European Journal for Philosophy of Science (2020) 10:4
https://doi.org/10.1007/s13194-019-0270-8

PAPER IN THE PHILOSOPHY OF THE SCIENCES OF MIND AND BRAIN

Construct validity in psychological tests – the case of implicit social cognition

Uljana Feest[1]

Received: 17 February 2019 / Accepted: 27 November 2019/Published online: 08 January 2020

Abstract
This paper looks at the question of what it means for a psychological test to have construct validity. I approach this topic by way of an analysis of recent debates about the measurement of implicit social cognition. After showing that there is little theoretical agreement about implicit social cognition, and that the predictive validity of implicit tests appears to be low, I turn to a debate about their construct validity. I show that there are two questions at stake: First, what level of detail and precision does a construct have to possess such that a test can in principle be valid relative to it? And second, what kind of evidence needs to be in place such that a test can be regarded as validated relative to a given construct? I argue that construct validity is not an all-or-nothing affair. It can come in degrees, because (a) both our constructs and our knowledge of the explanatory relation between constructs and data can vary in accuracy and level of detail, and (b) a test can fail to measure all of the features associated with a construct. I conclude by arguing in favor of greater philosophical attention to processes of construct development.

Keywords Implicit bias · Implicit tests · IAT · Construct validity · Epistemology of experimentation · Philosophy of psychology

1 Introduction

Empirical studies consistently demonstrate the existence of systematic discrimination on the basis of race, gender, age or sexual orientation and other attributes. For example, it is well known that there are many inequities between Blacks and Whites, ranging from employment and housing to credit (Pager and Shepherd 2008). Regarding

"Psychologists today should be concerned not with evaluating tests as if the tests were fixed and definitive, but rather with developing better tests." (Campbell and Fiske 1959, 103)

✉ Uljana Feest
feest@philos.uni-hannover.de

[1] Leibniz Universität Hannover, Institut für Philosophie, Im Moore 21, 30167 Hannover, Germany

academic employment, particular attention has been paid to inequities in employment and wages of black Americans as compared with white Americans and females as compared with males. In one study, science professors at American universities offered lower starting salaries to applicants with a female first name, even if the resumes were otherwise identical to those of male applicants (Moss-Racusin et al. 2012). Similar findings have been reported with respect to names that sound African-American (Bertrand and Mullainathan 2002). The accuracy of such findings will be presupposed in this article.

It has become common to use the expression "implicit bias" in relation to the type of discriminatory phenomena just described. The concept of implicit bias is taken from a current research area in social psychology, that of implicit social cognition (e.g., Payne and Gawronski 2010). It is typically taken to refer to a particular kind of prejudice, though some researchers distinguish between implicit prejudices and implicit stereotypes (e.g., Amodio and Devine 2006). Within philosophy, this concept has attracted some attention in the course of attempts to explain the gender gap in academic philosophy (e.g., Brownstein and Saul 2016). There is some sophisticated recent philosophical work that examines ontological and ethical questions in relation to this concept (e.g., Brownstein 2018).

While the expression "implicit bias" is still widely used in academic and societal discourse, there has also been a considerable backlash in the past few years, following several meta-studies, which throw doubt on the validity of one of the major investigative tools used to detect such biases, the IAT (Implicit Association Test). However, there is relatively little philosophical work on what test validity is, precisely, and what conditions must be met in order for a test to have validity. Moreover, while critics of the IAT have focused largely on its (lack of) *predictive* validity, there is little discussion about the test's *construct* validity. Nor is there much philosophical literature about what construct validity amounts to.[1] Lastly, it is often tacitly assumed that if a test of implicit social cognition has low validity, this should prompt us to abandon the corresponding construct altogether. By contrast, the case might be made that if a particular construct has high prima facie plausibility (or, as psychologist call it, "face validity") the phenomenon in question is worthy of further research regardless of whether we currently have adequate tests for it.

In this article, I will take up all of these points: I will use the IAT as an example to offer a principled discussion of the notion of (construct) validity in order to argue (a) that construct validity can come in degrees, and (b) that even if a test has low construct validity, it does not follow that the construct is not worth pursuing and developing. Taken together, these two points suggest that psychological constructs often start their lives as pre-theoretical or folk-psychological concepts, which can be – though by no means always are – improved as part of an iterative process, whereby the initial

[1] See Alexandrova and Haybron (2016) and Hood (2009) for two noteworthy exceptions to this general statement.

construct informs a test, which in turn can contribute to the further development of the construct, relative to which the construct validity of the test can be investigated, in turn enabling additional development/refinement of the constructs, etc. This is indeed an overarching thesis of this paper.[2]

After briefly laying out the tests in question (Section 2) I will highlight that there is a significant degree of disagreement within the scientific community as to what implicit social cognition is, precisely. Differently put, there is little agreement on the very construct of implicit social cognition. I will argue that this situation is typical of much psychological research, where the relevant constructs (or, as they are more commonly called in philosophy of science, "theoretical concepts") are underdeveloped and under construction (Section 3). This will prompt me to raise the question by what criteria it can judged that existing tests have (a certain degree of) validity. Section 4 will begin with a quick overview of the distinction between predictive validity and construct validity. While the low predictive validity of implicit tests has led some to become quite skeptical of the project of measuring implicit social cognition, I will argue that there is nevertheless good reason to inquire into the construct validity of such tests. I will emphasize that a test's *construct validity has to be specified relative to a particular construct*. This raises two questions, i.e. (1) whether a test can have construct validity relative to an underdeveloped constructs (more generally, what level of specificity a construct needs to have, such that a test could *in principle* be valid relative to it?), and (2) what kind of evidence needs to be in place to say of a test that it *in fact* has construct-validity relative to a given construct. These questions are addressed in Section 5. Regarding the first question, I will argue that construct validity is a matter of degree along two dimensions: (a) A test can have a certain degree of construct validity relative to a "lumpy" construct while failing to have such validity relative to a more fine-grained construct.[3] (b) A test can have a certain degree of construct validity relative to a "partial" construct, i.e., by measuring only *some* of the components specified by a construct. With regard to the second question, I will argue that statistical analyses can play a valuable role in construct development. Both of these points highlight the dynamic and ongoing research process. Section 6 will offer some concluding reflections about the implications of my analysis (1) for open questions about construct validity in the philosophy of psychology and (2) for discussions about the status and measurability of implicit social cognition.

2 Measuring attitudes by means of explicit and implicit tests

A fundamental concept of social psychology is that of *social attitude*. Social attitudes are attitudes towards human beings. This explanatory construct refers to a hypothetical mental state, process, or capacity that is believed to play a causal role in social behavior.

[2] This idea is, of course, familiar from Hasok Chang's work (e.g., Chang 2004); but see also Tal 2017, section 8.1. Note that while I am here imagining a scenario whereby a construct gets developed and refined in a gradual fashion, it is entirely compatible with my analysis that a construct might end up being eliminated in the course of the research process. In other words, the fact that a given folk psychological concept has a high face validity does not guarantee that it will ultimately be a viable scientific construct. Likewise, I take it that the validity of a test can also decline over time.

[3] By a "lumpy" construct I mean one that has in its extension two or more distinct phenomena.

In turn, when it comes to explaining discriminatory (e.g., racist or sexist) behavior, social attitudes are sometimes broken down into a cognitive and an evaluative component: a stereotype and a prejudice (Machery et al. 2010).

Both in practical and in research contexts there has been a long-standing interest in the measurement of attitudes (this is true for foundational research in social psychology as well as for polling and marketing research). For the better part of the twentieth century, investigations in these areas proceeded by asking people to describe or rank their preferences, mental states, or future behaviors with regard to specific objects. In this context, it is common to use quantitative questionnaires that allow people to rank their attitudes. Typical examples are Likert scales, feeling thermometers, and versions of the semantic differential (SD) test (Snyder and Osgood 1969).[4] In such tests, subjects are typically asked to indicate (e.g., on a scale from one to 7) the degree to which they agree with a given statement. In the realm of racial attitudes, a common test is McConahay's (1986) *Modern Racism Scale*, which has subjects rate a number of assertions about the status of black people in contemporary society (Blank et al. 2004). Notice that while subjects are explicitly asked to state/rank their attitude about a particular object, the data can also be used to make inferences beyond the stated attitudes. For example, a racism scale might ask subjects to state their attitude about a particular black person; yet the results might be used to make an inference about the person's attitude towards black people in general. Thus, we can distinguish between direct and indirect ways of inquiring into attitudes.

