Checking the Foundation: Implicit Assumptions About Implicit Association Testing

Corey F. Robinson

University of Pittsburgh

**“***If there exist, in human cerebral action, processes which seem fundamentally different or inexplicable in terms of our present construct of the elementary physiology of integration, then it is probable that that construct is incomplete or mistaken, even for the levels of behavior to which it is applied” –* Karl S. Lashley, 1951

***Introduction***

Over the past several decades, a great deal of attention has been focused, both within academic and public discourse, on the psychometric characteristics of a class of psychosocial instruments referred to as *implicit* measures. Grounded in dual-process theories of cognition, implicit measures, such as the Implicit Association Test (IAT), were developed for the purpose of measuring mental processes which are automatic, associative, or “implicit,” as opposed to controlled, propositional or “explicit.” Implicit measures rapidly generated interest among social scientists in particular, likely at least in part due to their potential immunity to two vulnerabilities of explicit, self-report measures: response biases andlack of introspective access (Gawronski, 2011). However, despite the IAT’s widespread impact, and despite the substantial number of papers discussing its psychometric properties, the validity of the IAT remains thoroughly unsettled.

Indeed, there exists a profusion of publications, which call into question the IAT’s construct validity (e.g. Tetlock & Mitchell, 2006; Brownstein, Madva, & Gawronski, 2019). Proponents of the IAT have responded to some of these concerns, but many remain unresolved. That such ambiguity remains as to what it is that the IAT *actually* measures is deeply concerning,considering the enormous implications of the IAT’s validity to a wide range of fields and subjects, including psychological theory and practice (Ostafin & Palfai, 2006), education (Hillard, Ryan, & Gervais, 2013), antidiscrimination law (Wang, 2006; Chamallas, 2001), workplace and employment discrimination (Sturm, 2001), healthcare practices (Byrne & Tanesini, 2015), and law enforcement training (Spencer, Charbonneau, & Glaser, 2016).

While construct validity is an inarguable necessity for the usage and interpretation of any psychosocial instrument, it is not entirely surprising that the IAT’s target construct is particularly slippery. After all, both the measure and the proposed interpretations of scores obtained through its administration are based upon unverified extrapolations from broad models of cognition (e.g. associative beliefs vs. propositional beliefs; implicit cognition vs. explicit cognitive). These high-level schemas of cognition are, effectively, putative proxies for multilevel networks of neural activations. It is worth noting here that such abstractions are not inherently problematic. Indeed, the consolidation of intricate, multidimensional phenomena into more comprehensive—and, importantly, more comprehensible—theories is integral to scientific inquiry. Nevertheless, there is persuasive evidence that the IAT’s theoretical underpinnings could be failing to account for the totality of cognitive structures and processes relevant to implicit biases (e.g. Amodio, Harmon-Jones, & Devine, 2003; Amodio & Devine, 2008).

A series of papers published in the early-2000s linking implicit race bias, as measured by the IAT, to amygdala activation provides a superbly relevant example of the ease with which ultimately erroneous inferences drawn from neural observations can at first appear perfectly reasonable (e.g. Cunningham et al., 2004; Phelps et al., 2000). In these papers, the observed relationship between IAT scores and amygdala activation was interpreted as evidence that implicit bias is a direct product of unconscious fear. However, there are two major issues with this conclusion. First, the amygdala is implicated in numerous cognitive functions and processes beyond fear processing, such as arousal, vigilance, learning, and decision-making. It is therefore possible that the relationship between IAT scores and amygdala activation arises from one or more of these alternative processes. Second, it makes little sense to implement the IAT as an instrument for investigating the neurological foundations of implicit attitudes, as the construct measured by the IAT remains itself ambiguous. By assuming the validity of the instrument for measuring implicit associations and proceeding to draw conclusions about the mechanisms underlying the *construct,* as opposed to the *instrument,* the authors have inadvertently committed a *petitio principii* (i.e. begged the question).

This error of interpretation does not imply, however, that the IAT is inappropriate for neuroscientific investigation. Rather, because the IAT is purported to measure implicit mental activity, the assessment of neural data in relation to IAT performance could provide important information about the validity of the instrument itself. It is surprising, considering the abundance of literature on the validity of the IAT’s usage as a measure of implicit bias, that seemingly little effort has been made to integrate the available literature on the systems and brain regions relevant to the IAT’s theoretical underpinnings into discussions of validity. Thus, in the present paper, I will *1)* review the IAT’s existing validity evidence, *2)* discuss the implications of relevant neuroscientific research to the IAT’s theoretical underpinnings and validity argument. In doing so, I hope to contribute to bridging the gap between the empirical evidence offered by cognitive neuroscientists and the theoretical explanations proposed by social psychologists for interpreting IAT results.

