Empirical Article: Registered Report

Advances in Methods and Practices in Psychological Science
1–13

DOI: 10.1177/2515245919870737
www.psychologicalscience.org/AMPPS

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

Hugh Rabagliati[1], Martin Corley[1], Benjamin Dering[2], Peter J. B. Hancock[2], Josiah P. J. King[1], Carmel A. Levitan[3], Jia E. Loy[4], and Ailsa E. Millen[2]

[1]Psychology, School of Philosophy, Psychology & Language Sciences, University of Edinburgh; [2]Psychology, University of Stirling; [3]Department of Cognitive Science, Occidental College; and [4]Linguistics & English Language, School of Philosophy, Psychology & Language Sciences, University of Edinburgh

Abstract

Crosby, Monin, and Richardson (2008) found that hearing an offensive remark caused subjects ($N = 25$) to look longer at a potentially offended person, but only if that person could hear the remark. On the basis of this result, they argued that people use social referencing to assess the offensiveness. However, in a direct replication in the Reproducibility Project: Psychology, the result for Crosby et al.'s key effect was not significant. In the current project, we tested whether the size of the social-referencing effect might be increased by a peer-reviewed and preregistered protocol manipulation in which some participants were given context to understand why the remark was potentially offensive. Three labs in Europe and the United States ($N = 283$) took part. The protocol manipulation did not affect the size of the social-referencing effect. However, we did replicate the original effect reported by Crosby et al., 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.

Keywords

eye tracking, offense, replication, Many Labs, open data, open materials, preregistered

Received 5/18/18; Revision accepted 7/15/19

On hearing an offensive remark, people often gaze directly toward 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 observers inspect the potentially aggrieved party's response so as to determine their own reaction (Crosby, 2006). Alternatively, such social gaze could reflect low-level semantic associations (Huettig & Altmann, 2005; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), which cause subjects to gaze at any visual stimulus that is related to the language they are hearing.

Crosby and her colleagues tested this latter hypothesis by tracking subjects' ($N = 25$) gaze during a "Hollywood Squares" task in which they viewed a film of three White men and one Black man participating in a video-conference conversation. 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). In a between-subjects manipulation, Crosby et al. varied whether the setup of the video conference allowed the Black discussant to hear this offensive remark. According to the association hypothesis, subjects should have involuntarily gazed toward the Black discussant whether or not he heard the remark, but according to the social-referencing hypothesis, subjects should have been more likely to gaze toward the Black discussant if they believed that he could hear the remark than if they believed he could not.

Corresponding Author:
Hugh Rabagliati, School of Philosophy, Psychology & Language Sciences, University of Edinburgh, Edinburgh, United Kingdom EH1 3NS
E-mail: hugh.rabagliati@ed.ac.uk

Results were consistent with the social-referencing hypothesis; subjects spent more time gazing at the Black discussant when they specifically believed that he could hear the offensive remark, compared with when they believed that the setup of the video conference meant that he could not hear the offensive remark (i.e., the interaction between duration of gaze across the discussants and video-conference setup was statistically significant). This pattern of behavior could not be explained by low-level differences between the two experimental conditions because all the subjects saw the same video of the offensive remark. Moreover, Crosby et al. also analyzed subjects' gaze during a second, nonoffensive remark (by a different White discussant) and found no interaction between the duration of gaze across the discussants and the video conference's setup.

As part of the Reproducibility Project: Psychology (RP:P; Open Science Collaboration, 2015), Jonas and Skorinko (2015) conducted a replication of this study at one laboratory in the United States of America and another in The Netherlands. Both teams used precisely the same materials as in the original experiment. When the laboratories' data were analyzed together ($N = 58$ subjects), the teams did not find a statistically significant interaction between duration of gaze across the discussants and video-conference setup; that is, they failed to replicate Crosby et al.'s critical finding. However, the American RP:P team also conducted a second replication ($N = 31$ subjects) using more situationally appropriate videos, in which reference to Stanford University (where the original research was conducted) was removed. In this replication, the relevant interaction was statistically marginal ($p = .07$), and the subjects' gaze duration followed the predicted pattern.

Thus, one potential explanation for why this method produced diverse results across the three studies is that processing of the offensive stimuli may vary importantly across cultural contexts, which would be consistent with the claim that many social-psychological phenomena are contextually sensitive (Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016). For example, in the first replication study, subjects may have been confused and suspicious as to why they were watching a video about Stanford, and that may have made them pay less attention to the offensive remark. Moreover, the Dutch subjects in particular may not even have been knowledgeable about what affirmative-action policies were. These factors would reduce the size of the social-referencing effect. We therefore developed a new replication protocol with the aim of mitigating cultural differences, ensuring that subjects were knowledgeable about the potentially offensive remark, and, thus, potentially enhancing the size of the social-referencing effect.

We did this in the most conservative way possible, replicating the Hollywood Squares task by using the same edited videos utilized in the second RP:P study (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 subjects' knowledge of affirmative action by having them watch a news report on the topic prior to completing the Hollywood Squares task (we also refer to this revised protocol as the *informed* condition). We then compared performance in this condition with performance in an *uninformed* condition mimicking the first RP:P study's protocol, in which subjects completed the Hollywood Squares task after completing an unrelated cognitive filler task (a flanker task).

