Article in press, *Journal of Experimental Social Psychology*

The Mind's "Aye"? Investigating Overlap in Findings Produced by Reverse Correlation Versus Self-Report

Jordan Axt

Nellie Siemers

Marie-Nicole Discepola

Paola Martinez

Zhenai Xiao

Emery Wehrli

McGill University

Word Count: Abstract = 145, Main Text (excluding footnotes) = 5844, References = 1028

Total Word Count = 7017

Corresponding Author Contact:
Department of Psychology, McGill University
2001 McGill College Ave
Montreal, Quebec, Canada, H3A 1G1
Email: jordan.axt@mcgill.ca
Phone: 514.398.4657

Authors' Note: The third, fourth, fifth and sixth authors contributed equally to this work and are listed alphabetically. All data and study materials are available at the project page on the Open Science Framework (https://osf.io/m2dn8/).

Abstract

Reverse correlation is an influential method for assessing mental representations. One benefit of reverse correlation is that the method may capture psychological content that individuals are unwilling to self-report due to social desirability concerns, particularly in domains like person perception or intergroup processes. To investigate the degree to which reverse correlation and self-report findings are aligned, 32 prior reverse correlation studies (totaling 148 analyses) were converted into comparable measures of self-report (total *N* = 3441). Despite only 13% of original studies containing a parallel self-report measure, 55% of research conclusions could be replicated using self-report, though effect sizes from the two methods were unrelated (*r* = -.06). The two methods were more likely to reach the same conclusion for findings that were rated as more intuitive. Future uses of reverse correlation will benefit from greater consideration to when the method is most necessary and informative.

Keywords: Self-report, Reverse correlation, Intergroup processes, Measurement, Stereotypes

The Mind's "Aye"? Investigating Overlap in Findings Produced by Reverse Correlation Versus Self-Report

Mental representations of social groups– internal images that come to mind when thinking of such groups (Brinkman et al., 2019)– are a psychological fact but a methodological challenge. While mental representations can be influential in guiding our beliefs and behaviors towards others, their content could be difficult or impossible to articulate or measure with any precision. Accurately measuring individuals' mental representations may be an invaluable method that could shed light on psychological constructs that are distinct from those assessed by either self-report or indirect measures of implicit associations (e.g., Bar-Anan & Vianello, 2018). This is precisely the goal of reverse correlation (Mangini & Biederman, 2004), an increasingly popular method in psychological research.

Reverse correlation can take several forms (e.g., Ahumada & Lovell, 1971; Ahumada, 1996). Broadly, the aim of the method is to find associations between features of a stimulus (such as the pixels that comprise a face on a screen) and participant responses. The method looks to highlight the visual inputs that are most relevant for determining certain information (e.g., what areas of a face are most important for perceiving different emotions). When applied to face perception, the method asks participants to make a perceptual judgment on a target face or faces. These faces have either been degraded in some way – such as by only showing certain regions (e.g., Schyns & Gosselin, 2001) or adding grayscale noise to certain pixels (Mangini & Biederman, 2004) – or are generated from computer graphics models that allow for fine-tuned control over what facial features are being combined in any target (e.g., Zhan et al., 2019). Through a variety of analytical techniques, researchers can then objectively determine what stimulus features drive different perceptions by applying a formal analysis to the images that have been viewed by participants (see Jack & Schyns, 2017 for review). For instance, van Rijsbergen, 2014 et al. (2014) had younger and older participants estimate the age of different visually degraded faces. The researchers then analyzed how strongly individual pixels on the face related to age judgments, finding that the area around the nose was the strongest determinant of age perceptions. Using similar analysis approaches, reverse correlation procedures have been used to identify the areas of a face that most differentiate judgments of gender (e.g., Nestor & Tarr, 2008), or reveal cultural differences in the degree to which various emotions have distinct facial representations (e.g., Jack et al., 2012).

However, the reverse correlation method has been applied more narrowly in many social psychology studies, particularly those investigating intergroup perception. In these cases, reverse correlation analyses focus more on participants' behavioral responses than on objective features of the face (e.g., pixel intensity). The procedure typically involves two phases (Brinkman, Todorov & Dotsch, 2017). In Phase I, participants complete a classification task (usually of at least 100 trials), where they choose from two images in selecting the better depiction of a certain feature or group (e.g., choosing the face of someone thought to be high or low in socioeconomic status). The images on screen are two versions of the same base image, one presented with a layer of random noise and the other presented with the inverse of that same noise. These many selections are then averaged into classification images (CIs), either at the group level (e.g., across the entire sample, making one average face for people perceived as having low socioeconomic status and one average face for people perceived as having high socioeconomic status) or the individual level (making average faces of high and low socioeconomic status for each participant). In Phase II, novel participants are then asked to rate these images on dimensions of interest (e.g., likability, threat) that have been selected by the researcher. Differences that emerge in ratings may indicate variation in how such groups are represented mentally, which may impact how group members are perceived and treated.

In recent years, reverse correlation has been used extensively in the person perception and intergroup literatures. One common justification for its use is that it captures associations that individuals may be unable or unwilling to report, particularly for reasons related to social desirability (Brown-Iannuzzi et al., 2021; Hinzman & Maddox, 2017; Karremans et al., 2011; Petsko et al., 2020). For example, participants may believe that welfare recipients are less competent than non-recipients, but would hesitate to admit such a clearly prejudiced belief on a self-report measure. However, in the context of a reverse correlation task, participants may lack the ability to regulate their performance on the task in the same manner as they can on a self-report scale, or may not understand how their responses are being used, and as a result aggregate performance may reveal a mental representation of welfare recipients that is indeed viewed as significantly more incompetent (Brown-Iannuzzi et al., 2017).

