggesting that subjects did have a persisting image of almost all the letters. According to the overflow argument, all or almost all of the 12 items are consciously represented, perhaps fragmentarily but well enough to distinguish among the 26 letters of the alphabet. However, only 3-4 of these items can be cognitively accessed, indicating a larger capacity in conscious phenomenology than in cognitive access. Importantly, the overflow argument does not claim that any of the items in the array are cognitively inaccessible, but rather that necessarily most are unaccessed. For comparison, consider the following: not everyone can win the lottery; however, this does not show that for any particular contestant the lottery is unwinnable. In other words, to say that necessarily most items in the array are not accessed is not to say that any are inaccessible.

Why does the argument for overflow appeal to memory rather than perception? It is not surprising that the capacity of the retina is greater than that of working memory. By relying on this form of memory, we isolate consciousness from the immediate feed of the world and the retina. The neural locus of the high memory capacity demonstrated in the Sperling phenomenon is in brain areas that are candidates for the neural basis of conscious perception rather than in the retina or early vision [2,8–10], and so the Sperling phenomenon may reveal the capacity of conscious phenomenology.

Many theorists have taken change 'blindness' (illustrated in Figure 1b) to show that the overflow argument is wrong. Unattended features of two pictures can differ without the perceivers noticing the difference, despite what seems to be clear perception of both pictures. According to the 'inattentional blindness' interpretation of this phenomenon [11–15], one does not notice the difference because one simply does not consciously see the specific aspect of the scene that constitutes the difference (for instance, the item that is present in one picture, absent in the other). According to the 'inattentional inaccessibility' interpretation [1,4–6,16–19], one normally consciously sees the item that constitutes the difference but fails to categorize or conceptualize it in a way that allows for comparison. Advocates of 'inattentional inaccessibility' think the term 'change blindness' is a misnomer.

Glossary

Access consciousness: a representation is access-conscious if it is made available to cognitive processing.
Change blindness: a misnomer for the phenomenon where people fail to identify changes in stimuli that are easy to notice if one attends to and conceptualizes the items that change.
Fragile visual short-term memory: a type of visual short-term memory, which consists in a persisting visual representation that is intermediate in capacity between retinally-based visual iconic memory and visual working memory and can last 4-5 seconds.
Introspective judgment: first-person judgment made about one's experience.
The overflow argument: an argument to the effect that the capacity of phenomenal consciousness exceeds that of cognitive access.
Phenomenal consciousness: what it is like for a subject to have an experience.
Visual iconic memory: a type of visual short-term memory in which visual representations persist after the stimulus has disappeared. Retinal persistence fuels what might be termed 'pure' visual iconic memory, lasting a few hundred milliseconds, whereas the weaker fragile visual short-term memory is based on persisting cortical activation.
Visual short-term memory: a term that is ambiguous between visual iconic memory, fragile visual short-term memory and visual working memory
Visual working memory: a form of short-term memory which can persist much longer and has a smaller capacity than any form of iconic memory and probably requires categorization. This form of memory is often regarded as the leading edge of long-term memory and is largely shared with other perceptual modalities (see Box 1).

Corresponding author: Block, N. (ned.block@nyu.edu)

1364-6613/$ – see front matter doi:10.1016/j.tics.2011.11.001 Trends in Cognitive Sciences, December 2011, Vol. 15, No. 12

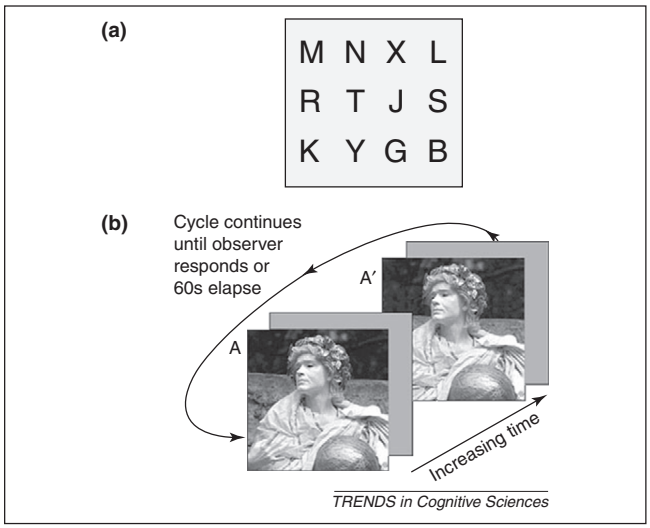

Figure 1. The Sperling paradigm and change 'blindness'. (**a**) An array of letters of the sort presented briefly in the Sperling experiment. Subjects can recall 3–4 items from the whole array but when a row is cued by a tone after the array has disappeared, subjects can recall 3–4 items from any given row. (**b**) An illustration of one kind of 'change blindness': a picture is presented followed by a blank followed by another picture that may differ somewhat from the first picture–and then the series repeats. Perceivers often find it difficult to see what changed, but when the change is pointed out it often seems incredible that one could have missed it. This can happen even when the pictures are arrayed side by side. Do the pictures in the figure differ or not? The answer appears at the end of this caption, but do not look at it until you think you have the answer from the figure itself. Figure courtesy of Ron Rensink. (Answer: in the picture on the left, the background changes at the level of the nose, whereas in the picture on the right, this change comes at the level of the chin.)

One issue in the debate has been the status of an often reported introspective judgment on the part of subjects in experiments with brief presentations [7,20–23] that one consciously sees more than one can cognitively grasp. This introspective judgment is regarded by some advocates of 'inattentional blindness' as the product of a cognitive illusion – the so-called 'refrigerator light illusion' [13,24–26] – according to which, one can see items whenever one attends to their location, thereupon assuming that they were already consciously represented [13,24]. Although there is no direct evidence for the refrigerator light illusion, it has been claimed recently that there is direct evidence that the introspection of richness is the product of one of a family of perceptual rather than cognitive illusions [27–34].

In short, the overflow argument has come under pressure both empirically [27–29,33,35], and conceptually [27–34] largely on the basis of 'change blindness'. At the same time, support for overflow has come from experiments employing other paradigms [8,9,36–42]. This article assesses the argument for overflow in light of these recent contributions, arguing that rich perceptual contents comport better with the evidence than sparse contents.

