To be presented at the 27th biennial Philosophy of Science Association meeting.

Epistemic Risk in the Triangulation Argument for Implicit Attitudes

Morgan Thompson

One important strategy for dealing with error in our methods is triangulation, or the use multiple methods to investigate the same hypothesis. Current accounts of triangulation focus the conditions under which it succeeds, but ignore the many ways it can fail in practice. Instead, I argue that an account of triangulation focused on epistemic risk is better able to describe how triangulation fails and to normatively guide future triangulation research.

In this paper, I defend the claim that a useful account of methodological triangulation needs to account for the ways triangulation is susceptible to failure in its practice rather than focusing primarily on how and why it succeeds in ideal cases. A theory or account of a practice should highlight potential failures in order to be useful. Consider some ethical theory that gives an account of right and wrong actions. In order to use this ethical theory to guide my actions, I need to know not just what makes an action right or wrong, but also some features of my moral psychology. What are the ways that I am likely to err? Should I be worried about having a weak will and lacking follow through for actions that I deem right? Knowledge of the ways in which I might err allows me to better use the ethical theory to guide my actions. Analogously, I argue that an account of triangulation that is useful in practice ought to explain not just why triangulation is successful in ideal cases, but also how it can fail in practice. To do so, I will appeal to the idea of epistemic risk from the literature on the types and roles of values in science, medicine, and technology. By identifying types of failure, this lays the groundwork for future normative work developing strategies to avoid or mitigate these risks in triangulation research.

1.1 Methodological Triangulation

Methodological triangulation involves the use of multiple methods to examine the same research question. Current accounts of triangulation are cashed out in terms of its success.[1] One view of triangulation sets out to: "identify at an abstract level the logic behind successful robustness arguments [and…] to determine what is required for a specific form of robustness analysis to be successful" (Kuorikoski and Marchionni 2016, 230). On another view, triangulation is defined as: "the use in empirical practice of multiple means of investigation to validate an experimental outcome" (Schickore and Coko 2013, 296). Current accounts agree on two success criteria: (i) the methods employed need to be sufficient diverse and (ii) the methods need to produce data about the same phenomenon.

How would this received view of triangulation account for cases of failure in practice? There is substantial discussion of the failure to have sufficiently diverse methods (i), which is what Wimsatt (1981) called "illusory robustness." Still these accounts of diversity are based on successful cases of triangulation (e.g., Schupbach 2018).

We can also consider the other success criterion in triangulation: that each method produces data about the same phenomenon (ii). While most philosophers working on triangulation recognize that this is a success criterion, relatively little has been said about how researchers can *know* they

[1] One exception is Stegenga (2009) who considers various problems with the use of triangulation as a strategy to deal with the problem of epistemic uncertainty in science. However, many of his critiques are not internal to the practice of triangulation. Stegenga's main concern is that philosophical accounts of triangulation provide no guidance when evidence both confirms and disconfirms the same hypothesis. But most centrally to this paper, Stegenga does not examine the epistemic risks triangulation arguments are subject to when they *appear* to be successful. These potential errors are all the more suspect because they mascaraed as successes.

have met this criterion.[2] Even less has been said about how researchers can fail to meet this success criterion.

1.1.1 Epistemic Risk

In order to flesh out an account of triangulation that explains how it can fail in practice, I appeal to the concept of epistemic risk, which is “any risk of epistemic error that arises anywhere during knowledge practices” (Biddle and Kukla 2017, 218). There are many types of epistemic risk that occur at different parts of the research process. The most discussed kind of epistemic risk is inductive risk (Douglas 2016), which is particularly predominant in discussion about the role of values in science, medicine, and technology. Although the name implies it is any risk in inductive inferences, it is a technical term that refers specifically to the risk in inductive inferences from evidence to acceptance or rejection of a hypothesis.

Following Biddle & Kukla (2017), I hold that focusing exclusively on inductive risk makes our philosophical accounts of epistemic risk deficient. Other types of epistemic risk include the risk in deciding whether to characterize some datum as evidence for a hypothesis, such as whether some particular slide contains tumors and whether the tumors were malignant (Biddle's (2016) interpretation of Douglas 2000, 569). Another example is risk in the inference from animal models to the target system of interest (usually in humans) as in research on exposure to bisphenol A in a particular rat model (Biddle's (2016) interpretation of Wilholt 2009).