Of course, there are advantages to using the reverse correlation method outside of downplaying concerns about social desirability. For one, the method is clearly preferable for research questions that are intrinsically focused on mental representations, meaning by definition other methods like self-report or various reaction-time based tasks are inappropriate. In addition, reverse correlation provides a “bottom-up” approach that can be highly useful; that is, researchers who are unsure of what dimensions or characteristics distinguish the mental representations of various groups can use the reverse correlation procedure to first generate images of each social group and then have Phase II participants rate such images using many possible traits. In this way, the reverse correlation procedure could reveal intergroup phenomena that are less influenced by a researcher’s own beliefs or assumptions.

These benefits aside, the reverse correlation method is still often used in an effort to avoid the potential influence of self-presentation when studying socially sensitive topics, which is a common concern in domains like intergroup relations or prejudice and stereotyping (Holtgraves, 2004). In these cases, researchers may not be interested in mental representations *themselves*, but rather use mental representations as an outcome and reverse correlation as a method for studying the larger topic of intergroup attitudes and beliefs. Specifically, reverse correlation may allow researchers to study these processes without having to worry about self-presentation concerns tainting participant responses, as may be the case with self-report.

However, recent work suggests that concerns about social desirability or self-presentation on self-report measures may be overstated (e.g., Axt, 2018). For instance, some people blatantly dehumanize others (Kteily et al., 2015) or hold a motivation to express prejudice (Forscher et al., 2015). As a result, it is possible that many of the conclusions gleamed from the relatively resource and computationally intensive reverse correlation method could also be produced with the simpler, faster and cheaper method of self-report. In this case, researchers could save time and resources by using self-report if the method was acceptable for their larger research questions. However, no prior study has investigated the correspondence between the two methods across more than a single domain or topic at a time.

We explored this issue directly by reviewing prior uses of the reverse correlation task in the person perception or intergroup relations literature, and converted each research finding into a measure of self-report. We first compared whether research conclusions generated from the reverse correlation method could be recreated using self-report, while also analyzing the relationship between the effect sizes generated from the two methods. Follow-up analyses sought to identify research contexts where reverse correlation findings were particularly likely to diverge from self-report (e.g., for targets that are comparatively hard or easy to mentally imagine, or for topics that are likely to be socially sensitive).

We correlated effect sizes from the original reverse correlation studies with our self-report replications to investigate any potential relationship between effects produced by the two methods. However, this analysis may be difficult to interpret for several reasons. For one, most prior research in reverse correlation has used group-level (rather than individual-level) classification images in Phase II. This approach fails to adequately convey the variance in responses among Phase I participants, which can inflate Type I errors and exaggerate reverse correlation effect sizes (Cone et al., 2020). In addition, like many domains in psychological research, the original reverse correlation findings may suffer from publication bias (Ferguson & Heene, 2012), as 84.7% of the outcomes we investigated from the reverse correlation literature produced reliable differences among Phase II participants, and the small number of null findings could indicate a separate source of inflation among original reverse correlation effect sizes. Finally, correlating the two effect sizes involves comparisons between different sample sources (e.g., lab versus online) and data were often collected several years apart from one another, introducing other factors that could weaken or remove any relationship between findings from the two methods. As a result, while we report the correlation between reverse correlation and self-report effect sizes, it may not be a particularly conclusive analysis concerning the extent to which the two methods rely on shared versus distinct psychological mechanisms.

To be clear, our method does not assume that self-report measures any objective truth, and that reverse correlation is then only useful in the extent to which it can reproduce conclusions from studies using self-report. Prior literature in psychology has done an effective job of highlighting some of the clear weaknesses of self-report, such as the potential for responses to be influenced by demand characteristics (i.e., responses that seek to satisfy the study's perceived goal; Brenner & DeLamater, 2016), extreme responding (i.e., when factors like emotional arousal or rapid responding lead to overreliance of extreme options on a rating scale; Paulhus & Vazire, 2007), or even a lack of introspective access into the processes guiding one's beliefs or behaviors (Nisbett & Wilson, 1977). Nor do we assume that findings from reverse correlation studies are only useful or interesting if they diverge from self-report. Rather, we believe that there are a variety of research questions in areas like intergroup relations and person perception where both reverse correlation and self-report are acceptable methods, and researchers currently do not know the extent to which the two approaches arrive at similar or differing conclusions when applied to the same topic. The present work then sought to provide an initial estimate of this level of agreement, with the hope of generating useful information to researchers using either method. This analysis may also spur future research looking to better understand where and why the two methods are likely to align versus diverge.

Methods

Identifying Eligible Articles

Articles published before September 22nd, 2020 were retrieved from PsycINFO, Scopus, and Web of Science with the keyword: "*reverse correlation*". Eligible articles were (1) published in English; (2) used a version of the reverse correlation paradigm that involved novel raters evaluating images from Phase I, and 3) either asked participants to completed the reverse correlation paradigm with the instruction to identify members of real-world social groups (e.g., liberals and conservatives; Tskhay & Rule, 2015) or recruited participants from differing demographic groups to complete Phase I (e.g., the degree to which German and Portuguese participants viewed German and Portuguese men as representative of the typical "European man"; Imhoff et al., 2011).

