Journal of Experimental Social Psychology 94 (2021) 104099

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Case Report

Large-scale field experiment shows null effects of team demographic diversity on outsiders' willingness to support the team

Edward H. Chang [a,*], Erika L. Kirgios [b], Rosanna K. Smith [c]

[a] Negotiation, Organizations and Markets Unit, Harvard Business School, Harvard University, Soldiers Field, Boston, MA 02163, United States
[b] Department of Operations, Information and Decisions, The Wharton School of the University of Pennsylvania, 3730 Walnut St., Philadelphia, PA 19104, United States
[c] Department of Marketing, Terry College of Business, University of Georgia, 630 S. Lumpkin St., Athens, GA 30602, United States

A B S T R A C T

Demographic diversity in the United States is rising, and increasingly, work is conducted in teams. These co-occurring phenomena suggest that it might be increasingly common for work to be conducted by demographically diverse teams. But to date, in spite of copious field experimental evidence documenting that individuals are treated differently based on their demographic identity, we have little evidence from field experiments to establish how and whether teams are treated differently based on their levels of demographic diversity. To answer this question, we present the results of a preregistered, large-scale (n=9496) field experiment testing whether team demographic diversity affects outsiders' responses to the team. Participants were asked via email to donate money to support the work of a team that was described and depicted as demographically diverse, or not. Even though the study was well-powered to detect even small effects (i.e., differences of less than 1.5 percentage points in donation rates), we found no significant differences in people's willingness to donate to a more diverse versus a less diverse team. We also did not find moderation by participant gender, racial diversity of the participant's zip code, or political leaning of the participant's zip code, suggesting that the lack of a main effect is not due to competing mechanisms cancelling out a main effect. These results suggest past research on the effects of demographic diversity on team support may not generalize to the field, highlighting the need for additional field experimental research on people's responses to demographically diverse teams.

1. Introduction

In recent years, public discourse about the benefits of diversity for teams has abounded. News articles state that diverse teams outperform homogeneous ones on business metrics (Holger, 2019), pundits claim that diverse teams can drive innovation and creativity (Hewlett, Marshall, & Sherbin, 2013), and scholars say that diverse teams are smarter than less diverse teams and make better decisions (Hoogendoorn, Oosterbeek, & Van Praag, 2013; Woolley, Chabris, Pentland, Hashmi, & Malone, 2010). Organizations are also taking note: Companies ranging from Apple to Zappos espouse visible commitments to diversity which emphasize the view that diversity is a value-add for their companies.

But does this public discourse about the benefits of diversity translate into actual willingness to support diverse teams? Understanding how people react to diversity in teams is important given that demographic diversity is increasing in the United States. Estimates predict that the United States will become a majority-minority country by 2044 (U.S. Census Bureau, 2015), with no single racial group composing more than 50% of the population. In addition, work of all sorts is increasingly done in teams (Kozlowski & Bell, 2003; Wuchty, Jones, & Uzzi, 2007). Together, these trends suggest that diverse teams will become increasingly common in the United States. Thus, it is important to understand responses to diversity in teams and whether teams that are high in demographic diversity elicit different reactions and treatment than teams that are homogeneous or low in demographic diversity do. Answers to these questions will give us a better understanding of the potential consequences of the increased prevalence of diverse teams in organizations.

Although substantial field experimental evidence has demonstrated that people respond differently to individuals based on those individuals' demographic identities (Bertrand & Duflo, 2017), we have relatively little understanding of people's behavioral responses to demographic diversity within teams and organizations in real-world contexts. Some research examines how people respond to diversity in teams in laboratory and online experiments (Avery, 2003; Lount, Sheldon, Rink, & Phillips, 2015; Van Dijk, Van Engen, & Van Knippenberg, 2012; Wilton, Sanchez, Unzueta, Kaiser, & Caluori, 2019), but to the best of our knowledge, there are no field experiments that measure real-world behavior towards diverse teams. This lack of field experimental evidence is problematic given research that shows that people's predictions about how they would behave in imagined scenarios can differ

* corresponding author.
E-mail addresses: ehchang@hbs.edu (E.H. Chang), ekirgios@wharton.upenn.edu (E.L. Kirgios), rosanna.smith@uga.edu (R.K. Smith).

