Methodological Issues in Predicting Discrimination from Attitudes, Prejudices, and Stereotypes
Methodological Issues in Predicting Discrimination from Attitudes, Prejudices, and Stereotypes

Rickard Carlsson   Jens Agerström
Linnaeus University   Linnaeus University

Corresponding author:
Rickard Carlsson
Department of psychology
Linnaeus University
391 82 Kalmar
Rickard.Carlsson@lnu.se
Methodological Issues in Predicting discrimination

Abstract

A central question for social psychologists is to what extent attitudes, prejudice, and stereotypes are precursors of ethnic and racial discrimination. Operationalized, this question can be framed as the extent measures of such constructs predict differential treatment of individuals from one group compared to a comparison group. Yet, in the literature, it is common to substitute this operationalization for a simpler one: measures predicting behavior toward a single group. We argue that this simpler operationalization lacks validity and yields uninformative effect sizes. We provide several suggestions on how to include, and make most use of, comparison groups, when predicting discrimination.
Race or ethnic disparate treatment discrimination occurs when individuals are treated differently only because they belong to different racial or ethnic groups (Gatewood & Field, 2001). For example, if black candidates are less likely to be invited to a job interview compared to white candidates, despite being equally qualified for the job (e.g., Bertrand & Mullainathan, 2004). A central theme in social psychology has been to view ethnic or racial discrimination as the behavioral consequence of negative attitudes, prejudices and stereotypes (hence referred to as A/P/S). Hence, the central question is whether A/P/S can predict the extent to which people treat individuals differently because of ethnicity or race.

Over the years, this core research question has become increasingly fine grained. For example, are attitudes, stereotypes or distinct emotional prejudices the strongest predictors of discrimination (e.g., Talaska, Fiske, & Chaiken, 2008)? Are people’s implicit A/P/S capable of predicting discriminatory behavior (e.g., Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013), and do they really matter for discrimination taking place outside the laboratory in noisy real-world settings such as the labor market (Rooth, 2010)? Are explicit A/P/S better at predicting controlled (e.g., jury decisions) discriminatory behaviors, whereas implicit measures are superior at predicting automatic (e.g., non-verbal behavior) discriminatory behaviors (Greenwald, Poehlman, Uhlmann, & Banaji, 2009)?

Glancing at the literature, one gets the impression that there is considerable research showing to what extent A/P/S, measured by both implicit and explicit measures, predicts ethnic/racial discrimination. Further, this literature also appears to have been neatly synthesized in recent meta-analytical reviews by Talaska et al. (2008) and Oswald et al. (2013). However, looking more closely, we find a clear divide in how this research question has been operationalized. Some researchers operationalize it as the extent to which an A/P/S measure predicts differential treatment of black people (or ethnic minority group) vs. white people (or ethnic majority group). Other researchers operationalize it as the extent to which an
A/P/S measure predicts treatment of individuals from a single specific group (e.g., behavior toward only black people). In Talaska et al. (2008), roughly two-thirds of the studies rely on this approach, and in Oswald et al. (2013) close to half of the studies do. This single-group operationalization has remained a common approach both in classical and contemporary published research. Although we do not wish to go into detail about the quality of specific studies, we note that studies using this approach are published in top journals of the field and become highly cited. Hence, our impression is that this single group approach is a widely accepted approach in this field. Yet, this way to operationalize discrimination is in stark contrast with how researchers studying discrimination (without the inclusion of A/P/S) tend to operationalize it. Here discrimination is almost exclusively operationalized as differential treatment, both in classical and contemporary research. This is true for lab studies, such as the Goldberg paradigm-studies (Goldberg, 1968), as well as for experiments in field settings (e.g., lost letter technique; Milgram, Mann, & Harter, 1965). Furthermore, this consensus is not confined to psychological research, as it seems to be standard procedure in the hiring discrimination literature in the field of economics too, where audit studies (Neumark, 1996) and correspondence testing studies (e.g., Riach & Rich, 2002) are very common.

The consensus in operationalization of discrimination among discrimination researchers is not the least surprising, since discrimination is inherently a relative concept. To discriminate is to distinguish between something compared to something else. In the case of ethnic and racial discrimination, it is to make a difference between people based on their ethnicity or race. Any operationalization that does not consider this inherent relative nature of discrimination has poor conceptual validity. In our view, such measures cannot be viewed as operationalizations of discrimination; they are proxies, at best.

