Chapter to appear in *Research Handbook on the Replication Crisis* Ed. David Trafimow*

Beyond reproducibility, making a psychological application safe to use

Magda Osman[1]

The replication crisis inevitably deepens issues of trust in the methods and findings that the scientific community uses to advance our collective understanding of human behavior. In turn, any doubts over psychological research undermines the use of any products that it develops (e.g. assessment tools, scales, frameworks) outside of academia. With the help of the Implicit association test (IAT) this chapter shows how measurement criteria (reliability, validity) are used to determine metrics for determining a sound evidence base for the test. However, escaping the replication crisis on the grounds that a body of work is highly replicable and has high predictive validity is not sufficient when it comes to the issue of applicability. The chapter argues that without a good theoretical basis to secure construct validity the same research would be ruled out of being safe for use outside of the sciences. The reason for being even more stringent about the concepts being investigated is that the stakes are higher when the decisions based on them carry consequence in practical settings.

1. Judge Business School, University of Cambridge, Trumpington Street, Cambridge, CB2 1AG, England. UK.

   Correspondence to: m.osman@jbs.cam.ac.uk

Introduction

Beyond the problem of reproducibility

Estimating the reproducibility of experimental studies is a bit like looking at a thermometer. The thermometer tells us that a suspected ill person has a high temperature, but not the underlying cause of the high temperature. By the same token, diagnosing the problem of reproducibility means deciding how much is down to validity as well as reliability, and whether there is a possible crisis of validity (Dennis, 2013). The latter is more profound because it means that the underlying concepts that measures are designed to expose are unsound. For instance, even if a measure produces replicable findings (e.g. Klein et al., 2014) it could arguably still fail on validity grounds (Trafimow, 2004, 2019a). The same problem concerns applied psychological research. A tool may produce consistent results on each administration, but if the concepts on which it is based on are unsound then the decisions made from it are also going to be unsound. The aim of this chapter is to explore this point in more depth using the IAT as a case study, with the aim of showing what is needed to limit a possible applicability crisis.

Why is the IAT a good case study?

The IAT is 30 years old, and it is one of the most famous measurement tools to come out of psychology. It is a way to indirectly measure an appraisal of something of interest (Greenwald, McGhee, & Schwartz, 1998). The appraisal can be expressed negatively or positively. By setting up different associations between the thing of interest and its valanced attributes, based on speed of responses to different combinations of associations, it elicits genuine "implicit" attitudes. A complex procedure is needed to induce genuine attitudes. The assumption is that there is a layer underneath public presentations of attitudes which more closely corresponding to an individual's genuine attitudes (Baumeister, 1992; Epstein, 1991). Second, we accumulate experiences, which are formative of our attitudes and behaviours. While they are inaccessible consciously, they exert influence without out conscious awareness (Greenwald & Banaji, 1995).

Because of its fame the IAT has been scrutinised closely. It was on the radar of the open science initiative when the problems of reproducibility come to the fore in the 2010s. Some might say that it came out of the replication crisis relatively unscathed (e.g., Conners et al., 2020; Klein et al., 2014). For reasons made clear later in this chapter, the IAT hasn't escaped the ongoing problems of its reliability. Relatedly, its predictive validity is still being contested, suggesting its limited ability to predict behavior outside of what it is designed to measure. The more profound problems it faces are on conceptual grounds, revealing concerns about its construct validity. The assumptions behind the IAT treats implicit and explicit cognitions as qualitatively distinct (Greenwald & Banaji, 1995). This type of conceptual framework has been met with significant challenges (e.g., Osman, 2004, 2010, 2014, 2016, 2018, 2021; Osman et al., 2017; Osman et al., 2020a; Osman et al., 2020b).

If the IAT remained a famous measure in the psychological science, then a crisis of replicability and validity would be an internal matter for psychological science to deal with. However, the IAT has (un)wittingly been instrumental to addressing socially problematic

behaviors through training regimes used in many public and private sector organisations (e.g., Calanchini, et al., 2021; Chang et al., 2019; FitzGerald, et al., 2019; Lai et al., 2014; Newkirk, 2019). In addition, to date, it used in some form of diagnostic capacity in clinical practice, policing, law, and advertising. Given this, the IAT is an ideal case study to explore the main argument of this chapter, which is: Even if a tool is shown to produce highly replicable findings, and has high predictive validity, so long as the concepts it rests on fail on construct validity grounds, there is no justification for the test's application in any diagnostic capacity, or to inform real world consequential decisions.

The chapter starts by exploring the problem the IAT was designed to solve. The discussion is followed by a consideration of the IAT's reliability and predictive validity. The final section focuses on construct validity, which is central to the argument proposed in this chapter. To reach a point where a psychological tool has real world applications, the question to ask is does it have construct validity? To answer this requires conventions that can be used to support effective theory development. One such example is Trafimow's (2019a) Taxonomy of assumptions: Theoretical, Auxiliary, Statistical, Inferential (TASI). Using a systematic approach that transparently exposes the relationship between all four types of assumptions is a productive way of advancing theory and practice, as well as exposing how to improve construct validity. If at the point of launching a tool for application the answer to the question is still 'not quite', then there is a problem. Neither reliability nor predictive validity can be used as sufficient grounds for launching a tool as an application to the outside of academia. With no sound foundation, the tool will invariably do more harm than good.

