American Psychologist
2017, Vol. 72, No. 9, 861–871

0003-066X/17/$12.00 http://dx.doi.org/10.1037/amp0000238

The Implicit Revolution: Reconceiving the Relation Between Conscious and Unconscious

Anthony G. Greenwald
University of Washington

Mahzarin R. Banaji
Harvard University

In the 1970s, memory researchers converged on interesting phenomena observed in Korsakoff-syndrome amnesic patients. These patients' performances on difficult tasks were reliably improved by practice sessions from which they could recall nothing. Related findings of indirect memory effects in college students triggered wide attention to phenomena that, in 1985, were first identified as *implicit memory*. Within a decade, the indirect measurement methods of implicit memory research had spread to social psychologists' studies of attitudes and stereotypes. After another two decades, the methods and findings of this developing revolution have revised understanding of how past learning, operating in ways that bypass conscious awareness, nevertheless shapes conscious judgment and perception. This revolution in psychological thinking is on the cusp of reconceiving the relation between unconscious and conscious mental process. Further, it demands researchers' careful attention to justification for many self-report measures that are now routinely treated as face-valid.

Keywords: implicit cognition, indirect measures, Implicit Association Test, unconscious inference, Implicit Revolution

Psychologists' recent increasing use of the word *implicit* started with the introduction of the label *implicit memory* by Peter Graf and Daniel Schacter (1985). Their article was titled "Implicit and Explicit Memory for New Associations in Normal and Amnesic Subjects." The article gave this description: "Implicit memory is revealed when performance on a task is facilitated in the absence of conscious recollection; explicit memory is revealed when performance on a task requires conscious recollection of previous experiences" (p. 501). In a subsequent review article, Schacter (1987) firmly established the concept of implicit memory, which he defined by connecting explicit–implicit to conscious–unconscious, writing: "Memory for a recent event can be expressed explicitly, as conscious recollection, or implicitly, as a facilitation of test performance without conscious recollection" (p. 501).

Since the mid-1980s, *implicit* has increasingly appeared as an indexing keyword in psychological publications. Figure 1 shows how this usage has grown, especially in the last decade. After introducing the implicit–explicit distinction as presently used, this article describes the history of *implicit,* including its development in cognitive psychology in the 1980s and its absorption into social psychology in the 1990s. The main events in this history were research method innovations. The often-surprising results produced by these methods have been incorporated into a collection of theories that incorporate dual-process conceptions of mental operations. This body of work is now prompting a reconceptualization of the relation between unconscious and conscious cognition.

The Implicit–Explicit Distinction

Implicit = Indirect; Explicit = Direct

In a word-fragment completion implicit memory task, subjects encounter multiple words on Day 1 (e.g., *BRAWN, MODAL*). On Day 2, when asked to complete the word fragment *BRA_ _* with a five-letter English word, *brawn*

Editor's note. Anthony G. Greenwald and Mahzarin R. Banaji received the 2017 APA Award for Distinguished Scientific Contributions for their collaborative research over a period of 30 years. This article is based on an invited presentation at the 125th Annual Convention of the American Psychological Association, held August 4–7, 2017, Washington, DC.

Authors' note. Anthony G. Greenwald, Department of Psychology, University of Washington; Mahzarin R. Banaji, Department of Psychology, Harvard University.

Anthony G. Greenwald is an officer of Project Implicit, Inc., a nonprofit organization with the scientific mission "to develop and deliver methods for investigating and applying phenomena of implicit social cognition, including especially phenomena of implicit bias based on age, race, gender or other factors."

Correspondence concerning this article should be addressed to Anthony G. Greenwald, Department of Psychology, University of Washington, Box 351525, Seattle, WA 98195-1525. E-mail: agg@u.washington.edu

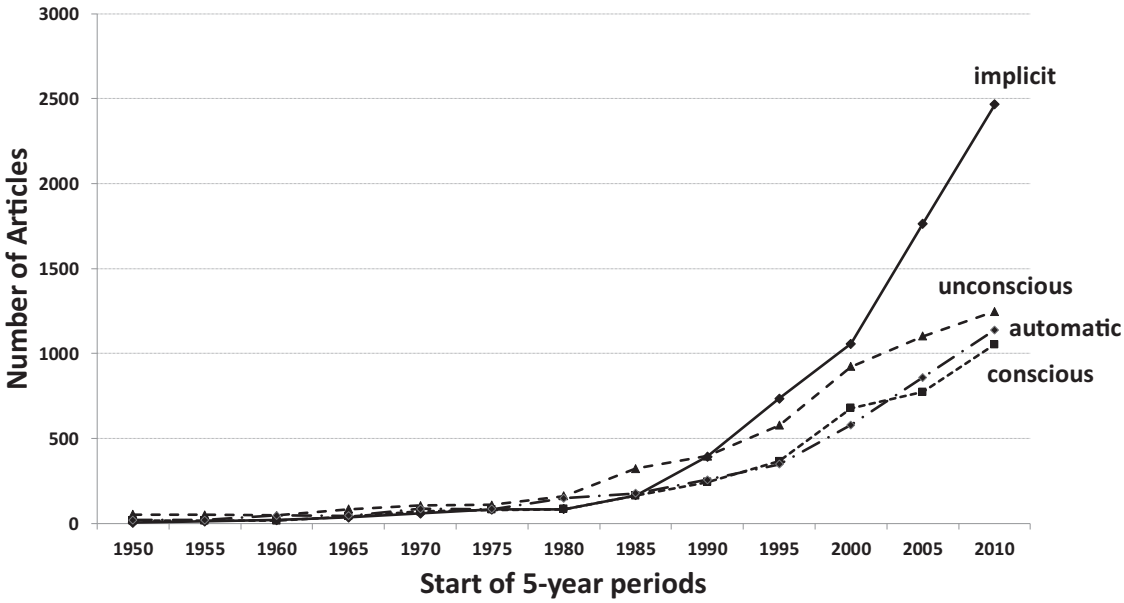

Figure 1. Use of words as keywords in the American Psychological Association (APA) PsycINFO database since 1950 (the search was limited to peer-reviewed journal articles).

may pop into thought. An implicit memory effect is apparent if the rate of the *brawn* completion is greater for those who saw *BRAWN* on Day 1 than for those who did not. Subjects who are asked whether they recognize *MODAL* as having been presented on Day 1 would be providing an explicit memory measure, in this case based on comparison of their "yes" responses with those of subjects who had not seen *MODAL* on Day 1. The recognition measure conforms to the definition of a *direct* measure of memory in that it is the response to a request to retrieve something known to have been previously encountered. The word-fragment-completion measure conforms to the definition of an *indirect* measure of memory, because it is the response to a request that includes no instruction to retrieve something previously encountered.