Disclosures

Preregistration

Prior to data collection, confirmatory analyses were preregistered on the Open Science Framework (http://osf.io/8ycn2/). Subsequently, we developed additional analyses (in response to peer review) that are as described in this article and that were also preregistered on the Open Science Framework (see the Results section at 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.

Ethical approval

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

Method

Sample

We recruited subjects at universities in the United Kingdom (University of Edinburgh and University of Stirling) and the United States of America (Occidental College). The goal at each testing site was to collect data from a

Table 1. Demographic Details of the Subjects at Each Test Site

Statistic	University of Edinburgh	University of Stirling	Occidental College
Sample *N*	93 (76 female)	105 (67 female)	119 (93 female)
Mean age (years)	22 (*SD* = 3.2)	22 (*SD* = 4.5)	20 (*SD* = 1.2)
Black subjects (*n*)	3	0	10
Included in analysis of the offensive remark (*n*)	88	100	88
Included in analysis of the nonoffensive remark (*n*)	88	99	89
Native English speakers (*n*)	68	75	105
Participated for payment (*n*)	92	75	56

total of 92 subjects, 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 the interaction of interest is somewhat smaller (f(u) = .37) than the effect size originally reported by Crosby et al. (f(u) = .47, or η_p^2 = .18). The original study involved 25 subjects, and the RP:P replications used approximately 30 subjects per site.

Data were collected from 317 English-speaking students, whose demographics are reported in Table 1. According to self-report, 248 were native English speakers, and 61 were fluent but nonnative English speakers; 8 did not report their language status. As in the original study, only individuals who did not identify as Black were included in the final sample; thus, data from 13 subjects were not analyzed further. At each site, some subjects were compensated with payment, and some were compensated with course credit.

Subjects were subsequently excluded from the analysis of a remark if their eye-tracking record during that remark had missing data on more than 40% of the samples (track loss, caused by, e.g., failure in the eye-tracking system or gaze directed away from the monitor). After exclusions, 277 subjects were included in the analysis of gaze during the offensive remark, and 277 subjects were included in the analysis of gaze during the nonoffensive remark; the total number of unique subjects included in the analysis was 283.

Materials and procedure

Subjects completed four tasks, in one of two orders, followed by 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 http://osf.io/w4h5x/. Table 2 summarizes the task order in each condition. In the uninformed condition, subjects completed the Hollywood Squares task after completing a flanker task. This procedure was intended to replicate the paradigm used in the RP:P, in which the Hollywood Squares task was performed along with a suite of cognitive distractor tasks that did not involve eye tracking. We assumed that the flanker task would not provide subjects with information about affirmative action. In our revised protocol (informed condition), subjects completed the Hollywood Squares task after freely viewing a series of news videos, one of which was on the topic of legal challenges to affirmative action. We reasoned that this video would provide subjects with necessary context to help them easily see why the critical remark in the Hollywood Squares task was offensive. In this condition, the flanker task followed the Hollywood Squares task, whereas in the uninformed condition, the news videos were presented following the Hollywood Squares task.

Before the Hollywood Squares task began, subjects were instructed to "please watch and attend to the following discussion on university admissions. At the conclusion of the discussion you will be asked questions about the discussion content and/or the discussion participants." Subjects' gaze was then 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 woman who was not seen announced

Table 2. Order of Tasks in the Uninformed Condition (Reproducibility Project: Psychology Protocol) and the Informed Condition (Revised Protocol)

Uninformed condition	Informed condition
Flanker task	Video task
Hollywood Squares task	Hollywood Squares task
Video task	Flanker task
Multiple-choice questions	Multiple-choice questions
Demographic survey	Demographic survey

that she was turning off the headphones of two of the discussants and asked the four discussants 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 (Speaker 1) made a nonoffensive remark (lasting 19 s), and then the White discussant in the top right (Speaker 2) made a potentially offensive remark (lasting 20.5 s). In the headphones-on condition, the woman announced that all four discussants would talk about the topic (and all four raised their hands to indicate that they could hear). From this point, the video was identical to the video in the headphones-off condition. As was done in the second replication in the RP:P, we edited the original video files, cutting out references to Stanford.[1] The entire task took approximately 1 min 45 s. 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.

A full transcript of each Hollywood Squares video can be found in the appendix. Subjects answered questions about the video they had watched at the conclusion of the study.

In the video task, subjects viewed four 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 four videos can be found at http://osf.io/weus5/. The four videos lasted a total of 6 min 30 s.

In the flanker task, subjects indicated the direction of an arrow that was surrounded by arrows that either matched or did not match its direction. Subjects completed 360 trials at their own pace; informal pilot testing suggested that the time taken to perform this number of trials matched the viewing time for the four news videos.

After these tasks, subjects responded to 10 multiple-choice questions concerning the Hollywood Squares video and 1 question asking them to rate their awareness of the political controversy around affirmative action prior to taking part in the study (scale from 1 to 7).

