Rabagliati H, Corley M, Dering B, Hancock PJB, King J, Levitan CA, Loy J & Millen AE, Many Labs 5: Registered Replication Report of Crosby, Monin & Richardson (2008), *Advances in Methods and Practices in Psychological Science* (Forthcoming).

Many Labs 5: Registered Replication Report of Crosby, Monin & Richardson (2008)

Hugh Rabagliati, School of Philosophy, Psychology & Language Sciences, University of Edinburgh

Martin Corley, School of Philosophy, Psychology & Language Sciences, University of Edinburgh

Benjamin Dering, Faculty of Natural Sciences, University of Stirling

Peter J.B. Hancock, Faculty of Natural Sciences, University of Stirling

Josiah King, School of Philosophy, Psychology & Language Sciences, University of Edinburgh

Carmel A. Levitan, Cognitive Science, Occidental College

Jia Loy, School of Philosophy, Psychology & Language Sciences, University of Edinburgh

Ailsa E. Millen, Faculty of Natural Sciences, University of Stirling

Multilab direct replication of: Crosby, J.R., Monin, B., and Richardson, D. (2008).Where do we look during potentially offensive behavior? *Psychological Science,19*, 226-228

Data and registered protocols: https://osf.io/weus5/

Keywords: Eye tracking, offense, replication, many-labs, preregistration

Address correspondence to:

Hugh Rabagliati

School of Philosophy, Psychology & Language Sciences

University of Edinburgh

Edinburgh, UK

EH1 3NS

hugh.rabagliati@ed.ac.uk

Abstract

Crosby, Monin and Richardson (2008) found that hearing an offensive remark caused participants (n=25) to look at a potentially offended person, but only if that person could themselves hear the remark. They thus argued that the computation of offense involves the coordinated processing of high level linguistic and interpersonal cues. Their key effect, however, was not replicated by Jonas and Skorinko (2015) as part of the Reproducibility Project: Psychology (Open Science Collaboration, 2015). Three labs from Europe and America (n=283) tested whether the size of that effect might be increased when the stimuli were modified to be more appropriate for a diverse range of participants, using a peer-reviewed and pre-registered protocol. We found that this manipulation of protocol did not affect the size of the social referencing effect but, interestingly, we did replicate the original effect reported by Crosby and colleagues, albeit with a much smaller effect size. We discuss these results in the context of ongoing debates about how replication attempts should treat statistical power and contextual sensitivity.

Many Labs 5: Registered Replication Report of Crosby, Monin & Richardson (2008)

On hearing an offensive remark, we often find ourselves gazing directly towards the potentially offended person. Crosby, Monin, and Richardson (2008) suggested two possible accounts of this behavior. It could be an act of *social referencing*, in which we inspect the potentially aggrieved party's response to determine our own reaction (Crosby, 2006). Alternately, such social gaze could reflect low-level semantic associations (Huettig & Altmann, 2005; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), which cause participants to gaze at any visual stimulus that is related to the language they are hearing.

Crosby and colleagues tested this low-level association hypothesis using an eye tracking task based on a "Hollywood Squares" task in which participants (n=25) viewed a film of a video-conferenced conversation between three White men and one Black man. Two of the White men discussed University admissions policies, and one specifically critiqued admissions policies based on race (typically known as affirmative action, an American policy providing special consideration for underrepresented minority groups), in a manner that might be offensive to the Black discussant (who was silent during this remark). Between-subjects, Crosby and colleagues manipulated whether the setup of the video-conference allowed the Black discussant to hear this offensive remark. Under the association hypothesis, participants should involuntarily gaze to the Black discussant whether or not he heard the remark, but under the social referencing hypothesis, participants should be more likely to gaze to the Black discussant if they believed that he could hear the remark.

Consistent with the social referencing hypothesis, participants spent more time gazing at the Black discussant when they specifically believed that he could hear the offensive remark, compared to when they believed that the setup of the video-conference meant that he could not hear the offensive remark (inferred from a statistically significant interaction between gaze across the discussants and video conference setup). Importantly, this pattern of behavior could

not be explained by low-level differences between the two experimental conditions, because participants all saw an identical video of the offensive remark. Moreover, Crosby and colleagues also analyzed participants' gaze during a second, non-offensive comment (by a different White discussant), and found no interaction between gaze across the discussants and video conference setup.

A replication of Crosby and colleagues' study was one component of the Reproducibility Project: Psychology (RP:P; Open Science Collaboration, 2015; Jonas and Skorinko, 2015, see https://osf.io/b98zw/), conducted by one laboratory in the United States of America and one in the Netherlands, and using precisely the same materials as Crosby and colleagues. When the laboratories' data were analyzed together (n=58 participants), they did not find a statistically significant interaction between gaze across the discussants and video-conference setup, i.e., they failed to replicate Crosby and colleagues' critical finding. However, the American RP:P group also conducted a second replication (n=31 participants) using more situationally-appropriate videos, in which reference to Stanford University (where the original research was conducted) was removed. Here, the relevant interaction was statistically marginal (p=.07), and the participants' gaze patterned in the predicted direction.

Thus, one potential explanation for why this method has produced diverse results across the three studies is that processing of the offensive stimuli may importantly vary across cultural contexts. For example, in the replication studies, participants may have been confused and suspicious as to why they were watching a video about Stanford, which made them pay less attention to the offensive remark. Moreover, the Dutch participants in particular may not even have been knowledgeable about what affirmative action policies were. We thus developed a new replication protocol that aimed to mitigate cultural differences and ensure that participants were knowledgeable about the potentially offensive remark.