Since the 1990s, a novel method of indirect measurements has gained in prominence: so-called *implicit tests*. These tests do not explicitly ask subjects about attitudes at all. Yet their results are used to make inferences about attitudes. Such tests have the obvious advantage of not relying on subjects' ability or willingness to accurately state their attitudes. Two kinds of widely used implicit tests are implicit association tests (IAT; Greenwald et al. 1998) and affective priming tasks (Fazio et al. 1995).[5] Both kinds of tests are premised on the assumption that the stereotypes and prejudices about a particular social group consist of networks of strongly associated attributes. For example, the stereotype of *female* contains a number of assumptions about biological, social, affective, and cognitive features of women. Both tests assume that such associations are indicative of social attitudes. The affective priming tests used in current social cognition studies ask subjects to categorize a given target word as positive or negative. It is possible to prime the categorization of words as positive or negative by the presentation of white and black faces, respectively, suggesting that many people associate 'bad' things with Blacks and 'good' things with Whites, and that this says something about their attitudes towards Blacks and Whites. In turn, the IAT relies on the assumption that it is easier to classify a given stimulus (such as the name "Emily") as belonging to a given category (e.g., "female") when that target category is presented alongside one with which it is associated (e.g., "family") as opposed to one with which it is less strongly associated (e.g., "career"). This test is used as part of a large-scale project conducted by a multitude of institutions, but situated at Harvard, called *Project*

[4] For more detail, see entries "Likert Scale" and "Feeling Thermometer" in the *Sage Encyclopedia of Research Methods* (Lavrakas 2008).

[5] There are many widely-used variants of either kind of implicit test (see Nosek et al. (2011) and de Houwer and Moors (2010) for overviews of implicit tests used in social psychology).

Implicit. This project studies a number of different attitudes by connecting specific categories (e.g., race, gender, sexual orientation) with specific attributes, some of which are cognitive (designed to establish, for example, whether people associate women with professional careers) and some of which are evaluative (designed to establish, for example, whether people have positive, negative, or neutral attitudes towards people of color).[6]

Of course, the existence of a novel class of tests does not imply the existence of a previously unknown class of phenomena. After all, implicit tests might simply be a new method for measuring what traditional attitude tests were measuring all along.[7] That said, if we find the results of implicit and explicit tests to be dissociated, this might be a reason to posit that traditional attitude tests and implicit tests do indeed measure different things (if they measure anything at all).[8] With respect to this question, we can note that the results of implicit and explicit attitude tests are correlated. However, the magnitude of the correlation depends on the circumstances of the testing situation and the object of the attitude. For example, the correlation is higher when subjects taking the explicit tests are asked to report their attitudes as quickly as possible, and it is lower when the attitudes in question concern "socially sensitive" topics (Greenwald et al. 2009). In the US, this latter finding is especially noticeable when it comes to attitudes of Whites towards Blacks, in that the results on explicit and implicit tests can be dissociated. It is the existence of such dissociations (i.e., the fact that people's explicitly reported attitudes can differ from those apparently measured by means of implicit tests) that lends plausibility to the hypothesis that implicit tests pick up on something not captured by explicit tests. But what might this be? And what are criteria of adequacy for implicit tests (or for any psychometric tests, for that matter)?

3 What do implicit tests (purport to) measure, and how do we know that they do?

At first blush, one answer that suggests itself is that the object of implicit tests is a particular kind of attitude, which is characterized by the fact that it is implicit. In the psychological and philosophical literature we find quite different accounts of what this means. Here I will briefly outline the main sources of disagreement to highlight that there is no unified or agreed on concept of implicit attitude in the current literature (see Brownstein 2017 for an overview), and to raise the question of what researchers do agree on.

Nosek and Banaji (2009), two prominent researchers in this field, start by providing a definition of the notion of *attitude* in general, characterizing it as "an association

[6] https://implicit.harvard.edu/implicit/

[7] This is emphasized by Schimmack (2019b), who accuses social psychologists of conflating "implicit measures of attitudes" with "measures of implicit attitudes". A similar concern is also behind de Houwer et al.'s (2009) suggestion that an indirect test should only be called an implicit test if it measures an implicit process (we will return to this debate in Section 5 below). For the purposes of this paper, I will stick with the expression "implicit tests," while not committing to a particular account of what they measure.

[8] In neuropsychology the term "dissociation" is commonly used to refer to cases where two operations that were thought to manipulate the same variable lead to divergent results, i.e., result that are not correlated, or at least not correlated as highly as expected.

between a concept and an evaluation" (ibid. 84) and distinguishing between associations that are deliberate, intentional and introspectively accessible and associations that are not. Both types of attitudes are conceived of as cognitive structures (C), where they take an "implicit C" to be "the introspectively unidentified (or inaccurately identified) trace of past experience that mediates R" (Greenwald and Banaji 1995, 5). Thus, an implicit attitude (on this account) is an association between a concept and an evaluation that is introspectively unidentified or inaccurately identified and that plays a causal role in the response (R) triggered by a particular stimulus. The authors posit that we can have two distinct attitudes about things and that these two attitudes can at times be in conflict (Nosek and Banaji 2009, 85).

Gawronski and Bodenhausen (2006), two other social psychologists, define "attitude" in theoretically slightly more open terms as "a psychological tendency to evaluate a given entity with some degree of favor or disfavor" (ibid. 693). Their APE model ("Associative-Propositional Evaluation") suggests that implicit and explicit attitudes are characterized by different types of underlying processes, namely associative and propositional processes, respectively, though they have recently emphasized that they do not intend this distinction to denote different types of storage in memory (Gawronski et al. 2017, 105). With this, the authors position themselves within a long tradition of "dual-process" approaches in social psychology.[9] As Brownstein (2017) points out, this psychological model has some parallels with one that has enjoyed some popularity within philosophy in recent years, namely Gendler's distinction between beliefs and aliefs, where the latter are clusters of automatically activated representations (Gendler 2011). Importantly, Gawronski & Bodenhausen understand "associative" not in terms of connections between attributes (which may or may not be introspectively accessed), but as a *type of process*, which they characterize as automatic and related to pattern activation.[10] This leads them to disagree with Greenwald and Banaji's account of *implicit* attitudes in one crucial respect: "We assume that people generally do have some degree of conscious access to their automatic affective reactions" (Gawronski and Bodenhausen 2006, 696).[11] Moreover, they take associative processes to be linked to associative learning at the stage of attitude formation. (Gawronski et al. 2017).

One other important figure in this literature, Jan De Houwer, shares Gawronski's focus on automatic, associative, retrieval processes (as opposed to the unconscious attitudes posited by Banaji). However, he argues that the thesis of automatic processes in retrieval should not be confused with the thesis that attitudes are *learned* in a purely automatic fashion. Drawing on his own work on evaluative conditioning (e.g., De Houwer 2014), he thereby distances himself from association formation models, which posit that association learning is automatic and stimulus driven. Instead he argues that learning about associated properties takes place always by way of the formation of propositions. This means that implicit attitudes, insofar as they reflect something that was previously learned, are propositional and can, thus, be true or false of the objects in

[9] For a classic point of reference see Chaiken and Trope (1999). Similar distinctions exist in other fields of psychology (see Frankish 2010). See Evans 2006, for a critique of the assumption that the two types of processes map onto two types of systems (e.g., in Kahneman's 2011 terms, system 1 and system 2).

[10] While I cannot pursue this topic here, Mandelbaum (2015) is spot-on when he remarks that psychologists and philosophers writing about implicit cognition have not taken enough care in distinguishing (and critically evaluating) various senses of the term "association."

