RICKARD CARLSSON & JENS AGERSTRÖM
2015:11

A closer look at the discrimination outcomes in the IAT literature

Linnæus University Centre
Labour Market and
Discrimination Studies

A closer look at the discrimination outcomes in the IAT literature

Rickard Carlsson
Linnaeus University

Jens Agerström
Linnaeus University

Corresponding author:
Rickard Carlsson
Department of psychology
Linnaeus University
391 82 Kalmar
Rickard.Carlsson@lnu.se

Author note: We thank Anthony Greenwald for his helpful comments on an early draft of this manuscript

Abstract

To what extent the IAT predicts racial and ethnic discrimination is a heavily debated issue. The latest meta-analysis by Oswald et al. (2013) suggests a very weak association. In the present meta-analysis, we took a closer look at the discrimination outcomes, and found that many of the outcomes were unsuitable operationalizations of discrimination. Furthermore, we found virtually no overall discrimination for the IAT to predict. Hence, the IAT has not yet been given a chance to prove its true worth. Indeed, evaluating the predictive validity of the IAT against these outcomes is similar to evaluating raincoats on sunny days; we should not be surprised if the raincoats receive a bad score, but this does not invalidate their usefulness in rainy weather. Given the current state of affairs, it would thus be premature if researchers, educators, and managers simply were to remove the IAT from their toolbox.

Keywords: implicit association test, ethnic discrimination, racial discrimination, meta-analysis

The Implicit Association Test (IAT; Greenwald, McGhee & Schwartz, 1998) has become an immensely popular tool for capturing implicit (or automatic) attitudes, prejudices and stereotypes. Naturally, there is an ongoing debate about the validity of the test, and one especially heated debate concerns whether the IAT can predict racial and ethnic discrimination (Blanton, Jaccard, Klick, Mellers, Mitchell, & Tetlock, 2009; Jost, Rudman, Blair, Carney, Dasgupta, Glaser, & Hardin, 2009a; Landy, 2008). When researchers cite evidence in favor of the IAT's ability to predict racial and ethnic discrimination, a meta-analysis by Greenwald, Poehlman, Uhlman, and Banaji (2009) is often referred to (see e.g., Allen, Sherman, & Klauer, 2010; Bergh, Akrami, & Ekehammar, 2012; Dunham, Baron, & Carey, 2011; Jost, Rudman, Blair, Carney, Dasgupta, Glaser, & Hardin, 2009b; Moss-Racusin, Phelan, & Rudman, 2010; Nier & Gaertner, 2012; Sabin & Greenwald, 2012; Van Bavel & Cunningham, 2011; Yogeeswaran & Dasgupta, 2010). This interpretation is based on the finding that the IAT correlates with race-related (r = .236) and intergroup (r = .201) relevant behavioral outcomes. Furthermore, in a subsequent meta-analysis, Oswald, Mitchell, Blanton, Jaccard, and Tetlock (2013) examined the predictive validity of the IAT in the specific context of ethnic and racial discrimination. They found a smaller overall correlation equivalent of r = .14, and thus concluded that the IAT is a poor predictor of discrimination. From an applied perspective, this weak link is, indeed, disappointing. However, in a reply, Greenwald, Banaji, and Nosek (in press) argued that this smaller correlation could be attributed to different inclusion criteria and techniques used in the two meta-analyses. Further, they argue that even small discrimination effects are readily accepted as important from an applied, including legal, perspective, insofar as they may have large cumulated effects on people's lives.

Although the exact effects obtained in the two previous meta-analyses differ somewhat, the overall conclusion is that the IAT is able to explain only a few percentage points of the

variation in the criterion outcomes. This raises, of course, the question of why this is the case. The reason may be conceptual: implicit bias may only have slight influence on discriminatory behavior. Further, the reason may be due to methodological problems with the IAT: perhaps the IAT is not a valid and reliable measure of implicit bias. For a moment, let us assume that there is a strong conceptual relationship between implicit bias and discriminatory behavior, and that the IAT validly and reliably captures such biases. In this case, the validity and reliability of the discrimination outcomes become the bottleneck. Indeed, if most of the variation in the criterion outcomes is due to other factors than discrimination, even a perfect conceptual link, and a perfect implicit measure, would result in only a few percentage points of the variation being predicted.

The question then is if there is any reason to suspect that there could be problems with the reliability and validity of these discrimination outcomes. Previous case-based analyses of some of the discrimination outcomes included in the meta-analysis suggest that there are. Blanton et al. (2009) re-analyzed two influential studies that had investigated the IAT discrimination link: McConnell and Leibold (2001) and Ziegert and Hanges (2005). They found that the main effect of discrimination in McConnell and Leibold (2001) study was ironic, with black people on average being favored relative to white people. A possible explanation for this unexpected result is that this study compared behavior toward a single black confederate and contrasted it with behavior toward a single white confederate, effectively rendering the discrimination due to race nested within individual confederates. In other words, the (reversed) discrimination effect may have been spuriously produced because of individual differences between the two confederates that have nothing to do with their respective race. This problem with a limited number of randomly selected, or matched, stimuli has received attention both in economics (Heckman, 1998) and psychology (Judd, Westfall & Kenny, 2012). In short, this type of design makes it impossible to know whether

(and how much of) the variance in differential treatment (the outcome) is due to discrimination or due to stimuli-specific effects. Hence, this type of outcome is both invalid and unreliable to use as an outcome variable when the goal is to predict discrimination.

Looking more closely at Heider and Skowronski (2007), Blanton and Mitchell (2011) found an ironic effect of discrimination (higher co-operation with a black individual compared to a white individual) in the first sub-study. For the second sub-study, however, there was some evidence of discrimination in the predicted direction, but as in McConnell and Leibold (2001), this result was based on a behavior towards a single black, and a single white, confederate, casting serious doubt about the internal validity of the study.