Indirect ≠ Unconscious; Direct ≠ Conscious

Early in the era of implicit memory research, competing usages of *implicit* were apparent. Some used it as a conceptual label for an unconscious form of memory. Others used it as an empirical term, referring to studies with indirect measures of memory.

Articles by Reingold and Merikle (1988) and Jacoby (1991) argued convincingly that indirect measures should not be treated as pure measures of unconscious processes; nor should direct measures be treated as pure measures of conscious processes. Reingold and Merikle analyzed the methods of experiments using visual masked priming (an indirect measure) in search of evidence for subliminal semantic activation (also known as subliminal priming). Jacoby analyzed the methods of implicit and explicit memory studies in similar fashion. The two articles reinforced each other, concluding that both direct and indirect measures must be assumed to combine influences of conscious and unconscious mental processes. The two articles also introduced data analysis methods that could appraise whether indirect measures revealed processes that could justifiably be identified as unconscious.

These two strong methodological contributions notwithstanding, many publications continue to use *implicit* and *explicit* with conceptual meaning (as unconscious vs. conscious), rather than with their better justified empirical (indirect vs. direct) meaning. Even though the present authors find themselves occasionally lapsing to use *implicit* and *explicit* as if they had conceptual meaning, they strongly endorse the empirical understanding of the implicit–explicit distinction.

What Is the Relation of Implicit Measures to Unconscious Mental Processes?

The foregoing description of the word-fragment-completion task observed that the subject's response may "pop into thought." Unanticipated appearance in conscious thought may be accepted by many as an indicator of unconscious process, especially if it occurs on Day 2 of a memory experiment in which the subject is unaware that the word popping into thought was encountered on Day 1.

Now consider the interpretation if the subject's response to the word fragment had been a word not encountered on Day 1. If, rather than *brawn, brace* had popped into thought on Day 2,

that would presumably have been equally the result of an unconscious process. For yet another comparison, consider an American high school student being asked to name the first president of the United States. This is a direct measure—a request to retrieve something known to have been previously encountered. *George Washington* will likely appear in the student's conscious thought without retrieval effort. Is this less automatic or less unconscious than *brawn* popping into mind in the fragment-completion task? Hopefully, this example reinforces the earlier point that neither indirect (implicit) nor direct (explicit) measures warrant treatment as pure indicators of unconscious or conscious mental process.

Harbingers and Precursors of the Implicit Revolution

Led by work of Elizabeth Warrington and Lawrence Weiskrantz (1970, 1974; Weiskrantz & Warrington, 1979), multiple researchers in the 1970s sought tasks on which memory performances of amnesic patients approximated those of nonpatient control subjects. These researchers' interest in memory abilities of amnesic patients was (distantly) presaged by Édouard Claparède's (1911/1951) report of an elderly female Korsakoff-syndrome amnesic patient. On one of Claparède's daily visits to his patient, as he was being reintroduced to her—something necessary each day because she had no recollection of his previous visits—Claparède surprised her by sticking her with a pin when he reached to shake her hand. The next day, when he again reached to greet her, she quickly withdrew her hand. When Claparède asked her why she withdrew, she was unable to link it to Claparède's behavior of the previous day. Claparède described his patient's hand withdrawal as the indicator of a memory that was separated from her conscious, psychological self.

Jumping ahead: For much of the 20th century, *unconscious* was effectively a taboo word in academic psychological publications (cf. Greenwald, 1992). This was in part a reaction against the great popularity of Sigmund Freud and psychoanalytic theory and in part also a reflection of American behaviorists' eschewal of both conscious and unconscious mentality in much of the 20th century. Figure 1 shows that scholarly uses of *unconscious* in post-1950 academic psychological publications were quite infrequent until about 1980. The subsequent increased reference to nonconscious mental function can be seen in the growing post-1980 use of three index terms: *unconscious, automatic,* and *implicit.* Only after 2005 did *implicit* separate itself from the other two.

Empirical Precursors

Multiple pre-1980 influences contributed to cognitive psychologists' seeing the value of investigating often ingenious indirect measures of memory. Developmental psychologists had long used inventive indirect measures to reveal mental content and process in young children. Their subjects included not only those too young to speak but also those young enough to lack language to describe mental processes. Since the early 1960s, social psychologists had been developing indirect measures to avoid experimental artifacts of the psychological laboratory, such as demand characteristics and impression management. Social psychologists' work on attributions and cognitive psychologists' research on heuristics and biases described many judgment distortions rooted in memory. Had all these findings appeared a few decades later—after the start of implicit memory research—some of them could easily have been described as revealing implicit memory phenomena.