Section 1: The Origin story of the IAT

As early as the 1940's, professionals drawing on psychological research (e.g., psychologists, Clinicians, Pollsters, Marketers) confronted a problem which is as relevant today. For certain direct measures of beliefs, attitudes, and emotions, there were considerable distortions in responses that suggested they could not be genuine. In the case of attitudes, Greenwald and Banaji's (1995) solution to uncovering genuine appraisals was an indirect measure, the IAT. By going back to origins of the problem, the follow section reveals that while it was borne out of indirect measures developed in the 1940's and 1950's, the theoretical concepts that it refers to are a departure from its origins.

Not Taking things at face value, the conscious version

Early research in psychology started to expose critical issues regarding measurement of psychological constructs; in particular, the concealment of "true" responses to direct measures of attitudes, beliefs, opinion, and feelings (Cantril, 1944a, 1944b, Edwards, 1957). This issue was most prevalent when asking people to express their views on controversial topics (e.g., general attitudes towards minority groups, voting preferences). Improvements to direct measures needed to consider external (e.g., social, political conditions) as well as internal factors as barriers to a genuine response. This is where concepts such as social

desirability, acquiescence and a whole host of others was borne[1] (Dickens, 1963; Greenwald[2] & Satow, 1970; Jackson & Messick, 1958, 1961; Messick & Jackson, 1957, 1958). To address this, the Marlowe-Crowne social desirability scale (Crowne & Marlowe, 1964) employed reversals, such as "I am sometimes irritated by people who ask favors of me", compared with "I never resent being asked to return a favor.". In forced choice set ups like this, it was possible to detect the socially desirable responses based on inconsistencies (i.e., false to the first, and true to the second). Following the same logic, simply agreeing (i.e., acquiescence) to any statement could reveal contradictory political positions. For example, in Messick and Jackson's (1957, 1958) authoritarian scale there were mirrored statements such as "Obedience and respect for authority are the most important virtues children should learn", and "A love of freedom and complete independence are the most important virtues children should learn." If people acquiesced, affirming the first statement indicated an authoritarian position, and affirming the second indicated a libertarian position. The inference from this would be that only one could be true for an individual at any one time, so acquiescence was a response set, rather than a genuine reflect of an individual's position on the topic.

In many studies examining responses to these types of scales there were no reliable negative correlations between reversed items, which highlighted two problems. The first is that reversals are hard to get right, that is, precise mirror opposites of statements are hard to construct. The second issue is that in the mind of the respondent[3], the mirrored statements are not necessarily psychologically valanced symmetrically. While not easily solving the first issue, rather than forced choice options (e.g., yes, no), a proposed solution to the second was to continuous response scales. A wider range of response options would be less confrontational and allow for degrees of acceptance or rejection of the statements. This solution exposed a host of other problems, such as extremity bias, and other response patterns where respondents worked around the scales to shield their "true" responses.

The theorising to support methodological approaches to survey design can be seen in standard textbooks on research methods today. This work highlights the pit falls of using certain types of items and response options, as well as explaining the psychological reasons

[1] This concept of acquiescence is a more nuanced interpretation of the type of response sets that have since been investigated in psychology that include: positivity bias, which essentially refers to response options that reflect positive outcome or evaluations, also related to social desirability, which is presenting your views in a light that is socially desirable, or experimenter bias, which is responding in ways that are going to be consistent with, or please the experimenter, and extremity bias, which is an aversion to select an option that is at the extreme end of a response scale (either positive or negative).

[2] A different Greenwald (Herbert) to the Greenwald (Anthony) of IAT fame.

[3] To illustrate, there may be hundreds of instances where I offer to help when asked. However, I can also call to mind when I was annoyed by someone asking me for help because I knew that they were capable and were wasting my time. My recall of the latter instance is the most recent one in memory, and/or because of biased sampling from memory negative instances come to mind quicker than the hundreds of accumulated positive instances (Rakow et al., 2008; Walasek, L., & Stewart, 2015). So, which is right? I "never resent being asked to do a favour" (which could be acquiescing), or "I am sometimes irritated by people who ask favors of me" (which might be viewed as a genuine response). Each item means different things in relation to the psychological salience of the sampling of examples I draw from memory, and how I evaluate those instances when responding to each question. I could respond yes to both, which may be a fair reflection of what I think, but an experimenter will interpret my responses as inconsistent.

why responses end up being distorted. Patterns in distortions reflect a complex set of biases that were internal to the respondent (e.g., extremity bias, acquiescence bias). The patterns reflected properties of the measurements themselves that needed improving (e.g., reversals, forced choice). The patterns reflected impositions informed by social, cultural, and political conditions that prevented honest responses (e.g., social desirability bias).

One path that psychological research took to solve the issue of distortions to direct measures was to improve rather than abandon them. As well as minimising the causes of distortions, this approach was instrumental in the development of the metrics in most psychological assessment scales used to date.

Not Taking things at face value, the unconscious version

An alternative path to improving direct measures of attitudes was to explore the use of indirect measures (e.g., Hammond, 1948). Hammond investigated ways to quantify the influence of prior experience on distortions in responses using indirect measures. To anticipate the direction and magnitude of distortions required introducing controlled distortions in measurement tools. Participants were presented force-choice response options, where for factual questions there was a correct response, but for "non-factual" questions, both options presented to people were incorrect. In attitude-non-factual versions respondents were told in advance that none of the options they were presented were correct, and they should respond to indicate their attitudes rather than a fact-based opinion. This could be compared with a group that were not told about the rigged set up. By comparing the two conditions, the direction and magnitude of distortions could be calculated.

