Remnants of Perception:
Comments on Block and the Function of Visual Working Memory

Jake Quilty-Dunn (Rutgers)

Word count: 3483 (text, incl. figure captions & footnotes) + 771 (references)

The Border Between Seeing and Thinking is an extraordinary achievement, the result of careful attention (and contribution) to both the science and philosophy of perception. The book offers some bold hypotheses. While the hypotheses themselves are worth the price of entry, Block's sustained defense of them grants the reader insight into countless fascinating experimental results and philosophical concepts. His unpretentious and accommodating exposition of the science—explaining rather than asserting, digging into specific results in detail rather than making summary judgments and demanding that readers take him at his word—is a model of how philosophers ought to engage with empirical evidence. It is simply not possible to read this book without learning something. It will surely play a foundational role in theoretical work on perception for many years to come.

1. The Perception–Cognition Border

At the center of the book's positive account of perception is Block's claim about the nature of visual representation, and how it differs from thought: the visual system represents the world the way an image does, and thought takes myriad forms, including forms more like language. More technically, vision is entirely *iconic* and cognition is paradigmatically (perhaps mostly?) *discursive*.[1] Block provides extensive elaboration and defense of this claim throughout the book (especially Chapters 4 through 8), so in what follows I will assume the basic idea and evidence base are well-understood.

For this sort of representational approach to the perception-cognition border, any *interface point* between perception and cognition provides a key test case. Does the way perception and cognition interact seem to suggest that they speak different languages and rely on some intermediating translation? Or do they use a common interlingua that allows for free transfer of information?

Of course, there are plainly significant limits on information transfer between perception and cognition. Your full belief that your friend is currently in Belgium doesn't prevent your visual system from perceiving their face on a street corner in Manhattan; your knowledge that footprints tend to be concave and that Figure 1 is lit from below doesn't (invariably, stably) prevent your visual system from using the "light-from-above prior" to see an unusual raised foot-shape; your belief that objects can

[1] For recent work on the discursive format of cognition, see Dehaene et al. 2022; Carcassi & Szymanik 2023; Kazanina & Poeppel forthcoming; Quilty-Dunn et al. forthcoming a, forthcoming b, and commentaries.

persist as separate individuals despite being connected by a dotted line doesn't allow you to track them efficiently as separate individuals (Scholl et al. 2001); and so on for many visual phenomena.