Finally, subjects completed a paper-and-pencil demographic survey that included questions about their ethnic and racial background. Each lab constructed its own questionnaire, using the demographic labels (e.g., African American, White British) in its national census.

Eye-tracking calibration always took place before subjects viewed the first video (i.e., before the Hollywood Squares discussion in the uninformed condition and before the first news video in the informed condition). In the informed condition, an additional drift correction was performed before the Hollywood Squares task. Data from the University of Edinburgh were collected on an EyeLink 1000 (SR Research, Ottawa, Ontario, Canada) sampling at 500 Hz (50 subjects) and an EyeLink 2000 sampling at 1000 Hz (43 subjects), and data from the University of Stirling and Occidental College were collected on an Eye Tribe eye tracker (iMotion, Copenhagen, Denmark) sampling at 30 Hz.

Analysis

Our analysis focused on subjects' gaze during the Hollywood Squares discussion. We processed these data by creating four equally sized areas of interest (AOIs), each corresponding to one of the discussants in the video (i.e., Speaker 2, who made the potentially offensive remark; the filler discussant; the Black discussant; and Speaker 1, who made the nonoffensive remark). Then, we calculated the total time that each subject spent gazing within each of the four AOIs during the offensive and the nonoffensive remarks.[2] Our preregistered plan at http://osf.io/tj6qh/ provides full details on how the data were processed.

Results

Confirmatory regression analyses were conducted using the R language (Version 3.5.3; R Core Team, 2018) and the *lme4* package (Version 1.1-19; Bates, Mächler, Bolker, & Walker, 2015); we fit the mixed-effects models using full maximum likelihood. Pseudo-R^2 statistics were calculated using the *MuMIn* package (Version 1.42.1; Bartoń, 2018). Degrees of freedom and p values were calculated using the *lmerTest* package (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 headphone-condition variable, the reference level was set as headphones off; for the protocol variable, the reference level was the RP:P protocol (uninformed condition).

Figure 1 shows results for gaze during the offensive remark, separately for each headphone condition in each protocol: Figure 1a shows total gaze time for each of the four AOIs, and Figure 1b shows gaze time for the two critical AOIs (the speaker who made the potentially offensive remark and the Black discussant) at each of the three sites. Figure 2 shows the analogous data for gaze during the nonoffensive remark.