Finally, Blanton et al. (2009) scrutinized the study of simulated hiring decisions by Ziegert and Hanges (2005). Blanton et al. (2009) note that although there is evidence of discrimination in the predicted direction (white candidates being favored over black candidates) the lack of a true experimental design for the discrimination outcome means that race and qualifications may have been confounded. Specifically, the difference observed could be due to the white job candidates' slightly higher qualifications. Again, the amount of variance that is due to discrimination becomes highly uncertain due to this methodological issue.

The present research

In the present research we will continue the work by Blanton et al. (2009) and Blanton and Mitchell (2011) in examining the discrimination outcomes of the literature, but rather than relying on a handful case examples, we will conduct a meta-analytical review that focuses on the validity and reliability of the discrimination outcomes. This allows us to draw conclusions about the general state of the discrimination outcomes that the IAT has been evaluated against. If these outcomes generally have validity and/or reliability issues, then their

overall weak correlation with the IAT has to be interpreted differently, compared to if the validity and reliability of the discrimination outcomes would not have been plagued by these issues. Specifically, our comprehensive meta-analysis will allow us to interpret the correlation between the IAT and the outcomes while taking in account the amount of variance that is potentially predictable. For example, if the data suggest that the total variance in the dependent variable that can be reliably attributed to discrimination is 6 %, it is quite impressive if the IAT explains 4 % of the variance (i.e., a correlation of $r = 0.2$). On the other hand, if 60 % of the variance can reliably be attributed to discrimination, explaining 4 % is not as impressive.

Since the meta-analysis of Oswald et al. (2013) is very recent, we will not conduct an entire new meta-analysis, but rather build on their work and add our new analysis to it. This is made possible thanks to their comprehensive supplementary material and their active encouragement for researchers to build on their meta-analysis with new findings.

The remainder of the present paper has the following outline. First, we present a conceptual definition of discrimination, and discuss the different ways of operationalizing discrimination. Next, we present the meta-analysis focusing on whether the outcomes are valid and reliable operationalizations of discrimination. We then provide an updated meta-analysis of the IAT's predictive ability of discrimination among the sub-set of the outcomes that offer the highest validity and reliability among the outcomes. A general discussion concludes the paper.

Conceptual and operational definition of discrimination

According to the Merriam-Webster dictionary (2015), the original meaning of the verb *discriminate* is to distinguish between [objects]. Commonly, the distinction concerns people. When it translates into the act of treating individual members of two or more groups

differently on unfair grounds, we get disparate treatment (Gatewood & Field, 2001). Disparate treatment is a direct form of discrimination against which various classes are protected under the current discrimination laws in many countries. Race or ethnic discrimination occurs, for example, when a black job applicant receives fewer job interviews than a white job applicant, despite having better or identical qualifications (Bertrand & Mullainathan, 2004). Note, however, that disparate treatment (or direct) discrimination has to be separated from structural forms of disparate impact (or indirect) discrimination which occurs when there are (sometimes facially neutral) structures (e.g., laws, policies) that on average favor one group over the other. However, because such structural discrimination rarely has one clearly defined actor (i.e., perpetrator), and the focus on of the paper is on the individual, psychological, component of discriminatory behavior, we will exclusively focus on direct discrimination (disparate treatment). We further limit our attention to race and ethnic discrimination. Hence, in the present research, we conceptually define race/ethnic discrimination as the behavioral act of treating individuals differently because of their race/ethnicity.

Valid operationalization of discriminations

In its operationalized form, the question of to what degree the IAT can predict discrimination becomes the question of to what degree IAT scores moderate an individual's tendency to treat individuals from a certain group (e.g., Black people) differently from individuals from another group (e.g., White people). A proper operationalization of race discrimination thus requires behavior measured towards individuals that are representative of white people and individuals that are representative of black people. As we will see, depending on the strictness of the interpretation of each inclusion criteria, what can be said to constitute a discrimination outcome will vary greatly.

First of all, since we conceptually defined discrimination as differential treatment, all operationalizations need to capture this relative difference on an individual level. As such, behavior towards black people, or white people, in isolation, cannot be operationalizations of discrimination, since they fail to capture *differential* treatment. Treating a black person (or white person) well, or badly, is not discrimination per se. In fact, such an act would not even constitute a violation of any discrimination law. As should be clear by now, the present paper only considers differential outcomes as valid operationalizations of discrimination.

Regarding the definition of behavior, a too strict definition would exclude nearly all studies in the literature. On the other hand, a very inclusive definition would include both the IAT and explicit measures of attitudes, prejudice and stereotypes, making the whole research question circular. Indeed, if we regard a key press in a computer paradigm as a behavior, then the IAT becomes a discrimination paradigm in that it measures the behavior (key press) in relation to representative individuals (often photo stimuli) of one group compared to another group. In an attempt to strike a balance between the two extremes, the present research will regard physical behavior (e.g., touching), verbal (e.g., insult) or non-verbal behavior (e.g., eye-contact), judgment of (e.g., ratings of candidates), or decisions (e.g., hiring decision) as discriminatory behavior, provided that it is directed toward an individual. Perhaps this distinction is best illustrated by an example. Imagine a participant who has a strong amygdala reaction when viewing photos of black individuals (neurological correlates), who states that s/he thinks that black people are unintelligent (stereotype), that she does not like them (explicit attitude or prejudice), and that she would certainly always choose a white candidate over a black one (behavioral intention). This participant is certainly explicitly prejudiced and has negative stereotypes of black people. Further, there are neurological correlates supporting that s/he is prejudiced. However, if upon meeting a black individual, the participant treats him or her in the same manner as when meeting a white individual, with no difference in verbal or

non-verbal behavior, and making the same judgments about intelligence, then the participant still has not shown any sign of discriminatory behavior.