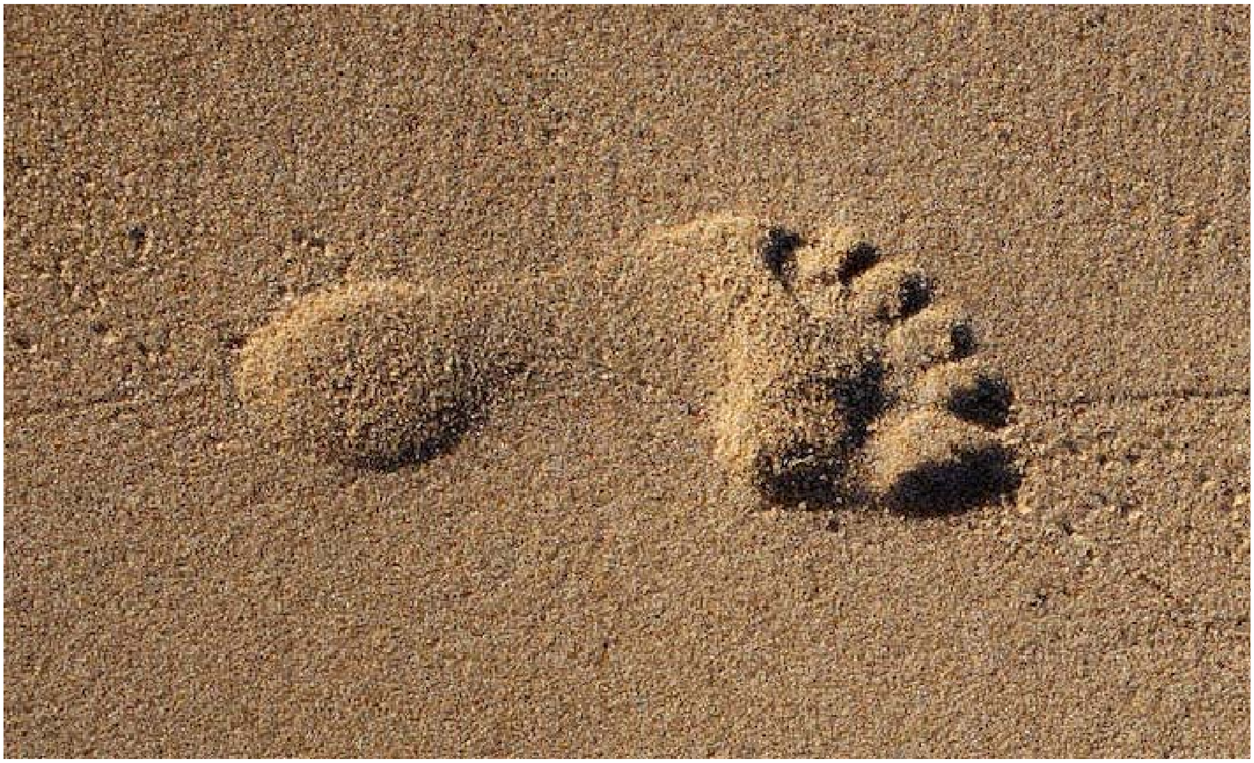

Figure 1. Concave footprint lit from below that might appear as a convex foot-shape lit from above; photo by Elviss Railijs Bitāns.

These failures of information transfer might stem from something other than a difference in representational format. They might instead stem from built-in limitations on information access. On modular approaches to perception, visual processes have access to their own stores of information, which might include the light-from-above prior but exclude cognitive information like your beliefs about illusions (Fodor 1983; Mandelbaum 2018; Quilty-Dunn 2020a). It's compatible with these approaches that vision represents in whatever format you like, including the very same format as propositional thought. In that case, it's not the fact that our beliefs are represented in a different format that prevents their free usage by perceptual processes—it's the fact that they're stored in the wrong place in memory. If we were able to move them to a memory location that a visual process had access to, then that visual process might be able to use them in the same way it uses the light-from-above prior.

Modular approaches are a prominent example of architectural approaches to the perception-cognition border. These approaches locate the distinction between seeing and thinking in architectural features of processing and place few restrictions on the form of perceptual representations. They include the view that perceptual processing is distinguished by its relationship to proximal stimulation (Beck 2018), by restriction in the features it computes over (Green 2020), and by other potential architectural features such as the use of special algorithms.

These views vary in the restrictions they place on information flow, with modular approaches typically positing the greatest restrictions. But they all allow (in principle) for pluralism about perceptual representation, which is arguably desirable given the many different computational demands of perception (Quilty-Dunn 2020b; Green 2023; Firestone & Phillips forthcoming). In particular, they allow that some perceptual processes might output discursive, conceptual representations. Thus the information flow from perception to cognition could proceed in those cases without translation. Representational approaches like Block's, on the other hand, must deny that the information transfer from perception to cognition works this way.

Promising test cases for the debate between these approaches are points at which perception hands information off to cognition. Our key question: does perception seem to hand information off in a way that suggests a common interlingua, or in a way that suggests a format discontinuity? We can ask more specific questions to probe this more general one, like: is the information transfer fast or slow? Are the same representations at the interface point usable for both cognitive and perceptual processes, or only ever one or the other?