We did this in the most conservative way possible, replicating the Hollywood Squares task by using the same edited videos utilized in the RP:P (i.e., we used the videos that did not

make reference to Stanford, but we did not develop new stimuli, a point that we return to in the General Discussion). In our Revised protocol, we enhanced participants' knowledge of affirmative action by having them first watch a news report on the topic, prior to completing the Hollywood Squares discussion task (we refer to this revised protocol as the Informed condition). We then compared performance in this condition to performance in an Uninformed condition, in which participants completed the Hollywood Squares task after completing an unrelated cognitive filler task (a flanker task), which mimicked the RP:P protocol from Jonas and Skorinko (2015).

Disclosures

Preregistration

Prior to data collection, confirmatory analyses were preregistered on the Open Science Framework (https://osf.io/tj6qh/). Subsequently, we developed additional analyses (in response to subsequent peer review) that are as described in the current manuscript and that were also preregistered on the Open Science Framework (https://osf.io/wfrt7/).

Data, materials, and online resources

All materials, data, and code are available on the Open Science Framework (https://osf.io/weus5/).

Reporting

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Subjects

Data were collected in accordance with the Declaration of Helsinki, and were approved by the following review boards: University of Edinburgh PPLS Research Ethics Committee (no. 275-1617/2); University of Stirling General University Ethics Panel (GUEP79), and by the Occidental College Human Subjects Research Review Committee (Levi-F16095 and Levi-F17057).

Conflicts of Interest

The authors declare that they have no conflicts of interest with respect to the authorship or the publication of this article.

Author Contributions

H.R., M.C., J.K and J.L developed the study protocol. J.K. and J.L created the study materials. All authors collected data or supervised data collection. H.R. and J.K. wrote the analysis code, and H.R. analyzed the data. H.R. drafted the manuscript and all authors critically edited, and approved submission.

Acknowledgments

We would like to particularly acknowledge the contribution of Thomas Scherndl to study design and data analysis. We would also like to acknowledge the research assistance of Hanna Järvinen, Erin Ball, Talitha Brown, Gordon Carmichael, Maria Costa Pinto Teixeira Dias, Judith Glikberg, Isabel Geddes, Lauren Henderson, Kirsten Laursen, Eleni Lekka, Neelam Mistry, Shannon Morgan, Vanessa Nguyen, Julie Strong, Brenda Guerrero Tates, Ellen McDermott, Sally Zhou, Hannah Wagner.

Prior versions

A registered, results-blind version of this manuscript can be found at (https://osf.io/wfrt7/).

Method

Sample

We recruited participants at universities in the United Kingdom (University of Edinburgh and University of Stirling) and the United States of America (Occidental College). Each testing site aimed to collect 92 participants total, 46 in the Informed condition and 46 in the Uninformed condition. This sample size was necessary to achieve 95% power (calculated using G*Power, Faul, Erdfelder, Lang, & Buchner, 2007), assuming that the true effect size for this paradigm was somewhat smaller than in the original report by Crosby and colleagues (an f(u) of 0.37 where the original study had an f(u) of 0.47, or partial eta squared of 0.18). The original study by Crosby and colleagues involved 25 participants, and subsequent replications used approximately 30 participants per site.

Data were collected from 317 English-speaking students, whose demographics are reported in Table 1. 248 reported being native English speakers, 59 reported being fluent but non-native English speakers, and 11 did not report their language status. As in the original study, only non-Black individuals were included in the final sample, meaning that data from 13 participants who identified as Black were not analyzed further. At each site, some participants were compensated with payment and some were compensated with course credit.

Participants were subsequently excluded from the analysis of each comment if their eye tracking record during that comment had missing data on more than 40% of samples (so-called trackloss, e.g., due to failure in the eye tracking system, or caused by the participant gazing away from the monitor). After exclusions, 277 participants were included in the analysis of the offensive remark, and 277 participants were also included in the analysis of the non-offensive remark; the total number of unique participants included in the analysis was 283.

Table 1. Demographic details of participants tested across institutions.

	Edinburgh	Stirling	Occidental College
n tested	92 [76 female]	105 [67]	119 [93]
Mean Age	22 (SD = 3)	22 (4)	20 (1.2)
Black participants	3	0	10
Included for Offensive/Non-Offensive remark	88/88	100/99	88/89
Native English speakers	71	76	110
Participated for payment	92	75	56

Materials and Procedure

Participants completed four tasks in one of two orders, and a demographic survey. The entire experimental sequence was implemented using OpenSesame software (Mathôt, Schreij, & Theeuwes, 2012) and the bundled packaged software can be found at https://osf.io/w4h5x/. Table 2 provides a schematic illustration of our procedure. In our replication of the RP:P protocol, participants completed Crosby and colleagues' Hollywood Squares paradigm after first completing a Flanker task. This procedure aimed to replicate the paradigm used in the RP:P, in which the Hollywood Squares task was intermingled with cognitive distractor tasks that did not involve eye tracking. In addition, we assumed that the Flanker task would not provide participants with information about affirmative action. In our Revised protocol, participants completed the Hollywood Squares paradigm after freely viewing a series of videos, one of which was on the topic of legal challenges to affirmative action. We reasoned that this video would

provide participants with necessary context to help them easily see why the critical remark in the Hollywood Squares task was offensive.

Table 2: Order of tasks in Uninformed (RP:P) and Informed (Revised) protocols.