[2] One exception is Kuorikoski and Marchionni (2016), who argue that triangulation primarily consists in justifying data-to-phenomena inferences. Relying on Bogen and Woodward (1988), Kuorikoski and Marchionni argue that researchers can use empirical reasoning to justify these inferences, such as intervening on the phenomenon to determine whether there are corresponding differences in the data. While I think their view is on the right track, it is (1) susceptible to the criticism of not explaining why triangulation sometimes fails and (2) does not provide a sufficiently developed account of the practice of triangulation. I aim to rectify these two issues here.

Current accounts of triangulation focused on success can only account for two types of epistemic risk: the failure to have sufficiently diverse methods (or Wimsatt's "illusory robustness") and, on my view, inductive risk. I will argue that an account of triangulation that explains failure will need to make use of epistemic risk more broadly as not all instances fall neatly under the risk of illusory robustness or inductive risk.

1.1.2 Schema for Triangulation in Practice

In order to develop an account of triangulation that highlights points of failure, I turn away from abstract success conditions and to the details of knowledge production via triangulation. I highlight important steps in the practice of triangulation from the causal production of data to its transition to playing an evidential role to the increased credence in some hypothesis. In this section I provide a schema for the practice of triangulation.

Let me first distinguish between data and phenomena (Bogen and Woodward 1988). Data are publicly observable reports that result from experimental or observational processes. They are not repeatable because they are the actual reports produced through experimentation or observation. Phenomena on the other hand are stable patterns in the world. Phenomena are often not directly observable and are characterized and explained by theory.

In the practice of triangulation, researchers identify multiple methods that are likely to produce data relevant to the same phenomenon. Each method may include some sources of error, such as random error from sampling or systematic error due to the instruments and procedures of the method. Unfortunately, researchers are often unaware of all sources of error in their methods. And these errors causally impact what data is produced. Yet, it is this data produced by imperfect methods that is the input for our inferential reasoning.

Here let me make a further distinction between data and evidence. Rather than thinking of evidence as a separate kind of entity, we can think of it as a role that data play in confirming or disconfirming some hypothesis. In some cases of triangulation, this step may not be trivial: when data is produced in radically different experimental and theoretical contexts, many assumptions may be required to get from these different datasets to evidence that bears on (some particular) hypothesis. This problem about the evidential role of data is what Stegenga (2009) calls this the problem of incongruity.

Consider also that the data may be used as evidence in relation to multiple hypotheses. That is, despite of the fact that it may have been collected with some particular purpose in mind, it can serve as evidence for or against other hypotheses. In the case of triangulation, we're interested only in data that can be used as evidence for the same hypothesis. I'll focus on hypotheses about the existence of a phenomenon, though triangulation can also be used to estimate parameters and constants (e.g., Avogadro's number). At this point in the practice of triangulation, it needs to be demonstrated that all of the diverse datasets can serve as evidence for or against the *same* hypothesis.

Then once the evidential role of the datasets with respect to the same hypothesis has been established, researchers can make an inference to accept or reject the hypothesis. Even if all of the datasets provide supporting evidence for the hypothesis, a judgement still needs to be made about whether sufficient evidence has been collected to accept the hypothesis.

Theory can help reduce the uncertainty for some cases of triangulation. If researchers are triangulating on a claim about the existence of a phenomenon, then they should use some theoretical characterization of that phenomenon that describes its features. Researchers need a sufficiently developed characterization of a phenomenon in order to distinguish between inferences to the phenomenon of interest from inferences to other phenomena.