2. Visual Working Memory

My test case will be arguably the central interface point between seeing and thinking: visual working memory (VWM). VWM is a functional memory location where information that bears some important relation to visual perception can be maintained for brief intervals (on the order of seconds) and manipulated. For example, if you look at your cat and wonder whether he can make the jump to the high bookshelf he's eyeing, your visual simulation of possible outcomes makes use of a visual representation of the scene that is stored in VWM. VWM is where focused thinking about the visual world begins.

I say VWM representations "bear some important relation" to visual representations because the question in the rest of this paper will be: what relation? One possibility is that VWM simply takes the outputs of visual processing—visual percepts—and stores them. Another possibility is that VWM cannot access visual percepts directly, and instead uses representations in another (perhaps discursive, cognition-friendly) format that are merely informed by visual information.

Consider the *function* of VWM. One view is that the function of VWM is largely for perception: it sustains information for use in perceptual computations that unfold over larger timescales. The subjective sense that we experience a unified visual scene across noticeable intervals of time—including movements of the eyes, body, and external objects that might unfold on the order of seconds rather than milliseconds—could be grounded in the functional interaction between VWM and online vision. Another view is that the function of VWM is largely for cognition: it transforms visual representations into discursive, conceptualized representations that can be broadcast to other cognitive systems.[2]

If VWM is primarily for perception, then we should expect it to sustain information in a format that perception uses. And if VWM is primarily for cognition, then we should expect it to represent in a conceptualized discursive format. Note that, if perception uses a discursive format, both these functions can be naturally accommodated: the discursive format is apt for cognitive use and, since it is native to perceptual processing, can feed back down into perception for slower perceptual computations.

One way into the function question, which I lack space to explore here, concerns nonhuman animals. If VWM is for perception, then it might evolve prior to the cognitive abilities humans happen to (re)use it for. The relation between perception and memory is ripe for philosophical analysis, and I hope more philosophers pursue it (see, e.g., Munton 2022; Green forthcoming).

The format of VWM and its relation to perception are discussed at length in *The Border Between Seeing and Thinking*. There is evidence that object-file representations, the representations we use to track objects and store information about their features, are discursive; thus, perception cannot be entirely iconic (Quilty-Dunn 2020b; Green & Quilty-Dunn 2021). Block argues, however, that object files are creatures of VWM and don't reflect the format of perception. For Block, VWM representations like object files are "conceptualized versions" (p.256) of perceptual representations (though see the final section below). VWM itself is a "cognitive scratch pad" (p.250). While he does not explicitly characterize the function of VWM, I read him as taking VWM to be primarily for cognition. *Perceptual* object representations, which Block sharply distinguishes from object files, are studied through characteristic effects such as apparent motion (discussed below) instead of effects that are generally agreed to tap into VWM, such as the object-specific preview benefit. We can thus use our questions about the structure and function of VWM as an inroad into debates about the format of perception itself.

3. Evidence

Here is a simple initial question: how *fast* is the consolidation of information from perception into VWM? If there is a format discontinuity between the two systems, then there will be a translation

[2] A related disagreement concerns whether VWM representations have a modality-specific format that relies on activation in visual areas (Harrison & Tong 2009; Carruthers 2015) or an amodal format that relies on frontal areas (Xu 2017). This debate somewhat crosscuts my concerns here, since I want to argue that visual perception itself generates discursive representations and these are held in VWM without alteration in their format—I'll leave neural implementation questions unaddressed here.

process that (*ceteris paribus*) will take extra time. So the quicker the consolidation, the stronger our credence should be that VWM represents in the same format as perception. Furthermore, Block notes (p.153) that the imagery literature suggests that transitions between discursive and iconic formats (e.g., going from the discursive symbol DOG to forming a mental image of a dog) are slow, around 1.5 seconds. We thus have a rare achievement in philosophy on our hands: an eminently testable question.