In their work on judgment and decision-making, Daniel Kahneman and Amos Tversky (especially Tversky & Kahneman, 1974) reported findings of subjects' responses to word problems that were inventively devised to reveal mental processes of which the subjects were unaware. An important subset of these findings challenged economists' long-standing conception of human rationality in decision-making and eventually led to work now identified as "behavioral economics." Tversky and Kahneman (1974) described their results as revealing "heuristics" and "biases," which they understood as mental shortcuts that could often deviate from rationality (p. 1124). Their *availability* heuristic was an effect of subjects' ease of retrieving category instances on their numerical estimates of probability or frequency. For example, one experiment had subjects review a list of names of well-known persons that included equal numbers of male and female names. When the women's names were selected to be more famous than those of the men in the list, subjects later judged that the list had more female than male names. Tversky and Kahneman identified as *anchoring* the effect on a numerical quantity estimate of exposure to an arbitrarily selected number. For example, when asked to judge whether 10 or 65 was too high or too low an estimate of the number of African nations in the United Nations and next to estimate the actual number of those nations, subjects who were initially asked to consider 10 produced numbers averaging 25, and subjects asked about 65 produced numbers averaging 45. These word problems devised by Tversky and Kahneman were effective indirect measures—they revealed effects of recent experiences while subjects answered questions that had no obvious relevance to those recent experiences.

Tory Higgins (Higgins, Rholes, & Jones, 1977) reported an innovative experiment in which subjects' first task obliged them to retain in memory a single word while completing each of a series of 10 perceptual judgments (a different word during each judgment). Four of the 10 memory words were trait adjectives. A subsequent, presumably separate "reading comprehension" experiment presented a paragraph that described "Donald" doing a series of activities. Each activity description was constructed to be am-

biguous in implications for Donald's characteristics. For example, Donald's activities of skydiving and crossing the Atlantic in a sailboat might be judged either as adventurous or as reckless. One of the trait adjectives held in memory in the initial perceptual task agreed with one of the ambiguous trait interpretations—some subjects had encountered "reckless," others "adventurous." Subjects' judgments of Donald, both immediately and after a delay of 10–14 days, were found to incorporate the (primed) traits at significantly above-chance rates. Higgins et al. (1977) credited their findings to the initial task's having made the relevant trait word *accessible* at the time of the judgment requested in the (presumably) unrelated second experiment.

Srull and Wyer (1979) devised a sentence-construction priming task in which subjects were asked to construct a series of sentences from sets of four words—for example: *leg break arm his* (which obliges a sentence describing a hostile act) or *the hug boy kiss* (which obliges a sentence describing a kind act). In two conditions, subjects received multiple four-word sets of either the hostile or kind type. They next proceeded to an experiment that included a task designed to reveal whether the quality embedded in the constructed sentences influenced trait judgments about the protagonist (Donald, again) of a series of vignettes that were ambiguous regarding hostility or kindness. Judgments of Donald's traits provided an indirect measure of the priming procedure's influence. Like Higgins et al.'s (1977) procedure, Srull and Wyer's procedure introduced an indirect measure that served to reveal one thing (a priming effect presumed to operate outside of awareness) by asking about something else (Donald's characteristics).

Theoretical Precursors

In Volume 3 of his *Handbook of Physiological Optics*, von Helmholtz (1867/1925) described "unconscious inference".[1] In 150 years' retrospect, Helmholtz's understanding of unconscious inference appears to capture the essence of the Implicit Revolution (Helmholtz's understanding is returned to at the end of this article). In 1895, Breuer and Freud (1895/1955) published *Studies on Hysteria,* which included five clinical case studies. Their first chapter contained the famous description "Hysterics suffer mainly from reminiscences" Breuer amplified that observation in his theoretical interpretation:

> If the memory of the psychical trauma must be regarded as operating as a contemporary agent, like a foreign body, long after its forcible entrance, and if nevertheless the patient has no consciousness of such memories or their emergence—then we must admit that *unconscious ideas exist and are operative* (emphasis in original). (p. 221)

This description connects readily to the modern concept of implicit memory. Breuer's reference to "psychical trauma "can be connected first with Claparède's pin-prick of his Korsakoff patient and then to the tamer experiences presented to subjects by Jacoby and Dallas (1981), Graf and Schacter (1985), and many others. Each produced an indirect indication of the impact of prior events, making use of behavior or verbal report that does not depend on memory for the past event(s).

More recent theoretical precursors in cognitive psychology date from the 1960s, including the work of Saul Sternberg (1966), Michael Posner and Ronald Mitchell (1967), and David Meyer (e.g., Meyer & Schvaneveldt, 1971). These were leaders in using decision latency measures to (indirectly) infer cognitive processes described in information-processing stage models (cf. Smith, 1968). These information-processing stage models grew out of investigations by Donald Broadbent and colleagues (e.g., Broadbent, 1958). Broadbent and numerous others investigated "preattentive" processes, which were understood to operate in parallel (meaning without capacity limitation) and without conscious awareness (meaning without ability to reveal via verbal report). Preattentive processes were contrasted with attentional processes, the latter being assumed to occur with limited mental capacity and with conscious awareness. In his review of the developing field of cognitive psychology, Ulric Neisser (1967) educated many developing researchers about the research by Broadbent and others that used indirect measures to reveal preattentive processing.

In the 1970s, frequent references by cognitive psychologists to "automatic" or preattentive processes indicated their readiness to investigate nonconscious mental phenomena at a time when *unconscious* was not yet a welcome term in academic psychology's scholarly publications (see Figure 1). The separation of preattentive processing from attentional processing evolved into a widely adopted two-process distinction first advanced by Posner and Snyder (1975), who called the two processes "automatic" versus "conscious" Shiffrin and Schneider's (1977) influential presentation of the distinction led to wide adoption of their labeling of *automatic* versus *controlled* processing. The automatic versus controlled distinction signaled an impending proliferation of "dual process" theories. A review of these by Stanovich, West, and Toplak (2014) listed 28 pairs of terms that had been introduced to describe two-process distinctions.

Start of the Revolution

A 1979 conference in which cognitive psychologists and clinical neuropsychologists discussed experimental investigations of amnesias led to the first publications presenting the type of work for which Graf and Schacter (1985) even-

[1] These few paragraphs only scratch the surface of a history that has been described more fully in portions of broader historical overviews of unconscious cognition by John Kihlstrom (1994, 2013) and in Schacter's (1987) historical review.

tually offered the *implicit memory* label. Researchers at the conference had identified an as-yet-unnamed class of memories that did not depend on recalling the experiences that established those memories. In the conference's published volume (Cermak, 1982), the chapter by Larry Jacoby most clearly developed and stated this conclusion: "As does the Korsakoff patient, normal subjects show effects in performance that are independent of recognition memory" (Jacoby, 1982, p. 112).