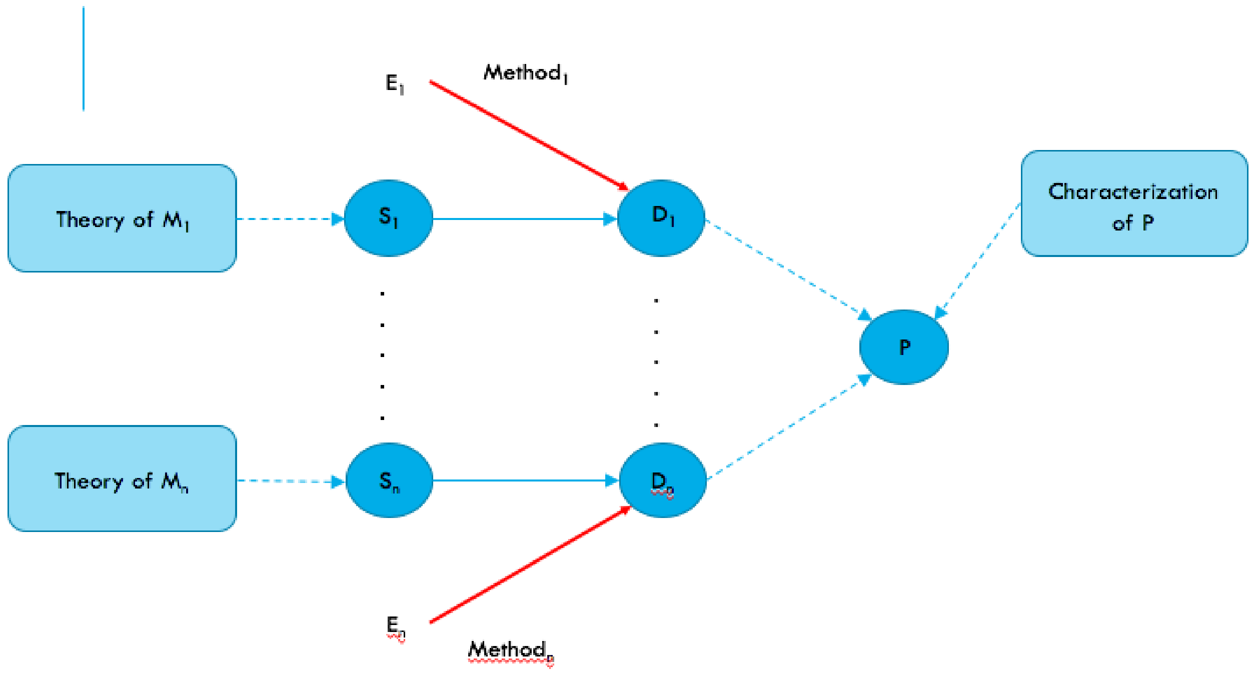

Figure 1. Schema of Triangulation

1.2 Triangulation in Implicit Social Cognition

Now that I've described the process of triangulation, I will demonstrate how it locates different types of epistemic risk. To do so, I will analyze the triangulation argument for implicit attitudes in social psychology.

By the mid-1990s, the majority of participants in psychology studies no longer self-reported holding explicitly racial attitudes (e.g., Dovidio and Gaertner 2000). In fact, many participants began to view racist acts as socially unacceptable and avoided committing racist actions themselves (Sue 2010). Yet, widespread racially discriminatory practices and racial disparities in economic, social, and health spheres persisted. Social psychologists posited that an explanation for these apparently contradictory features was that individuals still held racially biased attitudes, but that they were not reporting them when asked directly about their attitudes. So, researchers developed new techniques to control for the social desirability of appearing egalitarian (e.g., the "bogus pipeline" Jones and Sigall 1971). Indirect measures get around participants' ability and motivation to present themselves in a particular way to the researchers and instead measure their less controlled responses. As a result, researchers posited 'implicit attitudes' as a mental state or process. Implicit attitudes are automatically activated evaluative judgments about which participants are typically unaware or unable to control.

1.2.1 The IAT and the Evaluative Priming Task

The study of implicit attitudes bloomed. There are now nearly two dozen methods for measuring implicit attitudes. The two initial and most well-developed of these methods are the Implicit Association Test (IAT) (e.g., Greenwald, McGee, and Schwartz 1998) and the evaluative priming task (EPT) (e.g., Fazio et al. 1986). I discuss each in turn.

During a racial IAT, participants view stimuli from four categories: two racial groups and two evaluative groups. On any trial, each racial group is paired with a different evaluative category and these pairing are displayed on either side of the display screen. On typical racial IATs, two of the categories are stimuli related to two racial groups (e.g., faces of White and Black individuals) and two of the categories are evaluative stimuli (e.g., positive and negative words). Participants are asked to quickly categorize stimuli by pressing one of two keys on the corresponding to the disjunctive categories listed on the right and left sides of the display. Researchers can compare participants' reaction times on trials in which Black-positive and White-negative are paired to those in which Black-negative and White-positive are paired. A faster response time to the latter compared to the former is thought to indicate racial attitudes that more closely link Black people with negative concepts and White people with positive concepts (e.g., Mitchell, Nosek, and Banaji 2003).