The initial search provided 913 records, with 852 articles excluded after screening, primarily due to irrelevant use of the phrase "reverse correlation" or using the reverse correlation procedure in a manner unrelated to social psychology (e.g., psychophysics). Of the 61 remaining papers, three were removed for full text not being available, 30 were excluded for not investigating existing social groups (e.g., using individuals or novel groups, Young et al., 2014), three were excluded due to an inability to access the necessary target population (e.g., German students in a program for prospective teachers; Degner et al., 2019), one was excluded because translating into a self-report measure became tautological (i.e., the extent to which female bodies are more feminine; Lick et al., 2013), and one was excluded after authors failed to respond to a request for further necessary methodological information. Following these exclusions, 23 papers (totalling 32 studies) were deemed eligible (see the online supplement at https://osf.io/ar7uf/ for more detail behind why specific papers were excluded).

Conversion into Self-Report

Studies were first scanned for each reported analysis concerning Phase II participants' evaluations of Phase I images. Combined with the instructions provided to Phase I participants, this information was the source for translating each analysis into a self-report equivalent.

For example, in Kunst et al., (2018), Phase I participants completed a reverse correlation task where, across conditions, they selected the image that best depicted "a terrorist", "a terrorist driven by mental illness", or "a terrorist driven by ideological reasons". In the original study, the resulting three images were then rated by Phase II participants on a number of dimensions, with one result being that reverse-correlation generated images of terrorists driven by mental illness were perceived as having a higher socioeconomic background than the other two forms of terrorists. To translate this finding into a version of self-report, across three between-subjects conditions we had participants report the extent to which a “terrorist”, a “terrorist driven by mental illness” or “a terrorist driven by ideological reasons” was likely to come from a high or low socioeconomic background (1 = Much more likely to be low socio-economic status, 4 = Equally likely to have low socio-economic status or high socio-economic status, 7 = Much more likely to be high socio-economic status). See Table 1 for additional examples of converting reverse correlation findings into self-report measures, and full materials are available in the online supplement.

In all, studies contained 148 analyses that could be translated into self-report.[1] The online supplement contains a document detailing for each study: 1) demographics of the original samples, 2) original analyses and effect sizes, 3) the sample size needed to achieve 80% power for a self-report study, using original effect sizes, 4) any sample restrictions deemed necessary for the self-report analysis (e.g., only White participants), and 5) full text of all created self-report items. We maximized statistical power by using a within-subjects design when possible, specifically for research questions where we believed responses to one condition would not contaminate responses to the other condition (e.g., perceptions of depressed versus non-depressed people; Krendl & Freeman, 2019). However, we reverted to between-subjects designs when we believed exposure to multiple conditions would substantively influence responses (e.g.,

[1] One analysis was eligible (Kunst et al., 2019) but could not be completed due to an inability to recruit a large enough sample that matched the pre-specified criteria for identifying participants high in Social Dominance Orientation.

Table 1

Examples Converting Reverse Correlation Findings into Self-Report

Original Citation	Reverse Correlation Finding	Self-Report Conversion
Krendl & Freeman (2019)	CI of someone “suffering from depression” was rated as in less control of their health than the CI of someone that was healthy	Asking participants the degree to which a “depressed” person and “non-depressed” person was in charge of their health.
Lloyd et al. (2020), Study 2	CI of a “police officer” generated from Black participants was rated as more dominant than CI of a “police officer” generated from White participants.	Asking Black and White participants the extent to which they view police officers as dominant.
Brown-Iannuzzi et al. (2017), Study 2	CI of a “welfare recipient” was rated as more incompetent than the CI of unselected images meant to depict a non-welfare recipient.	Asking participants to what extent they view “welfare recipients” and “non-welfare recipients” as competent or incompetent.
Tskhay & Rule (2015), Study 1b	CI of a “liberal” was rated as happier than the CI of a “conservative”.	Asking participants the extent to which the typical liberal and typical conservative were happy.
Imhoff et al., (2013)	CI of a “nursery teacher” was rated as warmer than CI of a “manager”.	Asking participants the extent to which nursery teachers and managers are warm (i.e., kind or friendly).
Gundersen & Kunst (2019), Study 2	CI of a woman with “strong feminist attitudes” was rated as more threatening than CI of a woman with either “moderate gender-related attitudes” or CI of a woman with “strong pro-animal welfare attitudes”.	Asking participants the degree to which they feel threatened by a woman with either 1) strong pro-feminist attitudes, 2) moderate gender-related attitudes, or 3) strong pro-animal welfare attitudes.

Note. CI = Classification image.

reporting the extent to which a “poor person” versus a “middle-class person” was likely to be White; Lei & Bodenhausen, 2017).

Data Collection

Studies were grouped into 15 different online data collections using either Prolific Academic or Mechanical Turk. Each data collection included a demographics questionnaire and an attention check item. Studies were grouped based on sample sizes needed for achieving 80% power or for demographic requirements (e.g., needing only Black participants). As a result, data collections varied in length, and most data collections included responses to several different self-report translations of various reverse correlation studies. The table reported in the online supplement (https://osf.io/uqsfg?view_only=0bf5861053af4561adf8f695258f107d) details which items were included in each data collection. In cases where the same data collection included the self-report conversions of multiple studies, order of studies was randomized. In all, data collection involved 3441 participants (see online supplement for demographics). All measures, manipulations, and exclusions are disclosed.