Within a few years after the 1979 amnesias-centered conference, journal articles began to present related conclusions about memory function in normal subjects. Jacoby and Dallas (1981) wrote: "Effects in perceptual recognition can be largely independent of the level of recognition memory performance" (p. 316). Jacoby and Witherspoon (1982) focused on demonstrating a "dissociation of memory and awareness [which required showing] that performance of a memory task that requires awareness and performance on a test that does not require awareness are independent of one another" (p. 303).

Publications in the early 1980s took up a debate about whether findings of dissociations of memory from awareness demanded identification of two conceptually (and perhaps physiologically) distinct systems for memories that, respectively, did and did not depend on conscious recollection. Although vestiges of that never-resolved debate persist, the debate never undermined understanding that implicit memory was demonstrated by an *empirical* distinction between (a) implicit measures that indicated memory *indirectly* in performance and (b) explicit measures that indicated memory *directly* in recollections of memory-forming experiences.

Spread of the Revolution to Social Cognition

The authors of this article took advantage of conferences and colloquia to encounter the early work on dissociations of indirect evidences of memory from awareness of the memory-establishing experiences. A visit by Larry Jacoby to our (then) academic home, Ohio State University, was especially influential. Nevertheless, it took almost a decade of incubation before influences of this work were apparent in our social cognition research. In summarizing the influences that were shaping our research in the 1990s, our article "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes" (Greenwald & Banaji, 1995) widened the scope of a not-yet-apparent revolution. We mention here just a few of what, with more than two decades' additional retrospect, are now identifiable as the influences that shaped that article.

In "Telling More Than We Can Know: Verbal Reports on Mental Processes," Richard Nisbett and Timothy Wilson (1977) drew together earlier work in support of a conclusion for which they quoted George Miller's (1962) summary description: "It is the *result* of thinking, not the process of thinking, that appears spontaneously in consciousness" (p. 56). At about the same time, Robert Zajonc was developing the ideas that led to his article "Feeling and Thinking: Preferences Need No Inferences" (Zajonc, 1980).[2] Both of these, from social psychological perspectives, were promoting the role of automaticity in social cognition.

After exposure to both Nisbett and Zajonc in his graduate studies at the University of Michigan, John Bargh developed rapidly as a leading voice in advancing the role of automaticity in social perception and cognition. Bargh's work started with a study in which brief, visually masked presentations of hostile words influenced hostility judgments of the already-famous Donald (Bargh & Pietromonaco, 1982). Bargh then collaborated with Higgins on an influential review article (Higgins & Bargh, 1987) and later produced a chapter (Bargh, 1994) that gave a comprehensive review of both the importance and the complexities of the cumulating research on automaticity in social cognition.

In 1986, Russell Fazio adapted cognitive psychologists' semantic priming methods (e.g., Meyer & Schvaneveldt, 1971; Neely, 1977) to measure "automatic activation of attitudes" (Fazio, Sanbonmatsu, Powell, & Kardes, 1986). In Fazio's procedure, subjects rapidly categorized (with computer key-presses) each of a series of *target* words as pleasant (e.g., *appealing, delightful*) or unpleasant (e.g., *repulsive, awful*). A *prime* word that was presented 300 ms prior to each target word was either positively valenced (e.g., *gift, music, party, cake*) or negatively valenced (e.g., *death, hell, guns, crime*). Subjects were asked to remember the prime word so they could repeat it aloud after their valence-classification response to the target word. Fazio et al. (1986) found that pleasant-meaning target words were correctly classified more rapidly if they followed pleasant-meaning prime words than if they followed unpleasant-meaning prime words. They also found that more strongly valenced primes produced larger priming effects. Variations of this "sequential priming" method have been in regular use since their 1986 publication.

Studies of automatic activation of racial stereotypes soon followed. Jack Dovidio (Dovidio, Evans, & Tyler, 1986) described a study demonstrating automatic stereotype effects. Patricia Devine's (1989) dissertation research extended the previously mentioned subliminal priming methods of Bargh and Pietromonaco (1982) to automatic stereotypes. Devine's article brought attention to the possibility of dissociation between automatic stereotype activation and controlled inhibition of stereotype expression. These early adoptions of indirect measures in social cogni-

[2] In a few decades' retrospect, the possibly unconscious "feeling" processes that Zajonc (1980) labeled *affect* are difficult to distinguish from what more recent researchers (including the present authors) identify as *implicit attitudes.*

tion research between the mid-1970s and the early 1990s used variants of priming methods, examining the influence of a first (prime) stimulus on the response to a second (test) stimulus. Magnitudes of these effects were found to decrease with separation of the test stimulus from the prime (e.g., Srull & Wyer, 1979), to increase with applicability of the prime stimulus to the task instructed for the test stimulus (e.g., Higgins et al., 1977), and to decrease when subjects were aware, at the time of testing, of the relation of the test stimulus to a previously encountered prime (e.g., Lombardi, Higgins, & Bargh, 1987).