Generic consciousness combined with unconscious iconic memory

Since experimental subjects can perform the Sperling task successfully, information sufficient to determine 3-4 letters in each of 3 rows – that is, approximately 10.5 letters – must be instantiated in the brain. The fact that subjects in such experiments often observe that 'they saw more than they remembered' ([23], p. 39) motivates the premise of the overflow argument that the information is conscious. However, many critics starting with Kouider and Dehaene have claimed that this information is not conscious until attended and accessed [25,28–32,43,44]. As Cohen and Dennett put it: '[p]articipants can identify cued items because their identities are stored unconsciously until the cue brings them to the focus of attention' ([29] p. 359). Advocates of this view do not regard subjects' reports of a persisting conscious icon as wholly illusory; rather what some of them claim is that the conscious icon contains generic, non-specific, abstract or gist-like letter representations that are neutral between the shape of one letter and another [28–30,33,34,44–50]. They postulate what I will call the 'generic illusion', whereby subjects are said to confuse a generic representation of letter-likeness (parts of which can be made more specific via attention) with specific letter-shape representations that are already specified in consciousness. Cohen and Dennett emphasized gist and ensemble representations as well as generic representations [29]. One illustration of gist/ensemble representations was provided by Alvarez [51], who showed that when subjects track 4 moving disks they also are aware of the 'center of mass' of 4 distractor disks. For simplicity, I will lump gist-like and generic perception together under the heading of generic representation. The view that what is conscious before the cue is generic letter representations is depicted in Figure 2a.

A very different idea of what is in consciousness prior to the cue [28,29,33,34] is a scattering of some conscious features or fragments [27,28]. Attentional processes are supposed to combine the unconscious specific information from the cued row with conscious features or fragments to form conscious representations of specific letters in that one row. Kouider and colleagues have postulated a 'fragment illusion' to explain subjects' impression that they see a grid of specific letters when according to Kouider what they see is sparse fragments. What is common to both the supposed generic illusion and the fragment illusion is depicted in the lower left quadrant of Figure 2a, which indicates that the specific information on the basis of which subjects do the task is unconscious, only becoming conscious when amplified by the cue.

The fragment illusion vs. the generic illusion

Subjects in a Sperling-like experiment sometimes mistake a pseudo-letter, specifically a rotated or flipped letter, in a non-cued row for a real letter [27,28]. (Note that there are no illusions involving cued rows.) The procedure of this experiment is diagrammed in Figure 3.

The devil is in the details: first, the contrast of the displays in this experiment was reduced. Second, the displays were masked. (The authors say: '[i]mportantly, we also added a backward mask to the stimulus array, to eliminate possible retinal persistence of the visual information, which is not supposed to constitute phenomenal consciousness, but rather to input the phenomenal level (Block, 2007)' [27]). The effect of this manipulation was to significantly diminish the icon. How diminished? In standard Sperling paradigms (in which there is no mask), subjects recall approximately 3.5 items correctly from any cued row. That adds up to a total capacity of 10.5

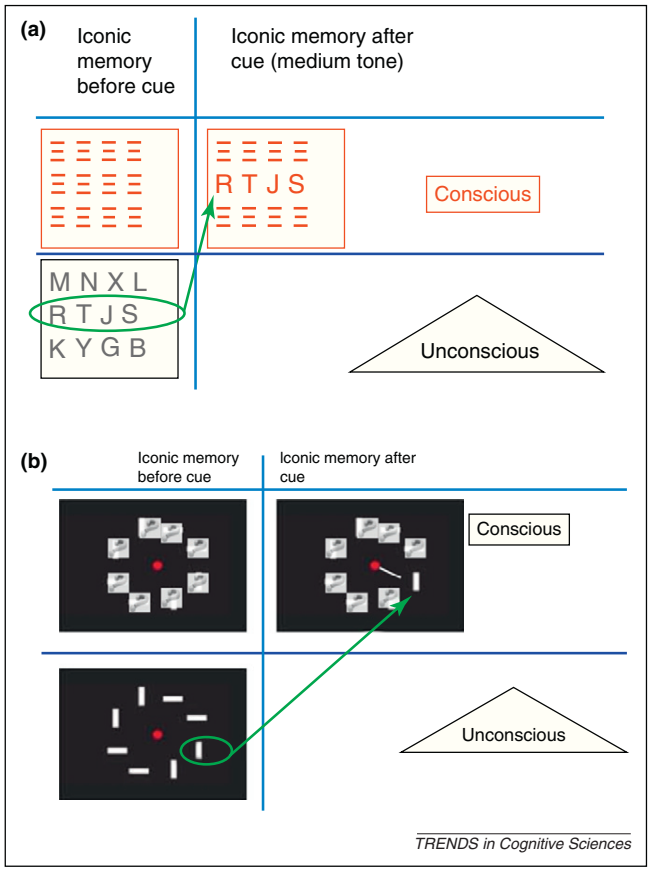

Figure 2. The generic illusion approach to the Sperling and Amsterdam group experiments. **(a)** The generic illusion approach to the Sperling experiment. The upper left quadrant represents what is conscious in iconic memory prior to the cue – the '☰' symbols are meant to indicate a phenomenal representation of a 'gist-like' or 'generic letter' that does not contain information sufficiently specific to decide among letters. The Sperling array is supposed to appear to the subject prior to the cue as an array of letters that don't look like any specific letter. The lower left quadrant indicates that the specific information necessary to do the Sperling task is in an unconscious representation. The green ellipse represents attentional capture and amplification that is supposed to move the unconscious specific information necessary to do the task into consciousness. The upper right quadrant represents the contents of consciousness after the cue, in which the cued row is conscious. The upshot is that the information on the basis of which a subject does the Sperling task is in unconscious iconic memory with consciousness representing only a geometric array of generic gists plus fragmentary sparse features before the cue. **(b)** The generic illusion approach to the Amsterdam group experiments. Question-marks in the conscious representation indicate generic unoriented rectangles. The way the initial Amsterdam array is supposed to appear to the subject is as an array of non-square rectangles with no specified orientation. As before, the green ellipse represents attentional capture and amplification that is supposed to move the unconscious specific information necessary to do the task into consciousness.