Discrimination researchers agree that capturing the difference in the treatment of group A relative to group B is an essential part of the operationalization of discrimination. Yet, a
large portion A/P/S researchers focus on predicting an outcome based solely on the behavior toward a single group (e.g., black people). Of course, the question of whether A/P/S can predict behavior toward individuals of certain races/ethnicities in isolation might be a relevant one in itself. However, it is not the same as whether A/P/S can predict discrimination. To illustrate this, we will make an analogy with consumer products. Let us assume that we have an explicit measure of attitudes toward Coca-Cola. Naturally, this attitude should predict one very important behavioral outcome: drinking Coca-Cola. Now, suppose we are specifically interested in whether people with positive attitudes toward Coca-Cola prefer drinking Coca-Cola to Pepsi-Cola when offered a choice. That is, if they discriminate between Coca-Cola and Pepsi-Cola. Clearly, this question of whether people prefer drinking Coca-Cola to Pepsi-Cola is not the same as whether they like drinking Coca-Cola. It is easy to imagine that positive attitudes toward Coca-Cola are positively related to drinking both Coca-Cola and Pepsi-Cola. This would be the case if the attitude measure also captures a general preference for soft drinks. Hence, the attitude measure that correlates most strongly with the tendency to drink Coca-Cola is not necessarily the attitude measure that correlates most strongly with the tendency to prefer drinking Coca-Cola instead of Pepsi-Cola.

As illustrated, a measure’s (A/P/S) ability to predict behavior towards a single object is not the same as its ability to predict differential behavior toward that object and another object. A measure’s ability to predict behavior toward black people is thus not always the measure that most strongly predicts differential behavior toward black versus white people (i.e., discrimination). Indeed, similar to the imaginary Coca-Cola attitude measure that had a component related to soft drinks more generally, the attitude measure toward black people might have a component that is related to behavior toward people in general, regardless of race. Suppose our behavioral outcome is how friendly a person is toward black people. An attitude measure that captures a latent construct of friendliness toward people in general,
Methodological Issues in Predicting discrimination

regardless of race, would spuriously increase the correlation with friendly behavior toward black people. However, it would also increase the correlation, in the same direction, with friendly behavior toward white people. In terms of predicting discrimination, this measure may end up just as good as, or worse than a measure that did not capture that general component.

We are not the first to acknowledge this problem, yet researchers in the field have generally not elaborated much on this issue. This is illustrated by Talaska et al.’s (2008) meta-analysis where the issue of measuring behavior toward a single group is briefly mentioned in a footnote:

...the size of the correlations in the studies that measured behavior toward outgroup targets only, without reference to behavior toward ingroup targets, may be inflated by this phenomenon. This finding questions the meaning of studies that measure behavior toward outgroup members only. Perhaps some attitude measures simply predict who will be more or less aggressive or conformist, rather than who will behave in a specifically prejudiced manner (Talaska et al., 2008, pp. 274).

However, a less obvious problem that to our knowledge has not received any attention in the literature is that the behavior toward single-group correlational operationalization may also seriously underestimate an A/P/S measure predictive validity. Imagine a young, and hopeful, A/P/S researcher who has constructed a new, promising, A/P/S measure. The hopeful researcher then decides to run a study to test its ability to predict discrimination. In addition to the new A/P/S measure, the hopeful researcher also includes a well-respected and well-established old measure of A/P/S for comparison. The results are devastating. There is zero correlation between the new measure and the discrimination outcome. On the other hand,
there is a weak correlation between the old measure and the outcome. The previously hopeful young researcher quietly hides the work deep down in a file drawer. However, doing so may be a bit premature. For a moment, let us imagine that this new A/P/S measure was, in fact, quite remarkable; it was perfect. Let us further imagine that the conceptual link with discrimination was also perfect. Then, why was there no correlation? The simple explanation may be that there was no discrimination to predict in the first place. The small amount of variance predicted by the old measure may simply be due to shared method-variance, or some other general component that inflated the correlation. After removing this variance, zero explained variance remains. However, if the researcher had included a comparison group then it would have been possible to use it as a manipulation check to ascertain how much of the variance that can be attributed to discrimination. Better yet, if the research instead had focused on a differential treatment operationalization, the researcher's effect size would be more directly related to the variance produced by individual differences in discriminatory treatment. In essence, the researcher would have been able to tell whether the new A/P/S measure had been given a fair chance, or not.