***What is the IAT?***

To measure the strength of these associations among concepts, participants are told to sort stimulus exemplars from four concepts—two concepts belonging to a target dimension, and two concepts belonging to an attribute dimension—using only two response keys. For each dimension, the two concepts should represent a contrasting pair (e.g. Black & White, Good & Bad). The premise underlying the procedure is elegantly simple: People should have more difficulty cognitively mapping two concepts onto a single response option when the two concepts are perceived as dissimilar from one another than when perceived as similar to one another. Accordingly, a person should be able to respond with little cognitive effort, and thus quickly, when the same response key is assigned for concept exemplars of “Hero” and exemplars of “Good.” When the same response key is assigned for exemplars of “Villain” and exemplars of “Good,” however, responses are expected to require more cognitive effort, and thus take longer to execute. Results are evaluated by first computing average response latencies, then comparing the average latency of the trials in which concepts *1a* and *2a* correspond to the response key *X* and concepts *1b* and *2b* correspond to response key *Y* with the average latency of the trials in which *1a* and *2b* correspond to *X* and concepts *1b* and *2a* correspond to *Y.* Substituting in the concepts found in the classic race-based IAT (i.e. White, Black, & Pleasant, Unpleasant), a simple representation of possible pairings is provided below:

*White + Pleasant; Black + Unpleasant || White + Unpleasant; Black + Pleasant*

***The Foundations of the IAT***

The scientific study of implicit cognition in modernity, depending upon one’s interpretation of the word “modern,” *could* be traced backwards to the theoretical contributions of Sigmund Freud and Edward Thorndike at the turn of the 20th. century. However, given the substantive differences between their respective concepts of implicit or unconscious cognition and the concept used today, it makes more sense in the present context to continue along the timeline. In the late-1950s and early-1960s, there was some interest into the notion of subliminal perception, or “subliminal encoding,” which was inspired to some extent by Freudian psychoanalysis (Eriksen, 1960). However, after many of the studies on the topic were subjected to heavy methodological criticisms (Dixon, 1971), interest quickly waned. Around the same time, several so-called “learning-without-awareness” studies, which investigated whether or not participants could acquire classically conditioned responses without being aware of conditioning contingencies, presented promising findings (Adams, 1957; Giddan & Eriksen, 1959).

Nonetheless, it was not until the discovery of priming phenomena (e.g., Higgins, Rholes, & Jones, 1977; Srull & Wyer, 1979) that the concept of implicit cognition, as it is recognized today, began to formalize. Some have argued (e.g. Harris et al., 2013) that it was from Meyer’s and Schvaneveldt’s analysis of semantic spreading activation in 1971 that the study of priming emerged. In truth, the concept of priming is more appropriately attributed to Karl Lashley (1951), who developed the concept in his prescient paper, *The Problem of Serial Order in Behavior*. Lashley, attempting to wrestle with the problem of how simple temporal sequences of reactions could support or explain the complexity and continuity of language production, invoked the term to describe pre- and post-excitatory associative activations by stimuli that inform other elements of some process. While more than twenty years passed before the term was retrieved, it was this very concept of priming as the process by which associative mental images are dynamically and spontaneously prepared, based on some input array, to facilitate some output array that served as the foundation for priming research. It is worth stating explicitly here that this specific concept of priming (as well as cognition in general) is also the bedrock upon which implicit social cognition, and thus the IAT, stands.

In 1977, Higgins et al. hypothesized, based both on Lashley’s work and on the findings of investigations into *crossover effects* (Steffy & Galbraith, 1974),that priming could not only induce or activate word production, but could even activate trait concepts to produce social judgments. It was a sensational finding, as it suggested that priming could influence impression formation within a specific, situational context. From here, theory on implicit social cognition blossomed. Priming has since been found to: influence perceptions of social norms (Aarts & Dijksterhuis, 2003), produce competitive or cooperative behaviors (Bargh, et al., 2001), manufacture temporary cultural values and political beliefs (Gardner et al., 1999), and more. One of the most significant developments within this branch of research was that priming, which had previously been understood as activating single trait concepts, could activate ingrained (but dormant) stereotypes of social groups (Devine, 1989). In this way, priming has transmogrified over time. Whereas priming initially served as a method for testing hypotheses regarding the *structure* of cognition and the activation of contextual impression formation, it gradually came to be used as a means to *reveal* unconscious beliefs and cultural artifacts. But, considering that we have yet to comprehend the system that yields priming effects, nor the dynamics among the system’s components, are such uses *justified*? Can such interpretations be verified? These questions should be kept in mind, as they foreshadow a central theme of the upcoming discussion about the IAT.

***Development of the IAT***

Inspired by prior research on implicit memory and priming (e.g. Meyer & Schvaneveldt, 1971, Tulving & Schacter, 1990), and by a 1995 paper by Greenwald and Banaji arguing for the extension of implicit and explicit memory concepts to sociocultural constructs, the IAT was first presented in 1998 as a means by which “socially significant automatic associations” (i.e. held attitudes and beliefs that may or may not be consciously or explicitly apprehended by the holder) could be objectively measured (Greenwald, McGhee, and Schwartz).