Hammond's (1948) error-choices technique was a popular compliment to direct measures and spawned an entire programme of research in psychology exploring the use of indirect measures. The indirect measures were alternatively referred to as "projection techniques", the most famous of which is the Rorschach ink blot test (Campbell, 1950; Haire, 1950; Meehl & Rosen, 1955[4]). The theoretical assumption behind these techniques[5] rested on a distinction between public and private psychological constructs (e.g., Rapaport, 1952). The theorists were at pains to distinguish themselves from psychodynamic conceptions of an unconscious mind. The theory of projective techniques (Frank, 1939; Rapaport, 1952) proposes that a public presentation of the self typically requires an effort to conform to social norms. The private self is structured around a personality that depends on dispositions. Projective techniques are designed to induce dispositions to be revealed by "projecting" them on to unstructured, abstract, non-verbal materials. Moreover, the projections can be treated as unfiltered because the materials that the projections are

[4] Then and now, the troublesome spectres of reliability and validity raised doubts about their usefulness (for discussion see, Meehl & Rosen, 1955), and are still employed to date (for discussion see Joffe & Elsey, 2014 and Lilienfeld, Wood, & Garb, 2000).

[5] See (Campbell (1950) and Bruner (1950) for detailed descriptions of, and analysis of projection techniques (e.g., Thematic Apperception Test, Human Relations Test, doll play techniques, movie story game, Sentence completion tests, Rosenzweig Picture frustration test).

displaced to are not value laden, unlike direct measures which only access the public presentation of the self.

While the theoretical basis of indirect measures designed in the 1940s and1950's was to delineate between a public and private psychological identity, this was not a dual framework account of cognition. Namely the theory didn't require making additional assumptions regarding conscious processes and unconscious processes. In fact, the theory provided an explanation for why external factors (e.g., society, politics) could pose problems when respondents were faced with direct measures. Distortions in response were inevitable expressions of inconsistencies between public and private reflections of the self. At the time of their development indirect tests were a way to remove these types of difficulties, and in turn the responses could be used as complements to, rather than replacements of, direct measures.

Section 2: Reliability and validity

Since the 1940s when the problem was first documented, psychological research still faces the presence of distortions in response to direct measures of peoples' mental contents. The theorising and experimental design at the time tells us what assumptions they had made to address the problem. One track followed the assumption that mental contents could be accessed directly so long as the methods were improved. The other track assumed that because direct measures were too value-laden to uncover accurate responses, indirect methods were needed. Crucially, neither approach depended on recourse to a dual framework of cognition that divides processes into conscious and unconscious ones.

What will be made clear in the detailed explication of the IAT is that it solves the problem of accessing honest representations of mental content by combining techniques from both tracks (e.g., reversals, projective techniques). However, it depends on a critical assumption that mental contents were comprised of implicit cognitions that were not consciously accessible. This means saddling itself with considerable conceptual baggage. To uncover the baggage, we need to inspect the tool itself, and why it is vulnerable to concerns regarding measurement criteria that it has been appraised against.

So, what is the IAT?

"Are you an introvert or an extrovert?". Despite knowing they're introverted one respondent's embarrassment leads them to say "extrovert". Another respondent who isn't especially reflective can't tell either way, so says "introvert". A third respondent has been battling with social anxiety and has been working on building on their confidence, and so they answer "extrovert" though neither option seems right to them. While extreme, the example here illustrates why going the route of a direct question for a topic that might be hard for people to give an accurate response to makes the case for alternative approaches. The IAT might be a good alternative, and this example has been used by researchers to show how (Schnabel, Asendorpf, & Greenwald, 2008).

To take the test a person needs to participate in an initial training which starts with learning the response to "me", using the left key, and "others" using the right key. Then training is

followed with learning the corresponding responses to “shy” as the left key, and “sociable” as the right key. The next training set involves combining the words, so that when the words “me” and “shy” appear the left key is pressed, and for “others” and “sociable” the right key is pressed. Then a reversal is included, so for the word “me” the right key is appropriate, and for “others” the left key is appropriate. In a final test people are tested with a reversal of the target, but not the attribute, so words “others” and “shy” are responded to with the left key, and “me” and “sociable” are responded to with the right key.

If people are genuinely shy, then they ought to respond faster to associations of “me” “shy” and by the same token respond faster to “others” “sociable” when compared to “me” and “sociable” and “others” and “shy”. The difference in speed to the different combinations reveals the tendency towards the “true” associations, where the difference between the pairs that are genuine associations are faster to respond to than their reversals. The bigger the difference between responding faster to authentic associations, and slower to inauthentic associations reflects how strongly held the genuine associations between target and attribute are.

The IAT combines indirect techniques (e.g., projections) with direct techniques (e.g., reversals), but the process is complex as are the assumptions behind the set up. As a measurement tool, we now need to consider the qualities of measurement. This allows us to say something about whether it systematically accesses authentic attitudes (i.e., reliable) which is what it was designed for. We can also then say something about its ability to predict other behavioural outcomes of interest (i.e., predictive validity). In both cases assumptions are made which touch on construct validity. Therefore, we need to know how sound the underlying concepts behind the tool are (i.e., construct validity). There is a dedicated section on this at the end given its criticality to the main argument proposed in this chapter.

So, is the IAT reliable?

Reliability in the context of the IAT means several things. The first is that the test has high test-retest reliability. This means that there is a high correlation in responses between a person taking the same IAT at two different time points. In fact, the reason why this concept was developed was precisely because of problems in inferring the stability of responses, most notably explored in psychometric tests and attitudinal questionnaires (Guttman, 1945). A test-retest reliability coefficient also requires an agreement of an acceptable level of variance between responses to the test taken at two different time points. In turn this allows for inferences about the stability of responses overtime as well as an accepted level of measurement error and noise.