[11] See Krickel (2018), for a recent critique of this assumption.

question. Moreover, De Houwer claims that association learning requires awareness. This has important implications: "It is assumed that a relation in the world can influence behavior only after a proposition about the relation has been consciously entertained to be true" (De Houwer 2014, 532), regardless of whether it is consciously entertained at the time of retrieval.[12]

A fourth theoretical model, the MODE-model by Fazio and others shares with the APE model the assumption of different types of processes (e.g., Fazio and Olson 2003, 301), "by which attitudes can affect judgments and behavior," i.e., processes of retrieval (Fazio and Olson 2014, 155). It also shares with Banaji's and De Houwer's approaches the assumption that attitudes are associations between objects and evaluations. Though not committing to any specific model about the format in which attitudes are represented, this model rejects the notion of two distinct systems, positing instead the existence of two kinds of retrievals. Based on this, the model ultimately rejects the distinction between implicit and explicit attitudes. The MODE model focuses on the circumstances under which either kind of retrieval come into action. MODE stands for "motivation and opportunity as determinants" and indicates the model's assumption that "the impact of attitudes will be reduced when individuals have both the motivation and the opportunity to deliberate about the available information and, in so doing, overcome the influence of any pre-existing attitude" (Fazio and Olson 2014, 157).

This brief overview of the current theoretical landscape in the field of implicit social cognition research shows that theoretical models differ along several dimensions: While most agree that attitudes are associations between concepts and evaluations, some take these associations (in the case of implicit attitudes) to be introspectively inaccessible (Banaji, Nosek, Greenwald), others think they are accessible, but that the retrieval is typically automatic (De Houwer, Gawronski). Some think that implicit attitudes are learned in an associative fashion (Gawronski) while others take all attitudes to be learned in a propositional format (De Houwer).[13] Lastly, while most of the above-mentioned authors assume as meaningful a distinction between implicit and explicit *attitudes*, others have called this very distinction into question (Fazio).

It is not my aim to argue in favor of any of these proposals, but rather to highlight that this field of study is characterized by a great deal of theoretical controversy, and that there is no generally agreed upon theoretical construct of implicit social cognition. My interest here is not in a philosophical analysis of *what implicit attitudes "really are,"* but rather in the question of how practicing researchers go about researching implicit attitudes, *given that they don't know what they are*, or rather, given theoretical disagreements about key features of implicit attitudes (such as the meanings of "attitude" and "implicit," respectively). One natural place to start is with the tests used to measure or investigate implicit social cognition, as these tests appear to provide a common frame of reference. The question that interests me here is whether – in the absence of theoretical agreement about the nature of implicit social cognition – there is

[12] A propositional model of implicit attitudes has recently also been endorsed in philosophy (cf. Mandelbaum 2016). For a recent analysis of the debate between associationist and propositional accounts see Byrd (2019).
[13] The distinction between associative and propositional learning is typically taken to consist in the fact that the former kind of learning occurs in an automatic fashion by mere empirical exposure to the relevant stimuli, whereas the latter involves "prior knowledge, instructions, intervention, and deductive reasoning." (de Houwer 2009, 1).

any way that the "goodness" of these tests can be established (or refuted, for that matter).

Within psychology, questions about the quality of a test are discussed as concerning the test's *validity*, a methodological topic that was most prominently discussed in the 1940s–1960s, but that has not received much philosophical attention since. In the remainder of this paper, I will treat the case of implicit social cognition as providing us with illuminating material with respect this issue, focusing my attention on a methodological discussion that took place in psychology (De Houwer et al. 2009 vs. Nosek and Greenwald 2009). My analysis will pinpoint, and shed light, on philosophical issues with regard to construct validity and construct development in general, while also putting us in a better position to evaluate current controversies about the status of implicit social attitudes.

4 Some existing debates about the validity of implicit tests

Within psychology, recent discussions about the validity of implicit tests have focused on the predictive and the construct validity of such tests, respectively. This distinction goes back to Cronbach and Meehl (1955), according to whom a test has predictive validity if its results correlate with some criterion variable thought to be relevant to the subject matter and construct validity if there is a theoretical construct that explains the test results.

4.1 Predictive and construct validity of implicit and explicit tests

A test has predictive validity if its results can predict a relevant criterion variable. In the case of attitude tests, such a criterion variable is typically the subjects' behavior towards instances of the object of the attitude. For example, if the IAT claims to be able to measure attitudes towards African Americans, then we would say of it that it had predictive validity if its results enabled us to predict whether subjects are likely to engage in discriminatory behavior towards African Americans. The predictive validity is quantified in terms of the correlation between the test score and the discriminatory behavior.[14] The attraction of predictive validity is that it allows us to remain neutral with respect to more thorny questions, such as whether we have a theoretically adequate conception of what implicit attitudes are.

Jost et al. (2009) have claimed that a number of studies reveal implicit tests to have a high degree of predictive validity.[15] However, this claim has been heavily debated and criticized in recent years. In a meta-analysis Greenwald et al. (2009) looked at 122 research reports to compare the predictive validity of explicit and implicit attitude tests, respectively. Two intriguing patterns are revealed by this study: (a) with two noteworthy exceptions ("White-Black race" and "other intergroup"), explicit tests are better

[14] The same criterion would be used to determine the predictive validity of explicit tests, such as the *Modern Racism Scale.*

[15] Jost et al.'s article is a reply to Tetlock and Mitchell's (2009) suggestion that the popularity of the notion of implicit biases is largely due to liberal guilt and is not backed up evidence about the predictive validity of implicit tests. In response they state that "truth is that such evidence is already strong, and it continues to grow in depth and breadth." (Jost et al. 2009, 46)

predictors of discriminatory behavior than implicit tests, and (b) in those two cases the predictive validity of both implicit and explicit tests is fairly low, as compared with (for example) those that test political preferences. While the predictive validity of implicit tests that measure attitudes of Whites towards Blacks is slightly higher than that of explicit tests, both have rather low predictive validities of roughly only .10 and .20, respectively. Another meta-analysis has called even these modest findings into question (Oswald et al. 2013). More recently, another meta-analysis has investigated the effectiveness of manipulations designed to change implicit biases (Forscher et al. 2017), reporting that even though it is possible to change implicit biases (as measured by existing tests), such changes do not translate into significant changes at the level of explicit biases or discriminatory behavior.[16] Mitchell and Tetlock (2017) therefore argue that "the implicit prejudice construct … should be retired" (Mitchell and Tetlock 2017, 165).

It should be noted that not everybody in the literature is convinced by these studies. For example, Michael Brownstein points to problems with some of the meta-analyses in question, noting that they lumped together data that were produced for different purposes and with different theoretical outlooks (Brownstein 2018, Appendix). In this paper, I do not wish to refute problems with the predictive validity of current tests of implicit social cognition. However, I would like to resist the conclusion that these problems make the construct of implicit social cognition obsolete. In the following, I will present two arguments for this. First, I will argue that the low predictive validity of (say) the IAT should not be taken to imply that such tests can make no epistemic contribution at all. Second, I will suggest that even if current tests for implicit social cognition are of questionable quality, this might simply mean that we currently don't have good tests for a phenomenon, which nonetheless warrants further conceptual and empirical work. The two points are related since if an argument for the epistemic merit of current implicit tests can be made (no matter how small these merits may be), this would also strengthen my point that we should work on an improved theoretical understanding of the phenomenon (and, correspondingly, we should work on devising better tests).

I will start with the first point by pointing out that the debate about the predictive validity of implicit tests has focused on one question, i.e., whether such tests can predict discriminatory behavior. But we often want to not merely *predict* behavior, but also to *explain* it. Indeed, the notion of an implicit bias is typically invoked as an *explanatory* construct. One important question here is therefore not merely whether implicit attitude tests predict discriminatory behavior better than explicit attitude tests, but whether what is being measured by implicit tests explains something about discriminatory behavior that is not explained by appeal to the properties or processes measured by explicit tests.[17] This may be of practical relevance as well. For example, even if the results on explicit tests of sexist attitudes are more highly correlated with (and perhaps also more

[16] See Bartlett (2017), for a brief popular write-up of this debate in the *Chronicle of Higher Education* (January 5, 2017). See also Singal's (2017) article in the *New York Magazine*, which offers a well-researched account of current debates about the validity of implicit measures.

[17] While mid-twentieth-century treated the concepts of prediction and explanation as closely related, not all predictive statements are also explanatory (as famously illustrated the example of the barometer and the storm). In a similar vein, even if it were to turn out that a particular implicit tests predict discriminatory behavior, it would not follow that the behavior in question is explained by implicit social cognition

explanatory of) discriminatory behavior against women than those of implicit tests, implicit tests might still be singling out a factor that makes a small but relevant additional contribution to the explanation of such behavior, and that perhaps even plays a substantial role in the coming about of a larger social phenomenon (see Greenwald and Banaji 2015, for an argument along these latter lines). In this regard, consider implicitly and explicitly measured political preferences: Even if implicit attitudes only explained a small part of the variation in voting behavior, they might still have a significant impact in a close election (see Nosek and Riskind 2012).