Power analyses were based on the effect sizes obtained from the original reverse correlation studies. Using effect sizes from the reverse-correlation analyses that rejected the null hypothesis, the median statistical power obtained for the self-report studies was greater than 99%, with 85% of analyses achieving at least 95% statistical power to detect an effect size equal to what was found in the reverse correlation analysis, though again original effect sizes may have been inflated due to reliance on group-level classification images (Cone et al., 2020). Twelve self-report analyses had less than 80% power to detect the original reverse correlation effect size, but of these analyses only two failed to reproduce the original conclusion (see online supplement table for the sample size and power for each analysis).

Open Practices

Materials, data, and analysis syntax for all studies is available at: https://osf.io/m2dn8/.

Results

Reproducing Reverse Correlation Conclusions

Studies were coded based on whether results from the self-report analysis reached the same conclusion as those in the original paper (i.e., rejecting or failing to reject the null hypothesis; see online supplement for full writeup of each analysis as well as table reporting effect sizes). Using a strict $p < .05$ criterion, conclusions matched for 55% (82/148) of the analyses, and was 59% (73/127) among analyses that originally rejected the null hypothesis. However, studies varied considerably in number of analyses, and it is possible that the above result could be influenced by studies including many analyses. To address this issue, we calculated a weighted average such that each study contributed equally to the overall rate. Using this approach, 59% of analyses reached the same conclusion when using self-report.

Despite a majority of analyses reaching the same conclusion, a correlation between the original and self-report effect sizes (converted into Pearson's *r*) found no reliable association, *r* (147) = -.060, *p* = .467, 95% CI [-.22, .10]. This finding suggests that though research conclusions produced from self-report and reverse correlation were frequently the same, the effect sizes obtained from the reverse correlation method were unrelated to those found using self-report. See Figure 1 for plot of reverse-correlation and self-report effects. In general, effect sizes were larger in the original reverse correlation studies (reverse correlation median *r* = .39, self-report median *r* = .24). In the General Discussion, we discuss some possible reasons behind

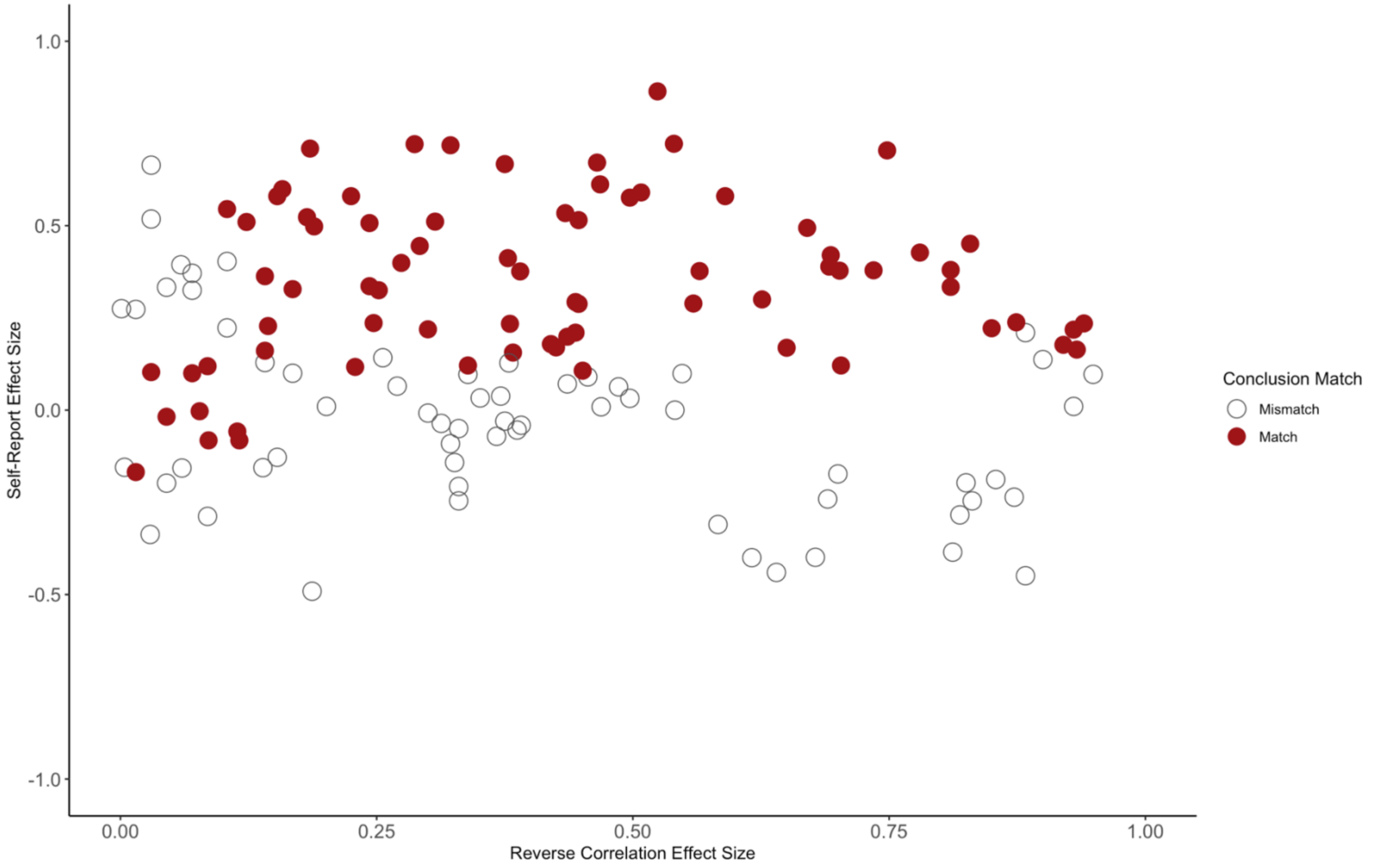

Figure 1. Scatterplot of effect sizes (Pearson's *r*) from 148 original analyses using reverse correlation versus when using self-report. Filled dots represent when the two methods reached the same conclusion (i.e., either rejecting or failing to reject the null hypothesis).