	Protocol
Uninformed (RP:P)	**Informed (Revised)**
Flanker task	Videos
Manual drift correction	Manual drift correction
Hollywood Squares	Hollywood Squares
Videos	Flanker task
Multiple-choice questions	Multiple-choice questions

In the Hollywood Squares task, participants were eye tracked as they watched one of two videos. These videos showed the four discussants (three White men and one Black man, in the bottom left corner) in a Hollywood Squares setup, such that each “talking head” took up one quarter of the screen. In the Headphones Off condition, the video began when a female voiceover announced that she was turning off the headphones of two of the discussants, and asked the participants to raise their hands if they could hear her. The two white discussants in the top row then raised their hands, and each was asked to discuss the topic of University admissions; first, the White discussant in the top left made a non-offensive remark (lasting 19s), and then the White discussant in the top right made a potentially offensive remark (lasting 20.5s). In the Headphones On condition, the female voiceover announced that all four participants would discuss the topic (and all four raised their hands to indicate that they could hear). From this point, the video was identical to the Headphones Off condition. Similarly to the

second replication in the RP:P, we edited the original video files in order to cut references to Stanford[1]. The entire task took approximately 1 minute and 45 seconds, and was always preceded by a drift correction (calibration took place prior to the first video which was presented to participants). The critical offensive remark ran as follows:

- *I think one problem with admissions is that too many qualified White students are not getting the spots they've earned. These students work hard all through school and then lose their spots to members of certain groups who have lower test scores and come from less challenging environments. They get an unfair advantage.*

In the Video task, participants viewed four video news reports, the last of which (from the American television channel New York 1) described a legal challenge to an affirmative action policy and the ensuing controversies. The three previous videos were on the topics of Neil Armstrong, bottled water, and clowns, in a random order. All videos can be found at https://osf.io/weus5/. The four videos in total lasted 6 minutes and 30 seconds.

In the flanker task, participants indicated the direction of an arrow that was surrounded by either direction-matched or non-matched arrows. Participants completed 360 trials at their own pace; informal piloting suggested that this matched the viewing time for the four news videos.

Before the Hollywood Squares discussion was presented, participants were informed that they would be subsequently asked a set of questions on the discussion which they were about to watch. Eleven multiple choice questions were presented at the end of the protocol, ten

[1] Our edits were slightly different from the second RP:P replication, in which the Headphones On video had reference to Stanford cut out, while the Headphones Off video had reference to Stanford muted.

assessed participants' attention to the videos used, and the final question asked participants to rate how aware they were of the politics of affirmative action prior to taking part in the study.

Finally, participants completed a paper and pencil survey of their demographic background, including questions about their ethnic and racial background. Each lab constructed their own questionnaire, using the question formats provided by their national census.

Eye tracking calibration always took place before participants viewed the first video (i.e., before the Hollywood Squares discussion for Uninformed participants, and before the first informational video for Informed participants). Participating labs used a variety of eye tracking systems. Data from University of Edinburgh were collected on an EyeLink 1000 sampling at 500Hz (50 participants) and an EyeLink 2000 at 1000Hz (43 participants), data from University of Stirling and Occidental College were collected on an EyeTribe sampling at 30Hz..

Analysis

Our analysis focused on participants' gaze during the Hollywood Squares discussion. We processed this data by creating four equally-sized areas of interest (AOIs), each corresponding to one of the discussants in the video. Then, for both the offensive and non-offensive remarks, we calculated the total time that each participant spent gazing within each of the four AOIs during each of the two remarks in the video.[2] The offensive remark lasted 20.5s total, and the non-offensive remark lasted 19s total. See our pre-registration at https://osf.io/tj6qh/ for full details on how data were processed.

Results

Confirmatory regression analyses were pre-registered at https://osf.io/tj6qh/. All data and analysis scripts can be found at https://osf.io/weus5/. Analyses were conducted using the R

[2] Note that a mistake in our replication protocol suggested that the screen resolution would be fixed to 1280 by 1024, but it was in fact fixed to 1024 * 768.

language (version 3.5.3, R Core Team, 2018) and the package lme4 (version 1.1-19, Bates, Maechler, Bolker, Walker, 2015); we fit the mixed-effects models using full maximum likelihood. Pseudo-R^2 statistics were calculated using the package *MuMIn* (version 1.42.1, Barton, 2018) and *p*-values were calculated using the package *lmerTest* (version 3.1-0, Kuznetsova, Brockhoff & Christensen, 2017). Predictor variables were dummy coded. For the AOI variable, the reference level was set as the Speaker of the remark; for the HeadphoneCondition variable, the reference level was set as Headphones off; for the Protocol variable, the reference level was the RP:P (Uninformed) protocol.

For the offensive remark, Figure 1a shows total gaze time to the four AOIs across both the Headphone conditions and Protocol manipulations, while Figure 1b shows gaze time to the two critical AOIs (the speaker and the minority discussant) for each of the three sites that provided data. Figure 2 shows the analogous data for the non-offensive remark.