The final aspect concerns that the behavior has to be directed toward individuals that are representative of the two groups. As such, measures capturing behavior that cannot be directly attributed to an individual's race and or ethnicity do not constitute discriminatory behavior. For example, when choosing between two political candidates, the candidates' race may certainly be a factor, but if the design of the experiment does not allow for race to be isolated from other factors, then the behavioral outcome (the decision) does not constitute discrimination. To illustrate, a study asking participants to choose between two real-life political candidates that differ with respect to race, is not a proper operationalization of discrimination suitable as an outcome variable, since there are likely to be other differences between the candidates. On the other hand, if race would have been experimentally manipulated (e.g., attached to the candidate's profiles) then the decision would constitute discrimination.

The quality of the manipulation deserves some special attention. Specifically, the manipulation of race (or ethnicity) has to be representative of the population it is drawn from (e.g., Black people vs. White people). Simply attaching a photo of one black candidate and another photo of a white candidate is highly problematic, since there may be other differences between the photos besides race. Further, a differential behavior against a single black confederate and a matched control is, in this regard, also a very weak design. It is simply too uncertain how much, if any, of the differential treatment can be attributed to the race (or ethnicity) of the confederate. In essence, such studies are not true experiments, but rather quasi-experimental manipulations. Yet, there can be no absolute line here. For example, a study that contrasts two CVs with a few matched names (e.g., Eric vs. Hassan) may work perfectly fine, if there is little reason (i.e., based on previous research) to suspect large

variation among the stimuli within each race or ethnicity. Similarly, photo manipulations of skin color may not require a large amount of stimuli, since they are true experiments (e.g., making a white person black or vice versa), but, there are, of course other threats to the validity of this approach (e.g., whether the photo manipulation was realistic or not).

To what extent the manipulation of race/ethnicity actually becomes representative of the two groups is a continuum and it is quite difficult to decide whether a certain study fulfill this criterion. Further, from a meta-analytical perspective, these problems should to some extent balance each other out. That is, although a single study that relies on a few matched confederates is inherently weak, the average of several such studies may still contain valuable information. This would hold true if there is no bias, but only measurement error, introduced in this manner. For example, Study A may include a much nicer black confederate and thus underestimate the true level of discrimination, whereas Study B may include a much nicer white candidate and thus overestimate the true level of discrimination. Accordingly, because of the inherent complexity in deciding whether a study is of sufficient quality in this manner, and because a strict cut-off would exclude a large proportion of studies, we will not consider this an exclusion criterion, but instead emphasize that only the meta-analytical average of such studies are relevant. Further, we will pay special attention to studies that have particularly weak designs in this regard (e.g., single matched confederate studies).

Estimating the reliability of the discrimination outcomes

So far we have discussed what types of operationalizations are valid to use as discrimination outcomes. Yet, even entirely valid measures will constitute poor outcome variables if they do not capture reliable amounts of discrimination. A proper estimation of this reliability would, of course, be based on the reliability indices of the outcomes and careful attention to what extent the correlation with the IAT becomes attenuated because of this. However, it is

exceedingly rare for such reliability indices to exist in the literature. Instead, we will estimate something slightly different: the main effect of discrimination in outcomes. With a main effect of discrimination in an outcome that has variance, we can be sure that at least *some* of the participants discriminated to a higher extent than others. In other words, we can be sure that at least part of the variance in the outcome is due to individual differences in discriminatory behavior.

When there is variance but no main of discrimination in the outcome, the situation becomes quite complicated, since there are several different processes that may have produced this outcome, besides individual differences in discrimination. A simple explanation is that the discrimination outcome failed to capture discrimination that, conceptually, occurs. For example, the outcome may lack sufficient precision to detect discrimination, or the participants may have become aware of the study's purpose and adjusted their behavior in order not to appear discriminating. Another explanation is that the participants did not discriminate. It is easy to imagine that there may be contexts where discrimination does not occur, or when it is reversed. For example, a researcher may only have used a group of black students as research participants, and predicted little discrimination against black targets, or even reversed discrimination. Another researcher may have designed an intervention that aims to eliminate discrimination. Clearly, studies that a priori are not expected to find discrimination, or reversed discrimination, are not good candidates when it comes to assessing the IAT's predictive validity. Rather, such studies should be analyzed as subgroups and in that regard prove extremely useful for assessing discriminant validity. For example, that the IAT should *not predict* any variance in these studies because there is none to predict in the first place, or that the direction of the effect should be reversed. However, such a detailed analysis is beyond the scope of the present research since there are too few studies of this type to allow for this kind of fine-grained sub-group analysis.

Recognizing that discrimination is likely a continuous phenomenon, even in the presence of strong main effects, there will typically be some people that are reversely discriminating (e.g., who prefer black people to white people). As such, the amount of individual differences in discrimination will typically be slightly higher than the size of the main effect of the discrimination. Of course, this suggests that there may be some level of individual differences in discrimination even when there is no main effect of discrimination. This happens when there is equally strong discrimination against the minority group as against the majority group. In other words, that some participants discriminated against white and some against black partners, ending up with equal treatment at the group level despite ample discrimination at the individual level. Unfortunately, confirming such mixed discrimination is difficult on a post-hoc basis. One cannot rely on the correlation with the IAT as proof of discrimination in the outcome, since this makes the whole research endeavor circular. In our view, one should instead carefully consider whether it is reasonable for a discrimination outcome of a predicted group to show a zero main effect but still contain meaningful amount of predictable variance due to discrimination. Simply – why do the findings sum to an average of zero? If this is due to, say, half of the participant belonging to a group that can be expected to exhibit reverse discrimination, then the effect size of this group difference can be used to determine the amount of predictable variance du to discrimination. On the other hand, if the lack of a main effect is due to some participants exhibiting reverse discrimination in order to appear non-prejudiced, this would instead suggest a serious validity problem with the study. In some cases, additional variables (e.g., ethnicity of participants) might be used to shed light on this issue. However, it is important to emphasize that the amount of variance in an outcome variable cannot alone be used as an estimate for individual differences in discrimination. Consider a study where the discrimination outcome is the choice between a black and a white partner. If each of the participants chooses a partner at

random, there will be equal treatment at the group level, no discrimination at the individual level, but still apparent variation in the outcome. Yet, this variation has nothing to do with individual differences in discrimination and is thus not suitable to predict by means of the IAT.