https://doi.org/10.1016/j.jesp.2020.104099
Received 9 June 2020; Received in revised form 17 December 2020; Accepted 22 December 2020
Available online 23 January 2021
0022-1031/

significantly from their actual behavior when they encounter those situations, particularly in race-related contexts (Kawakami, Dunn, Karmali, & Dovidio, 2009). Furthermore, studies on reactions to diverse teams have taken place in contexts where participants knew their behavior was being observed as part of an experiment. Given that people are often motivated to avoid appearing racist (Apfelbaum, Sommers, & Norton, 2008; Goff, Steele, & Davies, 2008; Monin & Miller, 2001), their behavior might differ depending on whether or not they know they are being observed. This suggests findings from existing research conducted with laboratory or online samples may not generalize to real-world behaviors.

How might people react to seeing diverse teams in the real world? One prediction is that outsider support will be greater for demographically diverse teams than for relatively homogeneous teams. In the United States, people say they largely view diversity positively (Bell & Hartmann, 2007). According to nationwide surveys, the majority of adults–particularly those who have more education–say that diversity makes the United States a better place to live (Fingerhut, 2018). Research has also shown that people espouse multiple rationales in favor of diversity in organizations (Ely & Thomas, 2001), and many organizations explicitly say they want to cultivate diversity (Jones & Donnelly, 2017). Thus, people may believe that diversity is good for teams–perhaps making teams more effective or more innovative–and therefore want to support more diverse teams.

In addition, there is some evidence that diversity can help improve the reputation of a group. For example, past research has shown that highlighting gender diversity in company marketing materials increases online participants' ratings of the prestige and reputation of the company (Wilton et al., 2019); racial diversity on a company's website can increase some participants' organizational attraction (Avery, 2003); and demographic diversity may help organizations escape negative scrutiny for lacking adequate diversity (Chang, Milkman, Chugh, & Akinola, 2019). These results suggest that people may view diversity in teams more favorably than they view homogeneity.

On the other hand, it is possible that this discourse on the benefits of diversity could merely be a form of diversity "happy talk" (Bell & Hartmann, 2007) such that people say they are willing to support diversity in teams in the abstract but do not actually support diversity when it comes to concrete behaviors. Prior work has shown that people typically discriminate against individuals from historically underrepresented backgrounds. For example, field experiments have shown that people discriminate against women and racial minorities–including Black Americans, Hispanic Americans, and Asian Americans–in a variety of contexts (Bertrand & Mullainathan, 2004; Milkman, Akinola, & Chugh, 2015; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012), and people commonly hold negative biases and stereotypes about women and racial minorities (Berdahl & Min, 2012; Fiske, Cuddy, Glick, & Xu, 2002; Forscher et al., 2019; Rosette, Leonardelli, & Phillips, 2008; Rudman & Glick, 2001). Prior work also suggests that these stereotypes and negative perceptions of women and racial minorities may extend to the groups and spaces they occupy (Bonam, Bergsieker, & Eberhardt, 2016; Chatman & O'Reilly, 2004). As a result, people may be biased against demographically diverse teams because they include more members of historically underrepresented populations that are negatively stereotyped. This bias may lead people to show less support for diverse teams than homogeneous teams in reality, despite expressing positive views towards diverse teams in the abstract.

People may also have lay theories about diversity that could lead them to penalize diverse teams. Research suggests that people perceive diverse teams as having greater relationship conflict as compared to homogeneous teams, even when behavior is held constant, and these biased perceptions of relationship conflict lead people to reduce their resource support of diverse teams in lab and online experiments (Lount et al., 2015). For example, Lount et al. (2015) asked participants to read identical transcripts of a team's discussions and manipulated whether the team (shown to participants in photographs) was racially homogeneous or diverse. They found that participants who evaluated a racially diverse team perceived the team as having more relationship conflict–even though the discussion transcripts were identical across conditions–and were less willing to provide resources to the racially diverse team as a result. A meta-analysis has also shown that demographic diversity in teams is correlated with negative rater biases: Relative to demographically homogeneous teams, demographically diverse teams are more likely to receive performance ratings from outsiders that are lower than their performance on objective benchmarks (Van Dijk et al., 2012). In other words, people may penalize demographically diverse teams both because of negative lay theories about diversity and because of bias against members of historically underrepresented populations within diverse teams.