the lack of a correlation between the effect sizes from the two methods, and highlight potential implications for future uses of the reverse correlation method.

Explaining Matches in Reverse Correlation and Self-Report

The results of the above analyses raise an obvious question: what factors explain when reverse correlation and self-report do or do not align? To investigate this question, we recruited three new samples of online participants to evaluate the reverse correlation findings on three dimensions: 1) social desirability of the obtained result, 2) ease of imagining the targeted groups, and 3) intuitiveness of the research finding. Each moderator could account for the modest rate of agreement between the two methods. For one, social desirability may impact self- report responses much more than reverse correlation responses, dampening the effect size of self-report studies but having little or no impact on reverse correlation studies. Conversely, intuitiveness of the finding may highlight domains where people are either willing or able to self-report their preferences, leading to greater alignment for more intuitive research questions. Finally, social groups that are difficult to mentally imagine may complicate responses much more on reverse correlation than self-report, adding noise to reverse correlation effect sizes that make results less likely to produce conclusions consistent with self-report.

In each sample, participants either reported on the social desirability, ease of imagination, or intuitiveness for all 148 analyses. For example, Hinzman and Maddox (2017) found that a reverse-correlation generated image of a “Black athlete” was rated as more racially prototypical than an image of a “Black businessman”. Assessing social desirability for this finding involved asking participants (*N* = 53), “Imagine someone admitted that they believed Black athletes are more racially prototypical (i.e., representative of Black people) than Black businessmen. How offended do you think others would be for admitting this belief?” with a ‘1 = Not at all’ to ‘5 =

Extremely' rating scale. In the ease of imagination sample, participants (*N* = 59) were asked to "Try to picture in your mind a "Black athlete" and a "Black businessman". How difficult was it for you to imagine these people?" using the same '1 = Not at all' to '5 = Extremely' scale (reverse scored so higher values meant greater ease). Finally, the intuitiveness sample (*N* = 60) asked raters to "Imagine that researchers found that Black athletes were viewed as more racially prototypical (i.e., more representative of Black people in general) than Black businessmen. How surprised would you be by this finding?" using the same response scale (reverse scored so higher values meant more intuitive findings). See online supplement for demographics and all items.[2]

A series of binary logistic regressions (*N* = 148) was used to predict whether or not the self-report analysis matched the conclusion of the reverse correlation analysis. When entered separately, the two methods were more likely to match when findings were 1) *lower* in average social desirability (B = -.78, *SE* = .25, *p* = .002, *r* = -.21, 95% CI [-.33, -.08], Nagelkerke R^2 = .09) and 2) judged as *more* intuitive (B = 2.41, *SE* = .26, *p* < .001, *r* = .55, 95% CI [.39, .67], Nagelkerke R^2 = .32), though there was also marginally significant evidence that findings were more likely to align when groups were *harder* to imagine (B = -.81, *SE* = .44, *p* = .064, *r* = -.22, 95% CI [-.42, .01], Nagelkerke R^2 = .03).

However, a correlational analyses found that all potential moderators were correlated with one another, such that outcomes that evoked more social desirability concerns contained groups that were easier to imagine (*r* = .290, *p* < .001, 95% CI [.14, .43]) and produced results that were seen as less intuitive (*r* = -.422, *p* < .001, 95% CI [-.55, -.28]), while outcomes that used more easily imagined groups also produced less intuitive results (*r* = -.344, *p* < .001, 95%

[2] One study (Brown-Iannuzzi et al., 2017, Study 2) was identified as eligible after moderator data were collected, so moderator variables were collected in a separate sample (*N* = 52).

CI [-.48, -.19]). Given this shared variance among moderators, a more informative analysis is likely to be a simultaneous logistic regression that uses all moderators at the same time. Indeed, entering all predictors into the logistic regression found that only intuitiveness of responses remained a reliable predictor of whether conclusions matched (Desirability B = -.33, *SE* = .30, *p* = .273, *r* = -.09, 95% CI [-.24, .07]; Imagination B = .21, *SE* = .52 *p* = .682, *r* = .06, 95% CI [-.22, .32]; Intuitiveness B = 2.33, *SE* = .49, *p* < .001, *r* = .54, 95% CI [.35, .67], Nagelkerke R^2 = .32). More intuitive findings were more likely to show alignment in research conclusions between the two methods.