The assumptions that need to be contended with here aren’t easy because they also carry implications for how to interpret high test re-test reliability or low-test re-test reliability (for a detailed account see, Polit, 2014). A feasible time interval between the administration of the tests needs to be determined. If it is too short, then critics might say it was inevitable that high test re-test reliability was found. If the interval is too long, then there is a greater risk of low-test re-test reliability. The latter presents a further problem because it requires theoretical assumptions about the phenomenon the IAT is indexing.

If we return to the illustration, when it comes to introversion and shyness, we might say that the IAT is accessing a trait, in which case we assume that what it is indexed should be stable. In turn, this also means that we should expect high test re-test reliability. But if the IAT is indexing attitudes about something that is novel, then it might be fair to assume that the attitude is a state. Thus, because a state is more labile because it will change with more experience, we might expect low test re-test reliability. Therefore, whatever the IAT is measuring (i.e., trait or state) needs to be specified in advance because of the implications it has for what level of variability is expected between the two administered tests. In fact, this is a moot point in attitudinal research in general (Steyer & Schmitt, 1990) as well as the IAT in particular (e.g., Jusepeitis & Rothermund, 2022).

Even when it comes to test-retest reliability the concept of validity matters. This is because there are theoretical assumptions that sit behind how to set the intervals between the two tests and how to interpret variance. One way to addressed this is to state from the outset the theoretical basis on which the constructs being used in each IAT should be treated (i.e., stable or labile). This avoid of the problem of only saying after the fact that high test-retest reliability is found that the IAT was indexing stable properties, or only after the fact that low test-retest reliability was found because the properties were labile.

However the test re-test reliability issues are resolved, the other matter of reliability is internally consistency (e.g., Trafimow, 2004). To understand this, it is easier to think about direct measures first to then see how the concept extends to the IAT. For direct measures we might have several items, or several scales developed to index the same underlying psychological construct. Good internal consistency would mean that the various scales claimed to index the same construct independently, and equally well, in turn correlate highly with each other. In this way the research community must make assumptions about the various scales available to decide which are worth including in a battery of tests to reveal the internal consistency of any one measure.

How does internal consistency apply to the IAT? To start with, other good independent measures need to be chosen to correlate scores with the IAT. But this presents another issue that needs to be resolved. Should the independent measures be other indirect measures like the IAT (option 1), or should they be direct measures (option 2)? As an alternative to this, which is a simpler approach, could be to split the responses in an IAT (e.g., odd and even trials), to look at the level of internal consistency within an individual IAT (option 3). In fact, all three options have been used to determine internal consistency of IATs (e.g., Banse et al., 2001; Conti et la., 2012; Hussey & Drake, 2020). Unfortunately, all three options carry issues.

Option 1 is problematic because it compounds several issues about which psychological constructs are being measured indirectly and how stable they are, and the added noise incurred from indirect methods. Option 2 seems perverse given that the rationale for the IAT was in response to problems revealed when using direct measures. Moreover, if there is low internal consistency, then there is an obvious ready-made explanation for this. The direct measures won't be able to access authentic attitudes, so there wouldn't necessarily be any compatibility with response to the IAT. Option 3 isn't straightforward either. There are several ways of implementing split-half reliability and they produce different outcomes

which have been critiqued on statistical grounds as far back as the 50's (Tryon, 1957) and recently (Pronk et al., 2022).

Given that research on the IAT has not been able to definitively resolve problems arising from test-retest reliability or internal consistency, for now it doesn't seem like the IAT could be considered reliable. Reliability speaks to the issue of the level of reproducibility of studies using the IAT, and in turn, ongoing discussing regarding the replication crisis in psychology. However, given the main tenet of this chapter, even if all problems with it IAT's reliability were resolved, it matters little if what it is measuring is based on unsound concepts.

So, is the IAT valid?

Understanding what validity means is no less tricky a concept than reliability (Clark & Watson, 2019). A test isn't validated, *per se*. A test is valid in as much as there is an acceptance about what can be inferred from what it measures (Cronbach & Meehl, 1955). Validity can be construed of in several ways (face validity, construct validity, concurrent validity, predictive validity, external validity) (Clark & Watson, 2019; Trafimow, 2004). Of the many ways that validity can be thought of, they essentially boil down to whether a test measures the phenomenon it set out to (construct validity) and whether the test predicts something else of interest (predictive validity). The next section focuses on construct validity in detail, so the focus here is to determine if the IAT has good predictive validity. An example of this would be that the IAT is taken by managers in an organisation. The responses to the IAT show negative attitudes towards women. A corresponding behavioral outcome is the managers' decisions regarding stymying career progress, but that this occurs only for women employed in the organisation. If there is a relationship that is stable, then the IAT can be used to predict similar behavioral outcomes in other organisations. The power of this type of predictive ability would be that a single test could be used to determine who would show specific types of positive or negative behavioral outcomes towards targets of interest.

In fact, predictive validity is a useful concept for the IAT because it can handle a range of social and cognitive phenomena, such as the attitude-behaviour gap. In brief, this is where people state a particular attitude, but that it does not correspond to actions they perform. For instance stating that one is not wasteful but does not recycle. By side stepping what is stated explicitly, the IAT should be a better predictor of actual behavior because the test indexes the genuine attitudes. So, does the IAT have high predictive validity? In short, a generous response would be that it is still hard to tell. There are those that reveal weak to no predictive validity of the IAT (Carlsson & Agerström, 2016; Kurdi et al., 2019; Meissner, et al., 2019; Oswald et al., 2013, 2015; Schimmack, 2021), and those that show reasonable predictive validity (Axt et al., 2022; Greenwald et al., 2009; Moore-Berg et al., 2019; Richeti et al., 2007; Richetin et al., 2010).