Finally, there may be no overall main effect of demographic diversity on outsiders' support of teams. This lack of a main effect of demographic diversity could occur either because teams high in diversity do not elicit different treatment than teams low in diversity or because of the competing theories explored above. But if it is true that competing mechanisms lead to overall null effects, the effects of team diversity on outsiders' support of the team should be moderated by characteristics of the outsider. For example, if having negative biases and stereotypes towards members of historically underrepresented groups is one of the mechanisms, we might expect the identity of the outsider to moderate any effects, as members of dominant groups in society may react differently to demographically diverse teams as compared to members of historically underrepresented groups (Craig & Richeson, 2014; Danbold & Huo, 2015; Plaut, Garnett, Buffardi, & Sanchez-Burks, 2011). On the other hand, if the mechanism is driven by people's lay beliefs regarding the benefits or costs of diversity, we might predict the effects to be moderated by the political orientation of the outsider. For example, surveys have shown that Democrats are significantly more likely than Republicans to have positive views of diversity in the United States (Fingerhut, 2018), so we might expect Democrats to be more supportive of demographically diverse teams as compared to Republicans.

To test these competing hypotheses and examine how demographic diversity affects outsiders' support of teams, we ran a preregistered, large-scale (n=9496) field experiment. In our experiment, we manipulated how we portrayed a team in terms of its demographic diversity by featuring a team that was either high or low in both racial and gender diversity and by explicitly labeling only the visibly diverse team as "diverse." We then solicited people to donate money to support the team's efforts. Even though the study was well-powered to detect even small effects (i.e., differences of less than 1.5 percentage points in donation rates), we found no evidence that our manipulation affected the rates at which people were willing to support the team, nor did we find moderation by participant gender, racial diversity of the participant's zip code, or political leaning of the participant's zip code. These results suggest past research on the effects of demographic diversity on team support may not generalize to the field, highlighting the need for additional field experimental research on the effects of diversity on people's responses to teams.

2. Study 1

In Study 1, we wanted to test the email stimuli used in our field experiment to assess whether people would differentiate between the diversity levels of the two teams presented in the study stimuli. This study was run in response to requests from reviewers, and data collection occurred after the field experiment was completed. This study was preregistered on AsPredicted.org (https://aspredicted.org/blind.php?x=y55xk6), and we report all measures, manipulations, and exclusions. Study data and code are available on OSF (https://osf.io/g5x7a/?view_only=65d79a3e57b248dcb20f57bfb59179e0).

We recruited 200 U.S. participants (56.0% identified as men) via Amazon Mechanical Turk. No participants were excluded from the data. Participants were randomly assigned to either a treatment condition or a

control condition. In both conditions, participants read an email that included photographs of a team of scientists. After reading the email, participants rated their perceptions of the team's diversity on a scale from 1 (Strongly Disagree) to 7 (Strongly Agree) using three items adapted from Danbold and Unzueta (2020): "This is a diverse team"; "This team is diverse enough"; and "This team has a diversity problem" (reverse scored). In the treatment condition, the team was explicitly described as "diverse" and was depicted using ten photos of scientists that included two White men, two Asian men, one Black man, three White women, one Asian woman, and one Black woman; in other words, the team of scientists depicted was composed of 50% women and 50% racial minorities. In the control condition, the team was not explicitly described as "diverse" and was depicted using ten photos of scientists that included eight White men, one White woman, and one Asian woman; in other words, the team of scientists depicted was composed of only 20% women and 10% racial minorities. The stimuli used in this study were identical to the stimuli used in the field experiment. Finally, participants answered demographic questions about their gender identity and racial identity. See Supplementary Materials for screenshots of the study stimuli.