General Discussion

Though a version of the reverse correlation method is often used in hopes of assessing beliefs or perceptions that individuals are unable or unwilling to report explicitly, a majority of research conclusions from literatures on person perception or intergroup processes could be reproduced using self-report, a methodology that requires fewer resources and is easier to implement. The moderately high rate of agreement between the two methods stands in contrast to the fact that only 13% of original papers included a parallel self-report measure to the reverse correlation task. Many research findings concerning intergroup perceptions, beliefs, or behaviors could have been obtained more quickly, and perhaps with more precision, using self-report. At the same time, effect sizes from the two approaches were unrelated. Though interpreting this analysis is difficult given evidence that effect sizes in prior reverse correlation studies are inflated, it does suggest that while many effects can be reproduced using either method, factors determining what elicits a stronger or weaker effect are specific to each methodology.

Follow-up analyses shed some light on characteristics that may determine contexts where reverse correlation findings diverge from self-report. Indeed, consistent with past perspectives on reverse correlation, the two methods were more likely to produce differing conclusions for topics high in social desirability, suggesting that reverse correlation may be effective at tapping into beliefs that individuals may not want to self-report. However, the strongest predictor of alignment in research conclusions was whether the finding was considered intuitive. Though it is hard to know beforehand whether one's research question is intuitive, such results still indicate that many uses of reverse correlation may benefit from more carefully considering whether the associations being studied are ones people are aware of and willing to report.

These same analyses may also shed light on the lack of an association between effect sizes for the two methods. It is unknown what the expected correlation should be between the two sets of studies, given that the current work is not a series of direct replications but rather a substitution of an entirely different method. Indeed, as mentioned previously, the reverse correlation method could be a useful approach regardless of whether results showed a strong or weak relationship with self-report. Still, the lack of a strong association could be surprising to some readers, and may suggest that for many topics, the two methods can arrive at the same conclusion, but the specific factors that create a strong or weak effect are not shared across methods. For example, it's possible that social desirability does have a dampening impact on self-report findings in socially sensitive areas, but little or no effect in reverse correlation. Social desirability would then weaken effect sizes in self-report studies but not in reverse correlation, thereby lowering any association between the two methods. Other factors are likely to impact effect sizes for reverse correlation more than self-report; for instance, it is possible that factors like quality of the base image used, or similarity between Phase I and Phase II samples are important features of reverse correlation that have no comparison in self-report studies. While the present work highlights that many conclusions from reverse correlation can be recreated through self-report, identifying those factors responsible for the lack of correlation in effect sizes is clearly a worthy topic of future research.

Indeed, we believe that there is some similarity between the present work and past investigations into indirect methods that rely on inferring the strength of associations from behavioral responses, such as the Implicit Association Test (IAT; Greenwald et al., 1998). In the years that followed the IAT’s introduction to the psychological literature, there was no systematic investigation into how performance on the measure correlated with responses on parallel measures of self-report across a range of topics. However, a subsequent study was able to explore this question using many attitudinal domains (Nosek, 2005), with results finding considerable variability across attitude objects, as some topics produced low IAT-self report correlations (e.g., short people versus tall people) and others produced correlations that were fifteen times as strong (e.g., Democrats versus Republicans). These data were crucial in identifying what factors moderate the strength of IAT-self report correlations.

The present work is unable to make conclusions at the same scope as this earlier study, but the modest rate of agreement in conclusions between the form of reverse correlation used in the target studies and the present self-report data highlights the need for future research into identifying the contexts or outcomes where the two methods are more or less likely to align. We believe such studies will bring practical and theoretical advances for the reverse correlation method. In short, these data are a large-scale attempt to identify whether a problem exists – specifically, whether there is reason to believe that agreement in reverse correlation and self-report findings depends on characteristics like attitudinal domain or type of trait used as the outcome measure – and future research can begin identifying answers to this problem.

At the same time, it is important to note that these results should not be interpreted as a point estimate for the rate at which reverse correlation findings can be reproduced with self-report. For one, these studies had the narrow eligibility of dealing with perception of social groups, and results may not generalize to other uses of the method (e.g., Ratner et al., 2014). In addition, we made several methodological decisions that departed from the original studies (e.g., moving from between to within-subjects designs depending on study content, often using online instead of undergraduate samples), and it is difficult to tell the extent to which such decisions influenced results, though if anything it's most likely that such discrepancies only reduced the level of correspondence between the reverse correlation and self-report methods. However, while these factors do create further departures between the original studies and the self-report data collected here, it is noteworthy that prior large-scale replication projects have failed to find strong effects of factors like sample demographics producing substantive levels of heterogeneity across replication attempts (e.g., Klein et al., 2018).

In addition, this work should not minimize other possible uses of the reverse correlation method. In particular, one clear benefit of the methodology is its data-driven approach, meaning that researchers are less constrained by their own pre-existing conceptions about how the targeted social groups may differ. To that end, reverse correlation may be a particularly useful approach when studying social groups that are relatively novel or have been understudied in psychological research. Similarly, maximizing the potential benefits of the data-driven aspect of reverse correlation could come from asking Phase II participants to freely describe the classification images produced from Phase I, and this more open-ended method may produce new insights or directions that would not have been otherwise apparent to researchers. This point holds even when some findings produced through reverse correlation could be considered intuitive, as the data-driven approach could still be used to highlight other dimensions through which the targeted social groups vary.

Moreover, there are many forms of reverse correlation paradigms (Jack & Schyns, 2017), and these alternative approaches and analysis strategies may continue to allow for novel insights. One recent study (Martinez & Todorov, 2022) investigated how reverse correlation images produced by Phase I participants can be analyzed via clustering mental representations by image similarity, and these clusters can reveal intriguing results; for example, participants who believed there were greater numbers of illegalized immigrants living in their neighborhoods produced mental representations that were on average *less* threatening than participants who reported lower levels of illegalized immigrants. Reverse correlation remains a useful method for social psychological research, but researchers will benefit from a greater understanding of the outcomes and contexts where results differ from those found using self-report.