Fortunately, there is clear evidence. Vogel et al. (2006) measured the time-course of VWM consolidation using *backward masks*—visual noise presented shortly after a stimulus to disrupt visual processing (Fig. 2). For example: if (i) a stimulus is presented, (ii) the mask appears 584ms later, and (iii) participants then fail to detect a color change in the item, then that is evidence they failed to get color information from perception to VWM within 584ms. Vogel et al. used different stimulus onset asynchronies (SOA) to see how much information was consolidated into VWM at different times. They then used change detection performance to estimate VWM capacity (represented with a K) at different SOAs.

They found that, in addition to the time it takes to form a perceptual representation of a colored square (note the y-intercept in Fig. 3), the SOA needed to increase by only 50ms per item for VWM storage. In other words, the time-course of consolidating one perceived item into VWM happens in 50ms. Furthermore, in an earlier paper using a different paradigm, Gegenfurtner and Sperling (1993) found that, when participants are spatially attending to a particular location in the display, items at that location can be consolidated in half that time (20-30ms). These durations are stunningly fast: an order of magnitude faster than a blink of an eye (100-400ms) and nearly two orders of magnitude faster than the cross-format translation processes Block cites (1000-1500ms). Thus the answer to our first question is that VWM consolidation is *fast*, likely too fast to involve cross-format translation.

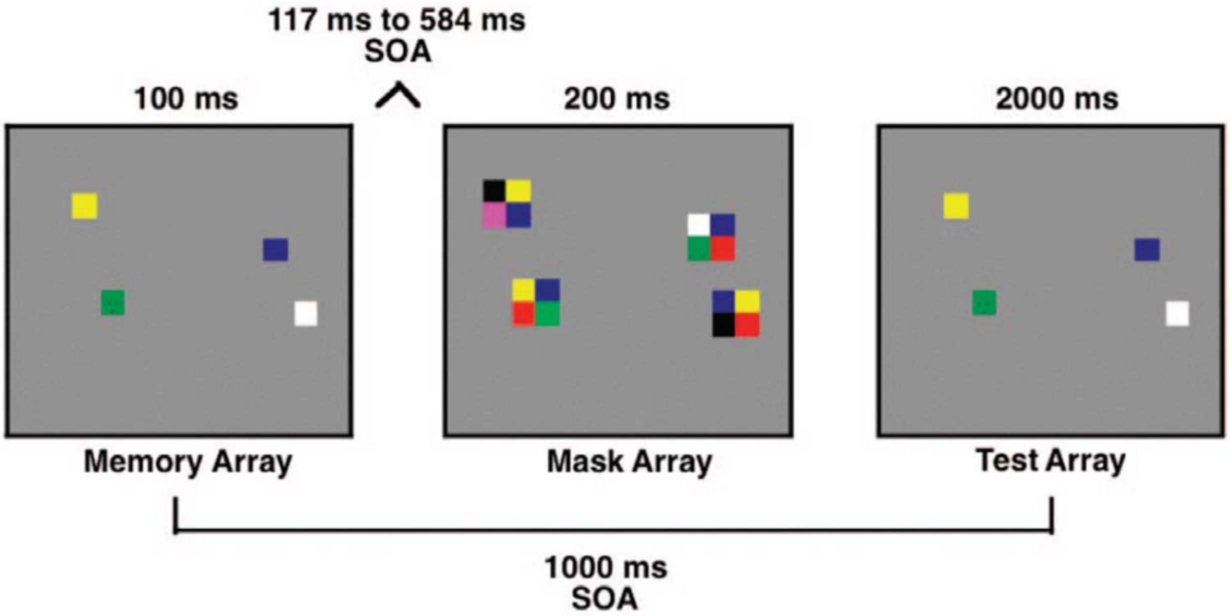

Figure 2—VWM task; Vogel et al. 2006.