Towards more rigorous prediction of discrimination

Our intention is not that this paper shall be a matter of academic nitpicking or remain of mere theoretical interest. Nor do we intend to spend the next pages complaining about the state of the literature. Rather, we intend it to be of practical importance for the pragmatic A/P/S researcher. To this end, the remaining paper will focus on how to make best use of the comparison group in the discrimination outcome, when investigating the relationship between A/P/S and discrimination.

First, we will highlight that focusing on moderating discrimination shifts the focus from predicting all variance in the outcome variable (e.g., aggressive behavior toward black
Methodological Issues in Predicting discrimination

people) to the variance that can be attributed to discrimination (e.g., difference in aggressive behavior toward black vs. white people). In relation to this, the main effect of group-level disparities, or better yet, the level of discrimination on the individual level, can be thought of as a manipulation check. Next, we will discuss the pros and cons of a within-participant approach, as well as the usefulness of some hybrid forms. We will further provide examples of discrimination outcomes where a known population parameter can be used as a comparison group, and discuss the quality of the comparison groups. The paper is concluded with a general discussion about the future for A/P/S-discrimination research, recognizing that the issues we have discussed are only but a few of the many challenges researchers face when studying this phenomenon. After all, not all problems are solved by means of comparison groups.

**Focus on moderation allows for a pragmatic manipulation check**

Predicting behavior toward a single group puts the A/P/S measure to the task of predicting all variance in this behavior. This is an absurd challenge for an A/P/S measure. Even among extremely racist persons, there are, of course, likely to be other factors influencing how they behave towards a black person besides their attitudes, prejudices, and stereotypes of black people (Ajzen, 1991). An effect size of $r = .2$, which equals to 4% explained variance, is generally considered to be weak (Cohen, 1988). Now, if the amount of variance due to discrimination in the outcome variable is 100%, then this is truly a weak effect with respect to predicting discrimination. But what if only 5% of the variance was due to discrimination? Then explaining 4% of *that* variance is a very impressive result. Unfortunately, we can rarely precisely know how much variance is due to discrimination and how much is due to (from this perspective) noise. Further, when correlating the A/P/S measure with behavior towards a
single group, we have no idea if it is the *right* variance that it correlates with and not some nuisance factor.

We can greatly improve this situation by calculating a main effect of discrimination and test if this main effect is moderated by the A/P/S measure. That is, to show that people in general discriminate to a certain amount, but that some people discriminate less, and some people discriminate more. In this regard, the main effect can be thought of as a manipulation effect: showing that discrimination exists at the group level. Of course, what types of manipulations checks that are available, and their precision, depend on the experimental paradigm used. For example, the gold standard in labor market discrimination research is the correspondence testing method (e.g., Bertrand and Mullainathan, 2004). In this paradigm, the researchers apply for real job openings through fictitious applications, focusing on job interview invitations (callbacks) as the outcome. A typical version is one where the employer receives two equivalent candidates that differ only with respect to their race and/or ethnicity signaled by their name, for example Eric and Hassan. In this case, the recruiter can make four different decisions:

1. Only Eric is invited
2. Only Hassan is invited
3. Both Hassan and Eric are invited
4. Neither Hassan or Eric are invited

Assuming that two candidates are truly equivalent on a case-for-case basis (i.e., not just through counterbalancing of nonequivalent applications), we can readily conclude that inviting both Hassan and Eric is evidence of equal treatment. Most of the cases where only Eric is invited are also likely due to discrimination relative to Hassan. Why not all of the cases? The reason is that there may, indeed, be other reasons for inviting only Eric. For
example, a recruiter may have had hundreds of applicants and, simply by chance, decided to stop reading applications after Eric's, and thus did not even read Hassan's. Perhaps Hassan's application was lost in the mail. To what extent this type of error variance is a nuisance factor or not can partly be estimated from the numbers of recruiters who invited only Hassan. If there is a large proportion inviting only Eric and a large proportion inviting only Hassan, then this could be due either to the presence of mixed discrimination of both Eric and Hassan, potentially driven by individual differences, or due to the presence of random noise. However, in this paradigm, there is a built-in control that should eliminate most type of undesired variance: none invited. In this case, neither Hassan nor Eric was invited. Although this is technically the same as equal treatment, it can also be understood as a missing value. Perhaps the job position was already filled, or perhaps both of the equivalent candidates where unsuitable for the job.