Notable Studies and Methodological Factors

While reviewing the data, one paper stood out for its departure from the overall pattern of results (Brown-Iannuzzi, McKee & Gervais, 2018). In the original paper, a sample of online Phase II participants judged the reverse-correlated image of a theist much more positively than that of an atheist (e.g., less trustworthy, more hostile), despite the sample including both theist and atheist participants. However, when translated into self-report (*N* = 231 in a between-subjects design) the findings were reversed; for 10 of 12 analyses, self-report results produced the opposite conclusion from the original paper; for example, theists were rated as *more* hostile and *less* trustworthy. These results were so striking that we ran a follow-up study using the same sample source (*N* = 193), now using a within-subjects design. Again, for 6 of 12 analyses, self-report results were reliably in the opposite direction from those found using reverse correlation in

Brown-Iannuzzi et al. (2018); for instance, in self-report, atheists were viewed as more trustworthy, competent, and likeable. In all, only one of the original analyses – religious people being viewed as more likely to be female– aligned with the results using reverse correlation. Removing Brown-Iannuzzi et al. (2018) from our primary analyses did not substantively alter our results or conclusions (see online supplement for analyses excluding the paper), but this work may be an effective case study for identifying a context where reverse correlation and self-report diverge.

In particular, one possible explanation is that people may view theists, or religious people generally, as having a pleasant outward appearance (which would be detected on reverse correlation) but a more sinister inner character (which would be detected on self-report measures assessing global perceptions). Some support for this perspective comes from a small follow-up study (*N* = 54), where 76% of participants agreed with the statement, "Religious people may look kind, helpful or trustworthy, but that often does not match their true personalities." In another follow-up (*N* = 122), 82% of participants disagreed with the statement "When it comes to religious people, what you see is what you get" (see online supplement for more details). Social groups that have a disconnect between perceived physical appearance and inner character may be fruitful areas for further understanding the strengths and weaknesses of each method. More generally, adding self-report items to reverse correlation studies will surely produce other domains where the two methods offer differing or opposing conclusions, and such cases will be fruitful for revealing when and why responses on the two methods are shaped by different psychological mechanisms.

Another paper may have complicated results given the Covid-19 pandemic. Michalak and Ackerman (2020) used reverse-correlation to study perceptions of "infected" and "germy" people, and participants' thoughts about such groups likely changed between the original study and when our data were collected in May 2021. In this paper, the rate of conclusion matches (48%) was comparable to that from the full sample, and results again did not substantively change when excluding this study from analyses, though removing both notable studies did create a positive -- albeit small -- correlation between the original reverse correlation effect sizes and the current self-report effect sizes (r = .205).

A final concern stems from the use of group-level or individual-level classification images in Phase II. As mentioned in the introduction, using group-level CI's has been shown to increase Type I errors and exaggerate effect sizes (Cone et al., 2020). Unfortunately, a large majority (91%) of the studies used group-level CI's, making any follow-up moderator analyses difficult, though the reliance on group-level CI's may be another factor in determining whether results from reverse correlation failed to match those found in self-report. The low number of studies with individual-level CI's (totaling six outcomes across three papers) also makes it uninformative to look at any possible correlation between these original, individual-level effect sizes and the effect sizes obtained when using self-report. In future work, we encourage researchers using reverse correlation to use the more unbiased individual-level approach when generating Phase I images in as well as adding parallel measures of self-report, and ideally with the same participants completing both phases of the reverse correlation method in addition to the self-report measures. Adopting these practices in future research will provide a more accurate estimate into the correlation between effect sizes produced from these two methods.

Conclusion

Reverse correlation is an increasingly popular method for studying intergroup processes. While the method has been used successfully to reveal differences in how others are represented, the same can be said for self-report. The study of intergroup and interpersonal phenomena -- and the use of reverse correlation -- will benefit from a deeper understanding of when the method will be particularly informative. The results reported here suggest that, for many domains, the perceptions in the back of minds are similar to those on the tips of our tongues.

References

Ahumada, A. J.Jr., & Lovell, J. (1971). Stimulus features in signal detection. *Journal of the Acoustical Society of America*, *49*, 1751-1756.

Ahumada, A. J. Jr. (1996). Perceptual classification images from Vernier acuity masked by noise. *Perception*, *25*, 2.

Axt, J. R. (2018). The best way to measure explicit racial attitudes is to ask about them. *Social Psychological and Personality Science*, *9*, 896-906.

Bar-Anan, Y., & Vianello, M. (2018). A multi-method multi-trait test of the dual-attitude perspective. *Journal of Experimental Psychology: General*, *147*, 1264-1272.

Brenner, P. S., & DeLamater, J. (2016). Lies, damned lies, and survey self-reports? Identity as a cause of measurement bias. *Social Psychology Quarterly*, *79*(4), 333-354.

Brinkman, L., Dotsch, R., Zondergeld, J., Koevoets, M. G., Aarts, H., & van Haren, N. E. (2019). Visualizing mental representations in schizophrenia patients: A reverse correlation approach. *Schizophrenia Research: Cognition*, *17*, 100138.

Brown-Iannuzzi, J. L., Dotsch, R., Cooley, E., & Payne, B. K. (2017). The relationship between mental representations of welfare recipients and attitudes toward welfare. *Psychological Science*, *28*, 92-103.