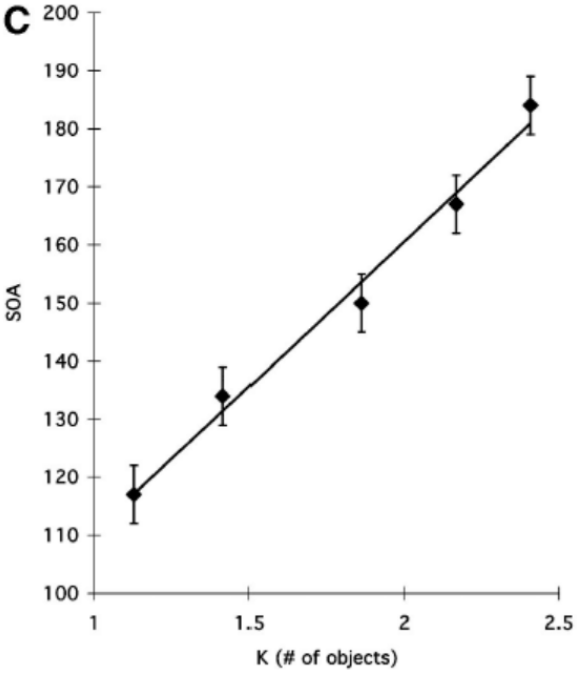

Figure 3—VWM capacity as a function of target-mask SOA; Vogel et al. 2006.

If there really is format continuity between VWM and vision, then we should not only see fast consolidation. We should also see fluid interaction between the two systems, making shared use of a common representational format. We therefore have another tractable question: does vision access VWM contents in its own computations, as if they share the same kinds of representations? Or does vision simply send information "upwards" to VWM and then lose access to that information? What we're looking for here are unambiguously perceptual effects that show direct sensitivity to VWM contents.

There is, perhaps unsurprisingly, a large body of experimental work studying the interaction between vision and VWM. For example, VWM contents that match visual perception (e.g., have the same orientation) enhance the precision of visual experience (Salahub & Emrich 2016). However, we want to carefully pull apart visual processing proper from manipulation of visual representations in VWM itself. The processes that underlie precision of feature report, for example, could use VWM resources. And according to Block, visual experience is often conflated with visual cognition, including on cognition-based theories of consciousness (e.g., Global Workspace Theory—Baars 1993; Dehaene 2014) that take VWM to bear a special relationship to consciousness.[3]

Fortunately, Block details genuinely perceptual effects (Ch.2), such as *popout*. In a popout effect, a salient feature captures attention in a way that is barely diminished at all by an increase in the number of distractor items (Fig. 4). Hyun et al. (2009) had participants hold a display in VWM and then detect a salient change in color or orientation; they found an increase in amplitude of the N2pc event-related potential that remained constant as set size increased, just as in simultaneous popout displays (Luck & Hillyard 1994). Another clearly perceptual effect (not on Block's list, but in the spirit of it) is *motion repulsion*: when you view two streams of dots moving in two different directions, the motions "repel" each other, and you see the angle separating them as larger than it actually is. Amazingly, holding one stream in VWM for two seconds is sufficient to cause participants to perceive the second stream as "repelled" away from the memorized stream (Kang et al. 2011).