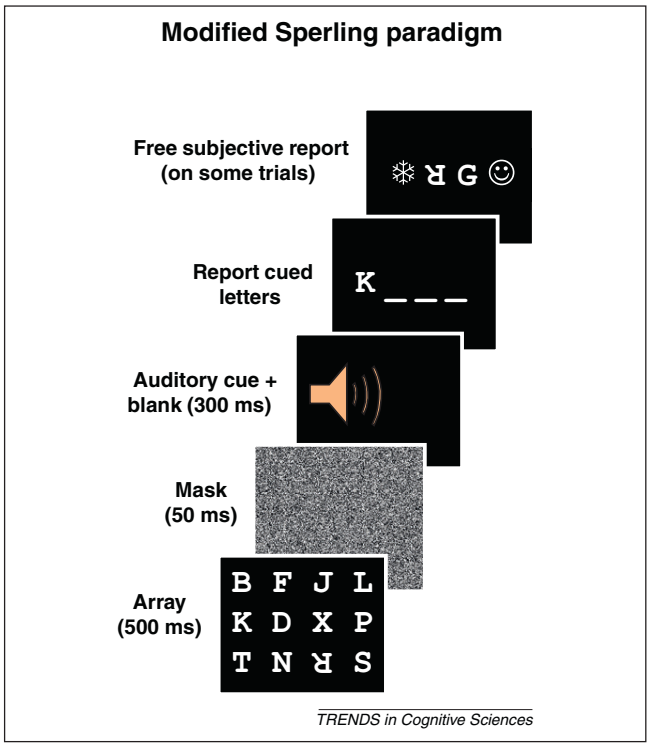

Figure 3. Kouider *et al.*'s modified Sperling paradigm [27,28]. An array of normal letters and other symbols (lower left) is presented for 500 ms and then 'masked', that is, followed by a pattern that makes them harder to see. (Masking was not used in the original Sperling experiment.) The array might have a rotated or flipped letter, for example the 'R' in the last row, or a 'wingding', such as the non-letters in the top panel above. An aural cue is presented, the pitch of which tells the subject to focus on a specific row. The subject then reports as many letters in the cued row as possible, as indicated in the second from top panel. In some trials, there is a 'free subjective report' procedure, during which the subject moves a cursor over a set of symbols, clicking when one of the symbols is thought to have occurred in the original array (top panel). Wingdings were recognized reliably whether they were in the cued or uncued rows, but rotated or flipped letters were only recognized reliably if they occurred in the cued rows – that is, sometimes subjects saw, for example, a rotated 'R' in an uncued row but in the free subjective report (top panel) reported a real 'R'. Reproduced, with permission, from [28].

out of 12 items in the array, the 'partial report advantage', which shows that iconic memory has a higher capacity than working memory (Box 1). In the experiment just described, the average recall is only 1.47 out of 4 items in each row, that is, 4.41 out of 12 (and only 3.9 out of 12 in cases where a pseudo-letter is present), not much above the capacity of working memory. It is usually said that the icon in the Sperling experiment has vanished by 500 ms because the average for any given row decreases to 1.5 letters [38] so by that criterion there is no icon in this experiment. The experiment by Kouider and colleagues [27] is taken to show that iconic memory in the Sperling experiment is fragmentary. However, the fragmentariness demonstrated may be due to the icon being diminished by low contrast stimuli and a mask. The experimenters were right to want to disrupt retinal persistence. However, there is a way to do that without disrupting iconic memory, by using an iso-luminant stimulus of the type used by Sligte *et al.* [8] (see Figure 4b), since it is invisible to (color blind) rods, which are the main source of retinal persistence.

Furthermore, the illusion effect is small. The authors calculate the rate of errors due to illusion (always in the uncued rows) to be in the vicinity of 10%-15%, not a surprising level of error from the perspective of a 'rich' view of perception that also allows that the shape representations contain features and feature fragments, so long as the fragments are specific enough to do the task with the observed degree of success [52]. This experiment makes it plausible that a robust icon of the sort in a normal Sperling-type experiment, even if fragmentary, does involve enough phenomenally conscious shape information to distinguish 3-4 out of the 26 possible letters in all three rows.

The putative 'fragment illusion' concerns the upper right quadrant of Figure 2a which depicts the conscious icon after the cue. (Recall that the 'free subjective report' procedure from Figure 3 in which subjects are supposed to click the mouse over items that were in the array occurs well after the cue.) The experiment is supposed to show

Box 1. Working memory

In the Sperling experiment, subjects recall about 3-4 items from an array of 12-16, and also from any cued row. The same number (though decreased slightly for infants) shows up in a variety of experiments that track memory of ongoing events [63,64]. For example, if a monkey sees an experimenter ostentatiously place a small number of pieces of apple in one bucket, one at a time, and a small number in another bucket, again one at a time, when the experimenter withdraws, the monkey reliably goes to the bucket with more pieces if the numbers of pieces in both buckets are 4 or fewer [65,66]. However, if the number of pieces in one of the buckets exceeds 4, the monkey's choices are at chance level: for instance, in the case of 3 vs. 4, monkeys reliably pick 4, but in the case of 3 vs. 8, they choose randomly. One might have guessed that more than 4 would be represented as 'a lot', but no; their working memory storage appears limited to 4 'chunks' [65,66]. Similar results were obtained with infants using this paradigm [67] and a completely different paradigm [68] and even for bees [69]. Using yet a different paradigm, a limit of 4 was also shown with multimodal stimuli–4 visual items or 3 visual items plus one spoken digit [70].

One of the basic controversies concerning working memory [71] is the question of whether the limit that shows up in so many paradigms is a basic feature of working memory [64,71–74] or a by-product of a shared pool of resources [75,76] and its mode of distribution to different representations. Alvarez and Cavanagh [77] showed (see Figure I) that how many object-representations are held in working memory and their resolution depends on their complexity, with a 4 item limit for colors but a 2 item limit for 3-D cubes [73,77], disconfirming a simple 'slot' model [78,79]. The work reported in Figure I suggests a limit between 4 and 5 for ideally simple items. More recent work suggests a more complex version of a model that combines something like slots with hierarchical structures representing objects and their properties [79,80]. The work discussed in the main text concerning three forms of visual short-term memory separately estimates variants of iconic memory and working memory. There are strong complexity effects in these experiments, but the capacity of fragile visual short-term memory discussed in the main text is always substantially greater than working memory for any given set of materials.

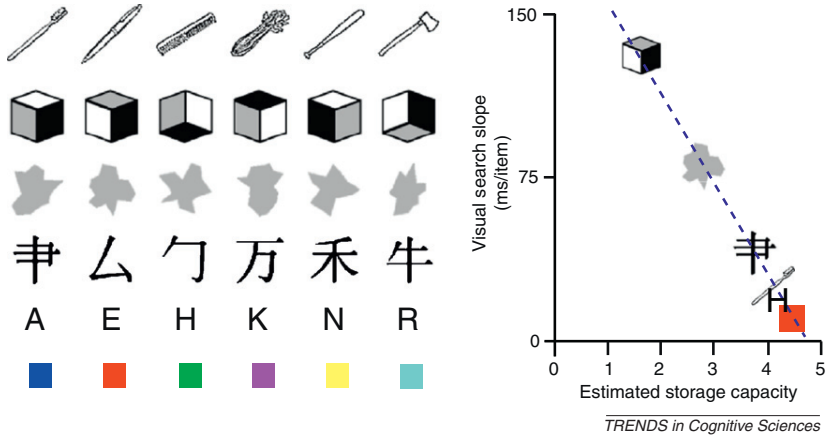

Figure I. Working memory capacity. Using a visual search procedure to estimate information load per item, Alvarez and Cavanagh [77] showed that the number of objects that can be stored in working memory depends on the complexity of the objects. The limit for cubes was under 2, whereas the limit for colors was slightly over 4. Extrapolating their data points suggests a limit of about 4.7 for theoretical stimuli with no information content. Both information load and number of objects limit visual working memory. This figure comes from an adaptation of Alvarez & Cavanagh's figure by S.J. Luck [71]. Reproduced, with permission, from [71].