According to the authors, the IAT could be used as a measure individual differences in *implicit attitudes*. Thatthe IAT was presented as a diagnostic inference tool and is persistently implemented as such (e.g. Gray et al., 2005) is perhaps its greatest distinguishing factor from other so-called implicit measures, such as the Go/No-Go test, which are primarily research tools for causal analyses. It should be noted that, in a later paper, the test’s developers did indicate that “it is premature to use the IAT as a diagnostic indicator for conclusions that have important, direct, and personal consequences” (Nosek, Greenwald, & Banaji, 2005). However, the prudence of this statement was swiftly abandoned when, in the following paragraph, the IAT was recommended as a clinical diagnostic tool.

In one of three experiments conducted and discussed within Greenwald et al.’s seminal 1998 paper, a race-based IAT asking participants to classify Black versus White names and pleasant-meaning words versus unpleasant-meaning words was administered to a sample (N=26) of White college students. Having found that participants responded significantly faster on White + Pleasant trials than on Black + Pleasant trials, the authors concluded that the task had uncovered “an implicit attitudinal preference for White over Black” people and indicated that “unconscious forms of prejudice” are ubiquitous.

However, numerous papers published in response to Greenwald et al. (1998) have contended that such an interpretation is not sufficiently supported by existing evidence or theory, proposed alternative explanations, and raised a range of concerns regarding the IAT’s psychometric properties (see: Brendl, Markman, & Messner, 2001; Karpinski, 2004; Klauer & Mierke, 2003; Fazio & Olson, 2003, Rothermund & Wentura, 2001, 2004, etc.). Perhaps an equal number of researchers have come to the IAT’s defense, arguing for its validity as a measure of “socially-significant” implicit attitudes (Greenwald, Nosek, Banaji, 2003; Egloff & Schmukle, 2002; Gawronski & Bodenhausen, 2017; Rae & Greenwald, 2017). This will be discussed further in the next section on the IAT’s validity and reliability.

***Internal Validity***

***Brief Aside on Reliability***

For an instrument to be useful, it must be shown to produce reliable outcomes. If a measure’s outcomes are not reliable, then the “signal” is effectively drowned out by the “noise.” Here, the word *signal* represents score variability produced by true inter-individual differences, while *noise* represents score variability produced by measurement error. Reliability, then, represents the fraction of total variation that is not attributable to error in measurement. The three commonly used reliability measurements are internal consistency, inter-rater reliability, and test-retest reliability.

For inferences drawn from IAT scores to be considered *valid,* it is necessary that the IAT also demonstrate reliable outcomes. If the IAT does not produce reliable outcomes, then the proposed inferences of its scores cannot be validated—it would be unrealistic to extricate the signal from the noise. Before discussing the existing evidence of internal consistency reliability and test-retest reliability, it is worth pausing to emphasize that, while reliability is a *necessary* condition for the validation of score-based inferences, it is not a *sufficient* condition. Thus, even if an instrument is shown to produce reliable outcomes, the interpretations of said outcomes can still be inappropriate.

***Internal Consistency Reliability***

One of the strongest features of the IAT is almost certainly its internal consistency. Whereas implicit measures, such as evaluative priming measures, are generally known to perform poorly in this area (Bosson, Swann, & Pennebaker, 2000), the IAT consistently demonstrates moderate-to-high internal consistency estimates (α=0.8; Banse et al., 2001). While the exact correlation coefficients (of course) differ between analyses, estimates consistently remain between 0.7 and 0.9, indicating an acceptable degree of inter-item homogeneity. This supports the developers’ claim that the IAT measures unitary constructs (Greenwald et al., 2006).

***Test-retest Reliability***

Estimates of the IAT’s test-retest reliability are far more varied than those of internal consistency reliability. In one study, the test-retest reliability was reported to be between .58 and .62 (Egloff, Schwerdtfeger, & Schmukle, 2005). Another study, the average test-retest reliability obtained across four time periods was merely .27 (Cunningham et al., 2001). Irrespective of which end of this spectrum one chooses to pay more heed, the IAT cannot be said to be temporally stable—at least according to the common prescription that test-retest reliability coefficients should be greater than or equal to .70. It is worth keeping in mind that what constitutes “acceptable” test-retest reliability estimate is dependent on a number of factors, such as the amount of time between administrations and the purpose of the test. That being said, in the paper by Egloff, Schwerdtfeger, and Schmukle (2005), follow-up testing was performed immediately. The small test-retest correlation observed suggests that IAT score outcomes are unreliable, and therefore that substantial random error is present in the composition of IAT scores.

***Other Threats to Internal Validity***

One potential confound to the proposed interpretation of IAT effects as a function of associative preference is the familiarity of stimulus items. To reinforce this as an alternative explanation, Brendl, Markman, and Messner (2001) conducted three experiments in which the IAT was administered using pseudo-words as “exemplars” for one of the two dimensions. In the first experiment, the dimension *not* comprised of pseudo-words used White names as exemplars. In experiments two and three, the names of insects were used in place of White names, for the reason that insects should be clearly more negative in valence than nonwords (which should have approximately neutral valence).

Across all three experiments, the average response latency was higher for the pseudo-word dimension than the real-word dimension. As Brendl, Markman, and Messner noted in their conclusions, it is “difficult to reconcile” the IAT as a measure of implicit attitudinal *preference* with the finding that a dimension that carries clearly negative valence produced faster responses than a dimension with theoretically neutral valence. Responding to this study, Greenwald and Nosek (2001) concluded that the IAT simply does not function when one or more of the dimensions are composed of nonwords. Surprisingly little research has explored the familiarity hypothesis further.