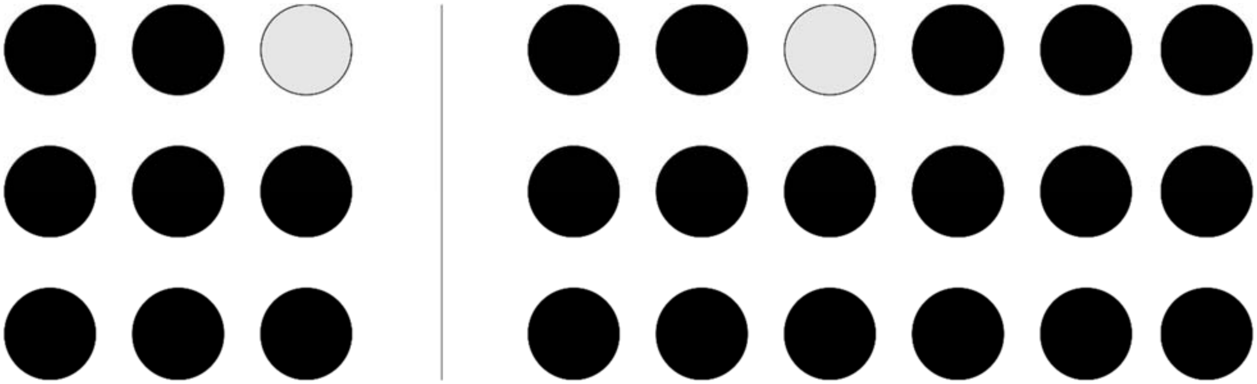

Figure 4—Popout: the light circle draws attention in both displays, despite the right display having twice as many items as the left.

VWM might often affect visual perception by guiding visual attention (e.g., in the popout case), which is different from direct computational access (Quilty-Dunn 2020a). But attentional shifts

[3] As Block puts it, "the global workspace model [is] a much better model of conceptualization of perception than of perceptual consciousness" (p.425).

are not the only source of these effects. Scocchia et al. (2013) found that perception of an ambiguous motion display is biased in the direction of a motion display held in VWM, but not in the direction of a recently attended motion display. Mendoza et al. (2011) had participants indicate whether motion of dot displays matched a sample; the sample was either concurrently presented (attention) or shown seconds earlier (VWM); finally, while participants performed that task, another dot display "pulsed" once with motion and participants had to indicate the direction of motion. When the motion pulse matched the attended/memorized sample, participants were better at detecting its direction. They were even better when the sample was *both* memorized and attended than when it was merely attended, showing that the effect of VWM is not due to attention alone. These vision-VWM interactions are fluid and bidirectional (see Teng & Kravitz 2019, which I lack space to discuss).

Recall that, for Block, object representations in perception should be sharply distinguished from VWM representations. We can therefore ask whether VWM contents drive object-based effects that Block agrees probe *perceptual* object representations. One such effect Block appeals to is apparent motion, the visual impression of an object moving from one location to another created simply by seeing two objects appear one after another in different locations. Hein et al. (2021) used a Ternus display, in which the apparent motion is ambiguous between three objects moving together (group motion) or two objects staying in place and a third object "leapfrogging" from one end to the other (element motion). They colored the objects such that one color was consistent with group motion (e.g., the green object is in the middle in both displays) and another was consistent with element motion (e.g., the pink object is on the left in the first display and on the right in the second). They then had subjects hold a color in VWM for an unrelated memory task. The VWM color drove apparent motion: if they memorized green and the green Ternus item suggested group(/element) motion, then they saw group(/element) motion. Thus VWM representations drive the very perceptual effects Block uses to distinguish perceptual object representations from VWM object representations.

Taking stock: VWM representations are consolidated from perception in tens of milliseconds and enter directly into perceptual computations. The evidence supports the hypothesis that the function of VWM is to sustain the outputs of perception without transforming their format, rendering them available for perceptual computations that take place over longer intervals of time than earlier sensory memory stores can robustly manage.

4. Hybrids in VWM?

As mentioned above, Block sometimes describes VWM representations as "conceptualized versions" (p.256) of perceptual representations. However, he also suggests that VWM representations are "perceptual" (p.113) and contain "perceptual materials" (p.258) or "remnants of perception" (p.260) enclosed in a "cognitive envelope" (p.249). In that case, perhaps his view is that some of the iconic outputs of perception are simply held as such in VWM (these are the perceptual materials), but are accompanied by discursive representations (this is the cognitive envelope). Many tractable questions are raised here, such as: are the iconic elements of VWM representations limited to a certain range of properties, e.g., shape and color?[4] Is there redundant representation of properties, such that some are encoded both iconically and discursively? How do the formats interact? Does memory consolidation work differently for both sorts of properties?