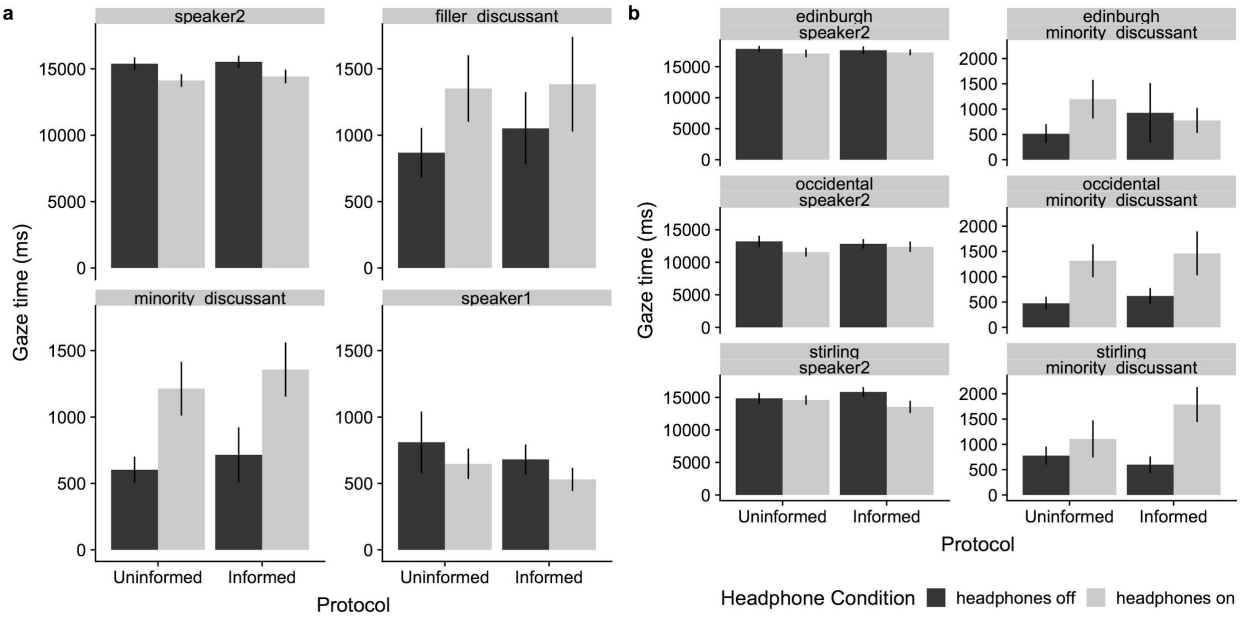

Figure 1. A. Mean gaze time to the four Areas of Interest (in ms) during the offensive remark, for each of the four conditions of the study. **B**. Mean gaze time to the speaker and minority

discussant during the offensive remark for each testing site. Error bars show +/- 1 standard error of the mean.

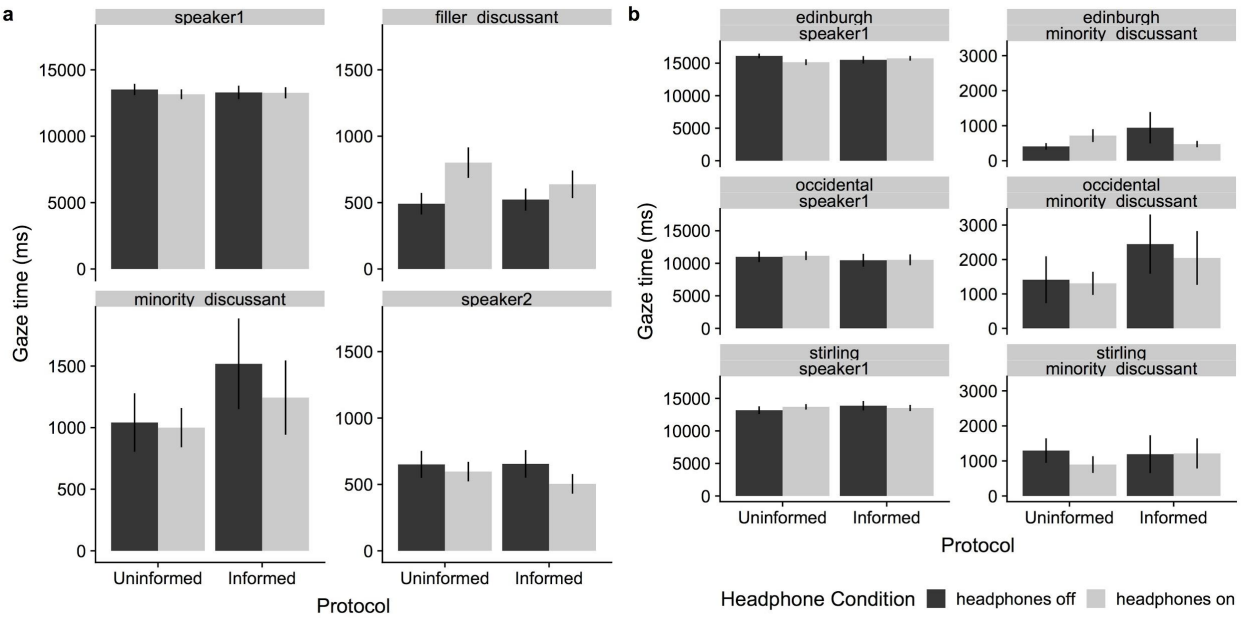

Figure 2. A. Mean gaze time to the four Areas of Interest (in ms) during the non-offensive remark, for each of the four conditions of the study. **B**. Mean gaze time to the speaker and minority discussant during the non-offensive remark for each testing site. Error bars show +/- 1 standard error of the mean.