Another threat to internal validity is the presence of block order effects. Several studies have found that IAT effects increase when participants begin with a response block in which dimension mappings are compatible (Nosek et al., 2005; Klauer & Mierke, 2005). It has also been demonstrated that IAT scores can be influenced by stimulus effects beyond dimensional valence. It is important that factors such as block order effects, which cause systematic error variance, be addressed.

Lastly, a number of authors (e.g. Mierke & Klauer, 2003) have observed that noteworthy correlations occur between IATs with presumptively unrelated content dimensions (e.g. self-esteem IAT and geometric shape IAT). Although it is perhaps possible that some relationship exists between self-esteem and geometric shapes, there is little evidence for this interpretation. A far more likely explanation is that the correlations observed between scores on theoretically unrelated IATs are artifacts of *shared method variance*, i.e. spurious variance that is imputable to the measurement method rather than to the underlying or latent construct(s) a given measure seeks to represent*.* Supporting the argument for the presence of shared method variance, several researchers have produced evidence that IAT effects are influenced by participants’ “general processing speed” (Blanton et al., 2006), as well as their cognitive ability to task-switch (Klauer et al., 2010). While such confounding cognitive skills have considerable implications for the *construct* validity of IAT-type tests, they are also significant to internal validity, as they introduce a threat to causal inference.

**On Construct Validity**

The greatest challenge to inferences made based on scores obtained by the IAT is the striking absence of clarity surrounding the true identity of the construct(s) measured by the IAT. According to the original definition presented by Cronbach & Meehl (1955), a construct is “some postulated attribute of people, assumed to be reflected in test performance.” The target attribute need not be highly systematized for the construct to be validated (Cronbach & Meehl, 1955). Thus, it is not necessarily a problem that the IAT’s proposed construct is considerably abstract—intelligence, for example, is an abstract construct. Attitudinal measures, too, aim to capture abstract constructs. What, then, *is* the problem?

One possibility is that the IAT does *not,* in fact, measure some postulated attribute of people. Rather, the IAT measures *associations.* This point, painfully obvious as it may be, is potentially important. The IAT is proposed to measure a “black-box” cognitive *process*. It is this cognitive process—whatever it may be in truth—that is then interpreted as being indicative of a trait or attitude. By the term “black-box,” I am primarily referring to a lack of insight into what is occurring—what is at play—when an association event transpires, and not to the quality of being abstract or *latent*.

Regardless of this, one of the best ways to evaluate an instrument for construct validity is to evaluate the extent to which its scores are capable of predicting relevant future behavior. Concerning this form of validity evidence, known as *predictive* validity, findings have been somewhat disappointing. In a meta-analysis of predictive validity, Greenwald et al. (2009) reported an average correlation between IAT scores and prediction criteria of .27. They noted, though, that “for socially sensitive topics, the predictive power of self-report measures was remarkably low and the incremental validity of IAT measures was relatively high” (Greenwald et al., 2009, p. 28). However, this claim is perhaps unfounded, as explicit measures showed higher correlations with prediction criteria than the IAT on three out of five socially sensitive domains (i.e. sexual orientation, substance abuse, and mental illness). In a later meta-analysis by Oswald et al. (2013), which included more rigorous coding for criterion measures than was used by Greenwald et al. in their 2009 analysis, the average correlation between IAT scores and prediction criteria was found to be lower than initially suggested (*r* = 0.14).

In addition to predictive validity evidence, another important form of obtaining construct validity evidence is to evaluate the extent to which its scores correlate highly with theoretically similar measures and do not correlate with theoretically different measures. These correlations, respectively, provide evidence for *convergent* validityand *discriminant* validity. Researchers investigating the relationship between the IAT and other implicit methods, such as evaluative priming methods, have reported extremely low correlations (.04 < *r* < .30; e.g. Fazio, 1999; Sherman et al., 2003). Interestingly, the relationship between IAT scores and scores obtained by explicit, self-report measures seems to differ substantially depending on which IAT is being used (e.g. race-based IAT vs. self-esteem IAT). Nosek and Smith (2007), using a multitrait-multimethod design, found correlations between IAT scores and equivalent self-report measure scores ranging from .12 (White-Black) to .56 (Creationism-Evolution).

It is possible, based on their findings, that the IAT is better at capturing associative preferences in some domains than in others. One explanation for this is that the IAT performs better on socially-sensitive domains (i.e. racial attitudes), and that this causes low correlations between IAT scores and self-report measures for such domains. If this were so, we would likely find that similar relationships arise within other socially-sensitive domains. Yet, the correlation between “straight-gay” preference as measured by the IAT and self-report was substantially higher (*r* = .39) than for “white-black” preference (*r* = .12).