Given the speed of VWM consolidation, Block could say (1) that consolidation actually has two phases: first, iconic outputs of perception are transferred to VWM in tens of milliseconds; second, a slower process adds discursive constituents. He could also argue (2) that only the iconic elements of VWM enter into perceptual computations, with the discursive elements playing a purely cognition-facing role. I know of no evidence that directly probes (1). However, it is problematic to suppose that the abstract, categorical properties represented in object files take significantly longer to form than other properties. Mandelbaum (2018) argues that visual categorization happens in tens of milliseconds, citing studies like Potter et al. 2014, in which images of scenes are shown one after the other for 13ms each and therefore masking each other. Block (pp.329-330) argues that these images are not effective masks, citing Maguire & Howe's (2016) follow-up study showing that lower-level stimuli (lines/edges) are better masks at these presentation times and eliminate evidence of rapid categorization. However, our question here is whether encoding conceptual categories is significantly slower than encoding low-level features. In a follow-up to the follow-up, Howe (2017) found that the minimal presentation time for categorization was about the same as that for detecting color and orientation (~35ms). It's implausible that consolidation of discursive representations into VWM is significantly slower than consolidating iconic elements.

What about (2), the prediction that vision only accesses iconic elements of VWM representations? The best evidence to the contrary concerns transsaccadic memory, i.e., the visual system's ability to maintain a coherent percept across eye movements without having to restart visual perception from scratch with each new fixation. As I've argued elsewhere, abstract categorical information in object files is preserved in transsaccadic memory (Quilty-Dunn 2020b, Section 5). Furthermore, the use of low-level properties like color in transaccadic memory is governed by these abstract categorical properties: e.g., color is used to track object identity across eye movements if the object category has a diagnostic color (*banana*) but not if it doesn't (*bucket*) (Gordon & Vollmer 2010).

Block argues that transsaccadic memory is essentially just VWM. But its significance for perception shouldn't be dismissed on these grounds. Visual perception—not just cognition, but visual perception itself—needs coherence across eye movements. And as we've seen, vision seems to access VWM contents regularly. So why deny that object files in transsaccadic memory/VWM play this foundational role in securing a coherent visual percept in spite of our constantly moving eyes? Block points to the possible role of long-term memory in transsaccadic memory, but this might simply be

[4] Block does argue for "deep differences between perception and the perceptual materials used in working memory" (p.16) even for color.

visual long-term memory, which is high-fidelity and arguably distinct from the "long-term memory" where beliefs and other cognitive states are held (Brady et al. 2008).

The following view appears increasingly plausible: our visual systems construct representations in various formats, including discursive object files; some of these (including object files) are held in VWM to be re-used in visual processing, including transsaccadic processes, and also to subserve cognition. If this view were true, many of the insights in Block's book about iconic format in perception would remain unaffected. The only cost would be the claim that there are no other formats in perception.

References

Baars, B.J. (1993). *A cognitive theory of consciousness*. Cambridge: Cambridge University Press.

Beck, J. (2018). Marking the perception–cognition boundary: The criterion of stimulus-dependence. *Australasian Journal of Philosophy*, *96*(2), 319–334.

Brady, T.F., Konkle, T., Alvarez, G.A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329.

Carruthers, P. (2015). *The centered mind: What the science of working memory shows us about the nature of human thought*. Oxford: OUP.

Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Penguin.

Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs: a hypothesis about human singularity. *Trends in Cognitive Sciences, 26*(9), 751–766.

Fodor, J.A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Gegenfurtner, K.R., & Sperling, G. (1993). Information transfer in iconic memory experiments. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(4), 845–866.

Gordon, R.D., & Vollmer, S.D. (2010). Episodic representation of diagnostic and nondiagnostic object colour. *Visual Cognition*, *18*(5), 728–750.

Green, E.J. (2020). The perception-cognition border: A case for architectural division. *Philosophical Review*, *129*(3), 323–393.

Green, E.J. (forthcoming). Can we perceive the past? In S. Aronowitz & L. Nadel (eds.), *Space, Time, and Memory*. Oxford: OUP.