The question, then, is whether current tests of implicit social cognition measure something that makes a genuine explanatory contribution to issues of general interest. This type of question is typically addressed under the label of "construct validity." Roughly speaking, to say of a psychometric test that it has construct validity is to assume that there is a construct (theoretical concept) that singles out the putative cognitive states, processes, or traits that explain (or at least partially explain) the behavioral phenomenon at stake. Now, given the theoretical disagreement about the very theoretical concept of implicit social cognition, this raises the question of whether it will in principle be possible to agree on how to construct-validate implicit tests. This question has two interrelated aspects: First, how detailed and accurate does a concept need to be in order to even be a *candidate concept* relative to which a test can be construct validated? Second, what kinds of theoretical and evidential standards have to be met in order to establish that a given test *in fact* has validity relative to a given concept?

In the literature about implicit social cognition there was an illuminating debate ten years ago about the construct validity of implicit tests, which turned (at least indirectly) on exactly these two questions. I will briefly introduce this debate in the remainder of this section in order to prepare the ground for a more in-depth analysis in Section 5 below.

4.2 A recent debate about the construct validity of implicit tests (a case study)

Let me begin by highlighting an important point: Even though construct validity concerns issues of the explanation (rather than the mere prediction) of social behavior in the real world, construct *validation* of a test proceeds by showing that the construct explains the behavioral data generated by the test, not the behavioral data that we might wish to predict on the basis of test results. This is important, because it suggests that a test might have construct validity within the confines of a given experimental validation study while failing to make a significant contribution to the prediction or explanation of behavior outside the lab. This touches on the issue of extrapolation, which has attracted some interest in recent philosophy of science.[18] While clearly highly relevant to the debate about implicit social cognition, this issue cannot be covered in this paper. In the following, I will therefore focus exclusively on the issue of how scientists attempt to construct-validate their tests within the lab.

In a 2009 article, De Houwer et al. propose that an implicit attitude test can be regarded as construct-validated only if there is evidence that its results are caused by the

[18] E.g., Steel (2008). Guala (2012), distinguishes between the internal and experimental validity of experiments. Psychologists sometimes refer to the latter as "ecological validity." (See also Feest and Steinle 2016)

attitude in question and that the processes leading to the results are implicit. But it seems that in order to implement these criteria of construct validation, one needs to already have a fairly detailed construct in place, which specifies what is meant by "attitude" and "implicit." However, it was precisely the absence of a general agreement on these terms that prompted us to turn to construct validation in the first place.[19] De Houwer et al. (2009) recognize this problem to some extent, settling for what they take to be the fairly theory-neutral idea that attitudes are associations between concepts in memory.[20] When it comes to the notion of *implicitness*, however, they keep with their own notion, arguing that this predicate (in the context of implicit attitude tests) describes automatic processes of retrieving and acting upon attitudes, rather than attitudes that cannot be introspectively identified. They characterize such processes in terms of "the absence of certain goals, awareness, substantial cognitive resources, or substantial time" (De Houwer et al. 2009, 350). For them, the construct-validation of an implicit test, thus, requires the following steps: one needs to (a) provide evidence that the test really measures attitudes (i.e., associations between concepts), (b) specify the causal mechanisms linking attitudes and test results, (c) specify which processes are thought to be automatic and (d) provide evidence that the process in question is indeed implicit *in just this sense*.

De Houwer et al. (2009) accordingly discuss existing research about the IAT and the AP to evaluate these questions. With regard to the question of whether existing tests measure attitudes, the authors cite studies that have experimentally manipulated associations between concepts and observed the effects of these manipulations on test results. Drawing on the results of these studies, the authors argue that IATs fare better than AP tests with respect to this question. On the other hand, they remark that the causal processes leading from attitudes to test results are better understood in the case of AP tests. Lastly, regarding the question of whether the processes in question are implicit, and if so, in what respect, they restate their own definition of "implicit" (focusing on "conditions involving proximal and distal goals, awareness, processing resources and time", De Houwer et al. 2009 357). They review the evidence that the relevant processes in question are implicit in any of these senses and state that "much of the work still needs to be done" (De Houwer et al. 2009, 262). The authors conclude: "For most measures it is not entirely clear what they measure, what processes produce the measure, and whether the processes are automatic in a certain manner" (ibid. 263). Thus, they claim that *most tests do not meet the normative criteria that they previously laid out for construct validity*, though they emphasize that this does not mean that implicit tests are useless as research tools (ibid.).

Other researchers have challenged the main points of the article by De Houwer et al. (2009). For example, Nosek and Greenwald (2009) take issue with De Houwer et al.'s construct of "implicit" as *automatic*. In a nutshell, they argue that de Houwer et al.'s concept of implicit social cognition is too narrow, given the current lack of agreement about this very issue. The question is whether there is a less narrow concept everyone

[19] For example, recall that De Houwer et al. conceptualize attitudes as propositionally structured and argue that such attitudes can be retrieved by particular (automatic) processes. By contrast, Banaji, Nosek and others conceptualize attitudes as associations between concepts and evaluations attributes, but argue that some (i.e., the implicit ones) are unconscious.

[20] I describe this idea as "theory-neutral" because several of the theoretical models introduced in Section 3 share it.

might agree on. While Nosek & Greenwald do not address this question directly, I argue that if we look closely at their practice, an indirect specification of the construct emerges: According to them, *a test* measures implicit attitudes if "[it] does not require awareness of the relation of the attribute to the response." (Nosek and Greenwald 2009, 374). However (and as the authors themselves point out), the fact that the IAT does not require awareness of the relationship between attitudes and response does not mean that the subjects are unaware of this relationship: Subjects may (a) be unaware of having the attitude in the first place, (b) be aware of the attitude, but not be aware of the fact that it causes a particular response, or (c) be aware of the attitude and of the fact that it causes a particular response, while not being able to control the response. The test itself does not differentiate between the two meanings of "implicit" outlined above (lack of introspective access and lack of control). Thus, it seems that in their conception of what is measured by implicit tests, the authors are happy to lump together several conceptually distinct possible states of affairs. The concept they tacitly endorse, hence, appears to be a "lumpy," or disjunctive, one. Indeed, they take this to be an advantage of their approach, pointing out that psychological concepts in the process of ongoing research are constantly changing. For this reason, it strikes them as counterproductive to commit to a narrow theoretical model of the phenomenon under study and to gear one's research activities toward trying to validate such an overly specific model. (Nosek and Greenwald 2009, 373).

This is a reasonable position to take. But what are its implication for the prospects of construct validation? Should we conclude that construct validation is simply not what we should be striving for at this point in time? Or is it possible to construct validate a test relative to a lumpy or otherwise theoretically open concept? Nosek & Greenwald are a little ambiguous on this question.[21] One point they bring across quite clearly, however is that De Houwer et al.'s analysis of construct validation is counterproductive for the progress of research: "[T]he most important considerations in appraising validity of psychological measures are those that speak to the measure's usefulness in research and application" (Nosek and Greenwald 2009, 375). De Houwer et al. (2009) acknowledge the importance of pragmatic considerations, indicating that a test can be useful for scientific purposes even if it does not meet their normative criteria of validity: "Even measures that are not perfect can still be useful" (348). Indeed, they go even further than this and say that "[t]he normative criteria should therefore not be interpreted as minimal conditions that must be met before a measurement outcome can be regarded as an implicit measure" (ibid.). Presumably this means that a test can be regarded as measuring implicit social cognition even if it does not pass De Houwer et al.'s strict criteria of construct validity. But then (assuming that they are right), the question is what *are* the minimal conditions that must be met before a measurement outcome can be regarded as an implicit measure, i.e., *as measuring implicit social cognition*? This was precisely the question that we were hoping to answer with this discussion of construct validity, but it is not explicitly answered by researchers on either side of the debate.

[21] E.g., on the one hand we find statements along the lines that De Houwer's criteria of construct validation are too narrow (implying that less narrow criteria might be available), but on the other hand, they also state that "scientific goals will more often be served by prioritizing pragmatic goals of establishing predictive validity" (Nosek and Greenwald 2009, abstract).