Confirmatory analysis 1 - Test of the association hypothesis

Our first planned analysis assessed whether these data confirmed Crosby and colleagues' finding that participants gazed longer to a potentially offended individual, if they believed that that individual had heard an offensive remark. We tested for a two-way interaction between AOI and Headphone condition during the offensive remark using the mixed model below, with statistical significance determined by model comparison. Our preregistered analysis had the form

*Total Looking Time ~ AOI * HeadphoneCondition + (1+AOI|Subject:TestingSite) + (1 + HeadphoneCondition | TestingSite)*

But we subsequently realised that incorporating a by-subject random slope of AOI was erroneous, because it resulted in the same number of regression parameters as datapoints (each subject provided one datapoint per AOI). Our final model thus had the form *Total Looking Time ~ AOI * HeadphoneCondition + (1|Subject:TestingSite) + (1 + HeadphoneCondition | TestingSite)*

Model comparison indicated that there was a significant interaction between AOI and Headphone condition (*X2* (3) = 22.11, $p < .001$, *pseudo R2* = .85). There was enhanced gaze to the Black discussant when he could hear the offensive remark ($M_{HeadphonesOn}$ = 1281ms, SD = 1711, $M_{HeadphonesOff}$ = 660ms, SD = 1352, *Beta* = 1814(*SE* = 420), t=4.3, p<.001).

Confirmatory analysis 2 - Test of protocol manipulation

Our second analysis tested whether the size of the interaction between AOI and Headphone condition was increased when participants were in the Informed (versus Uninformed) protocol. Our pre-registered regression had the form:

*Total Looking Time ~ Protocol * AOI * HeadphoneCondition + (1+AOI|Subject:TestingSite) + (1 + HeadphoneCondition | TestingSite)*

But in our analysis we again removed the by-subject random effect of AOI, for the reasons above.

There was not a significant interaction between AOI, Headphone condition, and protocol (*X2* (3) = 0.13, *p* = .99, *pseudo R2* = .85). The protocol did not significantly influence the degree to which participants gazed more at the black discussant when he could hear the potentially offensive remark (*Beta* = -123(*SE* = 841) , *t = -0.19, p* = .88).

Confirmatory analysis 3 - The non-offensive remark

To confirm that these findings were driven by the offensive remark, we repeated analyses 1 and 2 during the non-offensive remark. In Crosby and colleagues' original study, the Headphone condition did not affect gaze across the AOIs during this remark. We, also, found that there was not a significant interaction between AOI and Headphone condition (*X2* (3) = 1.55, *p* = .67, *pseudo R2* = .84), which was not accompanied by enhanced gaze to the Black discussant when he could hear the remark (*Beta* = 26(*SE* = 368), *t* = 0.07*, p* = .94). There also was not a significant interaction between AOI, Headphone condition and protocol (*X2* (3) = 0.73, *p* = .87, *pseudo R2* = .86), and gaze to the Black discussant in particular was not significantly affected by this combination of factors (*Beta* = -560(*SE* = 736), *t* = -0.76*, p* = .45).

Confirmatory analysis 4 - Awareness of Affirmative Action

To test whether participants who reported being more aware of affirmative action might be more sensitive to this manipulation, we used a regression of the form:

*Total Looking Time ~ AOI * HeadphoneCondition * Protocol * Awareness +*

(1+AOI|Subject:TestingSite) + (1 + HeadphoneCondition | TestingSite)

in which Awareness was mean centered and standardized. As above, we simplified this model by removing the by-subject random slope for AOI.

The four-way interaction between AOI, Headphone Condition,Protocol and Awareness revealed that awareness of affirmative action did not moderate the effect of protocol on gaze in this task (*X2* (3) =4.3, *p* = .23, *pseudo R2* = .86).

We also carried out an exploratory analysis assessing whether awareness of affirmative action moderated the significant interaction of AOI and Headphone Condition from Confirmatory Analysis 1, using a regression of the form:
*Total Looking Time ~ AOI * HeadphoneCondition * Awareness + (1+AOI|Subject:TestingSite) + (1 + HeadphoneCondition | TestingSite)*

The three-way interaction between AOI, Headphone Condition,Protocol and Awareness revealed that awareness of affirmative action did not significantly moderate the effect of protocol on gaze in this task (*X2* (3) =7.7, *p* = .051, *pseudo R2* = .86), although the *p* value was at a level often described as marginal.

Confirmatory analysis 5 - Differences between European and American samples

The stimuli used in this study are likely to have been more culturally appropriate for participants assessed in America than participants assessed in Europe, and so we compared the size of the critical effect between these groups. This was not part of the original pre-registered analysis plan, but was developed in response to reviews of this paper prior to the data being analyzed, and is thus confirmatory rather than exploratory.

We preregistered a regression of the following form:

*Total Looking Time ~ AOI * HeadphoneCondition * Protocol * Continent + (1+AOI|Subject:TestingSite) + (1 + HeadphoneCondition | TestingSite)*

But in our analysis we removed the by-subject random effect of AOI, for the reasons discussed in the previous section.

The four-way interaction between AOI, Headphone Condition, Protocol and Continent revealed that the European/American distinction did not significantly moderate behavior in this task ($X2$ (3) = 3.4, p = .34, *pseudo R2* = .87).

A subsequent exploratory analysis examined whether the original interaction of AOI and Headphone varied across the Continents, regardless of Protocol. This was tested using the regression

*Total Looking Time ~ AOI * HeadphoneCondition * Continent + (1|Subject:TestingSite) + (1 + HeadphoneCondition | TestingSite)*

The three-way interaction between AOI, Headphone Condition and Continent revealed that the continent of testing did not significantly influence gaze behavior in this task ($X2$ (3) = 0.5, p = .92, *pseudo R2* = .86).

Confirmatory analysis 6 - Logistic analysis

We conducted an additional analysis to account for the fact that, in some ways, eye tracking data is not suitable for analysis using a linear mixed effects model as above. In particular, models including the effects of AOI on Total Looking Time violate the independence assumption, since an increase in the time spent looking at one AOI necessitates decreases in the times spent looking at others. In this analysis, the dependent variable was the logit transformed proportion of time spent looking at the Black discussant over all discussants. Our regression had the form:

*Logit proportion ~ HeadphoneCondition * Remark * ProtocolManipulation + (1+Remark|Subject : TestingSite) + (1+HeadphoneCondition * ProtocolManipulation|TestingSite)*

but, following our earlier logic, we removed the by-subject random effect of remark because each participant only provided one datapoint per remark.