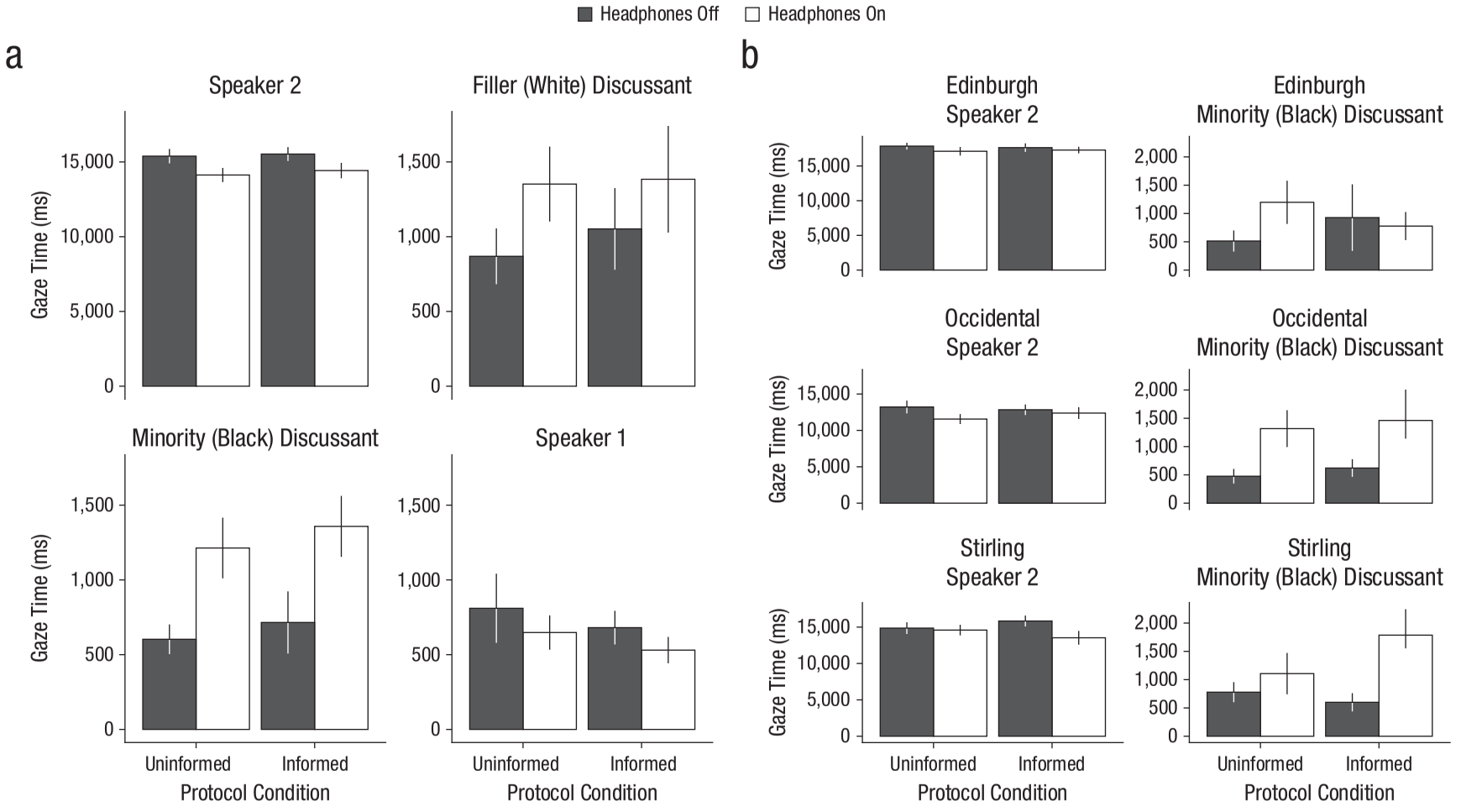

Fig. 1. Results for gaze during the offensive remark. The graphs in (a) show mean duration of gaze to each of the four areas of interest (Speaker 2, who made the potentially offensive remark; the filler discussant; the minority discussant; and Speaker 1, who made the nonoffensive remark) across the three testing sites. The graphs in (b) show mean duration of gaze to the two critical areas of interest (Speaker 2 and the minority discussant) separately for each testing site. All the graphs show results separately for the headphones-off and headphones-on conditions in each protocol condition. Error bars represent ±1 *SEM*.