To summarize, our focus on main effects of discrimination is limited in that it cannot distinguish between a poor measure and a proper measure used in a context (e.g., a study population or a specific situation) where discrimination simply does not occur, or when it is prevalent but perfectly mixed. Hence, it is important that the main effect of discrimination is only considered in studies where it is clear that a specific group has been predicted to be discriminated, and where it is little reason to suspect mixed discrimination. For the purposes of the present research, this limitation is not severe. Although it is possible that some outcomes that lack a main effect of discrimination could still reveal individual differences in discrimination that are predictable by the IAT, our view is that the IAT needs to be validated against outcomes which can provide a ballpark figure regarding what levels of correlation can be expected.

Method

Having conceptually and operationally defined discriminatory behavior, as well as explained our rationale for estimating the reliability of discrimination as an outcome variable to be predicted by the IAT, we now turn to our meta-analysis of the discrimination outcomes. We did so by evaluating each independent study (i.e., a sub-study or single experiment in a paper) through a series of questions:

1. Can the outcome variable be considered an operationalization of discriminatory behavior?

2. Is the outcome variable scored as differential treatment of a minority compared to a majority group?
3. Did the study examine a context and a population where discrimination of minority groups is predicted to occur?
4. How much of the variance in the outcome can be attributed to discrimination?

In deciding on questions 1 - 3, both authors of the present research read through the materials by Oswald et al. (2013) and all of the associated articles. We did not code our views independently of each other but rather had an open discussion throughout the process where we reached consensus in each case. A clear yes on these three questions was required for a study to be evaluated for question 4.

Rather than re-doing the coding and categorizations, we make every attempt to build on the earlier work of Oswald et al., and thus, as far as possible, retain their coding and categorization. In doing so, we are acknowledging the great amount of work already put down in categorizing this material, and aim to facilitate direct comparison between the present research and the original meta-analysis.

Can the outcome variable be considered an operationalization of discriminatory behavior?

Oswald et al. (2013) distinguish between six different types of discriminatory behavior; interpersonal behavior, person perception, micro behavior, policy preference, response time, and brain activity. The first three of these categories include behaviors that correspond well with our definition of disparate treatment discrimination. Interpersonal behavior includes behavior such as hiring decisions, person perceptions include measures such as judgments of trustworthiness of individuals, and micro behavior includes measures of non-verbal behavior

(e.g., smiling frequency). In contrast, policy preferences are not measures of discrimination per se. For example, although voting for McCain instead of Obama may certainly be related to prejudice, it does not, in itself, constitute a discriminatory behavior. Notice that if the choice had been between two candidates that *only* differed in race it would have been a discrimination measure. Furthermore, response time measures are conceptually similar to the IAT in itself, and a definition that is broad enough to include response time measures as discriminatory behavior would also by necessity include the IAT, making the whole research question circular. Finally, brain activity is an internal process rather than a discriminatory outcome. Indeed, other people are not necessarily directly affected by what is going on inside your skull.

Is the outcome variable scored as differential treatment of a minority compared to a majority group?

Oswald et al. (2013) distinguishes between four types of scoring methods: absolute scoring toward minority targets, absolute scoring of majority targets, relative ratings and differences scores. Importantly, only the relative ratings and the differences scores are measures of differential treatment discrimination.

Occasionally, the absolute ratings came from studies that contained absolute ratings of both groups, meaning that a re-analysis of those could be done in order to make the outcome a differential treatment variable instead. Since these studies use between-participant designs, this would necessitate a re-analysis of the raw data in order to properly test for the interaction effect between ethnicity/race and the IAT-score. Even with such data available, it would be a challenge to identify discrimination on the individual level in this type of between-participant design. We thus decided to not follow up on these categories.

When reading the literature, we discovered that three studies categorized as having difference or relative scores did in fact have behavioral outcomes scored against the minority groups without any majority group controls (Amodio and Devine, 2006; Florack, Scarabis, & Bless, 2001; Korn, Johnson, & Chun, 2012). Hence, these three studies were not included for further analysis.

Did the study examine a context and a population where discrimination of minority groups is predicted to occur?

Most studies examined discrimination toward a clearly definable minority group in a context where it was clear (from the introduction in the paper) that discrimination was to be expected. For example: discrimination of Black people relative to White people in a sample of predominately White college students. Two studies focused on the discriminatory behavior among participants who themselves belonged to minority groups. Ashburn-Nardo, Knowles, and Monteith (2003) tested discrimination of Black people vs. White people in a sample of Black students. Ma-Kellams, Spencer-Rodgers, and Peng (2011) tested discrimination of Chinese-Americans relative to European-Americans in a sample of European-Americans as well as a sample of Chinese-Americans. They further tested discrimination of Latinos vs. European-Americans in a sample of Latinos. The predictions concerning which groups should be the targets of discrimination in these two studies are not obvious, since ingroup favoritism may alter the direction entirely. In order to avoid such confusion we have chosen not to include these studies when assessing the main effect of discrimination. Ideally, these types of studies should be investigated in sub-group analysis in the meta-analysis, but they are simply too few for that to be possible. However, we did conduct some sensitivity analyses that suggested that inclusion of these studies (regardless of the direction they are scored when this is ambiguous to us) does not alter the overall findings.

Another important consideration is if some studies have included conditions where the IAT-discrimination correlation, or discrimination itself, was predicted to be eliminated or weak. After a thorough review of these studies, we found that Ziegert and Hanges (2005) hypothesized that the IAT would predict racial discrimination when the corporate climate promoted racial bias, but not when the corporate climate promoted racial equality. However, since the authors explicitly present and interpret the main effect of discrimination (across conditions) we found it appropriate to include both conditions in their averaged form.