All in all, these findings raise the question: Would a pure-process measure of implicit preferential associations differ dramatically between domains that theoretically invoke similar evaluation processes (i.e. prejudicial attitudes, automatic activation of culturally-ingrained stereotypes)? That there is not clarity on relatively basic questions of the appropriateness of a measure’s usage (Greenwald et al., 2005) is indicative of our present lack of understanding regarding the most fundamental properties of what an implicit association *is*, which structures or processes an implicit association invokes, and to which degree response latencies are able to capture an implicit association without contamination from other processes.

***Dual-Process Models and Bayesian Brains***

Inextricable from the IAT are the “dual-process” or “dual-system” models of cognition, which propose the existence of a delineation between *implicit* mental processes and *explicit* mental processes. The precise language used to describe these processes varies somewhat between theories, such that some refer to *associative-propositional* or *automatic-controlled* dichotomies, yet the fundamental message of each is the same: Within the human mind, there occurs some processes, which are unintentional, efficient, uncontrollable, and unconscious, and other processes, which are intentional, resource-intensive, controllable, and conscious. One need not be formally acquainted with the viscera of dual-process models to perceive significant similarity between them and the many, long-existent dualities generated in response to seeming inexplicabilities of human experience.

Indeed, the schematic of cognition presented within dual-process models has existed in many cultures and eras—albeit, with differing explanatory mechanisms. In the ontology of the antediluvian Vedic healers, consciousness is defined as “energy” that self-apprehends and, conversely, unconsciousness as energy that has forgotten itself. Plato, amidst his writings on virtue in *Meno*, asserts that the knowledge of all things is possessed by the soul even prior to one’s birth, but that one must engage in the practice of reminiscence (i.e. the act of remembering) to retrieve the dormant—read: unconscious—knowledge. Spinoza distinguished between conscious “Will” and unconscious “Appetite” when outlining his theory of moral consciousness. the seventeenth-century, in the eighteenth-century by Nietzsche, and in the nineteenth-century by Schopenhauer. Structuralist psychologists of the late-nineteenth and early-twentieth centuries (i.e. Wundt and Titchener) proposed that the components of consciousness may be unveiled through the methodical application of introspection. Arguably the most widely recognized and most complete treatment of the conscious-unconscious dichotomy, however, was developed by Sigmund Freud. Despite the many criticisms levied against the father of psychoanalysis (and against psychoanalysis itself), he was arguably the first to systematically codify mental activity according to deterministic principles, as well as to propose that human behavior may be wholly explained by mental processes and states that lie beyond the grasp of conscious awareness.

The implicit-explicit dichotomy reflects most broadly the popular delineation between “head” and “heart,” or reason and intuition. It is possible, and even probable, that the ubiquity of dualistic cognition as a cultural concept has contributed to the *prima facie* acceptance of dual-process models—much like the IAT—within many fields (Loewenstein & O’Donoghue, 2005; Fudenberg & Levine; 2006). Yet, despite possessing wonderful face validity, it remains yet to be empirically determined whether human mental functioning can or should be classified into two qualitatively distinct systems. Scientific nomenclature abounds, yet literature on dual-process systems consistently fails to clearly and cohesively define the conditions that distinguish the two modes of thought (Keren & Schul, 2009). The definitions that *are* provided seem to be little more than labels applied to outcome measures *post hoc,* as opposed to explanations of the phenomena underlying said outcomes.

To illustrate this error in reasoning, I will briefly describe an alternate reality in which spiritual beliefs guide modern scientific inquiry. In this alternate version of our world, perhaps known by its inhabitants as *Fearth* (i.e. **F**ake Earth), the Soul is considered to be scientifically evident and distinct from the Brain. Why? Well, for many reasons, but, for brevity’s sake, here are two. First, there is evidence of Soul Memory: In a study of anterograde amnesia, patients were asked to study a list of words (e.g. *CLEANSER, DARK,* etc.), after which free recall and recognition tests were administered. Compared to control subjects, the amnesic patients were significantly less likely to accurately recall or recognize list items when no additional stimuli were provided. However, amnesic patients were equally likely as controls to accurately recall list items when presented with word fragments (i.e. *CLE*\_\_\_). Second, it is possible to measure the strength of Soul Attitudes using a class of instruments referred to as “Soul measures” (as opposed to “Brain measures”), such as the Soul Association Test (SAT), which measures the strength of the Soul’s associations between concepts. We can be certain that Soul measures are measuring Soul processes because *1)* the Soul has been shown to exist by prior research on Soul Memory *2)* Soul processes are known to be unintentional, efficient, uncontrollable, and unconscious, whereas Brain processes are intentional, resource-intensive, controllable, and conscious Brain processes, thus low response latencies must be produced by Soul associations. As Soul-specific processes are observable and measurable, the existence of Soul is a scientific certitude. Further, we can also be certain that the Soul is a distinct system from the brain, because Soul-specific processes are known to be specific to the Soul.

It is worth taking a moment here to underscore the purpose of this parable, lest the lighthearted spirit in which it was written belie its significance. First and foremost, no implications about the truth or falsity of any phenomena should be inferred. Accordingly, the existence of some arrangement of cognitive processes, which have subsequently been observed *and categorized* under the epithet “implicit,” is not under dispute. Nor is the concept of unconscious cognition. Rather, the parable is intended to highlight several profound errors of inferential reasoning present within a substantive portion of literature on dual-process models—and, by extension, the IAT.