Evaluative priming tasks instead use stimuli from the categories of interest to prime participants before participants perform a categorization task on unrelated evaluative target stimuli. If researchers are interested in racial attitudes, they might use images of Black or White people to prime participants. Then during the categorization task, participants are asked to categorize positive- and negative-valence words (target stimulus). Researchers reason that reaction times on the categorization task will be influenced by the evaluative valence of the prime stimulus. If a participant holds negative attitudes towards White people, then after viewing a White stimulus prime, they will categorize negative target words more quickly than positive target words.

1.2.2 The Triangulation Argument for Implicit Attitudes

Social psychologists take indirect measures like the IAT and EPT to triangulate on the same phenomenon—implicit attitudes. Over time, theories about how to characterize implicit attitudes have changed, but the assumption that the triangulation argument for implicit attitudes is successful has remained. Here I will offer some evidence for this claim.

Discussing the views of the field at the time in a review article on the nature of implicit attitudes, Gawronski, Hofmann, and Wilber (2006, 486; citations removed) state:

> A widespread assumption underlying the application of indirect measures is that they provide access to unconscious mental associations that are difficult to assess with standard self-report measures. Specifically, it is often argued that self-reported (explicit) evaluations reflect conscious attitudes, whereas indirectly assessed (implicit) evaluations reflect unconscious attitudes.

While Gawronski and colleagues go on to critique this widespread assumption (at least, its attribution of ‘unconscious’ to implicit attitudes), this quote demonstrates the ubiquitous assumption among implicit attitude researchers that first-generation indirect methods measured implicit attitudes.

More recently social psychologists have developed a neutral characterization of implicit attitudes that does not commit to any particular view of ‘implicit’. This is to broadly accommodate issues that participants are able to predict the evaluative direction of their implicit attitudes (Hahn et al. 2014). As Greenwald and Lai write in a review article this year, “The currently dominant understanding of “implicit” among social cognition researchers is “indirectly measured.” The labels “indirectly measured attitude” and “implicit attitude” are used interchangeably in this review” (Greenwald and Lai 2020). Still the assumption remains: whatever indirect measures are measuring, it is the same phenomenon.

1.3 Two Epistemic Risks in Triangulation

In this section, I use my account of triangulation to highlight two examples of epistemic risks and where they arise in implicit attitude research. My account better explains what goes wrong in these cases than accounts of triangulation focused on success. That is, my account provides a better descriptive account of scientific practice, where triangulation does not always succeed. Here I identify two types of epistemic risk: (1) epistemic risk when data is taken to be evidence for some hypothesis and (2) inductive risk in determining a sufficient level of evidence for the acceptance or rejection of a hypothesis.

1.3.1 Moving from Data to Evidence

One major epistemic risk in triangulation is that we may mistakenly think that the different datasets can serve as evidence for the same hypothesis. We are particularly at risk of this error when we do not justify the claim that our methods measure aspects theoretically related to the same hypothesis. Data do not automatically bear on hypotheses. A datum can be an image from electron microscopy, a mark selecting an answer on a survey, or recorded video of a researcher interacting with participants. So, data needs to be interpreted in relation to the hypotheses for which they may serve as evidence. In doing this, researchers must infer on the basis of data and some assumptions to the confirmation or disconfirmation of a hypothesis.

I argue that this epistemic risk is relevant to the triangulation argument for implicit attitudes. The data produced and current assumptions in social psychology do not support the claim that the data produced by the IAT and EPT serve as evidence for the same hypothesis. In fact, according to some implicit attitude researchers, they serve as evidence for slightly different hypotheses.

In IAT studies, the categories of interest are made explicit to the participant as the categories must be identified and paired to perform the categorization task. Thus, IAT scores are thought to measure attitudes toward the general social category. Thus, they can serve as evidence for hypotheses about associations between evaluative categories and social categories.

In an evaluative priming task, on the other hand, the instructions do not explicitly determine the relevant categorical membership of the priming stimulus. It is generally accepted that due to this feature, evaluative priming tasks measure attitudes toward the stimuli rather than the category (Olson and Fazio 2003; Mitchell, Nosek, and Banaji 2003). Consider that the priming stimulus is often an image of a person's face. Researchers may wish to contrast Black and White faces as priming stimuli in an evaluative priming task; however, as a feature of the images individuals represented will also belong to other social categories (e.g., attractiveness, gender). Because the categorization task is only along the evaluative dimension, it is not made salient which of these categories a participant is responding to. Consider the case of a participant who when primed with a particular image of a Black face, categorizes positive stimuli more slowly than when primed with an image of a White face. The response discrepancy could be caused by a negative evaluations of the person-represented-in-the-image's perceived race, attractiveness, perceived gender, or any combination of these and other features.