In the paradigm above, the researcher can estimate the amount of variance that is due to discrimination by examining both the average level of callback difference between Hassan and Eric, the average level of callback (i.e., the base rates), and the individual results (i.e., recruiters inviting only one of them). Suppose that Hassan is invited in 20% of the cases, and Eric is invited in 30% of the cases, and that nearly all of the outcomes are either both invited, or only Eric invited, with the rare event of Hassan being the only one invited, then the researcher could conclude that nearly all variance in choice 1 – 3 can be moderated by an A/P/S measure. The researcher may include the none-invited and view this as a true equal treatment, or remove this category and view it as missing data, all depending on an informed understanding of what this category constitutes.

Now instead suppose that both Hassan and Eric were invited in 30% of the cases, but that there is large variation in that some invited only Hassan and some invited only Eric. In this case, the researcher cannot distinguish between individual variation in discriminating Eric
or in discriminating Hassan, and random noise in the outcome. The findings may still be very important and may guide future experiments that are designed to separate this specific, and unexpected, outcome. For example, perhaps there is some underlying moderating variable in the data that could be measured, or manipulated, in a future experiment.

Suppose that the researcher had focused on the outcome of Hassan alone as an operationalization of discrimination. In this case, the manipulation check had been impossible. Consequently, we could not have known how much of the individual difference in discrimination the A/P/S could moderate (i.e., variance due to discrimination) and how much it cannot (i.e., variance due to random noise). Recall the example where Hassan is invited in 20% of the cases. A naive view of the variable without a control group (Eric) would suggest that 20% did not discriminate, and 80% discriminated. However, had we examined the levels of callback for Eric, we would have known that 70% did not invite him either. In other words, nearly all of the observations that we would label as discriminating would be wrong. Perhaps even more importantly, we would have no way knowing how many of the recruiters labeled as discriminating are misclassified. It could be none of them, or it could be all of them.

Let us consider another type of experimental paradigm and the available manipulation check: a laboratory experiment where the participants get to choose whether to work with a (otherwise identical) black partner or a white partner. Half of the participants choose the black and the other half the white. In this case we have two extremes: either every single participant chooses their partner at random and there is no variance for the A/P/S measure to predict, or every single participant makes their choice based on racial preferences, meaning that the A/P/S measure could, in theory, predict 100% of the variance. Hence, in this case we do not know who, if any, actually discriminated. On the other hand, had the outcome shown a main effect of discrimination (30 vs. 70%) we would have known that at least the variance that
correspond to the main effect can be moderated by A/P/S. We would still not know which of
the 70% who choose a white partner actually discriminated, but we suspect that at least 20% of
them did. However, we also need to consider the statistical precision of this figure. In a small
sample this figure might have emerged although the participant truly chose partners at
random regardless of race. In this case, no participant would actually have discriminated and
there would be no variance to predict by means of A/P/S.

The astute reader will have noticed that thus far we have kept our reasoning in the realm
of discrete outcomes. However, we have done so for sake of simplifying the argument. The
logic remains for continuous ratings, but they are, perhaps, even more deceptive. For example, imagine a rating of friendliness toward black people that is made on a 10-point scale, with a mean of 5. It is tempting to assume that those with very low ratings (e.g., 1-3) are more likely to have discriminated than those with really high rating (e.g., 8-10). However, without any comparison, we cannot know. The discrimination of interest in this case is the differential rating, rather than the absolute value. Indeed, it may be that the seemingly very positive individual who rated the black confederate as a 7, is actually a discriminating individual, had we only known his/her inclination to rate white confederates as 10.

Although the check for a main effect of discrimination is a relevant manipulation check that gives the researcher a good start in knowing how strong a correlation to be expected with the A/P/S, it is important to emphasize that this type of manipulation check is not exhaustive. Consider a study where the participants consist of 50% white people and 50% black people and where the discrimination, is entirely driven by ingroup preferences. With black people discriminating white people, and white people discriminating black people, the true discrimination effect may be suppressed by this underlying interaction effect in the data. In
this case, it is necessary to analyze the relationship with A/P/S on the sub-group level for it to be meaningful.

To summarize, the main argument here is that it is important to consider just how much of the variance that is due to discrimination, before trying to predict it. A test of the main effect difference between the two groups (e.g., behavior toward black people and white people) is typically a good starting point, but researchers should, of course, make use of other indicators suggesting that there is variance besides this that is due to due to discrimination (e.g., because of mixed discrimination). Theory and data have to guide this investigation in order for the researcher to have an adequate understanding of what s/he aims to predict. Naively, assuming that the amount of discrimination variance that is predictable by means of A/P/S is equal to the variance available in the discrimination outcome variable, and that this remains the same between different experiments, may halt progress by holding new A/P/S measures to a standard they cannot accomplish. Further, researchers ambition to find results (even when there is none) is often referred to as a phishing. Continuing this analogy, if we know beforehand (due to a manipulation check) that there is likely no fish in this particular lake, we should be particularly skeptical about serving whatever we catch. Certainly, we should not blame the fishing rod for going hungry that day.