In the remainder of this paper, I will start to develop an account of construct validity and construct validation that (a) retains some of the insights of De Houwer et al.'s analysis while being weaker than theirs and (b) integrates some of the legitimate concerns articulated by Nosek & Greenwald, while (c) clarifying some underlying philosophical themes. In doing so I hope to do justice to the idea that implicit tests (if and insofar as they measure anything distinct from what is measured by explicit attitude tests) capture some feature of cognition that contributes to the explanation of the data generated by implicit tests. In addition, I hope to do justice to the dynamic nature of scientific research and its concepts, so rightly emphasized by Nosek & Greenwald. The bottom line of my analysis is going to be that construct validity is not an all-or-nothing affair, but can come in degrees, thereby making room for the possibility that current tests could in principle have some construct validity, despite the theoretical disagreements in this field of research.

5 Construct validity and the dynamics of concept development

In this section, I will provide three arguments for my thesis that construct validity can come in degrees. Two of them draw on the nature of constructs in the research process (they can be "lumpy," and they can be "partial") and one draws on the nature of evidential support for specific causal hypotheses in psychology. The underlying contention of this section is that construct validity is an epistemological concept, referring to the quality of a test, and that test-quality is not an all-or-nothing affair. Thus, while I share the realist intuition that a test can only be said to have construct-validity if its data are caused by some phenomena in the extension of a construct, I am not committed to the idea that in order for a test to have construct validity there has to be a construct that is, in its entirety, a faithful representation of an entity in the world. Nor do I believe that a test can only have construct validity if it successfully operationalizes a given construct in its entirety.

5.1 Lumpy constructs and the question of realism versus pragmatism

As mentioned above, to say of a test that it has construct validity is to assume that there is a construct (theoretical concept) that singles out the cognitive states, processes and/or traits that explain the behavioral phenomena of interest. This is determined by showing that the concept (or its referent) explains the *test data*. (Both of these canonical assumptions go back to Cronbach and Meehl 1955). But as we saw in Section 3, researchers disagree about the construct of implicit attitude. Hence, it is unclear how tests could be validated as long as there is no shared concept of the object of research. As we saw in Section 4, one response to the theoretical disagreements is to adopt a "lumpy" construct that does not differentiate between various conceivable interpretations of what is being measured by a test. The adoption of a lumpy construct reflects the recognition that there is still some uncertainty over which (if any) of the several, conceptually possible "implicit" processes indeed occur when implicit tests are administered, and (in case several of them occur) which of them, if any, will ultimately be regarded as processes of implicit social cognition proper. In the following I will

argue that this is a legitimate response, which does not undermine the basic presupposition about "real" causes that authors like De Houwer bring to the table.

At first glance, the thesis that construct validity can come in degrees might seem counterintuitive: If construct validity of tests requires that the test data be explained by the object of measurement, then this would seem to imply that the object of measurement exists. But existence is a matter of either/or, not of degrees. This objection confuses two separate issues, however: If a test has construct validity, this suggests that there is *something* in the extension of the relevant concept that is being measured by the test. However, this does not mean that the concept in its entirety maps onto one "kind" of thing in the world. After all, the concept might lump in a genuine part of the intended object with some confounders. Conversely, it might only describe a small part of the intended object.[22] This is why we need to distinguish between (a) the issue of realism with regard to the referent of a particular construct, and (b) the issue of the validity of a *test* relative to a given construct, where the former is an ontological issue, whereas the latter is an epistemological issue. Assessments about realism (about a construct) and construct validity (about a test) can diverge: While we cannot be realists about a lumpy construct, a test can nonetheless have a certain degree of validity relative to such a construct if the construct explains part of the test variance. I argue that even though the debate about construct validity is sometimes framed as a disagreement about realism, this is a counterproductive way of framing this, since both sides ought to be committed to the idea that if a construct explains any part of the variance of test behavior then it captures at least some aspect of reality that causally explains the data.

Let me briefly explain this by looking at De Houwer et al. (2009), who explicitly present themselves as scientific realists. In doing so they take themselves to be following Borsboom et al. (2004), who have forcefully argued against antirealist tendencies in the ways in which psychologists often talk about validity. Borsbooms's basic point is that if we go with a notion of construct validation as demanding that there is a construct that explains some of the variance of test result, then this implies the reality of the construct's referent such that it can *cause* the test results. In this vein, De Houwer et al. (2009) argue that "the claim that a measurement outcome provides a valid measure of a psychological attribute implies the ontological claim that the attribute exists in some form and influences behavior" (De Houwer et al. 2009). They elaborate that "it … seem[s] illogical to argue that the statement 'the outcome is a valid measure of attribute X' and the statement 'attribute X does not exist' are both true" (ibid.).

Nosek et al. (2012), in turn, push back with a pragmatist line, arguing that "[p]sychological constructs are unobservable … because they are not physical objects."[23] They even go so far as to say that "the definitions that describe those constructs are arbitrary. Their correctness is a function of how useful they are in connecting theory with evidence" (p. 32). I am sympathetic to the pragmatism expressed here: Concepts can be useful even if they contain some false assumptions about the subject matter. But, of course it does not follow that these assumptions are

[22] I refer to these two possibilities as that of "lumpy" vs. "partial" concepts, respectively. This section is devoted to lumpy concepts. I turn to partial concepts in Section 5.2 below.

[23] It seems to me that a more precise formulation of this point would be that the objects in the extension of psychological constructs are unobservable, but I will not go into this here.

arbitrary. Nor does it follow that there isn't something "real" in the concept's extension that is causally responsible for the test data. In this vein, I argue that we can reject a particular theoretical account of implicit social cognition as false or as prematurely specific (and hence be critical of attempts to construct validate implicit tests relative to that account), while at the same time entertaining the possibility that existing tests might have construct validity relative to a more lumpy construct of the subject matter. Likewise, we can reject a strong realism about the concept in question, while at the same time entertaining the possibility that at least part of the concept captures a mind-independent feature of interest, which in turn explains some of the variance of the test results.[24]

The crucial point is that as long as we are dealing with a subject matter characterized by a lot of epistemic uncertainty and conceptual openness, and as long as current tests empirically individuate their target phenomena in a lumpy fashion, these tests can at best be validated relative to lumpy constructs. There is no reason to suppose that the concepts will stay lumpy. For example, once it becomes possible to devise empirical tests that distinguish between cases where a subject is genuinely unaware of an attitude and cases where a subject is aware of an attitude but has no intentional control over whether or not she acts on it, we might well narrow down our initial construct of implicit social cognition.[25] Indeed, Nosek and his colleagues are well aware of this possibility, as expressed in an article entitled "Implicit Social Cognition: From Measures to Mechanisms" (Nosek et al. 2011). In it they argue that in the first stage of research on implicit social cognition it made sense to adopt a "lumping strategy," which used a variety of different tests and assumed them to be all measuring the same thing. By contrast, they argue that the emerging "age of mechanism" adopts "a 'splitting' strategy that, method by method, prioritizes the more slowly developing precision of terms and operative processes" (Nosek et al. 2011, 153). In other words, it seems that they anticipate the gradual development of more fine-grained and mechanistic concepts and explanations, a development, moreover, that they take to be constrained by mind-independent facts in the world.

Summing up, Nosek et al. put their fingers on an important feature of scientific practice, i.e., the changing and dynamic nature of scientific constructs. They are also right to emphasize the practical utility of tests even at points when the phenomenon at stake is not well understood and when the corresponding constructs are accordingly lumpy.[26] If we adopt the position that to construct validate a test is to show that the variance of data produced by the test is at least partially causally explained by the entities/processes singled out by a given construct, there is no *in principle* reason to suppose that the test cannot be validated relative to that construct, however lumpy it may still be.

[24] Apart from the question of whether there is ever sufficient evidence to be scientific realists about a specific psychological construct, I agree with Hood (2013) that we cannot make sense of measurement practices of psychologists without attributing to them what he calls a "minimal realism", by which he means the commitment that it is in principle possible to justify hypotheses that posit unobservable entities.

[25] This might be determined by using a process dissociation procedure (see Yonelinas and Jacoby 2012)

[26] In this respect their vision of the investigative process is quite compatible with my analyses in Feest (2010, 2011).