We found that perceived diversity of the team in the treatment condition ($M = 6.32$, $SD = 1.03$) was significantly higher than perceived diversity of the team in the control condition, ($M = 3.38$, $SD = 1.57$), $t(198) = 15.63$, $p < 0.001$, $d = 2.21$. This difference was not moderated by participant gender, $F(1, 196) = 1.02$, $p = 0.313$, or participant race, $F(1, 196) = 0.09$, $p = 0.767$. We used the criteria from Danbold and Unzueta (2020) to determine whether a team was perceived as sufficiently diverse and compared perceptions of diversity in each condition to the midpoint of the scale (i.e., 4). As predicted, participants in the treatment condition saw the team as sufficiently diverse, $t(98) = 22.38$, $p < 0.001$, $d = 2.25$, while participants in the control condition did not see the team as sufficiently diverse, $t(100) = 3.98$, $p < 0.001$, $d = 0.40$.

We conducted sensitivity power analyses using GPower (Faul, Erdfelder, Lang, & Buchner, 2007) using the sample sizes of 101 participants in the control condition and 99 participants in the treatment condition and the parameters $\alpha = 0.05$, two-tailed, and 80% power. These analyses suggest we had 80% power to detect an effect size of $d = 0.40$ between conditions and an effect size of $d = 0.28$ in the one-sample *t*-tests comparing each condition mean to a constant.

The results of this study thus provide evidence that people are able to distinguish between the diversity levels of the teams of scientists used as stimuli in our field experiment, and, as intended, people perceive the team in the treatment condition as sufficiently diverse and the team in the control condition as insufficiently diverse.[1]

3. Study 2: Field experiment

We conducted our field experiment in partnership with a program called StepUp, a 28-day digital rewards program designed to promote exercise among members of 24 Hour Fitness gyms. People who took part in StepUp earned small monetary rewards for going to the gym. At the end of the program, StepUp participants received an email from StepUp asking them to donate their earnings back to StepUp to support the program. Using the stimuli from Study 1, we experimentally manipulated whether the team of scientists behind StepUp was described and depicted as diverse in this email to see how team diversity affects outsiders' willingness to support the team. This study was preregistered on AsPredicted.org (http://aspredicted.org/blind.php?x=zu28g5), and we report all measures, manipulations, and exclusions.

3.1. Participants

Participants ($n = 9496$, 33.1% men) were determined by our field partner, StepUp. The desired sample size of 9000 to 9500 participants was preregistered, and we as researchers did not have control over the final number of participants recruited for the study. No participants were excluded from the data. Participants included every individual who joined the StepUp Program between May 8, 2018 and September 16, 2018. Our participants came from 2357 unique zip codes, and the majority (54.7%) came from California (the state with the largest number of 24 Hour Fitness gym locations in the U.S.). See Table 1 for summary statistics and Table 2 for correlations among variables in the dataset.

3.2. Procedure

After they completed the StepUp program, participants were randomly assigned to either a treatment condition or a control condition. All participants received an email that described StepUp as a not-for-profit program and included photos of ten scientists who helped run the program. The email asked participants to click through to a Qualtrics study to either claim the small monetary incentives they had earned for exercising or donate these earnings back to the StepUp program to fund the work of this team of scientists.

We used the same stimuli tested in Study 1 for this field experiment. In the treatment condition, we described the team of scientists behind StepUp as "diverse" (e.g., "StepUp is a not-for-profit program run by a diverse team of scientists"), and the ten photos of scientists included two White men, two Asian men, one Black man, three White women, one Asian woman, and one Black woman; in other words, the team of scientists depicted was composed of 50% women and 50% racial minorities. In the control condition, we did not use the term "diverse" to describe the team of scientists, and the ten photos of scientists included eight White men, one White woman, and one Asian woman; thus, this team was composed of only 20% women and 10% racial minorities. See Supplementary Materials for screenshots of study stimuli.

3.3. Dependent measure

Our dependent variable was whether participants donated their earnings back to StepUp to fund the work of the team of scientists behind StepUp. We coded this variable as 1 if participants donated their money and 0 if they claimed their money. If participants did not make an active decision (to either claim their money or to donate their money), this lack of decision was coded as a 1 (i.e., an implicit donation since they did not

Table 1
Summary statistics of study participants.