However, even if we reached a point that the IAT had high predictive validity, if we can't be sure that the concepts the IAT rests on are sound, we still face a problem. The problem being that we do not have a good basis on which to interpret the relationships between the responses to a test and the behavioral outcomes it predicts. Because of this, the IAT would

still fundamentally fail as a measure for wide use outside of the sciences because the consequences of decisions made from the results of the test cannot be confidently justified.

Section 3: Construct validity: Grounds for making an application safe to use

While it is unlikely that decisions are based solely on the IAT, it nonetheless is used to inform decision-making in a variety of domains: clinical practice (e.g. Tello et al., 2020), policing (e.g. Cesario, 2022), law (e.g. Lane et al., 2007), advertising (Maison et al., 2004), equality diversity and inclusion training (Greenwald & Lai, 2020). This means that there is some assignment of responsibility to the IAT because of the role it can play in decision-making. It is precisely for this reason that construct validity is crucial. There must be confidence that the measure is accessing what it is thought to. In the case of the IAT, strong assumptions are made about it accessing mental contents that are unknowable and are unverifiable by independent means. Therefore, this makes the strongest case for why construct validity is crucial when it comes to assessing whether a tool should be used to inform any consequential decisions in applied domains.

Analysing assumptions

To confidently say that a measurement tool is measuring the psychological construct intended we need to assess it according to its construct validity. The construct itself needs to be operationalised, and behind this is the theoretical machinery that structures the assumptions on which the construct can be defined. To illustrate the theoretical structures that support a construct we can used Trafimow's (2019a) TASI Taxonomy. To get to a position where a researcher empirically investigates a psychological phenomenon of interest are broad level assumptions. For instance, a broad assumption could be that people are able to give honest responses because mental contents are consciously available to people, and so good direct measures are needed to elicit them. In the TASI Taxonomy this assumption is a *theoretical assumption*. This type of assumption contains non-observational terms which cannot be directly tested empirically. So, to do this, we need to traverse the terms into observable ones which requires *auxiliary assumptions*. This could be something like the following, poor direct measures typically use dichotomous response options, and good direct measures use continuous scales. By doing this it is possible to specify properties that can be observed which can be translated into other types of assumptions. The aim is to get to a set of coherent theoretical and logically defensible auxiliary assumptions so that empirical hypotheses are generated. Statistical assumptions (i.e., sample statistics) are needed for hypothesis testing, and for further precision inferential assumptions can be specified (i.e., population parameters). In this way a set of assumptions from general to specific can be used to help articulate a construct and to operationalise it.

If we return to the IAT, our first port of call is the definition of implicit attitudes offered. "Implicit attitudes are manifest as actions or judgments that are under the control of automatically activated evaluation, without the performer's awareness of that causation" (Greenwald et al., 1998, p 1464). A broad level theoretical assumption from this is that there are cognitions that exist that are outside of conscious awareness but that have some

influence on observable behaviors[6]. This isn't enough to generate testable hypotheses yet. So, some auxiliary assumptions are needed to transform the concept 'automatically activated evaluation' into something that is observable, such as responses to an IAT. One of the auxiliary assumptions could be that an evaluation is an appraisal of a target, and this is valanced, so that the target is typically appraised either positively or negatively. Therefore, an association between a target and its attributes is an index of an evaluation. Another auxiliary assumption could be that for it to be implicit, the evaluation is made automatically, and it is on this basis that an expression of that evaluation, as an attitude, is later observed through an action or judgment.

Given that evaluations are described as automatic, and because evaluations are key to attitudes, we will also have to make a further assumption about the speed of the evaluation made. At this point we might specify in advance a statistical assumption about sample statistics when measuring response to an IAT to quantify what automatic is. The concept of automatic has its complement, controlled processes, either based on qualitative or quantitative distinctions. If the assumption is that they are qualitatively distinct then there are issues about where on a continuous measure (e.g., time) a division is made. Below a specified time interval would constitute automatic evaluations, and above the internal constitutes controlled evaluations[7]. If the assumption is that automatic and controlled processes are quantitatively different then a statistical assumption is still needed. For instance, this would require stating a typical response latency that can be classified as automatic and similarly for a response classified as controlled. Assumptions about variance in response latencies also need to be considered given individual differences. An auxiliary assumption here might be that variance of response latencies for automatic responses are minimal because the evaluations are much more stable. Another auxiliary assumption is that the strength of the evaluation is indexed by the difference in speed to an appraisal of the target when they are aligned or misaligned to the individual's actual automatic evaluation of the target. In turn this can be used to generate a further statistical assumption and in turn inferential assumptions can be used to help interpret the results of an IAT.

Does the IAT have construct validity?

In short, the response to the question is no. The psychological construct "implicit attitude" covers a multitude of interdependent assumptions across the TASI Taxonomy. The number of assumptions need not count against the validity of the construct. It is likely that conducting this type of analysis for any psychological construct would reveal several theoretical, auxiliary, statistical, and inferential assumptions. To determine construct validity we need to assess the soundness of the psychological construct. If in the process of reverse engineering

[6] We can see already that this broad theoretical point has clear implications for predictive validity because implicit attitudes are assumed to be causally efficacious, which is a deeply problematic assumption (Machery, 2022).

[7] By physics and mathematical modelling standards bifurcation in complex dynamical systems such as the neurological operations of the brain is not feasible in the way it is construed in dual process theories in psychology (e.g. Gross, 2021).

a pattern of responses from a measure back to what it is expected to measure we must commit several acts of faith along the way, then its construct validity isn't sound.