This regression did not reveal a significant interaction between Headphone condition, Remark and Protocol ($X2$ (1) = 1.6, p = .20, *pseudo R2* = .03).

We also conducted two exploratory follow-up regressions, modeled on Confirmatory Analyses 1 and 2. The first regression assessed the effect of Headphone Condition during the offensive remark, regardless of protocol. After dropping random slopes for testing site due to convergence issues, this had the form

Logit proportion ~ HeadphoneCondition + (1|TestingSite)

and revealed a significant effect of Headphone Condition, replicating Confirmatory Analysis 1 ($X2$ (1) = 10.25, p = .001, *pseudo R2* = .04).

The second regression assessed whether the Protocol manipulation interacted with Headphone Condition during the offensive remark alone. After simplification for non-convergence, this had the form:

*Logit proportion ~ HeadphoneCondition * ProtocolManipulation + (1+ HeadphoneCondition + ProtocolManipulation|TestingSite)*

There was not a significant interaction between Headphone Condition and Protocol Manipulation, replicating Confirmatory Analysis 2 ($X2$ (1) = 1.44, p = .23, *pseudo R2* = .04).

Discussion

Crosby, Monin and Richardson (2008) provided eye tracking evidence that social referencing occurs during offensive behavior, a result that failed to replicate in the original Reproducibility Project: Psychology. Here, we used a peer reviewed experimental manipulation of protocol to test whether Crosby and colleagues' paradigm might provide stronger evidence for social referencing when participants in the current sample were made more similar to the original sample, by explicitly providing them with some relevant background knowledge.

Interestingly, our overall results were consistent with Crosby and colleagues' original report. We found that, on hearing an offensive comment, participants were significantly more likely to gaze toward a potentially-offended person if they, too, could hear the offensive comment. However, we did not find that participants' behavior in the task was affected by the protocol to which they had been assigned: The effect of social referencing did not significantly differ between participants who were assigned to our new protocol versus assigned to the original protocol that Jonas and Skorinko (2015) used in the Replication Project: Psychology (2015). This raises the question of why the original Crosby et al. finding did not replicate in the Jonas and Skorinko

study, but did replicate here, given that our protocol manipulation did not appear to affect participants' behavior. One answer, we propose, lies in the observation that the key effect size found in our replication was considerably smaller than that found in Crosby and colleagues' original study.

To compare effect sizes, we calculated Cohen's *d* statistics for how gaze to the Black discussant during the offensive remark was affected by the Headphones On/Off manipulation. In the original study, participants gazed at the Black discussant for an average of 2588ms (SD=2085) in the Headphones On condition, and for 505ms (491) in the Headphones Off condition, resulting in a Cohen's *d* of 1.38. But in this replication, the relative figures were 1281ms (1711) in the Headphones On condition, and 660ms (1352) in the Headphones Off condition, resulting in a Cohen's *d* of 0.4, i.e., an effect that was a little more than one quarter the size of the original.

This discrepancy in effect sizes can be explained in one of two ways. The first possibility is that the original effect reported by Crosby and colleagues may be a so-called Type-M error (Gelman and Carlin, 2014), i.e., a report that correctly describes the existence of an effect, but that importantly mis-estimates its magnitude. Such mis-estimates are known to be more likely when studies have low statistical power, as was plausibly the case for Crosby and colleagues' original report, which used a between-subjects design (with a total of 25 participants across the two conditions), took only a single observation from each participant (i.e., gaze behavior during the offensive remark), and used a dependent measure that, one might expect, would be relatively noisy (time spent gazing across the participants, which could be affected by nuisance factors such as boredom, tiredness, etc). This Type-M error explanation can easily account for why Jonas and Skorinko (2015) failed to replicate the phenomenon: Their study would have been under-powered to detect the true underlying effect in this paradigm. In particular, if the original

report's effect size of 1.38 were correct, then 95% power could have been achieved with only 15 participants per group (according to an analysis conducted using G-Power, Faul et al.., 2007). But if the present estimate of 0.4 is closer to the truth, then the required sample size would be 164 participants per group, such that even the present study, with its final sample of 277, was likely underpowered.

The second possible explanation for the discrepancy in effect size comes from the notion of contextual sensitivity (Van Bavel, Mende-Siedlecki, Brady, and Reinero, 2016). In particular, behavior in a paradigm like this could importantly vary based on factors in a participant's background, in this case, their knowledge of affirmative action. For instance, participants in this replication may have been less familiar with the debate about affirmative action than participants from the original study, and thus fewer of our participants may have noticed that the potentially-offensive remark was, in fact, potentially offensive. The present stimuli were created by Crosby and colleagues to assess Stanford undergraduates in the mid-2000s, and so it is quite plausible that the stimuli would be better-matched to that cohort than to the present population, and so would elicit smaller effect sizes from our sample (for a related discussion, see Shafir, 2018).