Green, E. J. (2023). "The Perception-Cognition Border: Architecture or Format?" In B. P. McLaughlin & J. Cohen (eds.), Contemporary Debates in Philosophy of Mind. Oxford: Blackwell.

Green, E.J., & Quilty-Dunn, J. (2021). What is an object file?. *The British Journal for the Philosophy of Science, 72*(3).

Harrison, S.A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635.

Hein, E., Stepper, M.Y., Hollingworth, A., & Moore, C.M. (2021). Visual working memory content influences correspondence processes. *Journal of Experimental Psychology: Human Perception and Performance*, *47*(3), 331–343.

Howe, P.D. (2017). Natural scenes can be identified as rapidly as individual features. *Attention, Perception, & Psychophysics*, *79*, 1674–1681.

Hyun, J.S., Woodman, G.F., Vogel, E.K., Hollingworth, A., & Luck, S.J. (2009). The comparison of visual working memory representations with perceptual inputs. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(4), 1140–1160.

Kang, M.S., Hong, S.W., Blake, R., & Woodman, G.F. (2011). Visual working memory contaminates perception. *Psychonomic Bulletin & Review*, *18*, 860–869.

Kazanina, N., & Poeppel, D. (forthcoming). The neural ingredients for a language of thought are available. *Trends in Cognitive Sciences*.

Luck, S.J., & Hillyard, S.A. (1994). Electrophysiological correlates of feature analysis during visual search. *Psychophysiology, 31(3)*, 291–308.

Maguire, J. F., & Howe, P. D. (2016). Failure to detect meaning in RSVP at 27 ms per picture. *Attention, Perception, & Psychophysics*, *78*, 1405–1413.

Mandelbaum, E. (2018). Seeing and conceptualizing: Modularity and the shallow contents of perception. *Philosophy and Phenomenological Research, 97*(2): 267–283.

Mendoza, D., Schneiderman, M., Kaul, C., & Martinez-Trujillo, J. (2011). Combined effects of feature-based working memory and feature-based attention on the perception of visual motion direction. *Journal of Vision*, *11*(1), 1–15.

Munton, J. (2022). How to see invisible objects. *Noûs*, *56*(2), 343–365.

Phillips, I.B., & Firestone, C. (forthcoming). Visual adaptation and the purpose of perception. *Analysis*.

Potter, M. C., Wyble, B., Hagmann, C.E., & McCourt, E.S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics*, *76*, 270–279.

Quilty-Dunn, J. (2020a). Attention and encapsulation. *Mind & Language*, *35*(3), 335-349.

Quilty-Dunn, J. (2020b). Perceptual pluralism. *Noûs* 54(4), 807–838.

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (forthcoming a). The best game in town: The reemergence of the language of thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*.

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (forthcoming b). The language of thought hypothesis as a working hypothesis in cognitive science. *Behavioral and Brain Sciences.*

Salahub, C.M., & Emrich, S.M. (2016). Tuning perception: Visual working memory biases the quality of visual awareness. *Psychonomic Bulletin & Review*, *23*, 1854–1859.

Scholl, B.J., Pylyshyn, Z.W., & Feldman, J. (2001). What is a visual object? Evidence from target merging in multiple object tracking. *Cognition* 80, 159–177.

Scocchia, L., Valsecchi, M., Gegenfurtner, K.R., & Triesch, J. (2013). Visual working memory contents bias ambiguous structure from motion perception. *PloS one*, *8*(3), e59217.

Teng, C., & Kravitz, D.J. (2019). Visual working memory directly alters perception. *Nature human behaviour*, *3*(8), 827–836.

Vogel, E.K., Woodman, G.F., & Luck, S.J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(6), 1436–1451.

Xu, Y. (2017). Reevaluating the sensory account of visual working memory storage. *Trends in Cognitive Sciences*, *21*(10), 794–815.