Good task design will control for these differences as much as possible, but due to the design of the task, it is impossible to identify what features influence the participant's reaction times in the categorization task in any given case. The features that cause a response discrepancy may change over time even for the same participant because implicit attitudes are thought to be context dependent (Jost 2019) and the empirical findings that indirect measures generally have low test-retest validity (Bosson, Swann, and Pennebaker 2000).

In order to address this epistemic risk, researchers need to provide justification for the claim that the IAT and EPT produce data that can serve as evidence for the hypothesis that participants have a negative association with the social category of interest. For the IAT, this justification already exists. For the EPT, it is less obvious. So, using my account of triangulation, I have highlighted a particular weak point in the triangulation argument for implicit attitudes and emphasized a place for the development and elaboration of norms for successful triangulation. Note also that this epistemic risk does not fit neatly under the heading "illusory robustness" or inductive risk because the problem arises due to the differences in the methods and does not involve a judgement about accepting or rejecting a hypothesis.

1.3.2 Inductive Risk

Once we know data can serve as evidence for the same hypothesis, we can ask: How do researchers know there is sufficient evidence to accept the hypothesis? On my view, the epistemic risk of error here is best characterized as inductive risk. However, in the context of triangulation inductive risk takes a particular form. Specifically, researchers ought to be concerned about the risk of accepting the hypothesis when it is false. In cases where our hypothesis is about the existence of some phenomenon (as triangulation is often used), the inductive risk may be specifically sensitive to the error that data produced (and their evidential support) are actually for distinct phenomena. In other words, there is an inductive risk in accepting the hypothesis that some phenomenon of interest exists on the basis of triangulation, especially when we have not sufficiently ruled out the possible hypothesis that multiple phenomena are differentially driving the results.

Psychologists evaluate the validity of their tests using psychometrics. Relevant to my arguments, convergent validity is the extent to which two methods that are predicted to measure the same phenomenon are in fact measuring the same phenomenon. Low convergent validity suggests that two methods measure different phenomena. Psychologists often assess convergent validity by examining correlation coefficients.[3] If two methods measure the same phenomenon, they are expected to have high correlations in their scores. However, given that the two methods are distinct in some ways, there should not be a perfect correlation in their scores. There is no

[3] Other methods such as the multi-trait multi-method matrix (Campbell and Fiske 1959) have been used less frequently and less completely in the context of implicit attitudes.

well accepted threshold for what counts as sufficiently high convergent validity. But social psychologists hold that the IAT and EPT ought to have high convergent validity (e.g., Banaji 2001).

Unfortunately, researchers have found low correlations between the IAT and other implicit measures and thus, low convergent validity (Fazio and Olson 2003). The correlation in scores for the IAT and EPT range between r=.24 and r=.13. These are very low positive correlations. So, a participant's score on the IAT provides very little information about their EPT score, and vice versa.

One possible cause of the low correlations between IAT and EPT scores is the low reliability of EPT (De Houwer et al. 2009). Perhaps the scores do not correlate well due to noisiness in the data produced by unreliable methods rather than the methods measuring different phenomena. A recent comparison of seven indirect measures of attitudes Bar-Anan and Nosek (2014), the EPT had weak correlations with other indirect measures (including the IAT, r=.24).

However, there are two reasons to remain neutral with respect to these explanations. First, a measure need not be reliable for it to be valid (Borsboom, Mellenbergh, and Van Heerden 2004). The measure could track a context-dependent phenomenon, of which implicit attitudes is probably an example (Jost 2019). Second, as Bar-Anan and Nosek (2014, 677, original emphasis) suggest, low convergent validity and low reliability may *both* contribute to the low correlations of scores on indirect measures of attitudes:

> the most likely explanation for this pattern, coupled with the similar rank ordering for internal consistency, is that [Affective Misattribution Priming] and EPT are both relatively distinct, and *also* less effective in reliably assessing the target evaluation than are the other measures. […] it could still be the case that both measures assess unique components of evaluation that are not assessed by other indirect measures (including each other).