	Mean	SD	25th Percentile	75th Percentile
Gender (man = 1, woman or other = 0)	0.331	0.471	0.00	1.00
Cash earned during StepUp (in dollars)	1.82	2.90	0.29	2.43
Number of gym visits during StepUp	6.02	6.26	1.00	10.0
Percent of population in participant zip code that is White (2013 US Census data)	46.5	24.3	26.0	67.0
Percent of population in participant zip code that voted Republican in 2016 U.S. presidential election (based on cast ballots)	34.3	12.6	23.4	43.1

[1] In our Supplementary Materials, we present the results of an additional study showing further evidence that people distinguish between the two teams used as stimuli in our field experiment. Specifically, 81.5% of online participants presented with a choice between the two teams preferred to support the more diverse team. This was a significantly greater proportion than would be expected by chance, $z = 6.53$, $p < 0.001$, and it was not moderated by participant gender, $z = 1.49$, $p = 0.137$, or race, $z = 0.895$, $p = 0.371$.

Table 2
Field experiment correlation matrix.

	1	2	3	4	5	6	7
1. Condition (treatment = 1, control = 0)	1.00						
2. Donated back to StepUp	−0.009	1.00					
3. Participant gender (man = 1, woman or other = 0)	−0.012	0.010	1.00				
4. Cash earned during StepUp (in dollars)	−0.014	−0.300***	0.077***	1.00			
5. Number of gym visits during StepUp	−0.004	−0.264***	0.142***	0.585***	1.00		
6. Percent White in participant zip code	0.007	−0.010	0.007	0.026*	0.010	1.00	
7. Percent voting Republican in participant zip code	0.003	0.011	−0.008	0.040***	0.013	0.267***	1.00

*** *** $p < 0.001$.
* * $p < 0.05$.

claim their money), as per our preregistration. Although not preregistered, given that 84.3% of participants did not make an active decision, we also present all results limited to the 1493 participants who actively decided whether to claim or donate their incentives by clicking through to the Qualtrics survey linked in the email. While restricting our data in this way introduces selection bias, we present these results so interested readers can see that they are consistent with our results from the overall data.[2]

3.4. Moderator variables

3.4.1. Participant gender

For 95.6% of participants, we received participant gender data directly from our field partner. For the 4.4% of participants with missing gender information, we initially inferred their gender using a first name classifier from prior research (Morton, Zettelmeyer, & Silva-Risso, 2003). In this manner, we were able to categorize 99.1% of participants as men or women. For the remaining 0.9% of participants, a research assistant used a combination of online searches for first names to find commonly associated genders and Google searches of full names to find the participants' gender information. We labeled participant gender as "unknown" for the 3 participants who could not be classified by any of the above methods.

3.4.2. Racial diversity of participant's zip code

We used 24 Hour Fitness data to identify participants' zip codes. For 98.2% of participants, we used the zip code they listed as their home zip code when they signed up for a 24 Hour Fitness membership. For the 1.8% of participants missing this data, we used their gym check-in data to identify the zip code of their most frequently visited gym. Ultimately, five participants' zip codes remained unknown after classification.

We then used the "choroplethrZip" package in R to pull demographic data from the 2013 American Community Survey, a survey run by the U. S. Census Bureau. In particular, we were able to identify the percent of the population within each zip code that was White (non-Hispanic), Black (non-Hispanic), Asian (non-Hispanic), or Hispanic (all races). As per our preregistration, we operationalized racial diversity as the percent of the population within each zip code that was White according to this 2013 US census data.

3.4.3. Political leaning of participant's zip code

Using the same zip code data, as per our preregistration, we operationalized political leaning as the percent of ballots cast for the Republican candidate (Donald Trump) within each zip code in the 2016 U.S. presidential election.

3.5. Preregistered analysis plan

In our preregistration, our main analysis was to run an OLS regression with robust standard errors with an indicator for the treatment condition, a continuous control for the amount of money earned by the participant in the StepUp program, fixed effects for which version of StepUp the participant received,[3] and an indicator for participant gender.