A pattern of responses to an IAT requires us to accept several assumptions that involve acts of faith and/or are illogical, which I'll summarise here. The IAT accurately measures automatic evaluations. Automatic evaluations are more accurate reflections of attitudes compared to controlled evaluations. The difference between controlled and automatic evaluations is qualitative. Qualitative differences mean that the properties are structurally distinct. Our cognitive apparatus is housed in the brain, and so cognitively distinct properties are associated with neurologically distinct processes. Qualitative differences are observable through response patterns that are recorded as button presses. Qualitatively different response outcomes are detected on a cardinal scale (i.e., time).

A succinct way of capturing the enormity of what has been presented is to think about how you'd answer the following: "If a professional decision was made based on results from an IAT, would you accept the decision?". Given that the IAT is a case study, now imagine that the IAT could be replaced by mention of any type of tool, framework, scale, or assessment, the point should be clear. If the answer is no, the this should put into sharp focus the reality of the responsibility shouldered any product from scientific research that is presented as having practical use in the real world. Put in this pragmatic way, it should also help to demonstrate the value of having a criterion that could help place confidence in a positive response to a question like the one posed here. This is not to say that test re-test reliability, internal consistency and predictive validity aren't important, they are. An application would still need to show that it has met these criteria, but without construct validity, meeting the other criteria is not enough to respond yes to the question.

Concluding remarks

The replication crisis has exposed the need to integrate the criteria of reliability and validity of measurement into researchers day-to-day practice of theorising and empirical work. However, there has been limited discussion regarding the implications of the replication crisis for the products of research that have come out of psychological science. In turn, there has been no guidance as to what would be needed to avert any possible crisis of applicability. We already have good criteria for benchmarking measures in psychological science, and the analysis in this chapter shows which one is key to evaluating tools considered of value for practical application. Above and beyond several criteria (e.g., test re-test reliability, internal consistency, predictive validity), if a measure does not have construct validity, then there is no case for it to be used in the real world. The burden of responsibility for any decision that is informed by a tool means that the decision-maker needs to be assured that the tool measures what it is designed to. The phenomenon must be well understood, and the construct must be sound to enable transparent and accountable decisions. Therefore, the burden on construct validity must be higher still for applied psychological science, than basic psychological science.

References

Axt, J., Buttrick, N., & Feng, R. Y. (2022). A Comparative Investigation of the Predictive Validity of Four Indirect Measures of Bias and Prejudice. *Personality and Social Psychology Bulletin*, 01461672221150229.

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

Barber, T. X., & Silver, M. J. (1968). Fact, fiction, and the experimenter bias effect. *Psychological Bulletin, 70*(6, Pt.2), 1–29. https://doi.org/10.1037/h0026724

Baumeister, R. F. (1992). Neglected aspects of self-theory: motivation, interpersonal aspects, culture, escape, and existential value. *Psychological Inquiry*, *3*(1), 21-25. https://doi.org/10.1207/s15327965pli0301_3

Bruner, J. S. (1950). Social psychology and group processes. *Annual review of psychology*, *1*(1), 119-150.

Calanchini, J., Lai, C. K., & Klauer, K. C. (2021). Reducing implicit racial preferences: III. A process-level examination of changes in implicit preferences. *Journal of Personality and Social Psychology, 121*(4), 796–818. https://doi.org/10.1037/pspi0000339

Campbell, D. T. (1950). The indirect assessment of social attitudes. *Psychological Bulletin, 47*(1), 15–38. https://doi.org/10.1037/h0054114

Cantril, H. (1944a). The Issues—As Seen by the American People. *Public Opinion Quarterly*, *8*(3), 331-347. https://doi.org/10.1086/265693

Cantril, H. (1944b). Gauging public opinion. Princeton: University Press.

Carlsson, R., & Agerström, J. (2016). A closer look at the discrimination outcomes in the IAT literature. *Scandinavian journal of psychology*, *57*(4), 278-287. https://doi.org/10.1111/sjop.12288

Cesario J. (2022) What can experimental studies of bias tell us about realworld group disparities? Behavioral and Brain Sciences 45, e66: 1–71. doi:10.1017/ S0140525X21000017

Chang, E. H., Milkman, K. L., Gromet, D. M., Rebele, R. W., Massey, C., Duckworth, A. L., & Grant, A. M. (2019). The Mixed Effects of Online Diversity Training. Proceedings of the National Academy of Sciences of the United States of America, 116, 7778-7783. https://doi.org/10.1073/pnas.1816076116

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological assessment*, *31*(12), 1412. http://dx.doi.org/10.1037/pas0000626

Connors, S., Spangenberg, K., Perkins, A. W., & Forehand, M. (2020). Crowdsourcing the implicit association test: Limitations and best practices. *Journal of Advertising*, *49*(4), 495-503. https://doi.org/10.1080/00913367.2020.1806155

Conti, M. A., Jardim, A. P., Hearst, N., CORDáS, T. A., Tavares, H., & Abreu, C. N. D. (2012). Evaluation of semantic equivalence and internal consistency of a Portuguese version of the Internet Addiction Test (IAT). *Archives of Clinical Psychiatry (São Paulo)*, *39*, 106-110. https://doi.org/10.1590/S0101-60832012000300007

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281–302. http://dx.doi.org/10.1037/ h0040957

Crowne, D. P., & Marlowe, D. (1964). The approval motive. New York: Wiley.