that the conscious generic representations plus conscious fragments from the uncued rows cannot support letter identification, on the assumption that letter identifications are done on the basis of conscious representations. This reasoning is conceptually flawed, however, since the overflow argument is concerned with whatever it is in the conscious pre-cue icon (i.e., the upper left, not upper right quadrant) that allows subjects to do the task. The overflow argument takes no stand on whether or not the cue erases that conscious icon – and the cue has historically been pointed out as a source of decay of the icon. However, the fact that enough of the conscious representation does persist after the cue as indicated by the low (10%-15%) error rate – even with a very diminished icon – suggests that before the cue (Figure 2a, upper left quadrant) there was at least as much specific conscious information from all the rows as after the cue, since the cue cannot be expected to magnify conscious information in the uncued rows. Thus, contrary to what the authors argue, this result is consistent with the overflow argument and the hypothesis of rich phenomenology.

There is one factor in common to the supposed 'generic illusion' – the idea that the subjects' consciousness represents a grid of letters without any specific conscious information about what letters they are – and the supposed 'fragment illusion' – the idea that what is in the subjects' consciousness is fragments of letters or clusters of letter features. That common factor is the appeal to unconscious iconic memory that is specific enough to provide the information to do the task. However, there is no independent evidence for highly detailed unconscious iconic memory, and it goes counter to what subjects report about their own experience. As Baars notes, even after finding that they can report only 3 or 4 of the items, 'subjects – and experimenters serving as subjects – continue to insist that they are momentarily conscious of *all* the elements in the array' ([53], p. 15). Between the two illusions, the fragment illusion is somewhat more plausible, since it is difficult to understand what it would mean for subjects' consciousness prior to the cue to consist of conscious representations of a grid of instances of letterness without any specific shape representations. Critics of overflow [29,30,33] appeal to cases such as peripheral vision as an example of generic representation. However, even if it is true that we can see movement without shape in the periphery, this would probably stem from specialized motion-detection

circuits in cortical area MT/V5 and does not show that we can consciously see generic letterness. Stazicker [30] appeals to crowding as a case of conscious generic letter representation. However, subjects' reports of the experience of crowding do not sound anything like generic letter representation. Pelli's descriptions reflect a 'mess' of features, what he describes as 'excessive integration', whereby the visual system registers features clearly but does not make a clear assignment of features to individual items ([54], p. 1136).

Putting together the doubtfulness of conscious generic letterness with the evidence provided by Kouider *et al.* [27], the best view would be one in which the conscious fragments are sufficient to determine the differences among the letters of the alphabet for 3-4 items in all three rows, a total of about 10.5 specific shape representations in consciousness. That would explain the results in accord with what subjects say about their own experience while allowing for a minor illusion effect, as found by Kouider et al. [27]. And it would support rich phenomenology.

Theorizing about consciousness always depends in part on introspective judgments, but the reasoning I just gave in support of rich phenomenology makes use of aspects of these judgments that are accepted by both sides in the debate.

Three forms of visual short-term memory

A number of laboratories have shown that cues presented up to 12 seconds after an array has disappeared can enhance memory for the array [4,39,55–57]. This technique has been exploited most impressively by a group at the University of Amsterdam that has amassed evidence for a third form of memory, 'fragile visual short-term memory' (fragile VSTM) [8,9,36,37,60], in addition to iconic and working memory (see Figure 4a for a depiction of the basic experimental procedure employed by this group). Fragile VSTM appears to have a capacity intermediate between iconic and working memory: experiments along these lines have shown capacities for fragile VSTM of nearly double that of working memory using the paradigm of Figure 4a and even higher for the materials of Figure 4b.

Using the paradigm of Figure 4a, with initial arrays and cues as in 4b, subjects who receive the cue 10 msec after the initial array vanishes get 30 of 32 orientations right for the black-white display but only 20 right for the isoluminant red-gray display [8]. Since the isoluminant display is invisible to rods (which fire much longer than cones [58] but are color blind), it looks as if fragile VSTM may just be a weakened form of iconic memory. However, if the cue is presented 1 second or later after the first array, there is no difference in informational capacity between subjects who have seen the black-white display and those who have seen the red-gray display, demonstrating that fragile VSTM has no retinal component, unlike pure iconic memory. A second finding is that a flash of light – presented just before the cue – eliminates the difference between the black-white and red-gray displays when presented at 10 msec but has no effect when presented at 1 second or later. With cue delays up to 4 seconds, capacities for fragile VSTM are still well above the 4 items of working memory – 7 of 8 items at 1 second, 6 of 8 at 2.5 seconds and 5 of 8 at 4 seconds – while working memory remains steady at 4 items. Capacities for more complex stimuli are smaller [8,36], however, in a third finding that supports a difference between pure iconic memory and fragile VSTM, the proportionality between iconic and fragile VSTM remains about the same.

Importantly for the argument I will make below, Landman *et al.* [56] showed that subjects could make reports based on whether an item changed in either size or orientation, as well as size and orientation separately, which

Box 2. High-capacity mental imagery

A completely different paradigm [41,42] has provided further evidence in support of the overflow argument (Figure I). In these experiments, subjects were able to maintain a 12 dot image for up to 4-5 seconds with accuracies of about 90% as shown by the subjects' ability to superimpose the 12 dots on a 12-dot partial grid on the screen. This finding argues against sparse-consciousness theories, according to which one would expect much smaller capacities.