Still one promising finding is that unlike the Affective Misattribution Priming task, Bar-Anan and Nosek (2014) do not find a strong correlation between the EPT and direct measures of racial attitudes (i.e., self-report on surveys), which would have indicated the potential influence of deliberate evaluation in the indirect measurement. So, while some of the low correlations between the measures may be due to the low reliability of the EPT, it is possible that both low reliability and low convergent validity are part of the picture.

1.3.3 Why can't these be understood as a failure of diversity?

One potential objection is that the IAT and EPT are not sufficiently diverse methods. The basic idea is that whatever diversity criterion we accept (see Schupbach 2018), the IAT and EPT are too similar to count as distinct methods for the purposes of triangulation. I respond to this objection by clarifying that these methods historically descendant from different theories in psychology. In addition to my arguments that they produce data relevant to different hypotheses (section 1.3.1), this gives us some reason to think the methods are sufficiently diverse on any appropriate diversity criterion.

The two methods I discuss were developed out of different historical traditions in psychology (Payne and Gawronski 2015). Drawing on Shiffrin and Schneider's (Shiffrin and Schneider 1977) work on selective attention, Fazio and colleagues (Fazio, Jackson, Dunton, & Williams, 1995) developed the evaluative priming task to distinguish automatic and controlled processing. Controlled processing requires attention and can be altered voluntarily, whereas automatic processing takes place on memories stored in long-term memory, is automatically activated given the appropriate inputs, and is difficult to suppress.

Greenwald and Banaji's (1995) work on implicit attitudes came out of cognitive psychological research on implicit memory, which describes the way that earlier experiences can influence current performance on learned tasks without conscious awareness of the past experiences. Most famously, the patient H.M., who had a medial temporal lobectomy and thus lacked bilateral hippocampi and other structures, was unable to create new episodic memories. However, H.M. demonstrated the formation of new implicit memories through the time-savings in relearning motor skill tasks (Corkin 2002). As Greenwald et al. (1998) constructed it, the IAT is a measurement of implicit memory. So, both measures were designed based on different theories. In short, the evaluative priming task was designed to measure a construct that is typically uncontrolled or automatic while the IAT is designed to measure a construct that is typically unconscious or about which the individual is unaware.

1.4 Conclusion

In this paper, I have provided an account of triangulation that highlights locations and types of epistemic risk. In particular, I diagnosed two epistemic risks in implicit attitude research: (1) the risk that data do not serve as evidence for the same hypothesis, and (2) the particular inductive risk that there is insufficient evidence provided to conclude that there is a single phenomenon (given the plausibility of alternative hypotheses positing multiple phenomena). Neither is sufficiently described by illusory robustness and (1) is not a case of inductive risk either. Finally, I demonstrated that current accounts of triangulation focused on successful cases cannot provide explanations of why triangulation sometimes fails in practice and thus, do not develop sufficient norms to guide future triangulation research.

References

Banaji, Mahzarin R. 2001. “Implicit Attitudes Can Be Measured.” In *The Nature of Remembering: Essays in Honor of Robert G. Crowder*, edited by H.L. Roediger and J.S. Nairne, 117–50. Washington, DC: American Psychological Association.

Bar-Anan, Yoav, and Brian A. Nosek. 2014. “A Comparative Investigation of Seven Indirect Attitude Measures.” *Behavior Research Methods* 46(3):668–88.

Biddle, Justin B., and Rebecca Kukla. 2017. “The Geography of Epistemic Risk.” *Exploring Inductive Risk: Case Studies of Values in Science*, 215–38.

Biddle, Justin B. 2016. “Inductive Risk, Epistemic Risk, and Overdiagnosis of Disease.” *Perspectives on Science* 22(3):397–417.

Bogen, James, and James F Woodward. 1988. “Saving the Phenomena.” *The Philosophical Review* XCVII(3):303–52.

Borsboom, Denny, Gideon J. Mellenbergh, and Jaap Van Heerden. 2004. “The Concept of Validity.” *Psychological Review* 111(4):1061–71.

Bosson, J K, W B Swann, and J W Pennebaker. 2000. “Stalking the Perfect Measure of Implicit Self-Esteem: The Blind Men and the Elephant Revisited?” *Journal of Personality and Social Psychology* 79:631–643.