Hofmann, Gschwendner, Castelli, and Schmitt (2008) predicted that the IAT would be a weaker predictor of discriminatory behavior when participants' cognitive control resources were untaxed as opposed to taxed. Hence, these conditions need to be analyzed separately in the present meta-analysis. Unfortunately, this study could not be included due to missing data (see below).

Lastly, we carefully considered whether the original authors predicted mixed discrimination, or whether we our-selves found it plausible. Apart from the studies with minority participants discuss above, we found no such studies.

How much of the variance in the outcome can be attributed to discrimination?

In order to answer this question, we conducted a meta-analysis of the levels of discrimination in the studies that fulfill criteria 1 - 3. We decided to calculate a single averaged effect size along with 95 % confidence intervals for each independent sample used in the meta-analysis. This approach is different from Oswald et al. (2013) who disaggregated the data in as much detail as possible. Although we see the merits of their approach, our focus is on specific studies and disaggregating data would result in a less straightforward presentation of the results. For example, some experiments would include dozens of data points whereas others would only include one. Yet, they would still share many important characteristics. There are

different ways to statistically address this (e.g., multi-level modeling, clustering). We ultimately choose to aggregate on the level of independent samples. Importantly, this level does not necessarily correspond to a study in an article, but was chosen based on the criteria and for the statistical reasons of independence of the observations. In most cases, this aggregation was a fairly straightforward averaging of the results since the number of observations for each outcome was the same and nested in participants. However, in the study by Vanman, Saltz, Nathan, and Warren (2004), there were two discrimination outcome variables, but only a small portion of the sample actually completed both. In this case, we decided to focus only on the outcome that the entire sample had completed.

All studies used within-participants designs, and, unlike between-participants designs, there is a lack of consensus on how to best calculate effect sizes from such studies (see Lakens, 2013, for a full discussion on this). We choose to calculate d_z which is based on the *t*-statistic in the following manner: $d_z = t / \sqrt{} SD$. Thus, the effect size is based on the standard deviation from the difference scores, rather than a pooled standard deviation from the two measures (e.g., treatment of Blacks and Whites separately). Typically, the standard deviation of the difference scores are lower than the standard deviations of the measures themselves, and thus this effect size tends to be higher than that from a between-participant calculated Cohen's *d*.

There are several reasons behind our choice of d_z. First and foremost, we believe this to be the most relevant measure of discrimination for the present meta-analysis, since it focuses on the within, rather than between, participant variation in behavior towards two groups (i.e., discrimination). Second, we rarely had access to the data necessary to perform the other types of calculations. Third, this measure emphasizes the statistical effect in the data-material in relation to its statistical precision. Hence, this effect size statistic is most closely related to a manipulation check where the researcher makes sure that there is variance due to

discrimination that can be moderated by the IAT. Fourth, the most serious problem with this effect size statistic is that it is difficult to directly compare to between-participant studies, rendering it unsuitable for a meta-analysis that combines both within- and between-participant studies. Since we did not include between-participant studies, this was not a concern for the present meta-analysis.

No studies provided effect sizes that could directly be included in the meta-analysis. Yet, for most studies we could accurately calculate (Using an SPSS-script by Wuensch, 2012) Cohen's *d* and its 95 % CI based on reported t-values, reported exact *p*-values, or reported means and standard deviations. For studies that presented non-exact *p*-value we based our calculations on an average between this value (e.g., $p < .05$*)* and the next conventional significance level (e.g., $p < .01$). Unfortunately, for some studies meeting our three inclusion criteria, it was not possible to calculate reasonable precise estimates and these estimates are thus missing. When this was the case we contacted the authors and requested the missing data. We received the requested data in one case (Kang et al., 2010). In the case of Carney (2006) and Carney, Olson, Banaji, and Mendes (2006) it turned out to be a very time consuming process to produce the data in the form we required, as our analysis was not a priori to their research question. Further, since these two studies are not yet published, the authors have no obligation to share the data with us to begin with. In the case of Hugenberg and Bodenhausen (2004), the now over a decade old data was regrettably no longer available. Finally, Hofmann et al., (2008) did not respond in our short time frame, and we decided to move on with our analysis without pursuing this further. Hence, we would like to be clear that the reason for this missing data is due to practical circumstances and our short time frame, and does not suggest any lack in scholarship of the original authors.

In the end, twelve studies were included in the meta-analysis. A short description of the studies can be found in *Table 1*. Details regarding the coding for each study, the dataset as

well as the STATA-code necessary replicate the meta-analysis, is appended as electronic supplemental materials.

Results

We conducted a meta-analysis on the levels of discrimination using the *metan* command in STATA 12. We combined the effects based on random effects assumptions. Indeed, the very large heterogeneity (I-squared of 85 %) suggests that this approach is appropriate.

Looking at the forest plot (*Figure 1*), we first see that the overall effect size of discrimination is close to zero. The 95 % CI spans weak discrimination and weak reversed discrimination. Taken as a whole, then, the literature does not show robust levels of discrimination that can be predicted by the IAT. However, the level of heterogeneity among the studies is striking and is not driven by any specific outlier study. Large heterogeneity is not surprising given the large variations among the method used in the studies. However, it casts a doubt on the appropriateness to draw strong conclusions regarding the average level of discrimination, or, the average level of predicted discrimination from the IAT. Hence, our next step was to take a closer look at the individual studies.

The individual samples are plotted in ascending order on their level of discrimination main effect. It is thus easy to see which studies show a main effect of discrimination and which do not. When looking at the individual effect sizes, the first interesting finding is that the study of McConnell and Leibold (2001) does not appear to be representative of the studies. Indeed, close to strong levels of reversed (ironic) discrimination is an exception in the literature. Most studies, in contrast, show effect sizes very close to zero.

Another interesting finding is that the 95 % CI are very wide in the majority of the studies. In fact, for a typical study, the 95 % CI includes both a moderate effect of discrimination, and a moderate effect of ironic discrimination. Given that even small amounts

of discrimination can be an important phenomenon from an applied perspective (see Greenwald et al., 2015, for this argument), these studies are underpowered to test for discrimination.