Implicit Association Test

When the Greenwald and Banaji (1995) article appeared, the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) was receiving initial testing. In the IAT's procedure, subjects are asked to respond to four categories of stimuli with just two responses (usually two keys on a computer keyboard). In the first IAT, the four categories were flowers, insects, pleasant, and unpleasant, each represented by words. A surprisingly strong finding was that when one key was used for both flower names and pleasant-meaning words (and the other key for insect names and unpleasant-meaning words), responding averaged about 300 ms faster than if one key was for insects and pleasant and the other for flowers and unpleasant. The latency difference between these two combined tasks provided an indirect measure that Greenwald et al. (1998) described as measuring relative association strengths (e.g., p. 1477). Pairs of categories receiving faster responding were assumed to be more strongly associated than were those receiving slower responding. Greenwald et al.'s initial uses of the IAT measured associations of concepts with valence categories (attitudes) including the contrast of Black and White racial categories and the contrast of Korean and Japanese ethnic categories. Within a few years of its introduction, the IAT's method had also been adapted to measuring associations corresponding to stereotypes, identities, and self-esteem (Greenwald et al., 2002).[3]

The IAT's use as an indirect measure has roots in the traditions of semantic priming (Meyer & Schvaneveldt, 1971) and evaluative priming (Fazio et al., 1986). However, the IAT does not share the abstract priming structure of those earlier procedures—that is, it is not a measure of the effect of a first stimulus on the response to a second stimulus. Rather, it has a common conceptual ingredient with the Stroop (1935) effect, which also compares two tasks that differ in the way that stimuli are assigned to responses. In the Stroop task, subjects are asked to name colors of presented words. One condition of the Stroop task requires naming the colors in which words are printed when those words are names of other colors. Subjects respond slowly because of the difficulty of attending to the colors while inhibiting the overlearned response of reading the printed words that often name other colors. In the other condition, the words are the names of the colors in which they are printed (e.g., *red* printed in red) The latter is considerably easier, making use of overlearned pronunciation associations rather than requiring that the response-selection effects of those associations be inhibited.

The difference between speeds for the two versions of the Stroop task measures the mental effort to retain the instruction to respond to the word's print color rather than effects of the words themselves. The burden in the IAT is a similar burden on mental effort—for the IAT, the burden is to retain the instructions to give the same response to two nonassociated categories (e.g., flower and unpleasant). There is no such apparent burden when the two categories that require the same response are strongly associated (e.g., flower and pleasant).[4] The IAT allowed doing research on social psychology's traditional major topics indirectly—without asking research subjects to produce self-reports about their likes and dislikes (attitudes), their beliefs concerning characteristics of social groups (stereotypes), their self-concepts (identities), or their self-esteem.

Within a few years, research findings had established that IAT-measured attitudes and stereotypes were often either unwelcome to, or explicitly rejected by, research subjects. For example, an avowed feminist might discover via an IAT that he or she has a strong implicit (= indirectly measured) stereotype associating male with career and associating female with family, or a female scientist might discover an implicit stereotype associating male (more than female) with math and science, or a person with racially egalitarian beliefs might discover an implicit stereotype associating Black (more than White) with weapons or crime.

Implicit attitudes and stereotypes (i.e., attitudes and stereotypes revealed by indirect measures) may be established by formative experiences that occur outside of, and perhaps well prior to, formal educational experiences that plausibly play stronger roles in establishing their explicit (= directly measured) counterparts. Implicit attitudes and stereotypes may be acquired over many years from language and social experiences that cumulatively construct an overlearned repertoire of cultural expertise rooted in even more thousands of hours of experience than those invested by medical

[3] Demonstration versions of a variety of IAT measures are available at the Project Implicit Internet site (https://implicit.harvard.edu/).

[4] A relatively crude theoretical explanation of the IAT was first offered in 2002: "When the two concepts that share a response are strongly associated, the sorting task is considerably easier than when the two [are] weakly associated" (Greenwald et al., 2002, p. 8). The authors later discovered that, in 1973, David Meyer had proposed a serial information-processing stage model for a task in which subjects rapidly judged whether a stimulus word belonged to either of two categories (Meyer, 1973, p. 128). Meyer's (1973) model has several parameters that could explain greater speed when the two categories were semantically related than when they were unrelated.

doctors, virtuoso musicians, or world-class athletes in their professional training. This unsought cultural expertise, condensed into stereotypes and attitudes, may be as difficult or impossible to unlearn as are musical, medical, and athletic expertise.

Introspection: Conscious Judgment as Fallible Theory

Being aware of the long history of treatments of introspection (= self-inspection of thoughts and feelings) in philosophy, the present authors know better than to take up the topic of *truth* of introspective knowledge. They have no such qualms about considering the *validity* of introspection, which is amenable to empirical study. Most human perceivers intuit that conscious perceptions inform them validly about their surrounding physical and social world. When two objects appear to have different colors, perceivers assume they differ physically in color. Such intuitive faith in the validity of introspective knowledge no doubt extends to psychologists, many of whom know full well about perceptual illusions that provoke blatantly invalid perceptions, including seeing identically colored objects as being different in color.[5]

The same invalidity that is obvious in visual illusions is inherent in implicit attitudes and stereotypes but not nearly so obviously. Unlike those experiencing well-known visual illusions, ordinary social perceivers have no easy way to judge, from their social interactions, whether their stereotype-influenced perceptions may be invalid. However, Goldin and Rouse (2000) documented a stereotype-produced male = virtuoso illusion that became abundantly clear in retrospect, after major symphony orchestras in the United States adopted blind hiring auditions for open instrumental positions. Starting in the 1970s, applicants for open orchestra positions auditioned behind a screen, hiding everything but the auditionee's musical performance. The orchestra's selection committee could therefore not know whether the auditionee was male or female. The cumulative data of audition outcomes established that knowledge of an auditionee's gender must have unfairly influenced conscious judgments of women's performance prior to the innovation of blind auditions. Hiring of women instrumentalists increased from below 20% in 1975 to about 40% by 1990 (see Figure 3 in Goldin & Rouse's, 2000 article).[6] It could be concluded that, prior to the 1970s, women instrumentalists' performances were perceived (invalidly) as weaker when their gender was known than when it was not. They were (invalidly) disadvantaged by the male = virtuoso illusion.