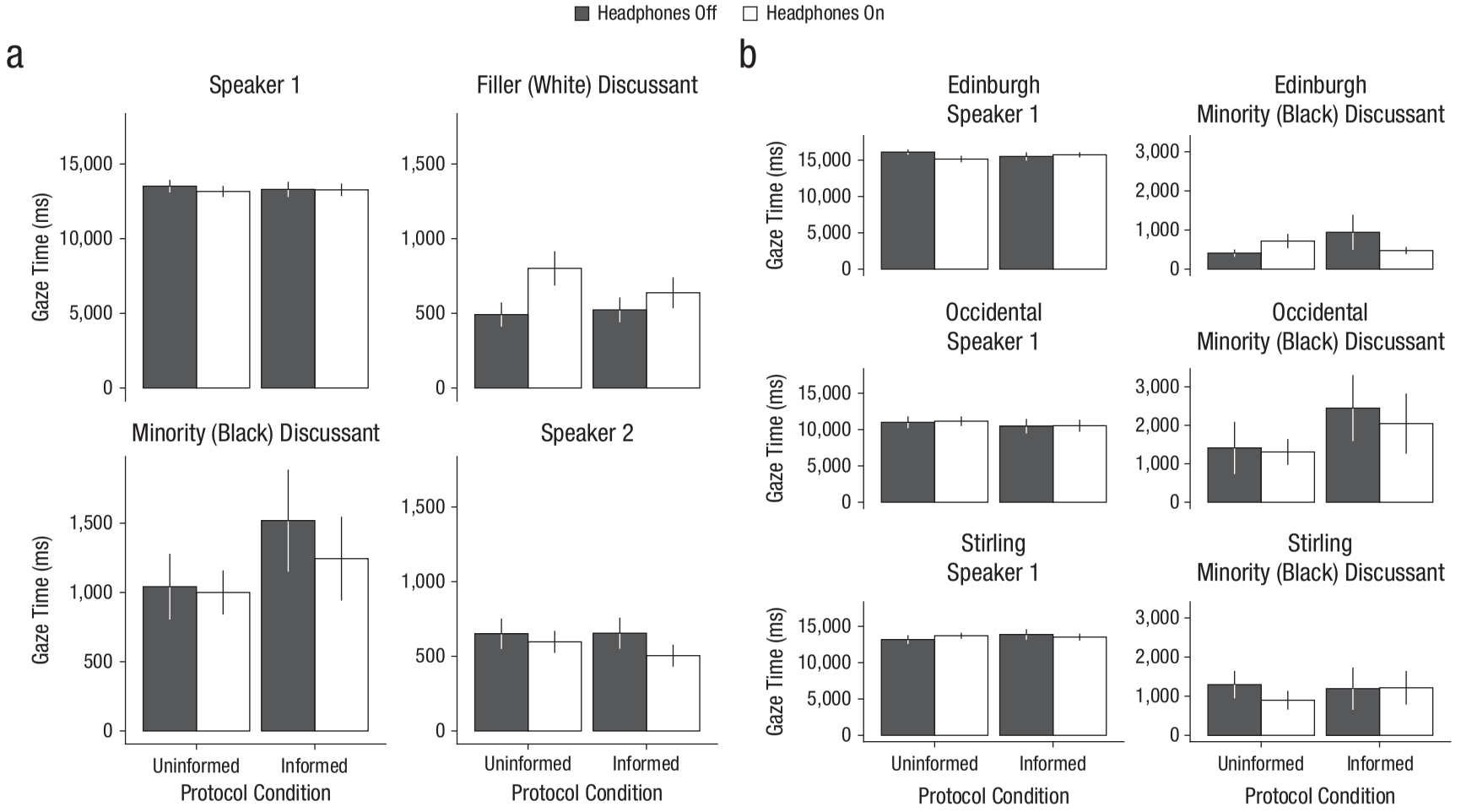

Fig. 2. Results for gaze during the nonoffensive remark. The graphs in (a) show mean duration of gaze to each of the four areas of interest (Speaker 1, who made the nonoffensive remark; the filler discussant; the minority discussant; and Speaker 2, who made the potentially offensive remark) across the three testing sites. The graphs in (b) show mean duration of gaze to the two critical areas of interest (Speaker 1 and the minority discussant) separately for each testing site. All the graphs show results separately for the headphones-off and headphones-on conditions in each protocol condition. Error bars represent ±1 *SEM*.

Confirmatory Analysis 1: test of the association hypothesis

Our first planned analysis assessed whether our data confirmed Crosby et al.'s finding that subjects gazed longer toward a potentially offended individual if they believed that he had heard an offensive remark than if they believed that he had not heard it. We tested for a two-way interaction between AOI and headphone condition during the offensive remark; statistical significance was determined by model comparison. Our preregistered analysis had the following form:

> Total Looking Time ~ AOI * Headphone Condition + (1 + AOI | Subject:Testing Site) + (1 + Headphone Condition | Testing Site)

However, we subsequently realized that incorporating a by-subject random slope of AOI was erroneous, because it resulted in the same number of regression parameters as data points (each subject provided one data point per AOI). Our final model thus had the form

> Total Looking Time ~ AOI * Headphone Condition + (1 | Subject:Testing Site) + (1 + Headphone Condition | Testing Site)

Model comparison indicated that there was a significant interaction between AOI and headphone condition, $\chi^2(3) = 22.11$, $p < .001$, pseudo-$R^2 = .85$. There was increased gaze toward the Black discussant when he could hear the offensive remark (headphones on: M = 1,281 ms, SD = 1,711; headphones off: M = 660 ms, SD = 1,352) $\beta = 1{,}814$ (SE = 420), $t(1105) = 4.3$, $p < .001$.

Confirmatory Analysis 2: test of the protocol manipulation

Our second analysis tested whether the interaction between AOI and headphone condition during the offensive remark was of a larger magnitude when subjects were in the informed (vs. uninformed) protocol. Our preregistered regression had the following form:

> Total Looking Time ~ Protocol * AOI * Headphone Condition + (1 + AOI | Subject:Testing Site) + (1 + Headphone Condition | Testing Site)

However, we again removed the by-subject random effect of AOI, for the reasons mentioned earlier.

There was not a significant interaction of AOI, headphone condition, and protocol, $\chi^2(3) = 0.13$, $p = .99$, pseudo-$R^2 = .85$. The protocol did not significantly influence the degree to which subjects gazed more at the Black discussant when he could hear the potentially offensive remark than when he could not, $\beta = -123$ (SE = 841), $t(1105) = -0.19$, $p = .88$.

Confirmatory Analysis 3: the nonoffensive remark

To confirm that these findings were driven by the offensive remark, we repeated Confirmatory Analyses 1 and 2 on the data for the nonoffensive remark. In Crosby et al.'s original study, headphone condition did not affect the distribution of gaze across the AOIs during this remark. We also found that there was not a significant interaction between AOI and headphone condition, $\chi^2(3) = 1.55$, $p = .67$, pseudo-$R^2 = .84$, and gaze toward the Black discussant did not increase when he could hear the remark, $\beta = 26$ (SE = 368), $t(1109) = 0.07$, $p = .94$. There also was not a significant interaction of AOI, headphone condition, and protocol, $\chi^2(3) = 0.73$, $p = .87$, pseudo-$R^2 = .86$, and gaze toward the Black discussant in particular was not significantly affected by this combination of factors, $\beta = -560$ (SE = 736), $t(1109) = -0.76$, $p = .45$.

Confirmatory Analysis 4: awareness of affirmative action

To test whether subjects who reported being more aware of affirmative action prior to the study might have been more sensitive to the protocol manipulation, we preregistered a regression of the following form to analyze gaze during the offensive remark:

> Total Looking Time ~ AOI * Headphone Condition * Protocol * Awareness + (1 + AOI | Subject:Testing Site) + (1 + Headphone Condition | Testing Site)

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

The results for the four-way interaction of AOI, headphone condition, protocol, and awareness revealed that awareness of affirmative action did not moderate the effect of protocol on gaze in this task, $\chi^2(3) = 4.3$, $p = .23$, pseudo-$R^2 = .86$.