Dennis, B. (2013). "Validity crisis" in qualitative research: Still? Movement toward a unified approach. In B. Dennis, L. Carspecken, & P. Carspecken (Eds.), Qualitative research: A reader in philosophy, core concepts, and practice (Series— Counter points) (pp. 3-37). New York, NY: Peter Lang.

Dicken, C. F. (1963) Good impression, social desirability, and acquiescence as suppressor variables. Educational and Psychological Measurement, 23, 699-720. https://doi.org/10.1177/001316446302300406

Edwards, A. L. (1957). Techniques of attitude scale construction. New York: Appleton-Century-Crofts.

Epstein, S. (1991). Cognitive-experiential self-theory: An integrative theory of personality. In R. Curtis (Ed.), The relational self: Convergences in psychoanalysis and social psychology (pp. 111-137). New York: Guilford Press.

Frank, L. K. (1939). Projective methods for the study of personality. *The Journal of psychology*, *8*(2), 389-413.

FitzGerald, C., Martin, A., Berner, D., & Hurst, S. (2019). Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: a systematic review. *BMC psychology*, *7*(1), 1-12. https://doi.org/10.1186/s40359-019-0299-7

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. G., & Banaji, M. R. (2017). The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist*, *72*(9), 86. http://dx.doi.org/10.1037/amp0000238

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. Annual review of psychology, 71, 419-445. https://doi.org/10.1146/annurev-psych-010419-050837

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17–41. https://doi.org/10.1037/a0015575

Greenwald, H. J., & Satow, Y. (1970). A short social desirability scale. *Psychological Reports*, *27*(1), 131-135. https://doi.org/10.2466/pr0.1970.27.1.131

Gross, T. (2021). Not one, but many critical states: A dynamical systems perspective. Frontiers in Neural Circuits, 15, 614268. https://doi.org/10.3389/fncir.2021.614268

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255-282. https://doi.org/10.1007/BF02288892

Haire, M. (1950). Projective techniques in marketing research. *Journal of marketing*, *14*(5), 649-656. https://doi.org/10.1177/002224295001400501

Hammond, K. R. (1948). Measuring attitudes by error-choice: an indirect method. *The Journal of Abnormal and Social Psychology, 43*(1), 38–48. https://doi.org/10.1037/h0059576

Hussey, I., & Drake, C. E. (2020). The Implicit Relational Assessment Procedure demonstrates poor internal consistency and test-retest reliability: A meta-analysis. https://doi.org/10.31234/osf.io/ge3k7

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin, 55*(4), 243–252. https://doi.org/10.1037/h0045996

Jackson, D. N., & Messick, S. (1960). ACQUIESCENCE AND DESIRABILITY AS RESPONSE DETERMINANTS ON THE MMPI 1. *ETS Research Bulletin Series*, *1960*(2), i-30. https://doi.org/10.1002/j.2333-8504.1960.tb00101.x

Joffe, H., & Elsey, J. W. (2014). Free association in psychology and the grid elaboration method. *Review of General Psychology*, *18*(3), 173-185. https://doi.org/10.1037/gpr0000014

Jusepeitis, A., & Rothermund, K. (2022). No elephant in the room: The incremental validity of implicit self-esteem measures. *Journal of Personality*, *90*(6), 916-936.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Nosek, B. A. (2014) Investigating variation in replicability. Social Psychology 45 (3):142–52. doi:10.1027/1864-9335/a000178

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist, 74*(5), 569–586. https://doi.org/10.1037/amp0000364

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., . . . Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*(4), 1765–1785. https://doi.org/10.1037/a0036260

Lane, K. A., Kang, J., & Banaji, M. R. (2007). Implicit social cognition and law. Annual Review of Law and Social Science, 3, 427-451. https://doi.org/10.1146/annurev.lawsocsci.3.081806.112748

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological science in the public interest*, *1*(2), 27-66. https://doi.org/10.1111/1529-1006.002

Lin, Y., Osman, M., & Ashcroft, R. (2017). Nudge: concept, effectiveness, and ethics. *Basic and Applied Social Psychology*, *39*(6), 293-306. https://doi.org/10.1080/01973533.2017.1356304

Machery, E. (2022). Anomalies in implicit attitudes research. *Wiley Interdisciplinary Reviews: Cognitive Science*, *13*(1), e1569. https://doi.org/10.1002/wcs.1569

Maison, D., Greenwald, A. G., & Bruin, R. H. (2004). Predictive validity of the Implicit Association Test in studies of brands, consumer attitudes, and behavior. *Journal of consumer psychology*, *14*(4), 405-415. https://doi.org/10.1207/s15327663jcp1404_9

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*(3), 194–216. https://doi.org/10.1037/h0048070

Meissner, F., Grigutsch, L. A., Koranyi, N., Müller, F., & Rothermund, K. (2019). Predicting behavior with implicit measures: Disillusioning findings, reasonable explanations, and sophisticated solutions. Frontiers in Psychology, 10, 2483. https://doi.org/10.3389/fpsyg.2019.02483

Messick, S., & Jackson, D. N. (1957). Authoritarianism or acquiescence in Bass's data. *The Journal of Abnormal and Social Psychology, 54*(3), 424–426. https://doi.org/10.1037/h0041682

Messick, S., & Jackson, D. N. (1958). The measurement of authoritarian attitudes. Educational and Psychological Measurement, 18, 241-253. https://doi.org/10.1177/001316445801800201

Moore-Berg, S. L., Briggs, J. C., & Karpinski, A. (2019). Predicting incidental and focal food consumption behaviors. *British Food Journal*, *121*(7), 1508-1520. https://doi.org/10.1108/BFJ-09-2018-0588

Newkirk, P. (2019). Diversity, Inc.: The Failed Promise of a Billion-Dollar Business. UK: Hachette.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic bulletin & review*, *11*(6), 988-1010. https://doi.org/10.3758/BF03196730

Osman, M. (2010). Controlling uncertainty: a review of human behavior in complex dynamic environments. *Psychological bulletin*, *136*(1), 65. ttps://doi.org/10.1037/a0017815

Osman, M. (2014). *Future-minded: The psychology of agency and control*. Bloomsbury Publishing.