Brown-Iannuzzi, J. L., McKee, S., & Gervais, W. M. (2018). Atheist horns and religious halos: Mental representations of atheists and theists. *Journal of Experimental Psychology: General*, *147*, 292-297.

Brown-Iannuzzi, J., Payne, K., Cooley, E., & Cipolli, W. (2021). Who gets to vote? Racialized mental images of legitimate and illegitimate voters. *Social Psychological and Personality Science*. Advance online publication.

Cone, J., Brown-Iannuzzi, J. L., Lei, R., & Dotsch, R. (2021). Type I error is inflated in the two-phase reverse correlation procedure. *Social Psychological and Personality Science, 12,* 760-768.

Degner, J., Mangels, J., & Zander, L. (2019). Visualizing gendered representations of male and female teachers using a reverse correlation paradigm. *Social Psychology*, *50*, 233-251.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science’s aversion to the null. *Perspectives on Psychological Science, 7,* 555-561.

Forscher, P. S., Cox, W. T., Graetz, N., & Devine, P. G. (2015). The motivation to express prejudice. *Journal of Personality and Social Psychology*, *109*, 791-812.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research*, *41*, 2261-2271.

Gundersen, A. B., & Kunst, J. R. (2019). Feminist≠ feminine? Feminist women are visually masculinized whereas feminist men are feminized. *Sex Roles*, *80*, 291-309.

Hinzman, L., & Maddox, K. B. (2017). Conceptual and visual representations of racial categories: Distinguishing subtypes from subgroups. *Journal of Experimental Social Psychology*, *70*, 95-109.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin*, *30*(2), 161-172.

Imhoff, R., Dotsch, R., Bianchi, M., Banse, R., & Wigboldus, D. H. (2011). Facing Europe: Visualizing spontaneous in-group projection. *Psychological Science*, *22*, 1583-1590.

Imhoff, R., Woelki, J., Hanke, S., & Dotsch, R. (2013). Warmth and competence in your face! Visual encoding of stereotype content. *Frontiers in Psychology*, *4,* 386.

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*, 7241-7244.

Jack, R. E., & Schyns, P. G. (2017). Toward a social psychophysics of face communication. *Annual Review of Psychology*, *68*, 269-297.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Sowden, W. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443-490.

Kteily, N., Bruneau, E., Waytz, A., & Cotterill, S. (2015). The ascent of man: Theoretical and empirical evidence for blatant dehumanization. *Journal of Personality and Social Psychology*, *109*(5), 901-931.

Kunst, J. R., Kteily, N., & Thomsen, L. (2019). "You little creep": Evidence of blatant dehumanization of short groups. *Social Psychological and Personality Science*, *10*(2), 160-171.

Kunst, J. R., Myhren, L. S., & Onyeador, I. N. (2018). Simply insane? Attributing terrorism to mental illness (versus ideology) affects mental representations of race. *Criminal Justice and Behavior*, *45,* 1888-1902.

Krendl, A. C., & Freeman, J. B. (2019). Are mental illnesses stigmatized for the same reasons? Identifying the stigma-related beliefs underlying common mental illnesses. *Journal of Mental Health*, *28*, 267-275.

Lei, R. F., & Bodenhausen, G. V. (2017). Racial assumptions color the mental representation of social class. *Frontiers in Psychology*, *8*, 519-519.

Lloyd, E. P., Sim, M., Smalley, E., Bernstein, M. J., & Hugenberg, K. (2020). Good cop, bad cop: Race-based differences in mental representations of police. *Personality and Social Psychology Bulletin*, *46*, 1205-1218.

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science*, *28*, 209-226.

Martinez, J.E & Todorov, A. (2022). Mapping varied mental representations: The case of representing illegalized immigrants. Manuscript submitted for publication. Accessed at https://psyarxiv.com/gc9rq

Michalak, N. M., & Ackerman, J. M. (2021). A multimethod approach to measuring mental representations of threatening others. *Journal of Experimental Psychology: General*, *150*, 114-134.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231-259.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*, 565-584.

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality* (pp. 224–239). London: Guilford.

Petsko, C. D., Lei, R. F., Kunst, J. R., Bruneau, E., & Kteily, N. (2020). Blatant dehumanization in the mind's eye: Prevalent even among those who explicitly reject it? *Journal of Experimental Psychology: General*. Advance online publication.

Ratner, K. G., Dotsch, R., Wigboldus, D. H., van Knippenberg, A., & Amodio, D. M. (2014). Visualizing minimal ingroup and outgroup faces: Implications for impressions, attitudes, and behavior. *Journal of Personality and Social Psychology*, *106*, 897-911.

Tskhay, K. O., & Rule, N. O. (2015). Emotions facilitate the communication of ambiguous group memberships. *Emotion*, *15*, 812-825.

Van Rijsbergen, N., Jaworska, K., Rousselet, G. A., & Schyns, P. G. (2014). With age comes representational wisdom in social signals. *Current Biology*, *24*, 2792-2796.

Young, A. I., Ratner, K. G., & Fazio, R. H. (2014). Political attitudes bias the mental representation of a presidential candidate's face. *Psychological Science*, *25*, 503-510.

Zhan, J., Garrod, O. G., van Rijsbergen, N., & Schyns, P. G. (2019). Modelling face memory reveals task-generalizable representations. *Nature Human Behaviour, 3*, 817-826.