**The benefits of within-participant designs**

A between-participant design is very common when studying discrimination. The reason for this is obvious. By presenting the participants with stimuli persons from two groups, there is a high risk that they will become aware of the research hypothesis. Yet, when the researcher’s goal is to predict discrimination from A/P/S, within-participant designs are more suitable. After all, the focus is on predicting differential behavior on the individual
level, which cannot be achieved with a between-participant design. Essentially, a within-participant is conceptually a better operationalization: if we are interested in individual differences in discrimination, we should measure this as directly as possible. This type of design is specifically well-suited for researchers who want to predict with any degree of precision how high an individual’s score on the A/P/S measure needs to be in order for him/her to be at substantial risk for performing discriminatory behavior.

Is it possible to design discrimination outcomes that enable high precision of detecting discrimination on the individual level, without using a true within-participant design that would arouse too much suspicion? A solution to this problem could be a between-participant design with multiple non-equivalent comparison groups. A fictive discrimination study where participants are randomly assigned to rate the perceived intelligence of either a black individuals or a white individual will serve as an example. These individuals are equivalent between participants, except for the race manipulation. The ratings of these individuals are then embedded among ratings of other individuals that vary on a large number of qualities, within participant, but remain constant between-participants. This would likely reduce suspicion about the specific focus of the study compared to presenting two equivalent individuals to the same participant, which may be quite unrealistic in many contexts. Despite using a between-participant design on the effect of interest, it is possible to calculate a differential treatment effect on the individual level relative the non-equivalent controls. Preferably, this design can be expanded to include ratings of several black (or white, manipulated between participant) individuals that vary among themselves (to avoid suspicion) to increase the reliability of the discrimination outcome. Of course, this type of design can be taken to the field through correspondence testing where the research sends out several applications to the same job.
Known population parameters as comparison groups

It is not always necessary to collect within-participant, or even between-participant data in order to have a comparison group. In some instances, it is perfectly defensible to rely on previously known population parameters. For example, the known mean and variance of behavior toward a comparison group. For example, consider a discrimination outcome in the form of signing a petition regarding an exemplary family belonging to an ethnic minority group moving in to a neighborhood. It may be reasonable to assume that the number of petitions against a comparison family belonging to an ethnic majority group would not receive any petitions at all, or at least very few petitions. Hence, the behavior towards this single group can be treated as a difference variable by subtracting the known value. Notice, however, that this technique only works on the individual level when the variance of the comparison group is close to zero. An a priori group value of perception of aggression against the comparison group, for example, will at best equal a between-participant design, since there is likely too much variance in such variables to be able to draw any conclusions in the individual case on how that particular individual would have behaved toward a comparison group had the data been collected. Still, if researchers are able to, through careful pre-testing, demonstrate that there is little variance in behavior toward a comparison group on a particular outcome, then it will be safe to treat the behavior towards the single group (e.g., black people) as differential treatment.

Not all comparison groups are created equal

Thus far, we have stressed the usefulness of including one or more comparison groups. However, not all comparison groups are of equal quality. A common approach for discrimination researchers is to manipulate the race or ethnicity by means of a few (or even single) stimuli pairs, such that participants’ behavior are measured toward a few black
confederates and a few white confederates (see e.g., Dovidio, Kawakami, & Gaertner, 2002; McConnell & Leibold, 2001). When testing for the difference between these two conditions (e.g., *t*-test), researchers almost exclusively treat race or ethnicity as a fixed factor, and participant as a random factor in their statistical analyses. When stimuli are statistically treated as a fixed effect, it cannot be taken for granted that the results generalize beyond the population of the particular stimuli used in a study, essentially meaning that nothing is known about how the participants would have reacted to another set of stimuli. As Judd, Westfall, and Kenny (2012) have recently shown, in order to be able to generalize a stimuli (e.g., black confederate) to the population from which it was drawn (e.g., black people), researchers need to statistically test this inference by treating it as a random factor. The question then is which of these inferences are crucial for discrimination research? Are we content that our findings cannot generalize beyond the included stimuli, or is this generalization actually a crucial part of the hypothesis testing and/or estimation?