5.2 "Partial" constructs

There is a second argument for why construct validity can come in degrees. This has to do with the fact that existing tests may only partially succeed in measuring the entities/processes in the extension of any given concept. Let me explain this by highlighting that in my discussion of lumpy constructs, I have made an unquestioned assumption. The assumption is that conceptual development in science mostly proceeds by way of splitting, where an initial construct "lumps together" several entities that later scientific developments recognize as distinct, and I have discussed the possibility of an empirical method (in this case, a test) as being valid relative to such a lumpy construct. Within the philosophical literature about natural kinds, this is a well-known assumption. Typical examples are the concepts of jade and gold, each of which used to have wider extensions than they do now, based on the fact that there are several substances with similar perceptual properties.[27] In a similar vein, Craver (2007) suggests that discovery in neuroscience often employs a splitting strategy. But does this assumption really do justice to the case of implicit social cognition? It may well describe some cases, but I don't think it's the whole story. Construct validity is a matter of degree for *two* reasons: The first one (which we have discussed extensively in the previous subsection) is that concepts can be lumpy. The second one (which we have not yet discussed) is that the test adequately measures only some of the processes that will ultimately be regarded as relevant to the construct. We may refer to this as a case where a test is construct validated only relative to a "partial" construct.

To explain what I mean, recall that the discussion of the lumpy construct in the previous section was not about the construct of social cognition as such, but only about one of the features associated with implicit social cognition, i.e., *implicitness*. But there are other features, namely that what implicit tests purport to measure are *attitudes*, which in turn are comprised of *cognitive*, *evaluational*, and *emotional* components. Differently put, implicit social cognitions are not simple states, but complex capacities comprised of a variety of distinct phenomena (Machery 2016). Machery argues that implicit attitudes are complex traits, i.e. broad-track dispositions to exhibit particular behaviors, emotions, beliefs etc., and he explicitly contrasts this view of attitudes with a picture that conceptualizes them as (implicit or explicit) mental states. I cannot go into a detailed discussion of this proposal here, but I suggest that Machery is right to highlight that many concepts in scientific psychology describe capacities that – when actualized – involve several distinct mental properties. If this is the case, it is quite conceivable that any given test might successfully measure a particular feature associated with implicit social cognition (albeit perhaps in a lumpy fashion), while doing a less successful job with others. Hence, it might have only low construct validity, because it measures one aspect of the purported entity (implicitness), but not others (e.g., beliefs about, or evaluations of, a given object).

There are two scenarios of how this might be the case: Current tests might fail to have a high degree of construct validity because either (a) there is a highly developed construct of implicit social cognition, but the test does not do a good job tracking all the properties posited by this construct, or (b) there is no highly developed, or generally

[27] See Hochstein (2016) for a critical discussion of this case, both in its own right and as it applies to psychology.

agreed upon, construct. In the case at hand, we see both: First, there are some highly developed constructs, but they cannot yet be validated at the level of grain they required (see de Houwer et al. 2009). And, second, there is disagreement about which constructs one should even attempt to validate (see the De Howuer/Nosek debate). What this highlights is that we should not narrow the debate to the construct validity of existing *tests*, but need to also engage in theoretical debates about what the tests are trying to measure, exactly (see Alexandrova and Haybron 2016, for a similar point). In all likelihood, both tests and constructs will continue to evolve. Focusing our exclusive attention on either whether existing tests have construct validity, or on whether existing constructs are adequate, overlooks that in scientific practice these two questions are closely intertwined. It follows that it would be short-sighted to reject a theoretical concept for the sole reason that current tests have low construct validity.

5.3 Construct validity and the causal explanation of test variance

In the previous two subsection I have argued that since constructs can be lumpy or partial, it is at least in principle possible that a test might have a certain degree of validity relative to those constructs. There is a certain sense in which the notion that (any kind of) test validity comes in degrees may seem obvious to practicing researchers in psychology since test validity is typically captured in statistical terms, i.e., in terms of correlation coefficients. Let me be clear that in this paper I am not aiming for a quantitative analysis of the notion of "degree." Nor am I committed to the idea that degrees of validity vary along one dimension only (quite on the contrary). What I hope to provide, rather, is a *qualitative analysis* of why (quantitative) degrees of construct validity are to be expected in practice, given the lumpy and partial constructs relative to which researchers attempt to develop and validate their tests.[28]

But we now need to address the question of how it can be demonstrated in practice that a given test indeed does have a certain degree of construct validity. Going back to the above-cited popular definition of construct validity, one simple answer is that a test has construct validity if the test results can be shown to be (at least partially) causally explained by the construct in question. But how can this be shown? With respect to this issue, the scientists introduced above (De Houwer and colleagues on the one hand, Nosek and colleagues on the other) have yet another instructive disagreement, which turns on the question of *what kind of evidence is required* to show that the results of a test can indeed be causally attributed to a unique set of entities or processes, which are distinct from those entities/processes responsible for the results of explicit tests. De Houwer et al. (2009) argue that in order to show that a test has construct validity, *experimental designs* are required that specifically test hypotheses at the level of detail and precision required by specific theoretical constructs (see Section 4.2 above). They contrast this with the "correlational methods", which they attribute to Nosek and his colleagues. The experimental studies they call for do not merely analyze the data generated by psychometric tests but are designed to probe specific assumptions about the object of measurement (implicit social cognition), including the mechanisms that bring about the test results.

[28] I would like to thank one referee for this journal for prompting me to clarify this point.

The distinction between experimental and statistical methods of validation is a little misleading since both approaches conduct experiments and both use statistical methods to analyze their data. The distinction is therefore better understood as one between scientists who argue that we cannot make a pronouncement about the construct validity of tests require more specific experiments and one group who think that existing data can be utilized to validate tests without further theoretical and experimental work. The basic point of this latter approach is that what is needed is statistical evidence in support of the hypothesis that implicit and explicit tests have *distinct sources of variance*. The basic idea is that if it can be shown that *at least some* of the test variance can be attributed to a unique latent factor, this suggests that a distinct explanatory construct is required. This is typically attempted by either one of two methods: Convergent-discriminant validation or the multitrait-multimethod approach. Convergent validation proceeds by establishing correlations between the results of tests purporting to measure the same trait (e.g., various implicit tests). Discriminant validation proceeds by demonstrating that there are only low (if any) correlations between tests purporting to measure different traits (e.g., implicit vs. explicit tests). A convergent-discriminant validation was attempted by Cunningham et al. (2001), though with a strong focus on convergent validation. Conducting two confirmatory factor analyses, they argued that despite initially low correlations between various implicit measures, there is a common latent variable (implicit social cognition) and noted that this presumed latent variable is weakly (r = .45) correlated with results on a particular explicit test (the Modern Racism Scale) while at the same time also being dissociated from such results. The latter finding obviously points to discriminant validation.

The method of convergent/discriminant validation faces the problem that it cannot rule out that correlations between the results on similar tests might be due to the instruments rather than the measured attribute. This problem was addressed by Campbell and Fiske's (1959) multitrait-multimethod approach (Campbell and Fiske 1959), which compares the results of similar and dissimilar tests, applied to different objects of measurement. Nosek and Smyth (2007) used this method on implicit tests, arguing that if the same test is used to test different attitudes (about race, gender, political candidates), the finding of merely low correlations between them will provide added confidence in the assumption that it is not the instrument itself that is causally responsible for the test variance, but the specific (presumably independent) attitudes. In a next step, the authors also (like Cunningham et al. 2001) use confirmatory factor analyses (CFA), designed to track the different sources of variance. They conclude that the data collected by explicit and implicit methods have distinct sources of variance.

Let us for the moment assume that Cunningham et al. (2001) and Nosek and Smyth (2007) are right and that implicit test have a unique source of variance. It seems clear that this would ultimately only be able to support a very weak thesis about the construct validity of implicit tests, since (as we discussed above) the construct does not have anything substantive to say about the nature of the two distinct sources of variance and does not claim to provide more than a partial explanation of implicit test variance. I argue that this is ok so long as we are willing to accept that construct validation can come in degrees and that a test can be construct validated relative to a lumpy and/or partial concept. In this vein, De Houwer et al. (2009) are perfectly justified in their assessment that in the long run it is not satisfactory to rest content with a construct of implicit social cognition that remains silent about specific features and mechanisms that

are distinctive of the phenomenon in question. However, according to the reading presented here, Nosek and his colleagues are ultimately not content with it either. As we saw in Section 4.2 above, they view implicit tests as valid relative to a relatively lumpy (and perhaps partial) construct, and they regard both the construct and the tests as mere stepping stones toward more fine-grained empirical knowledge about the phenomenon at hand. It seems, then, that proponents of the two sides of the debate should not disagree with each other.