This is not the place to review literature on stereotype validity and invalidity. For the present purposes, an alternative to such a review is a logic-based conclusion from the noncontroversial proposition that stereotypes are mental constructs that vary across persons. For example, a stereotype that associates male gender (more than female) with musical virtuosity varies in strength across people, including being reversed in direction (i.e., female = virtuoso) for some. Hypothesize that there exists a known and verifiably true population average of a measure of instrumental virtuosity difference between men and women. Assume also that perfectly accurate stereotype strength measures exist for individual persons (ignore that both assumptions are unrealistic—this is a thought experiment!). It would then, in principle, be possible to know the deviation of each person's measured stereotype level from the known true value of the virtuosity difference (if any). Necessarily, these individually measured deviations would be substantial for a great many people, meaning that the individual stereotypes held by those people were invalid. This conclusion that stereotypes can provoke invalid conscious judgments depends only on the (unarguable) assumption that there are individual differences in the stereotype.

Nisbett and Wilson (1977) concluded that human perceivers have valid introspective access to *results* of thought and perceptual processes, but that they lack introspective access to the mental processes that produce those thoughts and perceptions.[7] Nisbett and Wilson described this contrast as a distinction between mental *content* and mental *process*. Regarding mental content, they concluded that human perceivers have "private access" to content such as current sensations, emotions, evaluations, plans or intentions, and visual imagery (pp. 255–256). The problem, as they saw it, was with introspective access to the cognitive processes by which humans arrive at judgments and decisions:

> When people attempt to report on their cognitive processes, that is, on the processes mediating the effects of a stimulus on a response, they do not do so on the basis of any true introspection. Instead, their reports are based on a priori, implicit causal theories, or judgments about the extent to which a particular stimulus is a plausible cause of a given response. (p. 231)

Nisbett and Wilson concluded that subjects' "causal theories" about their mental processes could be valid if they

[5] See, for example, https://www.ted.com/talks/beau_lotto_optical_illusions_show_how_we_see.

[6] Goldin and Rouse (2000, p. 716) estimated that one third of the increase in hiring women was directly attributable to the effect of blind auditions. This is reasonably taken as a lower bound estimate because other candidate causes, especially increase in women instrumentalists graduating from music conservatories and changed attitudes of orchestral selection committees, could have been positively influenced by a culture changed by the innovation of blind auditions.

[7] Writing a year later in *Psychological Review*, John R. Anderson (1978) pointed to "a fundamental indeterminacy in deciding issues of [mental] representations" (p. 249). Referring to introspective reports of visual imagery, Anderson asserted that "skepticism about the[se] introspective reports is acceptable" (p. 271). This went beyond Nisbett and Wilson's (1977) skepticism about introspection, although Anderson still stopped short of doubting the validity of introspective reports of current perceptions and memories.

possessed causal theories that happened to be correct, but that this accuracy could not be attributed to introspective acumen.

The research that has accumulated since the mid-1980s suggests that Nisbett and Wilson's (1977) conclusion also applies to mental *content*. Here is a suggested extension of their just-quoted conclusion: *When people attempt to report on their conscious perceptions and judgments, they do so not based on valid introspection but by using traces of past (possibly biased) experience to construct (possibly invalid) theories of current data.*

This conclusion may appear novel, even radical, in relation to current psychological understanding of dissociation between (i.e., independence of) implicit and explicit measures. The present conclusion is that explicit measures are both informed by and (possibly) rendered invalid by unconscious cognition. This assertion is far from original here; it is only a restatement of the conclusion that Hermann von Helmholtz reached in articulating his conception of "unconscious inference" (*unbewusster Schluss*). In 1878, Helmholtz wrote,

> These inferences are unconscious insofar as their major premise . . . is formed from a series of experiences whose individual members have entered consciousness only in the form of sense impressions which have long since disappeared from memory. *Some fresh sense impression forms the minor premise, to which the rule impressed upon us by previous observations is applied* (emphasis added). (von Helmholtz, 1878/1971)

The "rule" to which Helmholtz referred can be the cause of a possibly invalid inference. In attributing unconscious inferences to a drama's audience members who judge that a skilled actor was experiencing the emotions portrayed on stage, Helmholtz showed his understanding that unconscious inference occurred for more than visual illusions; it extended to (conscious) judgments of others' mental states (von Helmholtz, 1867/1925, p. 28). These (extrospections?) are types of social judgments that occur in situations that allow operation of stereotypes, leading to judgments that can favorably or adversely affect others, such as in making hiring decisions or contributing to jury verdicts.

Concluding Observations and Continuing Agenda

Theoretical Understanding of "Introspection"

Perhaps many can properly say that they arrived long ago at the present authors' conclusion, which is that unconscious processes shape conscious judgments in often invalid ways. At the other extreme, many may reject the authors' view that validity of introspection should not be assumed. Although the Implicit Revolution has prompted a conclusion that Helmholtz reached in the mid-19th century, it has not produced a theoretical understanding of the mental structures that support the mental process called introspection. The authors label these presumed mental structures as *associations,* using the term *association* as a theory-uncommitted placeholder label (cf. Greenwald, Nosek, Banaji, & Klauer, 2005). Until this placeholder label is elaborated, perhaps with the aid of neuroscience methods, associations might be understood as mental pigments that operate in combination to construct rich mental images and judgments. A more psychological metaphor is that a mass of associative knowledge acts as a *cultural filter* that elaborates perception and judgment, in ways that can vary across persons when cultural environments have constructed the associative mass idiosyncratically. Some future replacements for these metaphors may provide testable theories of the mental processes that now have the placeholder label *introspection.*

Validity and Use of Self-Report Measures

In using implicit measures, social cognition researchers are often expected (by editors and reviewers) to demonstrate that their measures of attitude, stereotype, or self-esteem validly capture those constructs. Research of the last few decades, using these implicit measures in company with parallel self-report (explicit) measures, has made it increasingly clear that self-report measures often need similar validation.

The first meta-analysis of predictive validity of IAT measures (Greenwald, Poehlman, Uhlmann, & Banaji, 2009) examined the predictive validity of IAT and self-report attitude measures as a function of magnitude of correlation between the two. Although these implicit–explicit correlations were often of small to moderate magnitude (see also Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005; Nosek, 2005), the meta-analysis found that when these correlations were large, both types of measure were well correlated with behavioral indicators of the same attitudes.[8] The plausible interpretations of the more common pattern of weak implicit–explicit correlations are that (a) implicit and explicit measures tap distinct constructs or (b) they might be affected differently by situational influences in the research situation (cf. Fazio & Towles-Schwen, 1999; Greenwald et al., 2002) or (c) at least one of the measures, plausibly the self-report measure in many of these cases, lacks validity.