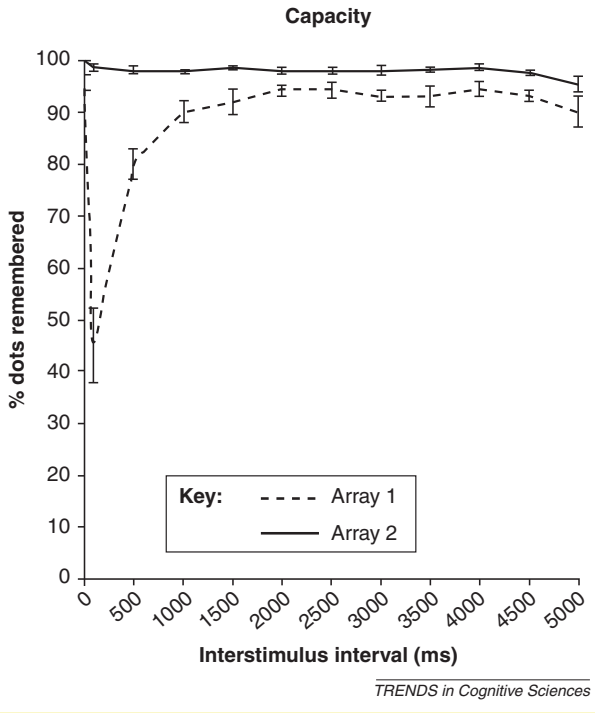

Figure I. Superimposing an imaged partial grid on a displayed partial grid. Brockmole *et al.* [41] used a procedure in which a 5 by 5 grid of dots with one missing dot is separated into two 12-dot partial grids. These partial grids, if superimposed, form a 5 by 5 grid with a missing dot. One such partial grid is presented on the screen briefly and subjects are encouraged to form an image of it. After a delay of 1-5 seconds, the second partial grid is presented and stays on the screen. The subjects' task is to identify the missing dot, something they can easily do if they can superimpose the imaged grid on the projected grid. The subjects' capacity for the imaged dots (percentage of dots remembered) can be computed by the type of errors made. The figure graphs this capacity measure against time. The subjects' retention of the initial grid drops rapidly immediately after the grid vanishes but after about 100 ms the percentage of dots remembered rises, reaching an asymptote at around 90% of the dots retained at a delay of about 1.5 seconds, a period that has been independently calculated by Kosslyn [81] to be the time it takes to generate a mental image. The dotted line in the figure represents the percentage of dots remembered from the first grid. (The solid line represents accuracy for the dots that are actually on the screen; of course the subjects very rarely say a dot in the display in front of them is missing.) The subjects report that they are generating an image and superimposing it on the partial grid on the screen, and their performance confirms their introspective judgments. This experiment has been replicated by Kosslyn [42] using memorized partial grids generated from commands. Reproduced, with permission, from [41].