Campbell, Donald T, and Donald W Fiske. 1959. “Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix.” *Psychological Bulletin* 56(2):81–105.

Corkin, Suzanne. 2002. “What’s New with the Amnesic Patient H.M.?” *Nature Reviews Neuroscience* 3(2):153–60.

Douglas, Heather. 2016. “Values in Science.” *Oxford Handbook in The Philosophy of Science*, 23.

Dovidio, John F, and Samuel L Gaertner. 2000. “Aversive Racism and Selection Decisions: 1989 and 1999.” *Psychological Science* 11(4):315–19.

Fazio, Russell H., Joni R. Jackson, Bridget C. Dunton, and Carol J. Williams. 1995. “Variability in Automatic Activation as an Unobtrusive Measure of Racial Attitudes: A Bona Fide Pipeline?” *Journal of Personality and Social Psychology* 69(6):1013–27.

Fazio, Russell H., and Michael A. Olson. 2003. “Attitudes: Foundations, Functions, and Consequences.” *The SAGE Handbook of Social Psychology*, no. January:123–45.

Fazio, Russell H, David M Sanbonmatsu, Martha C Powell, and Frank R Kardes. 1986. “On the Automatic Activation of Attitudes.” *Journal of Personality and Social Psychology* 50(2):229–38.

Gawronski, Bertram, Wilhelm Hofmann, and Christopher J. Wilbur. 2006. “Are ‘Implicit’ Attitudes Unconscious?” *Consciousness and Cognition* 15:485–99.

Greenwald, Anthony G., and Mahzarin R Banaji. 1995. “Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes.” *Psychological Review*.

Greenwald, Anthony G., D E McGee, and J L K Schwartz. 1998. “Measuring Individual Differences in Implicit Cognition: The Implicit Association Test.” *Journal of Personality and Social Psychology* 74:1464–1480.

Hahn, Adam, Charles M. Judd, Holen K. Hirsh, and Irene V. Blair. 2014. “Awareness of Implicit Attitudes.” *Journal of Experimental Psychology: General* 143(3):1369–92.

Houwer, Jan De, Sarah Teige-Mocigemba, Adriaan Spruyt, and Agnes Moors. 2009. “Implicit Measures: A Normative Analysis and Review.” *Psychological Bulletin* 135(3):347–68.

Jones, Edward E., and Harold Sigall. 1971. “The Bogus Pipeline: A New Paradigm for Measuring Affect and Attitude.” *Psychological Bulletin* 76(5):349–64.

Jost, John T. 2019. “The IAT Is Dead, Long Live the IAT: Context-Sensitive Measures of Implicit

Attitudes Are Indispensable to Social and Political Psychology." *Current Directions in Psychological Science* 28(1):10–19.

Kuorikoski, Jaakko, and Caterina Marchionni. 2016. "Evidential Diversity and the Triangulation of Phenomena." *Philosophy of Science* 83(2):227–47.

Mitchell, J P, Brian A. Nosek, and Mahzarin R. Banaji. 2003. "Contextual Variations in Implicit Evaluation." *Journal of Experimental Psychology: General* 132:455–469.

Olson, Michael A, and Russell H Fazio. 2003. "Relations between Implicit Measures of Prejudice: What Are We Measuring?" *Psychological Science* 14:636– 639.

Payne, B Keith, and B Gawronski. 2015. "A History of Implicit Social Cognition: Where Is It Coming From? Where Is It Now? Where Is It Going?" In *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications,* edited by B. Gawronski and B K Payne. New York: Guilford.

Schickore, Jutta, and Klodian Coko. 2013. "Using Multiple Means of Determination." *International Studies in the Philosophy of Science* 27(3):295–313.

Schupbach, Jonah N. 2018. "Robustness Analysis as Explanatory Reasoning." *British Journal for the Philosophy of Science* 69(1):275–300.

Shiffrin, Richard M., and Walter Schneider. 1977. "Controlled and Automatic Human Information Processing: II. Perceptual Learning, Automatic Attending and a General Theory." *Psychological Review* 84(2):127–90.

Stegenga, Jacob. 2009. "Robustness, Discordance, and Relevance." *Philosophy of Science* 76:650–61.

Wilholt, Torsten. 2009. "Bias and Values in Scientific Research." *Studies in History and Philosophy of Science Part A* 40(1):92–101.