The four experiments that appear to show reliable levels of discrimination deserve a closer look. Biernat et al. (2009) stand out from the rest in that their outcome variable is about shifting standards. In short, their outcome variable is the difference in difference of discrimination in objective and subjective ratings of Black and White people. Specifically, they find discrimination of Black people in objective ratings, but little difference in subjective ratings. This is an interesting finding because no other study contains discrimination on an objective rating scale.

Stanley et al. (2011; Study 1) is a study about trust ratings of faces of Black and White people. Importantly, the researcher has been rigorous in their selection of a large amount of stimuli, and we can see no reason to why the effect uncovered in this study should be due to something other than discrimination. On the other hand, the effect found in Heider and Skowronski (2007; Sample 2), may be due to the use of a single confederate approach. As such, their finding is best understood together with the similar study by McConnell and Leibold (2001). It could be that the Black confederate in McConnell and Leibold (2001) happened to be particularly nice, while the confederate in Heider and Skowronski (2007: Sample 2) happened to be not so nice. The final study that has a clear discrimination effect is the previously criticized (Blanton et al., 2009) study by Ziegert and Hanges (2005), which shows clear (close to strong) levels of discrimination. Possible, part of this this effect is due to a confounding effect of the applications the race manipulation was attached to.

To summarize, the overall effect of discrimination in the literature is virtually zero. There are only handful studies that in isolation demonstrate clear levels of discrimination, and even fewer that do so without serious methodological problems that may have produced the

result. Accordingly, there appears to be very small amount of variance that can reliably be predicted by the IAT. For completion sake we did a quick check of the correlation between IAT and the outcomes, and confirmed that it was similar for our subset: $r = 0.15$, 95 % CI [0.07, 0.22].

Discussion

Previous meta-analyses (Greenwald et al., 2009; Oswald et al., 2013) suggest that the IAT is a weak predictor of discrimination. This conclusion is based on the overall weak correlation between the IAT and discrimination outcome variables in the literature. Such weak correlations may stem from a weak conceptual link between the two constructs, or from measurement problems in either the IAT or the discrimination outcomes. Previous research has reviewed a handful of the discrimination outcomes and identified several concerns regarding their validity and reliability (Blanton et al. 2009, Blanton and Mitchell 2011). In the present research, we continue this work, but through a bird's eye meta-analytical approach. Building on the previous meta-analysis by Oswald et al. (2013), we evaluated the discrimination outcomes in terms of whether they are valid operationalizations of discriminatory behavior, and to what extent that they reliably detect discrimination that can be predicted by the IAT. Indeed, if there is no or little variance in the outcome variable that is due to discrimination, then the IAT has little or no variance to predict. Accordingly, the correlation coefficients can be expected to be very small, regardless of the underlying conceptual link or the quality of the IAT itself.

Our results suggest that only a small portion of the outcomes that were meta-analytically combined in Oswald et al. (2013) fits our conceptual and operational definition of discrimination. The remaining are other types of measures, such as brain activity, voting intentions, or behaviors directed only toward a single group without any comparison group. Although such studies may certainly be relevant outcomes in research studies, they are not

valid discrimination outcomes. We would like to emphasize that the original authors themselves do not always claim this to be the case, but rather had other research questions in mind when conducting the study than simply satisfying ad-hoc questions posed by meta-analysts. Hence, our evaluation of the studies is *not* an evaluation of their overall scientific merit. The evaluation only concerns the validity and reliability of the studies when it comes to discrimination outcomes.

The next question was whether the apparently valid outcomes detected reliable amounts of discrimination. A meta-analysis of the main effect of discrimination in the studies suggests that there is no reliable amount of discrimination to be predicted by the IAT. Indeed, the average effect of discrimination is zero, and the results are widely inconsistent between different studies. Attempting to meta-analytically test the correlation between IAT and discrimination thus appears futile. We are, essentially, chasing noise. We simply cannot expect any strong, or even moderate, correlations, based on the current literature. Thus, although our re-analysis showed highly similar correlation between IAT and discrimination (r = .15) we are reluctant to draw the conclusion of Oswald et al. (2013) that this suggest poor predictive validity of the IAT. In our view, this figure is as high as one could hope for given the lack of main effects in the discrimination outcomes. Further, there is little reason to suspect publication bias in either the discrimination effect or the correlation with the IAT figure. Indeed, most of the effects are actually non-significant. Still, the lack of clear levels of discrimination in the outcome makes the correlation between IAT and discrimination hard to interpret.

Limitations and suggestions for future research

We decided to re-analyze the meta-analysis by Oswald et al. (2013) because it is to our knowledge the most updated meta-analysis on the topic. This approach has the advantage of

facilitating a comparison with their original meta-analysis, thus avoiding conflating our specific point regarding the level of discrimination with other difference that inevitably will come up when conducting a meta-analysis on different dataset. The disadvantage is, however, that most recent research on the topic could not be included, and that we could not broaden the topic to also cover other types of discrimination (i.e., religion or sexual orientation).

The present meta-analysis is based on reported summary statistics. An alternative approach would be to base the meta-analysis on raw data. This would address the most serious limitation of the present study, in that between-participant experiments could also be analyzed. Further, a raw data re-analysis could make use of potentially available control variables (e.g., self-presentational issues), or more specific subgroup analyses, in order to achieve more precise estimates of the level of discrimination in the data. In some cases, it may even be possible to estimate reliabilities of the measures (e.g., when test-retest data are available), making it possible to calculate attenuated correlations. A more nuanced technique may be especially helpful in the case where a main effect test falls short (i.e., mixed discrimination with some discriminating the minority and others the majority group). Unfortunately, we do not find this approach realistic with the current literature. Our attempts to obtain raw data when summary statistics was not sufficient were quite unsuccessful. Our hope is that publicly sharing data (i.e., in repositories) will become more common in the future, and that this will enable meta-analysts to conduct this type of more fine-grained analysis.