Both of these accounts can potentially explain why subsequent work has observed a smaller effect size than Crosby and colleagues' study, and indeed it is quite possible that the observed discrepancy in effect sizes can be explained through a combination of statistical power and contextual sensitivity. However, we believe that there are good reasons for suspecting that statistical issues, rather than questions of context, shoulder more of the explanatory burden. First, as noted above, Crosby and colleagues' original study had a number of features that suggest low statistical power, such as small numbers of participants in a between-subjects experimental design. Under these conditions, tests of statistical significance will only find

positive results when either the tested effect is extremely large, or when it has been overestimated (i.e., the analysis results in a Type-M error). We suggest that the latter possibility is more likely, because the former possibility is implausible: If Crosby and colleagues' original effect size of 1.38 were correct, it would imply that the size of their measured social referencing effect was much greater than that of obvious and mundane effects that barely require statistical confirmation, e.g., that people who are more liberal tend to think that social equality is more important, that people who like eggs tend to eat more egg salad, and that men tend to weigh more than women (Simmons, Nelson & Simonsohn, 2013). Thus, if the originally reported effect size is implausibly high, then by implication it is likely to be a Type-M error.

Beyond this, the present results actually provide surprisingly little evidence to suggest that behavior in this task is sensitive to context and background. For example, we found no significant evidence that undergraduate participants behaved differently whether they were tested in Scotland or California (Confirmatory Analysis 5), even though one might expect the Californian participants to be much more similar to the original Stanford sample, and so to show a larger effect.[3] We also found no statistically significant evidence that participants behaved differently based on whether they were informed or uninformed about the social issues surrounding affirmative action, whether because of our experimental manipulation (Confirmatory Analysis 2) or because of their own background (see the exploratory analysis reported alongside Confirmatory Analysis 4, although note that the analysis produced a marginally significant *p* value). Both of these null results are unexpected under the view that the present

[3] It is possible that today's Californian participants are more similar to today's Scottish participants, than they are to the Californian participants tested in Crosby et al.. (2007); e.g., they may be less aware of controversies around affirmative action, and thus less likely to notice the offensive remark. That said, the present cohort were tested during an era in which, anecdotally, issues of social justice are prominent in the media (e.g., due to the efforts of groups such as Black Lives Matter), such that we find it unlikely that they would not perceive the offence.

task is highly contextually sensitive, such that it could generate an effect size of 1.38 in the original study, but only 0.4 in the present study.

Still, these null findings are not conclusive, and it remains possible that the present study might have produced effect size estimates that were as large as the original if we had used a somewhat different manipulation of experimental protocol. For example, rather than varying whether participants were deliberately informed about affirmative action, we could have varied whether participants viewed the original stimuli versus newly-created stimuli that were perhaps better-matched to the participants' background, and thus might have elicited stronger effects. This would be an intriguing direction for future work but, as we discovered when designing the present replication, such an approach does also lead to a number of difficulties with regard to experimental design and standardization, as well as comparability with prior work. For example, creating novel stimuli for this study would have required us to develop individualized videos for each testing site, that varied in terms of scripts, actors, accents, and languages, but were nevertheless still standardized in terms of offensiveness and believability. This would likely have proved a barrier to entry for other researchers aiming to join a collaborative project of this type. Moreover, in creating new stimuli, we would have needed to match them, in terms of offensiveness and believability, to the videos created for the original Crosby and colleagues' study; however, this would be impossible to do in practical terms because, to the best of our knowledge, the offensiveness and believability of the original stimuli were not normed for the original population. Thus, it is possible that the original stimuli were better-matched to the original population than to the populations tested here, but reconstructing that match is impossible, because we have no contemporaneous evidence as to the quality of that match.

Why did our manipulation of background knowledge not affect participants' behavior in this task? Given the discussion above, one strong possibility is that social context and background

knowledge simply have very small effects on behavior in this paradigm, which would be consistent with the small overall effect of social referencing that we uncovered -- if social referencing in this paradigm is itself a small effect, then we would not expect context and background to cause large fluctuations in this behavior. That said, it is also possible that our manipulation of background knowledge was simply not that effective, in that it may not have fully informed participants about the cultural context of affirmative action. Since we did not carry out a manipulation check, we have no way to confirm or deny this, which is an obvious flaw in the present methodology. Finally, as suggested by an anonymous reviewer, it is possible that both our treatment and control conditions acted to enhance the effect of social referencing in this paradigm; e.g., completing the Flanker task in the Uninformed condition may have enhanced participants' attentional control and magnified their ocular responses to the offensive remark. This remains possible, but we think it unlikely: Inspection of Figures 1 and 2 suggest no baseline differences in gaze behavior across the Informed and Uninformed conditions, and that learning to inhibit one specific distractor transfers to inhibition of other distractors (Kelley & Yantis, 2009).

Thus, we believe that, on the balance of possibilities, it is more likely that the original report's large effect size was a Type-M error, than that subsequent small effects are the result of failure to account for contextual sensitivity. But both possibilities remain viable hypotheses, and so we think it is helpful to consider what lessons can be learned from them for future work. Most obviously, statistical considerations suggest that future studies could improve on this paradigm by seeking ways to increase statistical power and measurement precision beyond testing additional participants, such as through having each participant provide more than one observation, or by using a within-subjects design. Considerations of contextual sensitivity, by contrast, suggest that future work will need to focus more seriously on measuring and quantifying the contextual fit between stimuli and participants. For example, we would be able to

draw stronger conclusions from our data if we could conduct an independent assessment of whether participants from the present populations, and participants from the original population, perceived the offensive remarks in similar ways. Future work on this question, therefore, might want to include either norming of the stimuli (e.g., explicit ratings of offensiveness) and/or a manipulation check, to measure the degree to which the stimuli induced the intended effect in participants.