We also preregistered that we would test whether these effects were moderated by participant gender, the racial diversity of the participant's zip code, and the political leaning of the participant's zip code. For these analyses, we preregistered that we would run the same regression we used for the main analysis but with interactions between the treatment indicator and the moderator of interest. We relied on OLS regression with robust standard errors instead of logistic regression because we preregistered three interaction analyses, interactions are not estimated without bias when using logistic regressions (Ai & Norton, 2003), and there are reasons to prefer linear regression to logistic regression when estimating causal effects on binary outcomes (Gomila, 2020). Although not preregistered, for completeness, we also report all results using logits and probits in footnotes (see also Table S3 in our Supplementary Materials for full results using logistic regression rather than OLS regression).

Study data and code are available on OSF (https://osf.io/g5x7a/?view_only=65d79a3e57b248dcb20f57bfb59179e0).[4]

3.6. Results

In the treatment condition, 92.56% of participants donated their earnings to support the team; in the control condition, 93.07% of participants donated their earnings to support the team, $z = 0.967$, $p = 0.334$. Using our preregistered OLS regression, we found no evidence that being in the treatment condition affected donation rates ($b = -0.007$, $SE = 0.005$, $p = 0.177$; 95% CI: [−0.017, 0.003]; see Table S1,

[2] We unfortunately did not track whether participants opened our emails, so the only way we can ensure that participants opened the email is if they made an active decision in our Qualtrics survey.

[3] The StepUp program included 54 different versions to test the effectiveness of various interventions that aimed to help people build lasting exercise habits. The program tested interventions like providing social norm information, varying incentives for exercise, and providing exercise advice. None of the interventions tested related to demographic diversity. Furthermore, participants received our email asking them to claim or donate their earnings an average of 34 days after they completed the StepUp program. Finally, all results are robust to the removal of this control (see Online Supplement Table S1).

[4] Given restrictions from our IRB and our field partner, we are unable to include participant zip codes or variables related to participant zip codes in the posted data. This means that several preregistered moderator variables inferred from participant zip code are not included in our publicly posted data to protect participant privacy: the percent White population in the participant's zip code and the percent of cast ballots for Donald Trump in the 2016 presidential election in the participant's zip code. Researchers interested in accessing these variables should contact the corresponding author who, upon request, will ask the IRB for approval to share the de-identified data with those moderator variables included.

Model 4 in the Supplementary Materials for complete regression results).[5]

Sensitivity power analyses conducted using GPower (Faul et al., 2007) and the narrow range of the confidence interval surrounding our estimate of the treatment effect suggest that we have adequate power to detect even very small effects. We adopted the conservative approach of calculating sensitivity to a basic proportions test, as this does not take into account the additional power afforded to us by our preregistered control variables. Given that 93.07% of participants donated in the control condition, sample sizes of 4806 participants in the control condition and 4690 participants in the treatment condition, and the parameters of $\alpha = 0.05$, two-tailed, and 80% power, we had 80% power to detect an absolute difference in donation rates of 1.39 percentage points between the treatment and control conditions.

Following a reviewer's recommendation, we also conducted a post hoc equivalence test to reject the presence of the smallest effect size of interest (Lakens, Scheel, & Isager, 2018). We set the smallest effect size of interest to $200, which would be the salary cost of tasking a hypothetical StepUp employee who makes $50,000 per year to spend one day making changes to the donation solicitation emails. Because the average donation amount was $1.82, 109.9 more people would have had to donate (or not donate) in the treatment condition to reach the $200 effect size of interest. Given that we had 4690 participants in our treatment condition, our treatment would have had to produce a 2.34 percentage point change in donation rates relative to the control for it to be worthwhile to implement one email over the other across the entire sample. However, using two one-sided tests, we find evidence that we can reject that the effect size is at least as large as 2.34 percentage points, $z = 4.32$, $p < 0.001$.