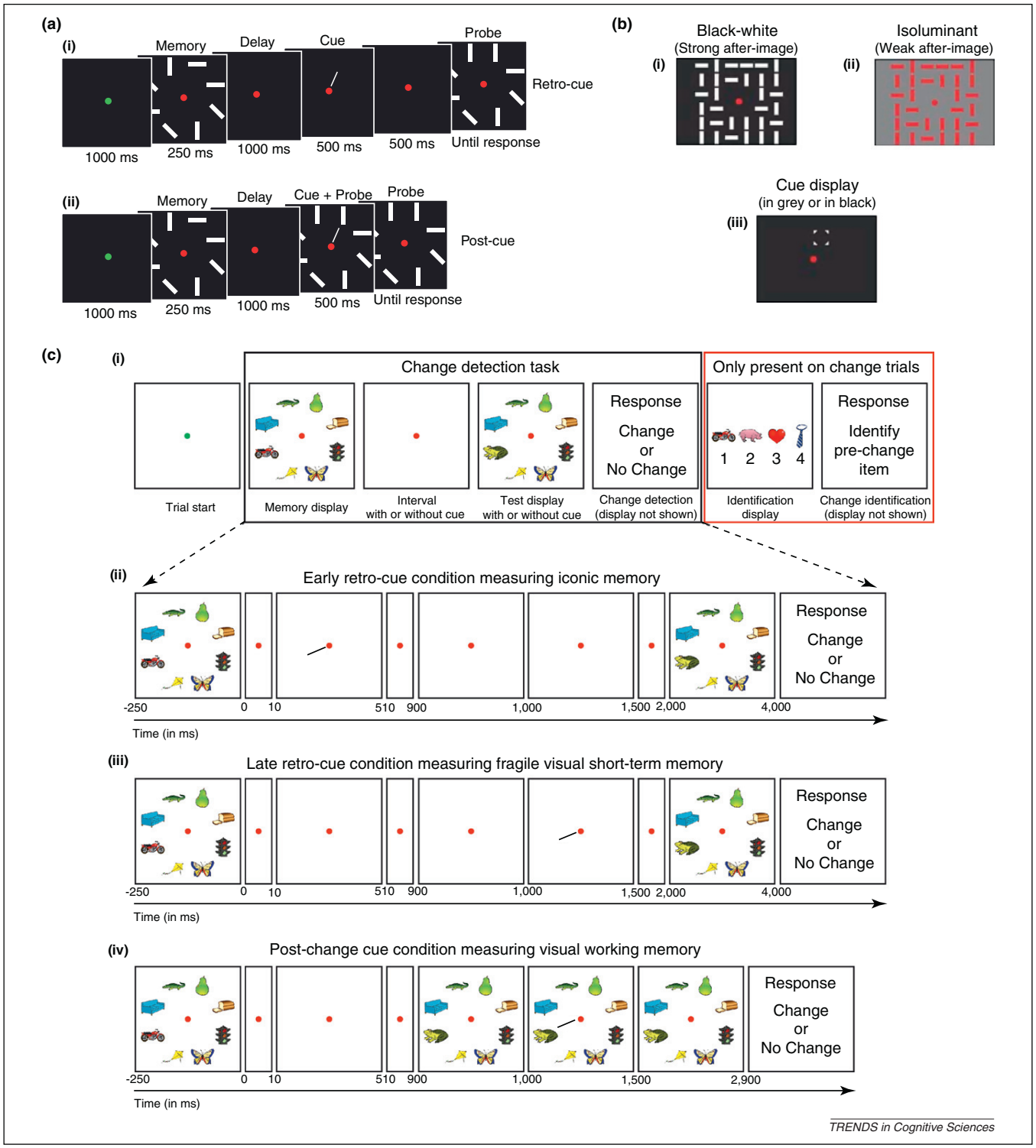

Figure 4. The Amsterdam group paradigms. **(a)** The basic Amsterdam paradigm [37]. An initial array is presented briefly after a fixation point (as depicted in the two leftmost panels). In this version, the initial array has 8 oriented rectangles and is presented for 250 msec. This is followed by a delay, then at the end of the sequence of both **(i)** and **(ii)** the probe array is presented, which is the same as the initial array, except one item may have changed orientation. At one or another point in this sequence a cue is presented that indicates to the subject that the indicated rectangle on the initial array is to be compared with the corresponding rectangle in the probe array. The subject makes a same/different judgment. The cue can come 10 ms after the memory array offset (not depicted in the figure), in which case it tests the capacity of iconic memory; or it can come at some point in the middle, as in **(i)**, usually between 1 and 4 seconds after the initial array offset, in which case it reflects the capacity of fragile visual short-term memory; or it can come with the probe array, as in **(ii)**, in which case it tests working memory. Reproduced, with permission, from [37]. **(b)** The prevention of retinal persistence paradigm [8]. The array on the left is black/white whereas the array on the right is isoluminant red/gray. The array on the left excites rods in the retina but the one on the right does not, since rods are color blind. Rods are mainly responsible for retinal persistence. The bottom item is the cue display used for these arrays instead of the line in **(a)**. Reproduced, with permission, from [8]. **(c)** High-resolution image paradigm [36]. The initial display consisted of 8 colored objects. As in the experiments described earlier, cues could be presented **(ii)** just 10 ms after the offset of the memory display (measuring pure iconic memory), **(iii)** 1 second after the offset (measuring fragile VSTM), or **(iv)** with the probe display (measuring working memory). Subjects were asked to report whether the indicated item had changed or not. In addition, on change trials, subjects were presented with the item from the memory display that had changed together with 3 other distractors that had not been involved in the displays (see rightmost set of panels in **(i)**). They were asked which of the items was the initial display item. In trials in which the subjects could not only detect the change but could

indicates that their representations included both dimensions, suggesting an image-like representation.

I mentioned that earlier work (summarized in [2,10]) showed that iconic memory is not based in early vision, but, in light of these findings, one sees that the so-called 'iconic memory' that was the subject of those Sperling experiments was an amalgam of two different kinds of memory: a rod-based 'pure iconic' memory lasting at most a few hundred msec and a much longer-lived fragile VSTM that is based higher up in the visual system and lasts up to 4-5 seconds. Indeed, Sligte *et al.* [9] found neuroimaging evidence of long lasting spatiotopic representations that correspond to fragile VSTM in visual area V4 but not in the lower areas V1, V2 or V3.

What is most relevant to the overflow argument is not that fragile VSTM differs from iconic memory but rather that it differs from working memory. As I mentioned, the capacities differ. Makovski and Jiang [38,57] have also found capacities comparable to the Amsterdam group's fragile VSTM that are greater than working memory though with a smaller advantage of fragile VSTM over working memory than the Amsterdam group: 30% for colors and 40% for odd shapes as compared with a minimum of 80% higher from the Amsterdam group. As Makovski *et al.* [38] point out, the difference is probably due to the fact that they used 9 colors and 10 odd shapes that are more confusable than the oriented lines used in most of the Amsterdam group's experiments. However, the similarities in these findings are more important to my case than the differences: both groups have found substantial capacity differences between working memory and fragile VSTM with cues presented 1 second or more after the initial display.

These capacity differences are also compatible with a view which assumes that fragile VSTM is just a weaker but richer form of working memory, attention being required to boost representations to a strength that would be manifested in standard working memory paradigms of the type mentioned in Box 1. A number of recent experiments, however, disconfirm the weaker/richer hypothesis.

The Amsterdam group used colored drawings in an experiment distinguishing between pure iconic memory, fragile VSTM and working memory, as before (see Figure 4c) [36]. All capacities were reduced significantly compared to the experiments with oriented rectangles described earlier. This experiment also involved a test of whether subjects remembered the original cued object from the memory array after the cue. In comparing memory for that original item, Sligte *et al.* found a nearly four-fold difference between fragile VSTM and working memory in high resolution representations, which suggests that fragile VSTM and working memory are at least partly based in distinct systems.

Vandenbroucke *et al.* [37] used three methods for decreasing attention to tasks of the sort I have been describing. The upshot was that all three methods of decreasing attention substantially decreased working memory scores, but only slightly decreased fragile VSTM. This differential effect of attention again suggests that fragile VSTM and working memory are partly based in distinct systems. Another experiment showed that a magnetic pulse delivered to right dorsolateral prefrontal cortex – an area known to play a primary role in working memory [59] – decreased working memory but not fragile VSTM [60].

In sum, there is evidence for a form of conscious memory that has a substantially higher capacity than working memory but, unlike iconic memory, is not based in the retina or early vision.

The generic illusion redux

Those anti-overflow theorists who have discussed the Amsterdam findings explain them by proposing generic representations of the initial array plus unconscious representations that are specific enough to do the task with the observed accuracy [29–32]. The cue is supposed to promote attentional amplification of the cued unconscious specific representation, which, when combined with the conscious generic representation, results in a conscious specific representation of the cued item. This account is depicted in Figure 2b. Phillips [31] supports this theory by noting that the forced-choice procedure of the Amsterdam group (in which the response required is to indicate whether or not the cued rectangle changed orientation between the initial and probe display) can reflect unconsciously driven guessing, as it does in blindsight. However, there is an important detail of one of the experiments described above that suggests otherwise. Recall that in the experiment depicted in 4c there was a four-fold difference between fragile VSTM and working memory in representations in which the subjects could choose the original cued object from 4 items, suggesting conscious representations of the original object playing a role. Could the choice of the original cued object be unconscious too? In principle, it could; however, we will not get anywhere in consciousness research by accepting the less plausible account.

There has been one direct test of the power of unconscious working memory (but see also [61]). Soto *et al.* [62] showed subjects a display that could either contain nothing or else a masked grid (a Gabor patch) tilted in one of six orientations, and asked subjects to judge whether it was tilted clockwise or counterclockwise relative to a highly visible second grid presented 2 or 5 seconds later. (The two grids always differed by 30°.) Subjects were then asked to rate the initial display for visibility of the grid. Those subjects who rated the initial display at the lowest level of visibility ('didn't see anything') could still compare it to the subsequent grid at a level above chance (slightly above 55%), even at 5 seconds later. The initial grid may have been seen with a mix of conscious and unconscious perception, as indicated by above chance sensitivity (d' = .297) to whether a grid was present or absent. (Subjects were more likely to give the lowest visibility score (.557) when there was no stimulus than when there was a stimulus (.441).) This task would be easy, compared to the Amsterdam tasks, if the initial grid was seen fully consciously, since there is only one item and so no need for a cue. Thus the

also say what the original item was, the representation was classified as high resolution. Detecting either the change or knowing the item that changed – but not both – was taken to indicate a low resolution representation. The number of high resolution representations in fragile VSTM was approximately 4 compared to slightly above 1 in working memory. Reproduced, with permission, from [36].

only direct evidence of the power of unconscious working memory suggests it is too weak to explain the memory of up to 5 oriented rectangles for up to 4-5 seconds in the Amsterdam paradigm.

One way in which the generic illusion account is more plausible for the Amsterdam group experiments than for the Sperling experiment is that the longer persistencies of the images in the Amsterdam group experiments do allow for the possibility that reports of reading the result off of the image could be due to attention moving from one rectangle to another, boosting them in turn from generic to specific. Variation of attention could in principle be tested by neuroimaging.