The most important finding of the present study is that the current literature is rather uninformative as to whether the IAT can predict discrimination or not, as there are few studies with discrimination as the outcome variable, and little evidence that there were any discrimination to predict. Hence, additional empirical work is an important task for future research. One route to go about this may be large-scale collaborations, such as the many-labs

replication project (Klein, Ratliff, Michelangelo, Adams, Bahník, Bernstein, et al. 2014), that focus on replicating some of these studies using samples large enough to allow precision to detect and moderate discrimination. In doing so, both empirical researchers as well as future meta-analysts need to (as argued by Greenwald et al., in press) realize that discrimination can be weak but still important, and design their paradigm so that they have high precision when determining how much of the variance can be attributed to discrimination, and subsequently be predicted by the IAT.

Conclusion

When assessing the IAT's ability to predict discrimination, we have to keep in mind the type of discrimination outcomes that the IAT is supposed to predict. The present research suggests that many of the outcomes are not valid operationalizations of discrimination, and among those that have apparent validity, there is little evidence of reliable amounts of discrimination that can be predicted. Hence, the IAT has been put up to the impossible task of predicting discrimination that is simply not there. In our view, evaluating the IAT's ability to predict discrimination based on the current literature is akin to testing out raincoats on sunny days. Unsurprisingly, the raincoats will receive a bad score, since they are particularly unsuitable on sunny days. Still, most of us would like to keep the raincoats in our wardrobe and use them when the rain starts pouring. Similarly, researchers should keep the IAT in their toolbox and use it to predict discrimination when it occurs.

References

References marked with an asterisk are included in the current meta-analysis of level of discrimination.

Allen, T. J, Sherman, J. W. ,& Klauer, K. C. (2010). Social context and the self-regulation of implicit bias. *Group Processes and Intergroup Relations*, *13*, 137-149. doi: 10.1177/1368430209353635

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91,* 652–661. Doi: 10.1037/0022-3514.91.4.652

Ashburn-Nardo, L., Knowles, M. L., & Monteith, M. J. (2003). Black Americans' implicit racial associations and their implications for intergroup judgment. *Social Cognition, 21,* 61–87. doi: 10.1521/soco.21.1.61.21192

Bergh, R., Akrami, N., & Ekehammar, B. (2012). The personality underpinnings of explicit and implicit generalized prejudice. *Social Psychological and Personality Science, 3,* 614-621. doi: 10.1177/1948550611432937

Bertrand M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination, *American Economic Review*, *94*, 991-1013. doi: 10.1257/0002828042002561

*Biernat, M., Collins, E. C., Katzarska-Miller, I., & Thompson, E. R. (2009). Race-based shifting standards and racial discrimination. *Personality and Social Psychology Bulletin*, *35*, 16-28. doi:10.1177/0146167208325195

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *Journal of Applied Psychology*, *94*, 567–582. doi: 10.1037/a0014665

Blanton, H., & Mitchell, G. (2011). Reassessing the predictive validity of the IAT II: Reassessing the predictive validity of Heider & Skowronski (2007). *North American Journal of Psychology*, *13*, 99 – 106.

Carney, D. R. (2006). The faces of prejudice: On the malleability of the attitude– behavior link. Unpublished manuscript, Harvard University.

Carney, D. R., Olson, K. R., Banaji, M. R., & Mendes, W. B. (2006). The faces of race-bias: Awareness of racial cues moderates the relation between bias and in-group facial mimicry. Unpublished manuscript, Harvard University.

Cumming, G. (2013). The new statistics: Why and how. *Psychological Science*. doi:10.1177/0956797613504966

Discriminate [Def. 1]. (n.d.). *Merriam-Webster Online.* In Merriam-Webster. Retrieved March 4, 2015, from http://www.merriam-webster.com/dictionary/citation.

Dunham, Y., Barrow, A. S., & Carey, S. (2011). Consequences of ‘‘minimal’’ group affiliations in children. *Child Development*, *82*, 793-811. doi: 10.1111/j.1467-8624.2011.01577.x

Florack, A., Scarabis, M., & Bless, H. (2001). Der Einfluß wahrgenommener Bedrohung auf die Nutzung automatischer Assoziationen bei der Personenbeurteilung. *Zeitschrift Für Sozialpsychologie*, *32*, 249-259. doi:10.1024//0044-3514.32.4.249

Gatewood, R. D., & Field, H. S. (2001). Human resource selection (5th ed.). Stamford, CT: Thomson Learning.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*, 553-561. doi: 10.1037/pspa0000016

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

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

Heckman, J. J. (1998). Detecting discrimination. Journal of Economic Perspectives, 12, 101–116. doi: 10.1257/jep. 12.2.101

*Heider, J. D., & Skowronski, J. J. (2007). Improving the predictive validity of the Implicit Association Test. *North American Journal of Psychology*, *9*, 53–76.

Hofmann, W., Gschwendner, T., Castelli, L., & Schmitt, M. (2008). Implicit and explicit attitudes and interracial interaction: The moderating role of situationally available control resources. *Group Processes and Intergroup Relations*, *11*, 69–87. doi:10.1177/1368430207084847

Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in social categorization: The role of prejudice and facial affect in race categorization. *Psychological Science*, *15*, 342–345. doi:10.1111/j.0956-7976.2004.00680.x

Jost, J. T., Rudman, L. A., Blair, I. V., Carney., D. R., Dasgupta., N., Glaser., J., & Hardin., D (2009a). An invitation to Tetlock and Mitchell to conduct empirical research on implicit bias with friends, ‘‘adversaries,’’or whomever they please. *Research in Organizational Behavior, 29*, 73 – 75. doi: 10.1016/j.riob.2009.06.009

Jost, J.T., Rudman, L.A., Blair, I.V., Carney, D., Dasgupta, N., Glaser, J. & Hardin, C.D. (2009b). 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. doi:10.1016/j.riob.2009.10.001