As is the case for our friends on Fearth, we possess clear, perhaps even irrefutable, evidence of processes that do not require conscious awareness. Yet, the presence of realevidence of unconscious processes does not inherently imply the existence of a second, distinct cognitive system. Similarly, the observance of neural and cognitive phenomena that possess the grouping of characteristics we understand as *implicit* need not be indicative of discrete attitudes and modes of thought. Perhaps, for example, cognitive events occur on a *continuum* of “implicitness-explicitness.” Or, perhaps cognitive events occur on *multiple* continua—one for each characteristic proposed to distinguish implicit cognition from explicit cognition. To put the point simply: We lack the information necessary to make substantive claims. It should be noted that the existence of dual-process models is not inherently problematic, as they are de factoheuristic devices for enabling understanding and communication with regard to what is arguably one of the most complex and abstruse systems known. Such devices can be invaluable for guiding inquiry and generating hypotheses. Dual-process models are not, however, treated as heuristic devices. Rather, they are presented and approached as scientific theory—as a foundational assumption to build upon (e.g. the IAT).

The human brain, though not homogenous by any means, is incredibly morphologically intricate (Swanson, 2003), interconnected (Bressler, 1995), and adaptable (Huttenlocher, 2009). In fact, modern neuroscience predominantly conceives of the brain as an interconnected Bayesian machine (e.g. Friston, 2011; Genon et al., 2018; Beck et al., 2008, etc.), insofar as neural structures represent computational units which function both individually *and* collectively to generate inferences about the causes of sensations using an adaptive causal model of relationships among world-states that produce sensory information. While each brain region is limited to some range of functions, the resultant output of a given specialized region is modulated by interaction or coactivity with other regions.

This understanding is compatible with the notion that the mind develops associations between concepts, as IAT researchers suggest. Under the Bayesian Brain model, the human brain is a “prediction machine” (Clark, 2013, pg. 181). As Lashley discussed, that humans use language and logic is evidence enough of automatic activations and associative structures. However, it seems that these associations occur within an *integrated* system. That is to say that, it is uncertain whether the “part” can be dissociated from the “sum.” It is perhaps this is has led to the underwhelming predictive ability of IAT scores. The known confounds of cognitive processing speed and executive function, as discussed above, only emphasize the possibility of this explanation.

**Conclusions**

Words can be fickle. Or, more precisely, we sometimes expect too much of words. We expect them to carry and preserve our most abstract concepts, and to deliver our signals without undue noise. Considering the ephemerality of mental representations (Dell, 1986), it could be considered a grand achievement that *more* words do not metamorphosize in meaning under our noses. Nevertheless, these representations effectively function as the interface by which our minds are able to comprehend and interact with the world (Lashley, 1951; Smith, 1994). In a powerful work by Lakoff and Johnson (1980), they argued that metaphors serve as the mechanism by which we are able to employ *experiential gestalts* (i.e. intricate conceptual structures, which emerge only through “constant negotiation with the environment and other people”) to structure experience in some other domain. Take, for example, the well-documented spatial bias effect on response to Likert scale items (Nicholls, et al., 2006). Research on this bias has shown that respondents whose first language is read from left-to-right (e.g. English) demonstrate systematic bias towards more rightward response options, whereas respondents whose first language is read from right-to-left (e.g. Arabic) demonstrate bias towards leftward response options (Shaki, Fischer, & Göbel, 2012). Many more examples exist of the ways in which the human mind codifies abstract concepts and representations by using already-existing scaffolding (e.g. *temporal construal*; Förster, Friedman, & Liberman, 2004), but the point need not be belabored.

Humanity’s use of metaphor is a phenomenal example of the multilevel network of computational units, which we call the brain, in action. We have Bayesian brains, and we are computational machines, but we also are *association* machines. It is worth being optimistic that the absence of a discrete “implicit” network does not imply that the measurement of associative processes cannot be achieved. It is also worth being pragmatic in establishing the path to such a measure. The issues of theory and validity described in this paper provide, I hope, ample rationale to begin considering alternative measurement paradigms. If a valid, diagnostic measure of associations is someday developed, it is reasonable to expect that it will occur through multidisciplinary or interdisciplinary collaboration, such that its theoretical and neural underpinnings are coherent. Intelligence testing is a wonderful example of an instrument being used to capture a latent construct that occurs through a multiplicity of cognitive processes and structures, which interact dynamically within the human brain (Deary, Penke, & Johnson, 2010). Perhaps, moving forward, proponents of associative measurement might consider ways in which associations can be measured by the “sum” of cognition, rather than the part.

.

Works Cited

Aarts, H., & Dijksterhuis, A. (2003). The silence of the library: environment, situational norm, and social behavior. *Journal of personality and social psychology*, *84*(1), 18.

Adams, J. K. (1957). Laboratory studies of behavior without awareness. *Psychological bulletin*, *54*(5), 383.