To sum up, the postulation of unconscious highly detailed iconic memory is unmotivated as noted in the discussion of the Sperling experiment and the evidence concerning unconscious perception suggests it is too weak to account for its results. In addition, Sligte *et al.*'s [9] evidence that persisting fragile VSTM is visible in V4 but not V1, V2 or V3 argues against highly detailed unconscious representations, since early visual areas would be the most obvious candidates for their locations. Finally, generic conscious representations of non-square rectangles that do not specify between horizontal and vertical orientations is difficult to accept. Although this issue is certainly still open, it is reasonable to tentatively accept the overflow interpretation and reject an interpretation of these experiments that appeals to the generic illusion.

A further type of experimental evidence for overflow is presented in Box 2.

Concluding remarks

There are two philosophical fallacies that may lead the anti-overflow forces astray. First, many critics of the overflow argument seem to think that a vote for overflow is a vote for inaccessible consciousness. For example, Cohen and Dennett [29] group the two views together as 'dissociative' theories that stand or fall together. However, as pointed out earlier, the fact that necessarily most items are not accessed does not entail inaccessibility of any items. A second mistake is to suppose that unconscious images are somehow the default view, something we have to be dislodged from by powerful evidence. As Phillips puts it, '[m]y claim is only this: for all that has been said, we have no reason to deny that performance in Landman and Sligte's change detection studies results simply from sub-personal, non-conscious informational persistence' ([31], p. 407). If instead we adopt the methodology of asking which hypothesis is better supported, I think we should prefer the overflow hypothesis.

Acknowledgements

I would like to thank the following colleagues for comments on an earlier draft: Tyler Burge, Richard Brown, David Chalmers, Susan Carey, Michael Cohen, Hakwan Lau, Ilja Sligte, David Rosenthal, James Stazicker, Annelinde Vandenbroucke and Frédérique de Vignemont.

References

1 Block, N. (1995) On a confusion about a function of consciousness. *Behav. Brain Sci.* 18, 227–247
2 Block, N. (2007) Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behav. Brain Sci.* 30, 481–548
3 Block, N. (2008) Consciousness and cognitive access. *Proc. Aristot. Soc.* 108, 289–317
4 Lamme, V. (2003) Why visual attention and awareness are different. *Trends Cogn. Sci.* 7, 12–18
5 Lamme, V. (2004) Separate neural definitions of visual consciousness and visual attention: a case for phenomenal awareness. *Neural Netw.* 17, 861–872
6 Lamme, V. (2006) Towards a true neural stance on consciousness. *Trends Cogn. Sci.* 10, 494–501
7 Sperling, G. (1960) The information available in brief visual presentations. *Psychol. Monogr.* 74, 1–29
8 Sligte, I.G. *et al.* (2008) Are there multiple visual *short-term* memory stores? *PLoS ONE* 3, e1699
9 Sligte, I.G. *et al.* (2009) V4 activity predicts the strength of visual short-term memory representations. *J. Neurosci.* 29, 7432–7438
10 Coltheart, M. (1980) Iconic memory and visible persistence. *Percept. Psychophys.* 27, 183–228
11 Tye, M. (2010) Attention, Seeing and Change Blindness. *Philos. Issues* 20
12 Rensink, R.A. *et al.* (1997) To see or not to see: the need for attention to perceive changes in scenes. *Psychol. Sci.* 8, 368–373
13 O'Regan, J.K. and Noe, A. (2001) A sensorimotor approach to vision and visual consciousness. *Behav. Brain Sci.* 24, 883–975
14 Simons, D. and Rensink, R. (2005) Change blindness: past, present and future. *Trends Cogn. Sci.* 9, 16–20
15 Noë, A. (2004) *Action in Perception*, MIT Press
16 Block, N. (2001) Paradox and cross purposes in recent work on consciousness. *Cognition* 79, 197–220
17 Block, N. (2007) *Functionalism, Consciousness and Representation*, MIT Press
18 Dretske, F. (2004) Change blindness. *Philos. Stud.* 120, 1–18
19 Dretske, F. (2007) What change blindness teaches about consciousness. *Philos. Perspect.* 21, 215–230
20 Gill, N.F. and Dallenbach, K.M. (1926) A preliminary study of the range of attention. *Am. J. Psychol.* 37, 247–256
21 Dallenbach, K.M. (1920) Attributive vs cognitive clearness. *J. Exp. Psychol.* 3, 183–230
22 James, W. (1890) *Principles of psychology*, Henry Holt
23 Sperling, G. (1983) Why we need iconic memory. *Behav. Brain Sci.* 6, 37–39
24 Thomas, N. (1999) Are theories of imagery theories of imagination? *Cogn. Sci.* 23, 207–245
25 Dehaene, S. *et al.* (2006) Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends Cogn. Sci.* 10, 204–211
26 O'Regan, J.K. (2011) *Why red doesn't sound like a bell: understanding the feel of consciousness*, Oxford University Press
27 de Gardelle, V. *et al.* (2009) Perceptual illusions in brief visual presentations. *Conscious. Cogn.* 18, 569–577
28 Kouider, S. *et al.* (2010) How rich is consciousness? The partial awareness hypothesis. *Trends Cogn. Sci.* 14, 301–307
29 Cohen, M.A. and Dennett, D. (2011) Consciousness cannot be separated from function. *Trends Cogn. Sci.* 15, 358–364
30 Stazicker, J. (2011) Attention, visual consciousness and indeterminacy. *Mind Lang.* 26, 156–184
31 Phillips, I.B. (2011) Perception and iconic memory: what Sperling doesn't show. *Mind Lang.* 26, 381–411
32 Phillips, I.B. (2011) Attention and iconic memory. In *Attention: philosophical and psychological essays* (Mole, C. *et al.*, eds), pp. 204–227, Oxford University Press
33 Lau, H. and Rosenthal, D. (2011) Empirical support for higher-order theories of conscious awareness. *Trends Cogn. Sci.* 15, 365–373
34 Brown, R. (2011) The myth of phenomenological overflow. *Conscious. Cogn.* DOI: 10.1016/j.concog.2011.06.005
35 Rahnev, D. *et al.* (2011) Attention induces conservative subjective biases in visual perception. *Nat. Neurosci.* DOI: 10.1038/nn.2948
36 Sligte, I.G. *et al.* (2010) Detailed sensory memory, sloppy working memory. *Front. Psychol.* 1, 1–10
37 Vandenbroucke, A.R.E. *et al.* (2011) Manipulations of attention dissociate fragile visual *short-term* memory from visual working memory. *Neuropsychologia* 49, 1559–1568
38 Makovski, T. *et al.* (2008) Orienting attention in visual working memory reduces interference from memory probes. *J. Exp. Psychol.: Learn. Mem. Cogn.* 34, 369–380

