policy) – often by assigning a points value to indicate each cue's weight. Conceptualising self-insight research as "double model recovery" emphasises some limitations in previous work that bear on N&S's *relevance* and *sensitivity* criteria for the adequate assessment of awareness (target article, Table 1), and points to avenues for developing better methods for assessing self-insight.

First, this double-model-recovery framework highlights a critical question: why – as seems standard – instantiate the implicit policy (from statistical model recovery) as the "correct" model, and therefore assume that any discrepancy between implicit and explicit policies represents the judge's failure to recover the "true" model? Model-recovery exercises in cognitive science usually consider multiple "families" of candidate model. Typically, lens model research considers a single family of models: compensatory linear rules that integrate a fixed number of cues – though it does consider different family members, which differ according to the number of cues used. Alternative families of non-linear (configural) or non-compensatory judgment models are less frequently considered in the multiple-cue judgment literature, even though several alternatives can be modelled, such as judgments made according to the similarity of each case to a prototype, judgments made following a non-exhaustive lexicographic search through cues, and judgments where cues are selected probabilistically and therefore different cues are used for different cases. In contrast, some research on multi-attribute *choice* does consider different families of models: for instance, comparing alternative models reflecting whether a compensatory or lexicographic decision rule is being applied (e.g., Bröder 2003). Additionally, this research on recovering choice processes highlights that different models reflecting quite distinct processes often fit the data similarly well. Therefore, even when a compensatory linear model fits the data, the judge may nonetheless have followed a quite different process in making his or her judgments. In such cases, any elicitation procedure that presupposes the compensatory linear combination of a fixed number of cues fails N&S's *relevance* criterion because the behaviours being probed are not those that drove the judgment. This is liable to generate a poor match between the implicit and explicit policies. Thus, by following a restricted approach to modelling the judge to dictate the constraints of that judge's self-description, we create an insensitive assessment of awareness and may misattribute poor modelling as poor self-insight.

Second, a double-model-recovery framework emphasises the potential for mis-recovery of the original judgment process by *either* recovery technique (statistical or human). As many others have done, I have pitted human judges against statistical rules in multiple-cue judgments and – as is typical – have found that "statistical judges" outperform their human competitors (Dawes et al. 1989). However, in one investigation (see Rakow et al. 2003), our statistical judge showed the same apparent lack of self-insight as its human counterparts. A seven-cue predictive model derived using logistic regression generated predicted probabilities (that an applicant would be offered a place at a given university) for a series of cases, each defined by multiple cues. Human judges also provided the same set of judgments. Using the same linear regression analysis applied to the human participants, the implicit policy for the statistical judge declared only five cues to be used reliably (i.e., significant). Thus the statistical judge showed the typical pattern of limited self-insight that human judges display, apparently overestimating the number of cues that it used! Thus, just as assessments of awareness may fail N&S's *sensitivity* criterion, so too, insufficiently sensitive model recovery via linear regression could contribute to an inappropriate conclusion of "limited self-insight" (for a technical discussion of this problem, see Beckstead 2007).

Third, we can consider strategies for assisting human judges in recovering (describing) their judgment policies, which may, also, influence the candidate models for the statistical element of the double recovery exercise. In a recent study, we asked mental health practitioners to self-identify with descriptions of alternative families of judgment models – (non-)compensatory and (non-)exhaustive – which drew on analogies to common decision aids such as "balance sheets" and "trouble-shooting guides." Many of our assessors selected those options that implied contingent information search or non-compensatory information integration in their own (triage) judgments. Thus, if required to describe themselves in terms of a compensatory model always using a fixed number of cues (as per most self-insight research – though arguably failing the *relevance* criterion), inevitably some participants were forced to misrepresent their policy. It would therefore be unsurprising if judges displayed "poor self-insight." Much work has been done on alternative strategies for eliciting the subjective weights for compensatory linear judgment policies (e.g., Cook & Stewart 1975). However, we need improved (i.e., *relevant* and *sensitive*) elicitation methods that allow for a wider range of information search and integration processes to be identified when judges are asked to describe their judgment policies. Fair assessment of a judge's self-insight requires that *both* the statistical exercise of deriving the implicit judgment policy *and* the elicitation exercise whereby the judge describes his or her own judgments allow – as far as possible – recovery of the processes by which the original judgments were made.