Amodio, D., & Devine, P. G. (2008). On the interpersonal functions of implicit stereotyping and evaluative race bias: Insights from social neuroscience. In *Attitudes* (pp. 213-246). Psychology Press.

Amodio, D. M., Harmon-Jones, E., & Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink response and self-report. *Journal of personality and social psychology*, *84*(4), 738.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für experimentelle Psychologie*, *48*(2), 145-160.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The automated will: nonconscious activation and pursuit of behavioral goals. *Journal of personality and social psychology*, *81*(6), 1014.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., ... & Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron*, *60*(6), 1142-1152.

Blanton, H., Jaccard, J., Christie, C., & Gonzales, P. M. (2007). Plausible assumptions, questionable assumptions and post hoc rationalizations: Will the real IAT, please stand up?. *Journal of Experimental Social Psychology*, *43*(3), 399-409.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, *42*(2), 192-212.

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *Journal of applied Psychology*, *94*(3), 567.

Bosson, J. K., Swann Jr, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited?. *Journal of personality and social psychology*, *79*(4), 631.

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of personality and social psychology*, *81*(5), 760.

Bressler, S. L. (1995). Large-scale cortical networks and cognition. *Brain Research Reviews*, *20*(3), 288-304.

Brownstein, M., Madva, A., & Gawronski, B. (2019). What do implicit measures measure?. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(5), e1501.

Byrne, A., & Tanesini, A. (2015). Instilling new habits: addressing implicit bias in healthcare professionals. *Advances in Health Sciences Education*, *20*(5), 1255-1262.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181-204

Chamallas, M. (2001). The market excuse.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological science*, *12*(2), 163-170.

Cunningham, W. A., Raye, C. L., & Johnson, M. K. (2004). Implicit and explicit evaluation: fMRI correlates of valence, emotional intensity, and control in the processing of attitudes. *Journal of cognitive neuroscience*, *16*(10), 1717-1729.

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology*, *37*(6), 443-451.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, *93*(3), 283.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology*, *56*(1), 5.

Deary, I. J., Penke, L., & Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature reviews neuroscience*, *11*(3), 201-211.

Dixon, N. F. (1971). Subliminal perception: The nature of a controversy

Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an implicit association test for assessing anxiety. *Journal of personality and social psychology*, *83*(6), 1441.

Egloff, B., Schwerdtfeger, A., & Schmukle, S. C. (2005). Temporal stability of the implicit association test-anxiety. *Journal of Personality Assessment*, *84*(1), 82-88.

Eriksen, C. W. (1960). Discrimination and learning without awareness: a methodological survey and evaluation. *Psychological review*, *67*(5), 279.

Fazio, R. H., & Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the “I”, the “A”, and the “T”: A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology*, *17*(1), 74-147.

Förster, J., Friedman, R. S., & Liberman, N. (2004). Temporal construal effects on abstract and concrete thinking: consequences for insight and creative cognition. *Journal of personality and social psychology*, *87*(2), 177.

Fudenberg, D., & Levine, D. K. (2006). A dual-self model of impulse control. *American economic review*, *96*(5), 1449-1476.

Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain connectivity*, *1*(1), 13-36.

Gardner, W. L., Gabriel, S., & Lee, A. Y. (1999). “I” value freedom, but “we” value relationships: Self-construal priming mirrors cultural differences in judgment. *Psychological Science*, *10*(4), 321-326.

Gawronski, B. (2002). What does the Implicit Association Test measure? A test of the convergent and discriminant validity of prejudice-related IATs. *Experimental psychology*, *49*(3), 171.

Gawronski, B., & Bodenhausen, G. V. (2011). The associative–propositional evaluation model: Theory, evidence, and open questions. In *Advances in experimental social psychology* (Vol. 44, pp. 59-127). Academic Press.

Gawronski, B., & Bodenhausen, G. V. (2017). Beyond persons and situations: An interactionist approach to understanding implicit bias. *Psychological Inquiry*, *28*(4), 268-272.

Genon, S., Reid, A., Langner, R., Amunts, K., & Eickhoff, S. B. (2018). How to characterize the function of a brain region. *Trends in cognitive sciences*, *22*(4), 350-364.

Giddan, N. S., & Eriksen, C. W. (1959). Generalization of response biases acquired with and without verbal awareness. *Journal of personality*.

Gray, N. S., Brown, A. S., MacCulloch, M. J., Smith, J., & Snowden, R. J. (2005). An implicit test of the associations between children and sex in pedophiles. *Journal of Abnormal Psychology*, *114*(2), 304.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, *102*(1), 4.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.

Greenwald, A. G., & Nosek, B. A. (2001). Health of the Implicit Association Test at age 3.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of personality and social psychology*, *85*(2), 197.

Greenwald, A. G., Nosek, B. A., & Sriram, N. (2006). Consequential validity of the implicit association test: comment on Blanton and Jaccard (2006).

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of personality and social psychology*, *97*(1), 17.

Higgins, E. T., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of experimental social psychology*, *13*(2), 141-154.