Summing up, then, I argue that statistical evidence can in principle be used to investigate constructs validity, where it is understood that construct validity comes in degrees and it can be relative to a lumpy and/or partial construct. Whether this is indeed the case for the implicit tests that are currently in use is a different matter. For example, the claim of discriminant validation of implicit tests has recently been questioned in a blog post by Schimmack (2019a) who re-analysed Cunningham et al.'s (2001) correlations and standard deviations) by means of structural equation modeling, but arrives at a different conclusion, according to which "a single factor model actually does fit the data better than the model reported in the original article." Thus, clearly the question of whether the results of implicit and explicit tests can be traced to distinct sources of variance (i.e., can be explained by means of a distinct construct) is still debated. It is not my aim in this article to resolve these debates as this would require a much more thoroughgoing investigation in the philosophy of statistics than can be provided here. It would moreover make prognoses about an ongoing scientific debate. However, I argue that we need to distinguish clearly between the question of whether this or that particular test (for example the IAT) has construct validity and the question of whether particular constructs (for example the construct of implicit social cognition) are worth pursuing and elaborating. While correlational data provide evidence in support of the construct validity of a particular *test*, failure to construct validate the test does not automatically invalidate the *construct*. Statistical models serve the important purpose of putting pressure on existing concepts and tests. They can prompt us to change and adjust the concepts, but they do not necessarily force us to abandon them altogether, as a naïve Popperianism might have us believe.

6 Conclusion

Let me conclude by summarizing the main point I have tried to make about construct validity, discussing the implications of my analysis for philosophy of science, and saying a few words about how this outlook might inform our thinking about implicit biases.

I have used the debate between De Houwer et al. (2009) and Nosek and Greenwald (2009) to highlight and disentangle a variety of questions to consider when analyzing the notion of construct validity: questions about realism and pragmatism, questions about concept development, and questions about the role of statistical vs. experimental evidence in construct validation. I argued that the two groups of authors come to their conflicting conclusions about the validity of implicit tests in part because they have in mind constructs at different levels of lumpiness and precision. Both should agree, however, or so I have claimed, that to construct-validate a test is to say that there is a construct that, at least partially, explains the variance of the test. In this vein, I have

suggested that the disagreement between them is not about scientific realism as such: Both sides agree that construct validation involves showing that test data are causally explained by some processes and entities in the extension of the concept. In this regard their practices of construct validation both involve the notion that in order for a test to be construct-validated, it needs to correspond to a construct that maps on some the mind-independent feature of interest. Where they differ is merely with regard to the question of how much detail and accuracy is required of the construct, and how adequately the test has to operationalizes the construct. In this paper I have argued that a test can have a certain degree of construct validity relative to a "lumpy" construct (i.e., a construct that conceivably lumps together two different possible states of affairs) and/or relative to a "partial" construct (i.e., if it only partially succeeds in operationalizing a given construct).

To say of a construct that it enjoys a certain degree of construct validity is to say that there is something in its extension that causally explains the variance of its results. While this presupposes the basic realist assumption explicated in the previous section, it does not require being a realist about the construct *in its entirety*. This is why test validity and construct realism can come apart: Test validity, I have emphasized, is an epistemological concept, and thus admits of degrees. A full-fledged realism about a specific constructs, by contrast, is an ontological commitment that does not admit of degrees.

I have suggested that existing tests for implicit attitudes may enjoy a certain (however small) degree of construct validity relative to a fairly lumpy and partial construct (though I have acknowledged that this is contentious in the literature). This argument relies on two crucial premises, namely (1) that causal explanations of some of the variation of the test data are satisfactory even if such explanations do not spell out the mechanisms by which the data come about and (2) that the correlational techniques of the multitrait-multimethod approach and factor analyses are adequate to the purpose of establishing such causal explanations. While I have not provided much detail to back up each of these premises, I hope to have laid out some themes on which the discussion should focus.

Lastly, I argued that while the issue of test validity is typically framed as concerning the quality of tests that are currently in use, the debate about the validity of implicit tests offers some insights into a dynamic field of ongoing research. The process of validation is not simply one where the validity of a given test is established once and for all and/or independently of theoretical input. In this vein, my analysis emphasizes that validation is a process in which test development and concept formation mutually inform each other, in a fashion not dissimilar to Hasok Chang's idea of research as an iterative process. (e.g., Chang 2004). As Sullivan (2016) points out, "Oftentimes, an investigator and/or his/her critics wonder whether the investigative procedures she has used in the laboratory satisfy the criterion of construct validity. Such worries prompt the processes of 'construct explication' and 'construct assessment' (668), thereby highlighting the conceptual openness characteristic of an ongoing field of research.

With the analysis presented in this paper I agree with this take, while also emphasizing that scientists may use particular tests as a research tool, because they have reason to believe that it has a certain degree of construct validity, even if the very aim of the research is to articulate the concept in question more adequately. In conclusion, let me emphasize that while moderate construct validity of existing tests may provide us

with good reasons for believing that existing tests are on to something (i.e., measure implicit social cognition, as opposed to "explicit" cognition), there may well be independent reasons for believing in the reality of whatever we are trying to measure. Such reasons can be either theoretically motivated or be rooted in the prima facie plausibility (or "face validity") of the phenomenon under investigation. This strikes me as an important point to make since recent debates about problems with implicit tests such as the IAT are sometimes treated as implying that there is no such thing as implicit bias (e.g., Mitchell and Tetlock 2017, cited above). I argue that even if existing implicit tests have very little construct validity, this should not necessarily deter us from investigating implicit social cognition (at least for the time being). This is a concept that strongly resonates with the everyday experiences of many people, especially women and members of minorities. As emphasized by a growing literature about epistemic injustice, such experiences should not be discarded off-hand. Needless to say, this does not establish the existence of implicit social cognition. Nor does it establish that current implicit tests have construct validity. But it suggests that when it comes to debates about implicit social cognition, the validity of current implicit tests is not the only factor to consider.

Acknowledgements Special thanks go to Dana Tulodziecki and John Carson for their helpful and encouraging comments on the first draft of this article. The author would also like to thank audiences in Hannover, Athens, Düsseldorf, Munich, St. Louis and Berlin. In particular, helpful conversations with Jolene Tan, Sheldon Chow, Caroline Stone, Natalia Washington, Eric Hochstein, Joe McCaffrey, Jackie Sullivan, Carl Craver, Philipp Haueis, and Anna Leuschner are gratefully acknowledged, as are the extremely thoughtful and constructive remarks and suggestions by several referees for this and one other journal.

References

Alexandrova, A., & Haybron, D. (2016). Is construct validity valid? *Philosophy of Science, 83*, 1098–1109.

Amodio, D., & Devine, P. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91*(4), 652.

Bartlett, T. (2017). Can we really measure implicit Bias? Maybe not. *Chronicle of Higher Education*. https://www.chronicle.com/article/Can-We-Really-Measure-Implicit/238807.

Bertrand, M., & Mullainathan, S. (2002). Are Emily and Brendan more employable than Lakisha and Jamal? A field experiment on labor market discrimination. http://www.chicagobooth.edu/pdf/bertrand.pdf

Blank, R., Dabady, M., & Citro, C. (Eds.). (2004). *Measuring racial discrimination. Panel on methods for assessing discrimination. Committee on National Statistics. Division of Behavioral and Social Sciences and Education*. Washington, CD: The National Academic Press.

Borsboom, D., Mellenbergh, G., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071.

Brownstein, M. (2017). Implicit bias. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2017 ed.). https://plato.stanford.edu/archives/spr2017/entries/implicit-bias/

Brownstein, M. (2018). *The implicit mind. Cognitive architecture, the self, and ethics*. New York: Oxford University Press.

Brownstein, M., & Saul, J. (Eds.). (2016). *Implicit Bias and philosophy*. Oxford: Oxford University Press.