Judd et al. (2012) argue that treating stimuli as a random effect and crossing them with participant as a random effect through mixed models, is essential for social psychological research. Failing to do so will increase type-I error by a rather large margin. For example, a significant difference found in the treatment of a single black confederate versus a single white confederate may not be representative at all. Perhaps the white confederate happened (by chance) to be nicer than the black confederate. In this case, the effect is confined to this stimuli pair, driven by a stimuli effect that is due to individual differences between the confederates, rather than their race per se.

Judd et al. (2012) provide a compelling case why social psychologist should routinely sample a large amount of stimuli and treat them as a random factor. Indeed, given the advances in statistical tests, achieving this is, analytically, no challenge. Our view is that this approach should become the future goal standard of social psychology research in general,
and discrimination research in particular. However, it certainly raises the bar for how to
design discrimination studies, and the researcher needs to have large samples of both stimuli
(i.e., confederates) and participants. To make such studies cost-efficient, researchers may
want to look into paradigms from behavioral economics (e.g., Mobius and Rosenblatt, 2006),
where the participants act as stimuli themselves, and the researchers focus on measuring their
interaction. Of course, if the research calls for video or photo stimuli, it will simply be a
matter of making use of a larger set of such stimuli.

From a pragmatic discrimination researcher perspective, we would like to emphasize
that the actual benefit of treating the stimuli as a random effect will vary greatly depending on
the expected variation of the stimuli in relation to the research question. If there is little
variation among different possible stimuli that can be sampled, there is no point in treating
this as a random factor when analyzing the data. Further, sampling a large set of stimuli that
we know, a priori, contain little variation, is wasteful of resources. In discrimination research,
the expected amount of variation among stimuli that can influence discrimination will vary
greatly for different types of stimuli and outcome combinations. Consider a study with
perceived aggression as the dependent variable. If the stimuli are confederates, then treating
this as a random factor becomes crucial. After all, we can expect variation among
confederates, regardless of race, on how aggressive they may appear. Indeed, there is a real
risk that any differential treatment of these confederates are not due to the population they
were sampled from (black individuals and white individuals) but rather due to variations
within these populations (variation among black individuals and variation among white
individuals). On the other hand, if the stimuli consist of photos, there is less room for
individual differences to shine through, but we should still be concerned about them in many
cases. If the stimuli are sets of black and white names, the amount of expected variance
plummets and researchers could rely on previous findings on this variation in relation to
Methodological Issues in Predicting discrimination

discrimination when deciding. Finally, suppose that the stimulus is a written statement on a medical record that simply indicates race. In this case, the sampling can be viewed as exhaustive (see Judd et al., 2012), and thus treating it as a random factor would not make any sense. Indeed, the factor is truly fixed.

To summarize, it is often necessary to treat the stimuli used for the ethnic/race manipulation as a random factor, in order to be able to generalize the findings to discrimination against the ethnic groups or races, rather than toward individual confederates or photos. However, when researchers carefully match stimuli, the variation may be reduced to levels where there is too little variation in the stimuli for random effects to be meaningful. Further, this makes for some quite fuzzy generalizations available from this approach in either case.

To strike a balance between good statistical inference, and pragmatic research practices, we propose the following. Whenever possible, a researcher should randomly sample stimuli and treat them as a random factor. This is readily possible with confederates (e.g., pool of participants) or photos (e.g., photo databases). However, researchers should not give up on carefully matched stimuli (e.g., photo manipulation), but rather pay attention to the variation among stimuli, when deciding whether a large set of stimuli treated as a random factor is more appropriate than a small set of stimuli treated as a fixed factor.

**Discussion**

In the present paper, we have argued that researchers aiming to predict discrimination by means of A/P/S should turn their attention away from the quite common approach of predicting behavior towards single groups (e.g., behavior toward black people), and instead focus on moderating individual differences in differential treatment. This approach is a better fit conceptually, since discrimination is inherently relative and cannot be operationalized as
an absolute behavior. It further reduces the risk of inflated (i.e., due to shared irrelevant variance) or attenuated (i.e., due to lack of discrimination in the outcome) correlations that can arise from a one-sided focus of predicting behavior towards a single group. Perhaps most importantly, careful use of comparison group will allow the A/P/S researcher to know how much (if any) of the variance is due to individual differences in discrimination, which is essential for being able to interpret the effect size of the correlation with A/P/S. In sum, we are essentially arguing that A/P/S researcher should hold discrimination outcomes to the same standard as they do their A/P/S measures. If the discrimination outcome is both invalid and unreliable, then the A/P/S measure cannot be faulted for not predicting it.