Osman, M. (2016). Nudge: How far have we come?. *Œconomia. History, Methodology, Philosophy*, (6-4), 557-570. https://doi.org/10.4000/oeconomia.2490

Osman, M. (2018). Persistent maladies: The case of two-mind syndrome. Trends in cognitive sciences, 22(4), 276-277. https://doi.org/10.1016/j.tics.2018.02.005

Osman, M. (2021). UK public understanding of unconscious bias and unconscious bias training. *Psychology*, *12*(7), 1058-1069. https://doi.org/10.4236/psych.2021.127063

Osman, M., Lin, Y., & Ashcroft, R. (2017). Nudging: A lesson in the theatrics of choice. Basic and Applied Social Psychology, 39(6), 311-316. https://doi.org/10.1080/01973533.2017.1375929

Osman, M., McLachlan, S., Fenton, N., Neil, M., Löfstedt, R., & Meder, B. (2020a). Learning from behavioural changes that fail. Trends in Cognitive Sciences, 24(12), 969-980. https://doi.org/10.1016/j.tics.2020.09.009

Osman, M., Radford, S., Lin, Y., Gold, N., Nelson, W., & Löfstedt, R. (2020b). Learning lessons: how to practice nudging around the world. *Journal of Risk Research*, *23*(1), 11-19. https://doi.org/10.1080/13669877.2018.1517127

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology, 105*(2), 171–192. https://doi.org/10.1037/a0032734

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2015). Using the IAT to predict ethnic and racial discrimination: Small effect sizes of unknown societal significance. *Journal of Personality and Social Psychology, 108*(4), 562–571. https://doi.org/10.1037/pspa0000023

Payne, B. K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going. *Handbook of implicit social cognition: Measurement, theory, and applications*, *1*, 1-15.

Polit, D. F. (2014). Getting serious about test–retest reliability: a critique of retest research and some recommendations. *Quality of Life Research*, *23*, 1713-1720. https://doi.org/10.1007/s11136-014-0632-9

Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2022). Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. *Psychonomic Bulletin & Review*, *29*(1), 44-54. https://doi.org/10.3758/s13423-021-01948-3

Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes*, *106*(2), 168-179. https://doi.org/10.1016/j.obhdp.2008.02.001

Rapaport, D. (1952). Projective techniques and the theory of thinking. *Journal of Projective Techniques*, *16*(3), 269-275. https://doi.org/10.1080/08853126.1952.10380430

Richetin, J., Perugini, M., Prestwich, A., & O'Gorman, R. (2007). The IAT as a predictor of food choice: The case of fruits versus snacks. *International Journal of Psychology*, *42*(3), 166-173. https://doi.org/10.1080/00207590601067078

Richetin, J., Richardson, D. S., & Mason, G. D. (2010). Predictive validity of IAT aggressiveness in the context of provocation. Social Psychology, 41(1), 27–34. doi:10.1027/1864-9335/a000005

Schnabel, K., Asendorpf, J. B., & Greenwald, A. G. (2008). Assessment of individual differences in implicit cognition: A review of IAT measures. *European Journal of Psychological Assessment*, *24*(4), 210-217. https://doi.org/10.1027/1015-5759.24.4.210

Schimmack, U. (2021). The Implicit Association Test: A method in search of a construct. *Perspectives on Psychological Science*, *16*(2), 396-414. https://doi.org/10.1177/1745691619863798

Steyer, R., & Schmitt, M. J. (1990). Latent state-trait models in attitude research. *Quality and Quantity*, *24*(4), 427-445. https://doi.org/10.1007/BF00152014

Tello, N., Harika-Germaneau, G., Serra, W., Jaafari, N., & Chatard, A. (2020). Forecasting a fatal decision: direct replication of the predictive validity of the suicide–implicit association test. *Psychological science*, *31*(1), 65-74. https://doi.org/10.1177/0956797619893

Trafimow, D. (2004). Attitude measurement. *Encyclopedia of applied psychology*, *1*, 233-244. https://doi.org/10.1016/B0-12-657410-3/00181-1

Trafimow, D. (2019a). A taxonomy of model assumptions on which P is based and implications for added benefit in the sciences. *International Journal of Social Research Methodology*, *22*(6), 571-583. https://doi.org/10.1080/13645579.2019.1610592

Trafimow, D. (2019b). Why successful replications across contexts and Operationalizations might not be good for theory building or testing. *Journal for the Theory of Social Behaviour*, *49*(3), 359-368. https://doi.org/10.1111/jtsb.12211

Trafimow, D. (2023). A new way to think about internal and external validity. *Perspectives on Psychological Science*, *18*(5), 1028-1046. https://doi.org/10.1177/17456916221136117

Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin, 54*(3), 229–249. https://doi.org/10.1037/h0047980

Walasek, L., & Stewart, N. (2015). How to make loss aversion disappear and reverse: Tests of the decision by sampling origin of loss aversion. *Journal of Experimental Psychology: General, 144*(1), 7–11. https://doi.org/10.1037/xge0000039