We also carried out an exploratory analysis assessing whether awareness of affirmative action moderated the significant interaction of AOI and headphone condition found in Confirmatory Analysis 1, using a regression of the following form:

> Total Looking Time ~ AOI * Headphone Condition * Awareness + (1 | Subject:Testing Site) + (1 + Headphone Condition | Testing Site)

The results for the three-way interaction of AOI, headphone condition, and awareness revealed that awareness of affirmative action did not significantly moderate the effect of headphone condition on gaze in this task, $\chi^2(3) = 7.7$, $p = .051$, pseudo-$R^2 = .86$, 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 American than for European subjects, so we compared the size of the critical effect between these groups. This analysis was not part of the original preregistered analysis plan, but was added in response to reviews of this project 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 * Headphone Condition *
Protocol * Continent + (1 + AOI | Subject:Testing
Site) + (1 + Headphone Condition | Testing Site)
```

Again, in our analysis we removed the by-subject random effect of AOI, for the reasons discussed previously.

The test of the four-way interaction of AOI, headphone condition, protocol, and continent revealed that the European/American distinction did not significantly moderate behavior in this task, $\chi^2(3) = 3.4$, $p = .34$, pseudo-$R^2 = .87$.

A subsequent exploratory analysis examined whether the interaction of AOI and headphone condition varied between the continents, regardless of protocol. This was tested using the following regression:

```
Total Looking Time ~ AOI * Headphone Condition *
Continent + (1 | Subject:Testing Site) + (1 +
Headphone Condition | Testing Site)
```

The results revealed that the three-way interaction of AOI, headphone condition, and continent did not significantly influence gaze behavior in this task, $\chi^2(3) = 0.5$, $p = .92$, pseudo-$R^2 = .86$.

Confirmatory Analysis 6: logistic analysis

We conducted an additional analysis to account for the fact that, in some ways, eye-tracking data are not suitable for analysis using a linear mixed-effects model. In particular, models including the effects of AOI on total looking time violate the independence assumption, because an increase in the time spent looking at one AOI necessitates decreases in the times spent looking at other AOIs. In this final confirmatory analysis, the dependent variable was the logit-transformed proportion of time spent looking at the Black discussant out of the total time spent looking at all four discussants. Our initially preregistered regression had the following form:

```
Logit Proportion ~ Headphone Condition * Remark *
Protocol + (1 + Remark | Subject:Testing Site) +
(1 + Headphone Condition * Protocol | Testing
Site)
```

Following our earlier logic, we removed the by-subject random effect of remark because each subject provided only one data point per remark.

This regression did not reveal a significant interaction of headphone condition, remark, and protocol, $\chi^2(1) = 1.6$, $p = .20$, pseudo-$R^2 = .03$.

Additional exploratory analyses

We also conducted two exploratory follow-up logistic regressions, modeled on Confirmatory Analyses 1 and 2. The first regression assessed the effect of headphone condition on the proportion of time spent looking at the Black discussant during the offensive remark, regardless of protocol. After we dropped random slopes for testing site because of convergence issues, this analysis had the following form:

```
Logit Proportion ~ Headphone Condition +
(1 | Testing Site)
```

The results revealed a significant effect of headphone condition, replicating Confirmatory Analysis 1, $\chi^2(1) = 10.25$, $p = .001$, pseudo-$R^2 = .04$.

The second regression assessed whether the protocol manipulation interacted with headphone condition during the offensive remark. After simplification for nonconvergence, this regression had the following form:

```
Logit Proportion ~ Headphone Condition *
Protocol + (1 + Headphone Condition +
Protocol | Testing Site)
```

There was not a significant interaction between headphone condition and protocol, $\chi^2(1) = 1.44$, $p = .23$, pseudo-$R^2 = .04$. Thus, Confirmatory Analysis 2 was replicated.

General Discussion

Crosby et al. (2008) provided eye-tracking evidence that social referencing occurs during offensive behavior, a result that failed to be replicated in the RP:P (Jonas &

Skorinko, 2015). In the current replication study, we used a peer-reviewed experimental manipulation of the protocol to test whether the original paradigm might provide stronger evidence for social referencing if subjects were explicitly provided with relevant background knowledge to help them interpret the potentially offensive remark and its consequences (in this case, information about debates concerning affirmative action).

Our overall results were consistent with Crosby et al.'s original report. We found that, on hearing an offensive remark, subjects were significantly more likely to gaze toward a potentially offended person if that person could hear the offensive remark than if he could not. However, we did not find that subjects' behavior in the task was affected by the protocol to which they had been assigned: The social-referencing effect did not differ significantly between subjects assigned to our new protocol and those assigned to the protocol that Jonas and Skorinko (2015) used in the RP:P. This raises the question of why the original finding was not replicated in Jonas and Skorinko's study, but was replicated in ours, given that our protocol manipulation did not appear to affect subjects' 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 et al.'s original study.

To compare effect sizes, we calculated Cohen's *d* statistics for the effect of the headphones manipulation on the duration of gaze to the Black discussant during the offensive remark. In the original study, subjects gazed at the Black discussant for an average of 2,588 ms (*SD* = 2,085) in the headphones-on condition and 505 ms (*SD* = 491) in the headphones-off condition; the resulting Cohen's *d* was 1.38. But in the present replication, the corresponding averages were 1,281 ms (*SD* = 1,711) in the headphones-on condition and 660 ms (*SD* = 1,352) in the headphones-off condition, and the Cohen's *d* was 0.40. Thus, the effect we observed was a little more than one quarter the size of the original.

This discrepancy in effect sizes can be explained in two ways. The first possibility is that the original effect reported by Crosby et al. may be a Type M error (Gelman & Carlin, 2014), that is, an incorrect estimate of the magnitude of a true effect. Such misestimates are known to be more likely when studies have relatively low statistical power, as was plausibly the case for the original study, which used a between-subjects design (with a total of 25 subjects across the two conditions), took only a single observation from each subject (i.e., gaze behavior during the offensive remark), and used a dependent measure that one might expect to be relatively noisy (time spent gazing at the individual discussants could be affected by nuisance factors such as boredom, tiredness, etc.). This explanation can easily account for why Jonas and Skorinko (2015) failed to replicate the phenomenon: Their study would have been underpowered to detect the true underlying effect in this paradigm. If the original report's effect size of 1.38 was correct, then 95% power could have been achieved with only 15 subjects per group (according to an analysis conducted using G*Power; Faul et al., 2007). But if the present estimate of 0.40 is closer to the truth, then the required sample size would be 164 subjects 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 sizes comes from the notion of contextual sensitivity (Van Bavel et al., 2016). In particular, behavior in a paradigm like this one could vary importantly with subjects' background, in this case, their knowledge of affirmative action. For instance, subjects in this replication may have been less familiar with the debate about affirmative action than were subjects in the original study, and thus fewer of our subjects may have noticed that the potentially offensive remark was, in fact, potentially offensive. Our stimuli were created by Crosby et al. 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 the observed effect size has been smaller in subsequent work than in Crosby et al.'s study, and indeed it is quite possible that the 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 earlier, a number of the original study's features, such as small numbers of subjects in a between-subjects experimental design, suggest low statistical power. Under these conditions, tests of statistical significance will yield positive results only when the tested effect is extremely large or is 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 the original effect size of 1.38 were correct, it would imply that the size of the measured social-referencing effect was much greater than that of obvious and mundane phenomena (e.g., people who are more liberal tend to think that social equality is more important, people who like eggs more tend to eat more egg salad, and men tend to weigh more than women) that barely require statistical confirmation (Simmons, Nelson, & Simonsohn, 2013). 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 subjects behaved differently whether they were tested in Scotland or California (Confirmatory Analysis 5), even though one might expect the Californian subjects to be much more similar to the original Stanford sample, and therefore to show a larger effect.[3] We also found no statistically significant evidence that subjects behaved differently depending on whether they were informed or uninformed about the social issues surrounding affirmative action, whether because of our manipulation of the protocol (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 Hollywood Squares task is highly contextually sensitive, such that it could generate an effect size of 1.38 in the original study, but only 0.40 in the present study.

Still, these null findings are not conclusive, and it remains possible that we might have obtained effect-size estimates that were as large as the original estimate if we had used a somewhat different manipulation of the experimental protocol. For example, rather than varying whether subjects were deliberately informed about affirmative action, we could have varied whether they viewed the original stimuli or newly created stimuli that were perhaps better matched to the subjects' 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 leads 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 videos that were individualized for each testing site (so that scripts, actors, accents, and languages varied across sites) 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 their offensiveness and believability to the videos created for Crosby et al.'s original 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 we tested, but reconstructing that match is impossible because there is no contemporaneous evidence as to the quality of that match.

Why did our manipulation of background knowledge not affect subjects' behavior in this task? The preceding discussion suggests that 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 social-referencing effect that we uncovered: We would not expect context and background to cause large fluctuations in an effect that is itself small. That said, it is also possible that our manipulation of background knowledge was simply not very effective, in that it may not have fully informed subjects about the cultural context of affirmative action. Because we did not carry out a manipulation check, we have no way to confirm or deny this possibility, which is an obvious flaw in our methodology. Finally, as suggested by an anonymous reviewer, it is possible that our treatment and control conditions both acted to enhance the effect of social referencing in this paradigm; for example, completing the flanker task in the uninformed condition may have enhanced subjects' attentional control and magnified their ocular responses to the offensive remark. This is possible, but we think it unlikely: First, inspection of Figures 1 and 2 suggests no baseline differences in gaze behavior between the informed and uninformed conditions. Second, there is no evidence that learning to inhibit one specific distractor transfers to inhibition of other distractors (Kelley & Yantis, 2009).