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. Journal of Personality and Social Psychology, 103, 54–69. doi: 10.1037/a0028347

*Kang, J., Dasgupta, N., Yogeeswaran, K., & Blasi, G. (2010). Are ideal litigators white? Measuring the myth of colorblindness. *Journal of Empirical Legal Studies*, *7*, 886 - 915. doi:10.1111/j.1740-1461.2010.01199.x

Klein, R. A., Ratliff. K. A., Michelangelo, V., Adams, R., Bahník, S., Bernstein, M. J., Bocian, K., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*, 142-152. http://dx.doi.org/10.1027/1864-9335/a000178

Korn, H., Johnson, M. A., & Chun, M. M. (2012). Neurolaw: Differential brain activity for Black and White faces predicts damage awards in hypothetical employment discrimination cases. *Social Neuroscience*, 7, 398-409. doi:10.1080/17470919.2011.631739

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863, doi:10.3389/fpsyg.2013.00863

Landy, F. J. (2008). Stereotypes, bias, and personnel decisions: Strange and stranger. *Industrial and Organizational Psychology*, *1*, 379-392. doi: 10.1111/j.1754-9434.2008.00071.x

*Ma-Kellams, C., Spencer-Rodgers, J., & Peng, K. (2011). I am against us? Unpacking cultural differences in ingroup favoritism via dialecticism. *Personality and Social Psychology Bulletin*, *37*, 15-27. doi:10.1177/0146167210388193

*McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology*, *37*, 435–442. doi:10.1006/jesp.2000.1470

Moss-Racusin, C., Phelan, J., & Rudman, L. (2010). "I'm Not Prejudiced, but . . .": Compensatory Egalitarianism in the 2008 Democratic Presidential Primary. *Political Psychology*, *31*, 543-561. doi: 10.1111/j.1467-9221.2010.00773.x

Nier, J. A., & Gartner, S. L. (2012). The Challenge of Detecting Contemporary Forms of Discrimination. *Journal of Social Issues*, *68*, 207-220. doi: 10.1111/j.1540-4560.2012.01745.x

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

Sabin, J. A., & Greenwald, A. G. (2012). The Influence of Implicit Bias on Treatment Recommendations for 4 Common Pediatric Conditions: Pain, Urinary Tract Infection, Attention Deficit Hyperactivity Disorder, and Asthma. *American Journal of Public Health*, *102*, 988–995. doi:10.2105/AJPH.2011.300621.

*Sabin, J. A., Rivara, F.P., & Greenwald, A. G. (2008). Physician implicit attitudes and stereotypes about race and quality of medical care. *Medical Care*, *46*, 678-685. doi:10.1097/MLR.0b013e3181653d58

*Sargent, M. J., & Theil, A. (2001). When do implicit racial attitudes predict behavior? On the moderating role of attributional ambiguity. Unpublished manuscript, Bates College.

*Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 7710-7775. doi:10.1073/pnas.1014345108

Van Bavel, J. J., & Cunningham, W. A. (2011). A social neuroscience approach to self and social categorisation: A new look at an old issue. *European Review of Social Psychology*, *21*, 237–284. doi:10.1080/10463283.2010.543314

*Vanman, E. J., Saltz, J. L., Nathan, L. R., & Warren, J. A. (2004). Racial discrimination by low-prejudiced Whites: Facial movements as implicit measures of attitudes related to behavior. *Psychological Science*, *15*, 711–714. doi:10.1111/j.0956-7976.2004.00746.x

Wuensch, K. L. (2012). Using SPSS to obtain a confidence interval for Cohen's d. http://core.ecu.edu/psyc/wuenschk/SPSS/CI-d-SPSS.pdf.

Yogeeswaran, K. & Dasgupta, N. (2010) Will the "real" American please stand up? The effect of implicit national prototypes on discriminatory behavior and judgments. *Personality and Social Psychology Bulletin*, *36*, 1332-1345. doi:10.1177/0146167210380928

*Ziegert, J. C., & Hanges, P. J. (2005). Employment discrimination: The role of implicit attitudes, motivation, and a climate for racial bias. *Journal of Applied Psychology*, *90*, 554–562. doi:10.1037/0021-9010.90.3.553

Table 1. Characteristics of the 12 individual samples that were included in the current meta-analysis of discrimination.

Study	Sample	*N*	Discrimination measure
Biernat et al. (2009)	1	136	Ratings (objective - subjective) of Black versus White students' academic ability
Heider & Skowronski (2007)	1	140	Cooperative behavior toward African American versus Caucasian confederates in a prisoner's dilemma game
Heider & Skowronski (2007)	2	55	Verbal and nonverbal friendliness behaviors toward African American versus Caucasian confederates
Kang et al. (2010)	1	68	Evaluation of an Asian versus a White litigator's deposition
Ma-Kellams et al. (2011	2	60	Attributions after reading scenario where European-American vs. Latinos and Chinese behaved negatively.
McConnell & Leibold (2001)	1	41	Nonverbal behaviors (e.g. speaking time) toward Black versus White experimenters
Sabin et al. (2008)	1	60	Recommended treatment for Black versus White patients appearing in case vignettes
Sabin et al. (2008)	2	33	Recommended treatment for Black versus White patients appearing in case vignettes
Sargent & Theil (2001)	1	38	Choice of partner (Black versus unspecified race) for an upcoming intellectual task
Stanley et al. (2011)	1	50	Trust ratings of Black versus White faces
Stanley et al. (2011)	2	43	Money offers to Black and White partners in a trust game
Vanman et al. (2004)	1	80	Choice between Black versus White applicants for a teaching fellowship
Ziegert & Hanges (2005)	1	99	Evaluations of Black versus White job applicants during a role-play exercise

Figure 1. Effect sizes (Cohen's *d*) and confidence intervals (95 %) for the main effect of discrimination in the independent samples. The independent samples are plotted in ascending order on their level of discrimination.