What we (don't) know about what we know

doi:10.1017/S0140525X13000836

Shlomi Sher[a] and Piotr Winkielman[b,c]
[a]*Department of Psychology, Pomona College, Claremont, CA 91711;*
[b]*Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109;* [c]*University of Social Sciences and Humanities, 03-815 Warsaw, Poland.*
Shlomi.Sher@pomona.edu pwinkielman@ucsd.edu
http://psy2.ucsd.edu/~pwinkiel/

Abstract: The hypothesis of unconscious influences on complex behavior is observationally equivalent to the dissociability of cognition and metacognition (reportability). The target article convincingly argues that evidence for unconscious influence is limited by the quality of the metacognitive measure used. However, it understates the empirical evidence for unconscious influences and overlooks considerations of cognitive architecture that make cognitive/metacognitive dissociations likely.

In their target article, Newell & Shanks (N&S) identify methodological problems in the study of unconscious influences on decision making. Because awareness is indexed by subjects' reports about what they know and how they know it, such studies seek, in effect, to demonstrate dissociations of cognition and metacognition: One component of the design shows that information of some kind is influencing behavior in some way; a second component elicits subjects' reports about the information they possess and the manner in which they are using it. Evidence for unconscious influence is obtained when the (relatively indirect) cognitive measures and the (relatively direct) metacognitive measures paint inconsistent pictures of the underlying mental process.

As the authors note, such studies are only as compelling as the metacognitive measures they use – and measures lacking in reliability, relevance, immediacy, and sensitivity are often employed. Indeed, research on unconscious influence suffers from its own distinctive array of perverse incentives. As in other areas of psychology, the researcher is driven to obtain evidence for a clear effect on the cognitive measure. This incentive fosters practices that make Type I errors more likely (Simmons et al. 2011). In the study of unconscious influences, the researcher typically has a second incentive – to *fail* to find evidence of a significant effect on the metacognitive measure. This incentive may foster practices that make *Type II* errors more likely in the assessment of awareness. One effective way of failing to detect awareness is the use of unreliable, irrelevant, insensitive, and/or belated probes. (Ironically, such bias in the choice of probes might itself be unconscious.)

Complicating matters, it is not always easy to formulate, let alone satisfy, the central criterion of relevance – that "assessments should target only information relevant to the behavior" (target article, Table 1). What is relevant to a behavior depends on what causes it. As a result, subtly flawed or imprecise causal theories of behavior can lead even the well-meaning and careful researcher to misidentify relevance in designing a measure of awareness.

Although these methodological problems are important, and although they challenge some influential findings, we believe that the target article understates the full empirical and theoretical case for unconscious influences on complex behavior. To make the empirical case adequately would require a counter-review rather than a commentary. Recent reviews of unconscious cognition that are more comprehensive and, in our view, also more balanced are provided by Kouider and Dehaene (2007) and by Simons et al. (2007). The empirical study of unconscious cognition has its share of murky bathwater, but we believe the outlines of a baby are distinctly discernible within it.

Critically, the target article also understates the *theoretical* case for unconscious influences in complex behavior. N&S suggest that such influences make for "good stories," and that they confirm "strong ex ante beliefs" about mental causation that soften the critical judgment of researchers and journal editors. The explanatory role of unconscious influences is otherwise dismissed, as when the authors state that we do not "need to posit 'magical' unconscious processes producing answers from thin air" (sect. 6.2). Are unconscious processes mere explanatory magic?

As we noted above, the hypothesis of unconscious influence is observationally equivalent to the claim that cognition and metacognition are imperfectly coupled and sometimes strongly dissociate (because "conscious awareness" is measured by metacognitive report). In this regard, unconscious processes are no more "magical" than any other functional dissociation in cognition. Such processes are predicted by any cognitive architecture that represents metacognition as a limited subset or partial aspect of the mind.