Hillard, A. L., Ryan, C. S., & Gervais, S. J. (2013). Reactions to the implicit association test as an educational tool: A mixed methods study. *Social Psychology of Education*, *16*(3), 495-516.

Huttenlocher, P. R. (2009). *Neural plasticity*. Harvard University Press.

Karpinski, A. (2004). Measuring self-esteem using the Implicit Association Test: The role of the other. *Personality and Social Psychology Bulletin*, *30*(1), 22-34.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the implicit association test. *Journal of personality and social psychology*, *81*(5), 774.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on psychological science*, *4*(6), 533-550.

Klauer, K. C., Schmitz, F., Teige-Mocigemba, S., & Voss, A. (2010). Understanding the role of executive control in the Implicit Association Test: Why flexible people have small IAT effects. *The Quarterly Journal of Experimental Psychology*, *63*(3), 595-619.

Klauer, K. C., & Mierke, J. (2005). Task-set inertia, attitude accessibility, and compatibility-order effects: New evidence for a task-set switching account of the Implicit Association Test effect. *Personality and Social Psychology Bulletin*, *31*(2), 208-217.

Mierke, J., & Klauer, K. C. (2003). Method-specific variance in the implicit association test. *Journal of personality and social psychology*, *85*(6), 1180.

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the implicit association test: IV. *Implicit measures of attitudes*, 59-102.

Lashley, K. S. (1951). The problem of serial order in behavior. Cerebral mechanisms in behavior. Jeffress, LA New York.

Loewenstein, G., & O’Donoghue, T. (2005). Animal spirits: Affective and deliberative processes in human behavior. *Unpublished manuscript, Cornell University*.

Mitchell, G., & Tetlock, P. E. (2006). Antidiscrimination law and the perils of mindreading. *Ohio St. LJ*, *67*, 1023.

Meyer, D. E., & Schvaneveldt, R. W. Facilitation in recognizing pairs of words: Evidence of a dependence

between retrieval operations. Journal of Experimental Psychology, 1971, 90, 227-234.

Nosek, B. A., & Smyth, F. L. (2007). A multitrait-multimethod validation of the implicit association test. *Experimental psychology*, *54*(1), 14-29.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin*, *31*(2), 166-180.

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring?. *Psychological Science*, *14*(6), 636-639.

Ostafin, B. D., & Palfai, T. P. (2006). Compelled to consume: The Implicit Association Test and automatic alcohol motivation. *Psychology of Addictive Behaviors*, *20*(3), 322.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *Journal of personality and social psychology*, *105*(2), 171.

Nicholls, M. E., Orr, C. A., Okubo, M., & Loftus, A. (2006). Satisfaction guaranteed: the effect of spatial biases on responses to Likert scales. *Psychological Science*.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of cognitive neuroscience*, *12*(5), 729-738.

Rae, J. R., & Greenwald, A. G. (2017). Persons or situations? Individual differences explain variance in aggregated implicit race attitudes. *Psychological Inquiry*, *28*(4), 297-300.

Rothermund, K., Wentura, D., & Bak, P. M. (2001). Automatic attention to stimuli signalling chances and dangers: Moderating effects of positive and negative goal and action contexts. *Cognition & Emotion*, *15*(2), 231-248.

Rothermund, K., & Wentura, D. (2004). Underlying processes in the implicit association test: dissociating salience from associations. *Journal of Experimental Psychology: General*, *133*(2), 139.

Schnabel, K., Asendorpf, J. B., & Greenwald, A. G. (2008). Understanding and using the implicit association test: V. measuring semantic aspects of trait self‐concepts. *European Journal of Personality: Published for the European Association of Personality Psychology*, *22*(8), 695-706.

Shaki, S., Fischer, M. H., & Göbel, S. M. (2012). Direction counts: A comparative study of spatially directional counting biases in cultures with different reading directions. *Journal of Experimental Child Psychology*, *112*(2), 275-281.

Sherman, S. J., Rose, J. S., Koch, K., Presson, C. C., & Chassin, L. (2003). Implicit and explicit attitudes toward cigarette smoking: The effects of context and motivation. *Journal of Social and Clinical psychology*, *22*(1), 13-39.

Smith, C. B. (1994). *Whole language: The debate*. ERIC Clearinghouse on Reading, English, and Communication, Indiana University, 2805 E. 10th St., Smith Research Center, Suite 150, Bloomington, IN 47408-2698.

Spencer, K. B., Charbonneau, A. K., & Glaser, J. (2016). Implicit bias and policing. *Social and Personality Psychology Compass*, *10*(1), 50-63.

Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social psychology*, *37*(10), 1660.

Steffy, R. A., & Galbraith, K. (1974). A comparison of segmental set and inhibitory deficit explanations of the crossover pattern in process schizophrenic reaction time. *Journal of Abnormal Psychology*, *83*(3), 227.

Swanson, L. W. (2003). The amygdala and its place in the cerebral hemisphere. *Annals of the New York Academy of Sciences*, *985*(1), 174-184.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science*, *247*(4940), 301-306.

Wang, L.-i. (2006). Discrimination by default: How racism becomes routine. New York University Press.