[8] Although the topic is not of direct concern in this article, readers may be interested in a debate about magnitudes of IAT measures' magnitudes of correlation with discrimination-relevant behavior. In their meta-analysis, Oswald, Mitchell, Blanton, Jaccard, and Tetlock (2013) found smaller correlations than those found by Greenwald et al. (2009). As pointed out by Greenwald, Banaji, and Nosek (2015), the magnitudes of correlation found in both meta-analyses "were large enough to explain discriminatory impacts that are societally significant either because they can affect many people simultaneously or because they can repeatedly affect single persons" (Greenwald et al., 2015, p. 553)

What Is Changing in the Understanding of "Unconscious"?

For the last 30 years, researchers have used indirect measures to observe effects of prior experiences for which subjects often had no memory on conscious judgments and actions. This still rapidly growing body of research calls for a revised conception of the relation between unconscious cognition and conscious action. Helmholtz's *unbewusster Schluss* (unconscious inference or unconscious conclusion) is a good starting place for this revised conception. Although Helmholtz did use the language of logical reasoning to explain his concept (see the earlier quote from von Helmholtz, 1867/1925), neither then nor now is there any indication of logic-like reasoning associated with the phenomena Helmholtz was explaining. Some other account of the contribution of the unconscious process to conscious judgment is needed.

Theories of the relationship between unconscious and conscious cognition fall into three categories: (a) *superordinate* unconscious cognition, a high level of unconsciously operating intelligence, capable of guiding the person's thought and actions (e.g., Freud, 1923/1961); (b) *coordinate* unconscious cognition, operating in parallel with but independently of conscious cognition; and (c) *subordinate* unconscious cognition, occurring prior to conscious attention, as in information-processing stage theories (see discussion of these in Greenwald, 1997).

Phenomena of implicit memory and implicit social cognition do not readily fit with any of the three just-described types of existing conceptions of unconscious–conscious relationship. Nevertheless, the subordinate-unconscious conception, which theorizes relatively simple unconscious processes preceding conscious attention and action, can be a useful starting point. At the same time, the well-known serial-stage information-processing form of the subordinate view offers no conception of how unconscious preattentive processes can produce complex conscious judgments. Existing theory therefore does not yet explain (a) what types of varied mental structures are produced by varied histories of experience with social stimuli or (b) how those (still-to-be-identified) structures shape variations across persons in the conscious perception of a given stimulus situation.

For perceptual illusions, one can assume that all perceivers have experienced similar perceptual histories with the stimuli used in the illusion; the illusion can then be explained entirely as an effect of some well-designed change in stimulus context relative to that perceptual history. For social illusions, such as the male = virtuoso illusion described earlier, it is just the reverse. Social experience histories are presumed to vary among people, producing attitudes, stereotypes, and self-concepts that differ among persons; variations in conscious perception or judgment in response to the same social stimulus can therefore be attributed to variations in the mental structures resulting from the differing past histories of social experiences.

Final Words

The sequence of introduction of innovative research methods producing previously unobservable findings that in turn demand new theory occurs regularly in scientific work (cf. Greenwald, 2012). The sequence can be judged revolutionary when the resulting new theory deviates sufficiently from existing understanding. This article's proposition that a conceptual revolution is now in progress may not be fully convincing until the theoretical challenges described in this article are effectively addressed.

References

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review, 85,* 249–277. http://dx.doi.org/10.1037/0033-295X.85.4.249

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer, Jr., & T. K. Srull (Eds.), *Handbook of social cognition: Basic processes; Applications* (Vols. 1–2; 2nd ed.; pp. 1–40). Hillsdale, NJ: Erlbaum.

Bargh, J. A., & Pietromonaco, P. (1982). Automatic information processing and social perception: The influence of trait information presented outside of conscious awareness on impression formation. *Journal of Personality and Social Psychology, 43,* 437–449. http://dx.doi.org/10.1037/0022-3514.43.3.437

Breuer, J., & Freud, S. (1955). Studies on hysteria. In J. Strachey (Trans.), *The standard edition of the complete psychological works of Sigmund Freud, Vol. II (1893–1895).* London, United Kingdom: Hogarth Press. (original German publication, 1895)

Broadbent, D. E. (1958). *Perception and communication*. London, United Kingdom: Pergamon.

Cermak, L. (Ed.). (1982). *Human memory and amnesia*. Hillsdale, NJ: Erlbaum.

Claparède, E. (1951). Recognition and "me-ness" In D. Rapaport, *Organization and pathology of thought: Selected sources* (pp. 58–75). http://dx.doi.org/10.1037/10584-003 (Original work published 1911)

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56,* 5–18.

Dovidio, J. F., Evans, N., & Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology, 22,* 22–37.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50,* 229–238. http://dx.doi.org/10.1037/0022-3514.50.2.229

Fazio, R. H., & Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 97–116). New York, NY: Guilford Press.

Freud, S. (1961). The ego and the id. In J. Strachey (Trans. & Ed.) *The standard edition of the complete psychological works of Sigmund Freud* (Vol. 19, pp. 3–66). London, United Kingdom: Hogarth. (Original work published 1923)

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review, 90,* 715–741. http://dx.doi.org/10.1257/aer.90.4.715

Graf, P., & Schacter, D. L. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, 11, 501–518. http://dx.doi.org/10.1037/0278-7393.11.3.501

Greenwald, A. G. (1992). New Look 3: Unconscious cognition reclaimed. *American Psychologist, 47,* 766–779. http://dx.doi.org/10.1037/0003-066X.47.6.766

Greenwald, A. G. (1997). Self-knowledge and self-deception: Further consideration. In M. S. Myslobodsky (Ed.), *The mythomanias: An inquiry into the nature of deception and self-deception* (pp. 51–71). Mahwah, NJ: Erlbaum.

Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science, 7,* 99–108. http://dx.doi.org/10.1177/1745691611434210

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4–27. http://dx.doi.org/10.1037/0033-295X.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. *Journal of Personality and Social Psychology, 108,* 553–561. http://dx.doi.org/10.1037/pspa0000016

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109,* 3–25. http://dx.doi.org/10.1037/0033-295X.109.1.3

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74,* 1464–1480. http://dx.doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., Banaji, M. R., & Klauer, K. C. (2005). Validity of the salience asymmetry interpretation of the Implicit Association Test: Comment on Rothermund and Wentura (2004). *Journal of Experimental Psychology: General, 134,* 420–425. http://dx.doi.org/10.1037/0096-3445.134.3.420

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,* 17–41. http://dx.doi.org/10.1037/a0015575

Higgins, E. T., & Bargh, J. A. (1987). Social cognition and social perception. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (Vol. 38, pp. 369–425). Palo Alto, CA: Annual Reviews.

Higgins, E. T., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology, 13,* 141–154. http://dx.doi.org/10.1016/S0022-1031(77)80007-3

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31,* 1369–1385. http://dx.doi.org/10.1177/0146167205275613

Jacoby, L. L. (1982). Knowing and remembering: Some parallels in the behavior of Korsakoff patients and normals. In L. Cermak (Ed.), *Human memory and amnesia* (pp. 97–122). Hillsdale, NJ: Erlbaum.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541. http://dx.doi.org/10.1016/0749-596X(91)90025-F

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110,* 306–340. http://dx.doi.org/10.1037/0096-3445.110.3.306

Jacoby, L. L., & Witherspoon, D. (1982). Remembering without awareness. *Canadian Journal of Psychology, 36,* 300–324. http://dx.doi.org/10.1037/h0080638

Kihlstrom, J. F. (1994). The rediscovery of the unconscious. In H. Morowitz & J. L. Singer (Eds.), *The mind, the brain, and complex adaptive systems* (pp. 123–143). Reading, MA: Addison Wesley.

Kihlstrom, J. F. (2013). Unconscious processes. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 176–186). http://dx.doi.org/10.1093/oxfordhb/9780195376746.001.0001

Lombardi, W. J., Higgins, E. T., & Bargh, J. A. (1987). The role of consciousness in priming effects on categorization: Assimilation versus contrast as a function of awareness of the priming task. *Personality and Social Psychology Bulletin, 13,* 411–429. http://dx.doi.org/10.1177/0146167287133009

Meyer, D. E. (1973). Correlated operations in searching stored semantic categories. *Journal of Experimental Psychology, 99,* 124–133. http://dx.doi.org/10.1037/h0034763

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90,* 227–234. http://dx.doi.org/10.1037/h0031564

Miller, G. A. (1962). *Psychology: The science of mental life*. New York, NY: Harper & Row.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General, 106,* 226–254. http://dx.doi.org/10.1037/0096-3445.106.3.226

Neisser, U. (1967). *Cognitive psychology*. New York, NY: Appleton-Century-Crofts.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259. http://dx.doi.org/10.1037/0033-295X.84.3.231

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General, 134,* 565–584. http://dx.doi.org/10.1037/0096-3445.134.4.565

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,* 171–192. http://dx.doi.org/10.1037/a0032734

Posner, M. I., & Mitchell, R. F. (1967). Chronometric analysis of classification. *Psychological Review, 74,* 392–409. http://dx.doi.org/10.1037/h0024913

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 55–85). Hillsdale, NJ: Erlbaum.

Reingold, E. M., & Merikle, P. M. (1988). Using direct and indirect measures to study perception without awareness. *Perception & Psychophysics, 44,* 563–575. http://dx.doi.org/10.3758/BF03207490

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 501–518. http://dx.doi.org/10.1037/0278-7393.13.3.501

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review, 84,* 127–190. http://dx.doi.org/10.1037/0033-295X.84.2.127

Smith, E. E. (1968). Choice reaction time: An analysis of the major theoretical positions. *Psychological Bulletin, 69,* 77–110. http://dx.doi.org/10.1037/h0020189

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,* 1660–1672. http://dx.doi.org/10.1037/0022-3514.37.10.1660

Stanovich, K. E., West, R. F., & Toplak, M. E. (2014). Rationality, intelligence, and the defining features of Type 1 and Type 2 processing. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 80–91). New York, NY: Guilford Press.

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153,* 652–654. http://dx.doi.org/10.1126/science.153.3736.652

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643–662. http://dx.doi.org/10.1037/h0054651

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131. http://dx.doi.org/10.1126/science.185.4157.1124

von Helmholtz, H. (1925). *Handbook of physiological optics* (Vol. 3; J. P. C. Southall, Trans.). New York, NY: Optical Society of America. (Original work published 1867)

von Helmholtz, H. (1971). The facts of perception. In R. Kahl (Ed.), *Selected writings of Hermann Helmholtz.* Middleton, CT: Wesleyan University Press. (original German publication, 1878)

Warrington, E. K., & Weiskrantz, L. (1970). Amnesic syndrome: Consolidation or retrieval? *Nature, 228,* 628–630. http://dx.doi.org/10.1038/228628a0

Warrington, E. K., & Weiskrantz, L. (1974). The effect of prior learning on subsequent retention in amnesic patients. *Neuropsychologia, 12,* 419–428. http://dx.doi.org/10.1016/0028-3932(74)90072-4

Weiskrantz, L., & Warrington, E. K. (1979). Conditioning in amnesic patients. *Neuropsychologia, 17,* 187–194. http://dx.doi.org/10.1016/0028-3932(79)90009-5

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35,* 151–175. http://dx.doi.org/10.1037/0003-066X.35.2.151

Received July 7, 2017
Revision received October 3, 2017
Accepted October 4, 2017 ■