Byrd, N. (2019). What we can (and can't) infer about implicit bias from debiasing experiments. *Synthese* (online first). https://doi.org/10.1007/s11229-019-02128-6.

Campbell, D., & Fiske, S. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. New York: Guilford Press.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress. Oxford studies in the philosophy of science*. New York: Oxford University Press.

Craver, C. (2007). *Explaining the brain: mechanisms and the mosaic unity of neuroscience.* New York: Oxford University Press.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

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

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior, 37*(1), 1–20.

De Houwer, J. (2014). Why a propositional single-process model of associative learning deserves to be defended. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual processes theories of the social lmind* (pp. 530–541). New York: Guilford.

De Houwer, J., & Moors, A. (2010). Implicit measures: Similarities and differences. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 176–193). New York: Guilford Press.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 135*(3), 347–368.

Evans, J. S. (2006). *Dual system theories of cognition: Some issues*. Proceedings of the Annual Meeting of the Cognitive Science Society, 28. Retrieved from https://escholarship.org/uc/item/76d4d629

Fazio, R., & Olson, M. (2003). Implicit measures in social cognition research. Their Meaning and Use. *Annual Review of Psychology, 54*, 297–327.

Fazio, R., & Olson, M. (2014). The MODE model: Attitude-behavior processes as a function of motivation and opportunity. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind* (pp. 155–171). New York: Guilford Press.

Fazio, R., Jackson, J., Dunton, B., & Williams, C. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69*(6), 1013–1027.

Feest, U. (2010). Concepts as tools in the experimental generation of knowledge in cognitive neuropsychology. *Spontaneous Generations: A Journal for the History and Philosophy of Science, 4*(1), 173–190.

Feest, U. (2011). Remembering (short-term) memory. Oscillations of an epistemic thing. *Erkenntnis, 75*, 391–411.

Feest, U., & Steinle, F. (2016). Experiment. In P. Humphreys (Ed.), *Oxford handbook of philosophy of science* (pp. 274–295). Oxford: Oxford University Press.

Forscher, P., Lai, C., Axt, J., Ebersol, C., Herman, M., Devine, P., & Brian Nosek, B. (2017). A meta-analysis of change in implicit bias. https://www.researchgate.net/publication/308926636_A_Meta-Analysis_of_Change_in_Implicit_Bias

Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass, 5*(10), 914–926. https://doi.org/10.1111/j.1747-9991.2010.00330.x.

Gawronski, B., & Bodenhausen, G. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692–731.

Gawronski, B., Brannon, S., & Bodenhausen, G. (2017). The associative-propositional duality in the representation, formation, and expression of attitudes. In R. Deutsch, B. Gawronski, & W. Hofmann (Eds.), *Reflective and impulsive determinants of human behavior* (pp. 103–118). New York: Psychology Press.

Gendler, T. (2011). On the epistemic costs of implicit bias. *Philosophical Studies, 156*, 33–63.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4.

Greenwald, A., & Banaji, M. (2015). Statistically small effects of the implicit association test can have societally large effects. *Journal of Personality and Social Psychology, 108*(4), 553–561.

Greenwald, A., McGhee, D., & Schwartz, J. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480.

Greenwald, A., Andrew Poehlman, T., Uhlmann, E. L., & Banaji, M. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17–41.

Guala, F. (2012). Experimentation in economics. In U. Mäki (Ed.), *Handbook of the philosophy of science, Philosophy of economics* (Vol. 13, pp. 597–640). Boston: Elsevier/Academic Press.

Hochstein, E. (2016). Categorizing the mental. *The Philosophical Quarterly, 66*, 745–759.

Hood, B. (2009). Validity in psychological testing and scientific realism. *Theory & Psychology, 19*(4), 451–473.

Hood, B. (2013). Psychological measurement and methodological realism. *Erkenntnis, 78*, 739–761.
Jost, J., Rudman, L., Blair, I., Carney, D., Dasgupta, N., Glaser, J., & Hardin, C. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in Organizational Behavior, 29*, 39–69.
Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Strauss & Giroux.
Krickel, B. (2018). Are the states underlying implicit biases unconscious? – A Neo-Freudian answer. *Philosophical Psychology*. https://doi.org/10.1080/09515089.2018.1470323.
Lavrakas, P. (Ed.). (2008). *Encyclopedia of survey research methods*. London: Sage.
Machery, E. (2016). De-Freuding implicit attitudes. In M. Brownstein & J. Saul (Eds.), *Implicit bias in philosophy. Volume I. Metaphysics and epistemology* (pp. 104–129). Oxford: Oxford University Press.
Machery, E., Faucher, L., & Kelly, D. (2010). On the alleged inadequacy of psychological explanations of racism. *The Monist, 93*(2), 228–255.
Mandelbaum, E. (2015). Associationist theories of thought. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2016 ed.). https://plato.stanford.edu/archives/sum2016/entries/associationist-thought/
Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *NOUS, 50*(3), 629–658.
McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination and racism* (pp. 91–126). New York: Academic.
Mitchell, G., & Tetlock, P. (2017). Popularity as a poor proxy for utility. The case of implicit prejudice. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions*. Hoboken: Wiley.
Moss-Racusin, C., Dovidiob, J. F., Brescollc, V., Grahama, M., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences of the United States of America, 109*(41), 16474–16479. https://doi.org/10.1073/pnas.1211286109.
Nosek, B., & Banaji, M. (2009). Implicit attitude. In P. Wilken, T. Bayne, & A. Cleeremans (Eds.), *Oxford companion to consciousness* (pp. 84–85). Oxford: Oxford University Press.
Nosek, B., & Greenwald, A. (2009). (Part of) the case for a pragmatic approach to validity: Comment on De Houwer, Teige-Mocigemba, Spruyt, and Moors (2009). *Psychological Bulletin, 135*(3), 373–376.
Nosek, B., & Riskind, R. (2012). Policy implications of implicit social cognition. *Social Issues and Policy Review, 6*(1), 113–147.
Nosek, B., & Smyth, F. (2007). A multitrait-multimethod validation of the implicit association test. Implicit and explicit attitudes are related but distinct constructs. *Experimental Psychology, 54*(1), 14–29.
Nosek, B., Hawkins, C., & Frazier, R. (2011). Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Sciences, 15*(4), 152–159.
Nosek, B., Hawkins, C., & Frazier, R. (2012). Implicit social cognition. In S. Fiske & C. N. Macrae (Eds.), *Handbook of social cognition* (pp. 31–53). New York: Sage.
Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/a0032734 Advance online publication.
Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology, 34*, 181–209.
Payne, B. K., & Gawronski, B. (2010). A history of social cognition. Where is it coming from? Where is it now? Where is it going? In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition. Measurement, theory, applications* (pp. 1–15). New York/London: The Guilford Press.
Schimmack, U. (2019a). No discriminant validity of implicit and explicit prejudice measures. Replication Index. Retrieved August 13, 2019, from https://replicationindex.com/2019/02/02/no-discriminant-validity-of-implicit-and-explicit-prejudice-measures
Schimmack, U. (2019b). No incremental predictive validity of implicit attitude measures. Replication Index. Retrieved August 13, 2019, from https://replicationindex.com/2019/02/04/no-incremental-predictive-validity-of-implicit-attitude-measures
Singal, J. (2017). The creators of the implicit association test should get their story straight. https://nymag.com/intelligencer/2017/12/iat-behavior-problem.html.
Snyder, J.G. & Osgood, C.E. (Eds.). (1969). *Semantic differential technique: a sourcebook*. Chicago: Aldine.
Steel, D. (2008). *Across the boundaries: Extrapolation in biology and social science*. Oxford: Oxford University Press.
Sullivan, J. (2016). Construct stabilization and the unity of the mind-brain sciences. *Philosophy of Science, 83*, 662–673.

Tal, E. (2017). Measurement in science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2017 ed.). https://plato.stanford.edu/archives/fall2017/entries/measurement-science/

Tetlock, P., & Mitchell, G. (2009). Implicit bias and accountability systems: What must organizations do to prevent discrimination? In B. M. Staw & A. Brief (Eds.), *Research in organizational behavior* (Vol. 29, pp. 3–38). New York: Elsevier.

Yonelinas, A., & Jacoby, L. (2012). The process-dissociation approach two decades later: Convergence, boundary conditions, and new directions. *Memory & Cognition, 40*, 663–680.