Thus, we believe that it is more likely that the original study's large effect size was a Type M error than that the subsequent small effects in replication studies were the result of failure to account for contextual sensitivity. But both possibilities remain viable hypotheses, so we think it is helpful to consider their implications for future work. Most obviously, statistical considerations suggest that this paradigm could be improved in future studies by seeking ways to increase statistical power and measurement precision that go beyond testing additional subjects, such as by having each subject 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 subjects. For example, we would be able to draw stronger conclusions from our data if we could conduct an independent assessment of whether our subjects and the original subjects perceived the offensive remarks in similar ways. Future work on this effect, therefore,

might include norming the stimuli (e.g., obtaining explicit ratings of offensiveness) or conducting a manipulation check, to measure the degree to which the stimuli induce the intended effect in subjects.

Finally, in considering lessons learned, we note one additional factor that may have supported our replication of the key result from Crosby et al.: 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 et al., 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 than in the previous replication attempts, and that the potential for differences in methodology to affect measurement error across testing locations was minimized.

Conclusion

In the current project, we sought to replicate Crosby et al.'s (2008) finding of a social-referencing effect, as well as test for moderation of this effect by varying the protocol to make our sample's background knowledge more similar to that of the original sample. We did not find evidence for moderation by protocol: Subjects' gaze behavior did not significantly vary depending on whether they received the manipulation of their background knowledge. However, in contrast to Jonas and Skorinko (2015), we did replicate Crosby et al.'s original finding of social referencing, although the effect size measured in this project was considerably smaller than that reported for the original study.

Appendix: Transcripts of the Videos in the Hollywood Squares Task

Before the Hollywood Squares task began, subjects read: "Please watch and attend to the following discussion on university admissions. At the conclusion of the discussion you will be asked questions about the discussion content and/or the discussion participants." They then watched one of two videos:

Headphones-off video

Woman's voice: For this part of the discussion, only Participants 1 and 2 [referring to Speakers 1 and 2] will discuss a topic. Participants 3 and 4 [referring to Speakers 3 and 4], 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?

Speaker 1 (White): So it's just two of us now?

Woman's voice: Yes. [Speaker 1 gives a small nod] The other participants will be brought into the conversation later. Can you two hear each other?

Speaker 1: Yes.

Speaker 2: Er, yes.

Woman's voice: The first question is, "What changes or improvements would you make?" Participant 1, your response?

Speaker 1: 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 2.

Speaker 2: 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 video

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?

Speaker 1 (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 1, your response?

Speaker 1: 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 2.

Speaker 2: 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.]

Transparency

Action Editor: Daniel J. Simons
Editor: Daniel J. Simons
Author Contributions

H. Rabagliati, M. Corley, J. P. J. King, and J. E. Loy developed the study protocol. J. P. J. King and J. E. Loy created the study materials. All the authors collected data or supervised data collection. H. Rabagliati and J. P. J. King wrote the analysis code, and H. Rabagliati analyzed the data. H. Rabagliati drafted the manuscript, and all the authors critically edited it and approved its submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Open Practices

Open Data: https://osf.io/weus5/registrations
Open Materials: https://osf.io/weus5/registrations
Preregistration: https://osf.io/weus5/registrations
All data, code, and materials have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/weus5/registrations. The protocol and analysis plans were preregistered at the Open Science Framework and can be accessed at https://osf.io/weus5/registrations. Changes to the preregistered analyses are described in the text. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245919870737. This article has received badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

TC

ORCID iDs

Hugh Rabagliati https://orcid.org/0000-0001-9828-5857
Carmel A. Levitan https://orcid.org/0000-0001-5403-444X
Ailsa E. Millen https://orcid.org/0000-0001-7112-0841

Acknowledgments

We would like to particularly acknowledge the contribution of Thomas Scherndl to the study's 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, and Hannah Wagner.

Prior Versions

A registered, results-blind version of this manuscript can be found at the Open Science Framework (https://osf.io/b26js/).

Notes

1. Our edits were slightly different from those in the second RP:P replication, in which references to Stanford were cut in the headphones-on video and muted in the headphones-off video.
2. Note that a mistake in our replication protocol suggested that the screen resolution would be fixed to 1280 × 1024 pixels, but it was in fact fixed to 1024 × 768 pixels.
3. It is possible that our Californian subjects were more similar to our Scottish subjects than to the Californian subjects Crosby et al. tested; for example, they may have been less aware of controversies concerning affirmative action, and thus less likely to notice the offensive remark, than Crosby et al.'s subjects were. That said, the present cohort were tested during an era in which, anecdotally, issues of social justice are prominent in the media (e.g., because of the efforts of groups such as Black Lives Matter), so we consider it unlikely that they would not perceive the offense.

References

Bartoń, K. (2018). MuMIn: Multi-model inference (R package Version 1.42.1) [Computer software]. Retrieved from http://CRAN.R-project.org/package=MuMIn

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). doi:10.18637/jss.v067.i01

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

Crosby, J. R., Monin, B., & Richardson, D. (2008). Where do we look during potentially offensive behavior?

Psychological Science, *19*, 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*, 641–651. doi:10.1177/1745691614551642

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

Jonas, K., & Skorinko, J. (2015). *Replication of "Where do we look during potentially offensive behavior?" by JR Crosby, B Monin, D Richardson (2008,* Psychological Science*)*. Retrieved from https://osf.io/nkaw4/

Kelley, T. A., & Yantis, S. (2009). Learning to attend: Effects of practice on information selection. *Journal of Vision*, *9*(7), Article 16. doi:10.1167/9.7.16

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). doi: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*, 314–324. doi:10.3758/s13428-011-0168-7

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

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Shafir, E. (2018). The workings of choosing and rejecting: Commentary on Many Labs 2. *Advances in Methods and Practices in Psychological Science*, *1*, 495–496.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). *Life after P-hacking*. Retrieved from http://ssrn.com/abstract=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, USA*, *113*, 6454–6459. doi:10.1073/pnas.1521897113