Finally, in the spirit of lessons learned, we note one additional factor that may have supported the present study's replication of the key result from Crosby, Monin and Richardson (2008), which was that recent advances in open source software allowed us to easily standardize a complicated eye tracking experiment across labs. In particular, we created this study using the open source experiment builder Open Sesame (Mathôt, Schreij, & Theeuwes, 2012) and were then able to bundle it as an executable file so that all participating labs used precisely the same testing parameters. This meant that data quality could be better standardized, thus minimising the potential for differences in methodology to affect measurement error across testing locations.

Conclusion

The current project sought to replicate the finding of Crosby, Monin and Richardson (2008), as well as test for moderation of this effect by varying the protocol to make the current sample's background knowledge more similar to that of the original sample. We did not find evidence for moderation by protocol: Participants' gaze behavior did not significantly vary based on manipulation of their background knowledge. However, in contrast to the Replication Project: Psychology, we did replicate Crosby and colleagues' original finding about social referencing,

although the effect size measured in this project was considerably smaller than that reported by the original study.

References

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Bartoń, K (2018). MuMIn: Multi-Model Inference. R package version 1.42.1. https://CRAN.R-project.org/package=MuMIn

Crosby, J.R. (2006). Targeted social referencing and the perception of discrimination. *Dissertation Abstracts International: B. The Sciences and Engineering, 67*, 2874.

Crosby, J. R., Monin, B., & Richardson, D. (2008). Where do we look during potentially offensive behavior? *Psychological Science, 19(3),* 226-228. doi: 10.1111/j.1467-9280.2008.02072.x

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191. doi: 10.3758/BF03193146

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. Perspectives on Psychological Science, 9(6), 641-651. doi: 10.1177/1745691614551642

Huettig, F., & Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition, 96(1)*, B23-B32. doi: 10.1016/j.cognition.2004.10.003

Jonas, K. & Skorinko, J. (2015). Replication Report of Crosby, Monin & Richardson (2008, Psych. Science). Unpublished manuscript, accessed from: https://osf.io/b98zw/.

Kelley, T. A., & Yantis, S. (2009). Learning to attend: Effects of practice on information selection. Journal of Vision, 9(7), 16-16.

Kuznetsova A, Brockhoff PB, Christensen RHB (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1-26. doi: 10.18637/jss.v082.i13 (URL: http://doi.org/10.18637/jss.v082.i13).

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44(2)*, 314-324. doi:10.3758/s13428-011-0168-7

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349(6251),* aac4716. doi: 10.1126/science.aac4716

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: Looking at things that aren't there anymore. *Cognition, 76(3),* 269-295. doi: doi.org/10.1016/S0010-0277(00)00084-6

Shafir, E. (2018). The Workings of Choosing and Rejecting: Commentary on Many Labs 2. Advances in Methods and Practices in Psychological Science, 1(4), 495-496.

Simmons, J. P., and Nelson, L. D., and Simonsohn, U. (2013). Life after P-Hacking. Talk presented at the Meeting of the Society for Personality and Social Psychology, New Orleans, LA, 17-19 January 2013. Available at SSRN: https://ssrn.com/abstract=2205186 or http://dx.doi.org/10.2139/ssrn.2205186

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., & Sedivy, J.C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268,* 1632–1634. doi: 10.1126/science.7777863

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences,* 113(23), 6454-6459. doi: 10.1073/pnas.1521897113

Appendix

Transcripts of Videos: Edited

Headphones Off (edited)

Woman's voice: For this part of the discussion, only participants one and two will discuss a topic. Participants three and four, I'm turning off your microphones and headphones now. Okay. Can you raise your hand if you can hear me?

The two (White) participants on the top half of the screen raise their hands.

Woman's voice: Great. For this first part, each of you has been asked to think about several possible questions. I will choose one of the questions, each of you will give your initial response, then the two of you will have time to discuss it between you. Do either of you have any questions?

Participant one (White): So it's just two of us now?

Woman's voice: Yes, *[participant one gives a small nod]* the other participants will be brought into the conversation later. Can you two hear each other?

Participant one: Yes.

Participant two: Er, yes.

**Woman's voice:* The first question is, "What changes or improvements would you make?" Participant one, your response?

Participant one: I think we should consider having admission interviews. I know there are some downsides, but I think some students might benefit from being able to present themselves in person rather than just on paper. They would also have a chance to learn more, and get some of their initial questions answered.

Woman's voice: Okay, participant two.

Participant two: I think one problem with admissions is that too many qualified White students are not getting the spots they've earned. These students work hard all through school and then lose their spots to members of certain groups who have lower test scores and come from less challenging environments. They get an unfair advantage.

Clip ends.

Headphones On (edited)

Woman's voice: For this part of the discussion, all four of you will discuss a topic. Each one of you should be able to hear me and hear each other through your headphones now. Okay. Can you raise your hand if you can hear me?

All four participants raise their hands.

Woman's voice: Great. For this first part, each of you has been asked to think about several possible questions. I will choose one of the questions, each of you will give your initial response, then the four of you will have time to discuss it among you. Do any of you have any questions?

Participant one (White): So it's all four of us now?

Woman's voice: Yes, all four of you will be participating in this part of the conversation. Can you all hear each other?

All four participants: Yes.

**Woman's voice:* The first question is "What changes or improvements would you make?" Participant one, your response.

Participant one: I think we should consider having admission interviews. I know there are some downsides, but I think some students might benefit from being able to present themselves in person rather than just on paper. They would also have a chance to learn more, and get some of their initial questions answered.

Woman's voice: Okay, participant two.

Participant two: I think one problem with admissions is that too many qualified White students are not getting the spots they've earned. These students work hard all through school and then lose their spots to members of certain groups who have lower test scores and come from less challenging environments. They get an unfair advantage.

Clip Ends.