For example, consider Baars's (2005) global workspace theory (GWT). Contrary to the target article's cursory account of it, this model is motivated by basic computational problems in cognition. Behavioral and neurophysiological investigations suggest the existence of multiple semi-independent "modules" specialized for different facets of information processing. This division of cognitive labor solves some problems, but it also *creates* a problem. Specifically, information from different modules needs to be integrated to represent arbitrary perceptual combinations, solve unfamiliar problems not handled by any one module, organize motor programs around a coherent action plan, and build an internal model of "the self." To address this problem, GWT proposes that only a small subset of relevant information is selected for "broadcasting" across the network. Integration is thus obtained, but it is incomplete and comes at the expense of reduced information bandwidth and processing speed. The theory makes sense of the *local* patterns of neural activity (with relative *inactivity* in globally connected association areas) that are observed in subliminal priming experiments and in the behavioral automatisms of sleepwalking, epilepsy, and the vegetative state (Baars 2005). This architecture implies that non-selected information can bias behavior, but without flexible integration or accessibility to report.

Similar predictions are made by other models that, in principle, distinguish the process of metacognitive report from other processes. For example, well-known models of learning and memory distinguish between procedural and declarative systems (Squire 1992). While the systems are thought to interact in the control of complex behavior, declarative knowledge of procedural mechanisms is at best indirect. Other models highlight constraints on the (coarse-grained) format of metacognitive representations. These constraints may limit the kinds of information that are available to report (Winkielman & Schooler 2011). Notably, dissociations are even possible within metacognition, as when we overtly report that certain states are ineffable – we experience more than we can overtly describe or express.

In any event, the claim that some cognitive operations are inaccessible to metacognition is not magical, but conceptually coherent and consistent with current knowledge. It predicts systematic mismatches between cognitive processes and subjects' overt reports about those processes – even when probes of awareness are reliable, relevant, immediate, and sensitive. Indeed, we would be interested to see the authors propose a principled sketch of a cognitive architecture in which cognition and metacognition are inseparable. To us, such a panpsychic architecture sounds like magic.

Extremely rigorous subliminal paradigms demonstrate unconscious influences on simple decisions

doi:10.1017/S0140525X13000848

Michael Snodgrass, Howard Shevrin, and James A. Abelson
Department of Psychiatry, University of Michigan Medical Center, Ann Arbor, MI 48105.
jmsnodgr@med.umich.edu shevrin@med.umich.edu
jabelson@med.umich.edu

Abstract: While showing unconscious influences on complex decisions is indeed difficult, relevant awareness in relatively simpler subliminal paradigms is more easily assessed. Utilizing objective *detection* (vs. more typical identification or classification) tasks to assess awareness overcomes longstanding residual methodological problems, and prior work using such methods (e.g., Snodgrass & Shevrin 2006) clearly shows unconscious influences on simple decisions.

Newell and Shanks (N&S) marshal impressive evidence that claims for unconscious influences on decision making are likely premature in the three areas they primarily discuss. Especially in such complex paradigms (e.g., multiple-cue judgment), we agree that it is very difficult to adequately assess relevant awareness, and hence that the jury is (or should be) still out. In contrast, however, subliminal paradigms are simpler, making assessing relevant awareness easier. Of course, such paradigms also face methodological hurdles, and we agree that much subliminal work does not overcome these difficulties. Still, contra N&S, we argue that subliminal paradigms can demonstrate unconscious influences on simple decisions under certain conditions. We first summarize our methodological analysis (cf. Snodgrass et al. 2004a, pp. 849–53), and then briefly describe some prior work that meets these extremely stringent methodological criteria. We focus on objective threshold paradigms, wherein performance on awareness assessment tasks does not exceed chance (i.e., $d'=0$). Skeptical interpretations are more plausible in subjective threshold paradigms, wherein performance exceeds chance and/or awareness assessment is often weak (e.g., post-experimental inquiries).

How should relevant awareness be assessed? All stimulus-related effects (e.g., semantic priming), whether conscious or unconscious, require at least partial stimulus identification. Accordingly, forced-choice prime identification tasks (e.g., "Was it word A or word B that was just presented?") adequately assess awareness in principle, because any conscious partial identification will raise performance above chance. For example, given "happy" and "terror" as response alternatives, perceiving the letter "t" would favor the latter response. Consequently, demonstrating