Although we have in this paper chosen to specifically focus on the role of the comparison group in discrimination outcomes, we think that it is important to put this into perspective by also discussing other issues when predicting discrimination. One such issue concerns whether the discrimination reflects actual behavior, rather than, for example, self-reported behavioral intentions. Indeed, the relative paucity of research studying behavioral outcomes within the field of personality and social psychology is not a new observation and this issue was thoroughly discussed by Baumeister, Vohs, and Funder (2007) in their article “Psychology as the science of self-reports and finger-movements”. They concluded in their article that the direct observation of behavior is rare and that it has been increasingly replaced by introspective self-reports, hypothetical scenarios, and questionnaire ratings.

Our impression is that the A/P/S-discrimination literature is somewhat better in this regard. Following the definition set by Baumeister et al. (2007), roughly a third of the discrimination outcomes studies would be classified as having captured behavior, such as partner choice, seating distance or laughter at jokes. Yet, the majority of the literature relies on questionnaire ratings (e.g., person perception), introspective self-reports, and hypothetical responses. In our view, in the case of discrimination, it is quite difficult to draw a clear-cut
Methodological Issues in Predicting discrimination

line between what is a behavior and what is not. For example, a judgment of an essay's quality may have higher ecological validity in relation to teacher's discrimination of students, than a rather artificial study about non-verbal behavior. Yet, the latter may seem more behavior-like than the former. Simply preferring behavior-like outcomes might only increase the mundane validity of the outcomes; making them look valid although they do not really capture anything of actual relevance. Our suggestion is thus that researchers should not worry too much about whether their outcomes can be strictly classified as behavior or not, but rather of the internal, external and ecological validity of the measures.

We are using the ecological validity term to emphasize the concern that discrimination outcomes captured in artificial settings may (often without the researcher’s knowledge) remove the incentives to discriminate. For example, a recruiter that is hiring people to a real job position has a high-stake decision to make. To discriminate when making a real hiring decision is a psychologically very different situation than one in which the same recruiter evaluates candidates in a laboratory. Indeed, in that setting, the incentives may instead be to appear socially desirable. Discrimination outcomes that have been affected by self-presentational issues may seriously inflate, or attenuate the correlation with A/P/S. Importantly, the inflation/attenuation may be different for different types of A/P/S measures. Suppose we have an A/P/S measure that is not affected by self-presentational bias and that predicts a discrimination outcome reflecting high levels of self-presentational bias. In this case, there may be very low levels of discrimination and thus a weak correlation. On the other hand, had the A/P/S also reflected a strong component of self-presentational bias, this component may artificially inflate the correlation. This issue becomes especially likely if there is a high degree of shared method variance between the A/P/S and the discrimination outcomes. For example, if participants provide ratings of "Black people in general" when completing the A/P/S measures and differential ratings of black vs. white individuals as the
discrimination outcome. As we have stressed in this paper, a practical solution to this problem is to make proper use of manipulation checks of the main effect of discrimination, ascertaining that there is substantial discrimination, rather than simply nuisance, to predict.

One solution to this self-presentational problem, which also allows for behavioral outcomes, is to take the research to the field, relying on measuring real-life discrimination outcomes unobtrusively. However, there are severe limits to what can be studied in such paradigms. For example, a researcher may want to test whether an A/P/S measure’s predictive ability is moderated by some factor (e.g., time pressure), which is rarely feasible in a field experiment. Hence, we are not suggesting that researchers should give up on laboratory-based research in favor of capturing real-world behavior data. Rather, researchers should rely on both approaches, but pay careful attention to the size of the correlation with A/P/S in relation to the type of discrimination outcome it has been related to. A correlation of $r = .20$ may seem unimpressive in the context of a carefully conducted laboratory study, but may be as good as it gets when correlating the A/P/S measure with an inherently noisy, real-life discriminatory behavioral outcome. Further, a laboratory study plagued by self-presentational issues may still provide relevant information, provided that researchers realize these problems and consider the correlation with A/P/S in light of this.