39 Lepsien, J. *et al.* (2005) Directing spatial attention in mental representations: interactions between attentional orienting and working memory load. *Neuroimage* 26, 733–743
40 Lepsien, J. and Nobre, A.C. (2007) Attentional modulation of object representations in working memory. *Cereb. Cortex* 17, 2072–2083
41 Brockmole, J.R. *et al.* (2002) Temporal integration between visual images and visual percepts. *J. Exp. Psychol.: Hum. Percept. Perform.* 28, 315–334
42 Lewis, K. *et al.* (2011) Integrating images and percepts: new evidence for depictive representation. *Psychol. Res.* 75, 259–271
43 Naccache, L. and Dehaene, S. (2007) Reportability and illusions of phenomenality in the light of the global neuronal workspace model. *Behav. Brain Sci.* 30, 518–519
44 Grush, R. (2007) A plug for generic phenomenology. *Behav. Brain Sci.* 30, 504–505
45 Levine, J. (2007) Two kinds of access. *Behav. Brain Sci.* 30, 514–515
46 Papineau, D. (2007) Reuniting (scene) phenomenology with (scene) access. *Behav. Brain Sci.* 30, 521
47 Sergent, C. and Rees, G. (2007) Conscious access overflows overt report. *Behav. Brain Sci.* 30, 523–524
48 Burge, T. (2007) Psychology supports independence of phenomenal consciousness. *Behav. Brain Sci.* 30, 500–501
49 Byrne, A. *et al.* (2007) Do we see more than we can access? *Behav. Brain Sci.* 30, 501–502
50 Kouider, S. *et al.* (2007) Partial awareness and the illusion of phenomenal consciousness. *Behav. Brain Sci.* 30, 510–511
51 Alvarez, G. and Oliva, A. (2008) The representation of simple ensemble visual features outside the focus of attentiion. *Psychol. Sci.* 19, 392–398
52 Block, N. (2007) Overflow, access and attention. *Behav. Brain Sci.* 30, 530–542
53 Baars, B. (1988) *A Cognitive theory of consciousness*, Cambridge University Press
54 Pelli, D. *et al.* (2004) Crowding is unlike ordinary masking: distinguishing feature integration from detection. *J. Vis.* 4, 1136–1169
55 Griffin, I.C. and Nobre, A.C. (2003) Orienting attention to locations in internal representations. *J. Cogn. Neurosci.* 15, 1176–1194
56 Landman, R. *et al.* (2003) Large capacity storage of integrated objects before change blindness. *Vis. Res.* 43, 149–164
57 Makovski, T. and Jiang, Y.V. (2007) Distributing versus focusing attention in visual short-term memory. *Psychon. Bull. Rev.* 14, 1072–1078
58 Adelson, E.H. (1978) Iconic storage: the role of rods. *Science* 201, 544–546
59 Curtis, C. and D'Esposito, M. (2003) Persistent activity in the prefrontal cortex during working memory. *Trends Cogn. Sci.* 7, 415–423
60 Sligte, I.G. *et al.* (2011) Magnetic stimulation of the dorsolateral prefrontal cortex dissociates fragile visual short tem memory from visual working memory. *Neuropsychologia* 49, 1578–1588
61 Hassin, R. *et al.* (2009) Implicit working memory. *Conscious. Cogn.* 18, 665–678
62 Soto, D. *et al.* (2011) Working memory without consciousness. *Curr. Biol.* 21
63 Cowan, N. *et al.* (2007) The legend of the magical number seven. In *Tall Tales about the brain: things we think we know about the mind, but ain't so* (Della Sala, S., ed.), pp. 45–59, Oxford University Press
64 Luck, S.J. and Vogel, E.K. (1997) The capacity of visual working memory for features and conjunctions. *Nature* 390, 279–281
65 Hauser, M. *et al.* (2000) Spontaneous number representation in semi-free ranging rhesus monkeys. *Proc. R. Soc. London: Biol. Sci.* 267, 829–833
66 Wood, J. *et al.* (2008) Free-ranging rhesus monkeys spontaneously individuate and enumerate small numbers of non-solid portions. *Cognition* 106, 207–221
67 Feigenson, L. *et al.* (2002) The representations underlying infants' choice of more: Object files vs. analog magnitudes. *Psychol. Sci.* 13, 150–156
68 Feigenson, L. and Carey, S. (2003) Tracking individuals via object files: evidence from infants' manual search. *Dev. Sci.* 6, 568–584
69 Gross, H.J. *et al.* (2009) Number-based visual generalization in the honeybee. *PLoS ONE* 4, 1–9
70 Saults, J.S. and Cowan, N. (2007) A central capacity limit to the simultaneous storage of visual and auditory arrays in working memory. *J. Exp. Psychol.: Gen.* 136, 663–684
71 Luck, S.J. (2008) Visual short-term memory. In *Visual memory* (Luck, S.J. and Hollingworth, A., eds), pp. 43–85, Oxford University Press
72 Vogel, E.K. *et al.* (2001) Storage of features, conjunctions and objects in visual working memory. *J. Exp. Psychol.: Hum. Percept. Perform.* 27, 92–114
73 Zhang, W. and Luck, S.J. (2008) Discrete fixed-resolution representations in visual working memory. *Nature* 453, 233–237
74 Awh, E. *et al.* (2007) Visual working memory represents a fixed number of items regardless of complexity. *Psychol. Sci.* 12, 329–334
75 Wilken, P. and Ma, W.J. (2004) A detection theory account of change detection. *J. Vis.* 4, 1120–1135
76 Bays, P.M. *et al.* (2011) Storage and binding of object features in visual working memory. *Neuropsychologia* 49, 1622–1631
77 Alvarez, G. and Cavanagh, P. (2004) The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychol. Sci.* 15, 106–111
78 Zosh, J.M. and Feigenson, L. (2009) Beyond 'what' and 'how many': capacity, complexity and resolution of infants' object representations. In *The origins of object knowledge* (Santos, L. and Hood, B., eds), pp. 25–51, Oxford University Press
79 Brady, T.F. *et al.* (2011) A review of visual memory capacity: beyond individual items and toward structured representations. *J. Vis.* 11, 1–34
80 Wheeler, M. and Treisman, A. (2002) Binding in *short-term* visual memory. *J. Exp. Psychol.: Gen.* 131, 48–64
81 Kosslyn, S.M. *et al.* (2006) *The Case for Mental Imagery*, Oxford University Press