The core question which A/P/S-discrimination researchers want to address is to what extent stable individual differences are prognostic of future discriminatory behaviors. As with a person’s A/P/S, which are assumed to show rather high chronicity, we would expect that a person’s inclination to discriminate would show some degree of consistency. In light of this, it is surprising that virtually the whole literature consists of studies measuring one-shot discriminatory outcome(s) during one specific study session. After all, what we really want to know is how well A/P/S can predict stable individual differences in the tendency to discriminate. Thus, if the field is to move forward, longitudinal studies are required that
Methodological Issues in Predicting discrimination

measure people’s discriminatory tendencies on repeated occasions over an extended period of time. Although costly and time consuming, such studies are feasible to conduct. One could capitalize on the fact that that university education normally takes years to complete and that many programs include substantial practical training. Thus, it would be possible, for example, to follow medical students throughout their medical training and assess their treatment recommendations made in response to simulated patient case exercises where the race or ethnicity of the patient has been manipulated. Such exercises are likely to be an integral part of their medical training, and therefore do they not only allow for repeated testing, they are also realistic exercises on which the students are motivated to do well. Further, participant pools (e.g., student samples or Mturk) can be another way to implement this. The participants can initially consent to researchers correlating findings between different individual studies they participate in.

The role of meta-analysis

In the start of the paper we mentioned that the literature appears to have been neatly synthesized by Talaska et al. (2008) and Oswald et al. (2013). Although most of our criticism so far has been directly toward individual studies, it has implications for meta-analytic research too. Meta-analysts should not draw conclusions based on aggregations of studies that have valid operationalizations of discrimination and studies that do not (i.e., no comparison group). Further, the meta-analysts need to carefully consider the possibility of both inflated and attenuated correlations for the reasons we have reviewed in this paper. This is especially important when comparing different types of A/P/S measures, as what may inflate the correlation for one measure, may attenuate it for another measure. For example, self-presentational issues may inflate correlations between explicit scales and discrimination outcomes, but attenuate the correlations between implicit measures and discrimination
Methodological Issues in Predicting discrimination outcomes. Minimally, the meta-analysts need to consider the level of discrimination in the outcomes that the studies set out to predict. Even better, the meta-analysis should be conducted on aggregation of moderation analysis. At present, this may not be realistic. We do realize that meta-analysts have to make the best of the available literature. Setting the bar too high will likely exclude almost every study and make the meta-analysis meaningless. Yet, doing so will send a strong signal that more research of higher quality is indeed needed. Continuing meta-analyzing discrimination outcomes without proper attention to their quality will further contribute to the unfair evaluation of A/P/S measures.

Conclusion

So far we have argued that without rigorous measures of discrimination we cannot make much sense of the correlation between A/P/S and discrimination outcomes. It is now time to admit that our reasoning is a slight exaggeration. Even a poor proxy outcome will tell us something. In fact, we do not always need to calculate a correlation coefficient in order to gain insight into the A/P/S-discrimination link. In his (1934) seminal study, LaPiere recorded how often two individuals from a single group (a Chinese couple) either received or were declined accommodation or service by proprietors of various establishments. He found that virtually all establishments (127/128) offered accommodation or service to the Chinese couple. This was, of course, very surprising news to LaPiere as 92% the same establishments later reported that they would not accept Chinese people as guests. LaPiere’s design, which lacked a comparison group, was good enough to answer the question whether of attitudes and behavior always correlate. Further, thanks to the near perfect acceptance rate of the Chinese couple, the need for the comparison group became moot. Similarly, researchers predicting behavior toward a single group on the basis of A/P/S may be able to answer the question whether there is any relationship between A/P/S and discrimination. In this case, even a poor
proxy outcome may be good enough, especially if the correlation is sizeable, since it is unlikely for a sizeable correlation to be entirely spurious. However, 80 years after La Pierre’s study, it is time for A/P/S-discrimination research to move from simple binary questions, to precise estimations of how much discrimination A/P/S can predict. To be able to fulfill this research endeavor, researchers need to adopt more rigorous methods and take both the conceptualization and measurement of discrimination as seriously as they do with their A/P/S measures.

References


Methodological Issues in Predicting discrimination

doi: 10.1177/0956797611430953

doi:10.1037/a0028347


doi:10.1177/1368430207078696


\[\text{In essence, the link between A/P/S and race/ethnic discrimination is a question about predicting individual differences in direct (or disparate treatment) discrimination. The focus is on the (social) psychology of the perpetrator-victim interaction. This type of disparate treatment (direct) discrimination is distinct from disparate impact (indirect) discrimination where the source of the discrimination is not in a perpetrator (or his or her A/P/S), but in structures, policies or laws that disfavor certain groups on average. For example, disparate impact occurs if white job candidates are more likely to be hired when an employer hire staff based on the results from an aptitude test that typically produces race differences in performance. Throughout the present research, when we use the term discrimination, we always refer to disparate treatment.}\]