Next, we tested whether participant gender moderated the effect of team diversity on donation rates. While we found that men donated significantly more often than women did ($b = 0.023$, $SE = 0.007$, $p = 0.001$; 95% CI: [0.008, 0.038]), we found no interaction between the treatment condition and participant gender ($b = -0.006$, $SE = 0.010$, $p = 0.536$; 95% CI: [−0.027, 0.014]; see Table S2, Model 2 for complete regression results), suggesting that participant gender did not moderate the effect of team diversity on outsiders' willingness to support the team.[6]

We also found no evidence of moderation by racial diversity of the participant's zip code ($b = 0.024$, $SE = 0.020$, $p = 0.228$, 95% CI: [−0.016, 0.065][7]; see Table S2, Model 3 for complete regression results), nor evidence of moderation by political leaning of the participant's zip code ($b = 0.030$, $SE = 0.039$, $p = 0.447$, 95% CI: [−0.048, 0.108][8]; see Table S2, Model 4 for complete regression results).

When restricting our data to only those participants who made an active decision, we found results consistent with those from our full data. For this subset of the data, donation rates were lower but still did not differ significantly across conditions: 54.83% of those in the treatment condition chose to donate their earnings, while 56.63% of those in the control condition chose to donate their earnings, $z = 0.648$, $p = 0.517$. Applying our preregistered OLS regression to this subset of the data, we again found no effect of our treatment on donation rates ($b = -0.016$, $SE = 0.025$, $p = 0.530$, 95% CI: [−0.064, 0.033]; see Table S4, Model 1 in the Supplementary Materials for complete results). Similarly, when we tested our three moderation hypotheses, we found results consistent with our analyses of the full data: We found no evidence of moderation by participant gender ($b = -0.046$, $SE = 0.055$, $p = 0.410$, 95% CI: [−0.151, 0.060]; see Table S4, Model 2 in the Supplementary Materials for complete results), racial diversity of the participant's zip code ($b = 0.152$, $SE = 0.103$, $p = 0.140$, 95% CI: [−0.052, 0.356]; see Table S4, Model 3 in the Supplementary Materials for complete results), or political leaning of the participant's zip code ($b = 0.186$, $SE = 0.201$, $p = 0.355$, 95% CI: [−0.207, 0.579]; see Table S4, Model 4 in the Supplementary Materials for complete results).

4. General discussion

We present the results of the first–to our knowledge–field experiment testing how outsiders respond to demographic diversity in teams. In spite of being well-powered to detect even small effects (i.e., differences of less than 1.5 percentage points between conditions), we found no evidence that our manipulation affected the rates at which people were willing to financially support the team, nor did we find moderation by participant gender, racial diversity of the participant's zip code, or political leaning of the participant's zip code.

Our results could be interpreted in a positive light given that we do not detect significant discrimination against demographically diverse teams. Our findings suggest that there is not a diversity penalty for teams, even in contexts dominated by White men, which should reassure organizations worried they might receive backlash for diversifying. On the other hand, these results could also be interpreted in a negative light, as people do not seem to penalize teams that *lack* demographic diversity either. These results could suggest that one barrier to diversifying teams is that people may not care about homogeneity once there are meager levels of gender and racial diversity, or they may not notice homogeneity in the first place.

Interpreting null results is always challenging, as there are undoubtedly many limitations to any study. For example, in our experiment, the control condition depicted a team that was not completely homogeneous. It is possible that we would have gotten different results if we had compared a team composed entirely of White men to a gender- and racially-diverse team; even though people did not perceive the team in the control condition as sufficiently diverse in Study 1, people still could have been relatively satisfied with the minimal levels of diversity present (Chang et al., 2019; Dezső, Ross, & Uribe, 2016).

In addition, we find large rates of inaction in our study. This is a consequence of our field setting: Participants may never have opened their email, may not have read it closely, or may have simply chosen not to respond. This feature of our field setting is common–in most everyday situations, people are not forced into making active choices, particularly in the realm of charitable giving. That being said, given our large sample size, we still had a sizable number of participants who made active decisions to donate or claim their earnings. Although limiting our analyses to participants who made an active choice introduces selection bias, as our treatment may have influenced their decision to respond to the email, we find largely identical results on this subsample of our participants.

We must also be cautious about interpreting the results of our moderation analyses by participant zip code. Using zip codes as proxies for participant-level characteristics is inherently a noisy way to measure underlying theoretical constructs. For example, the percent of the population that identifies as White within a zip code could be a noisy proxy for whether a participant is White, but it could also be a noisy proxy for opportunities for intergroup contact (MacInnis, Page-Gould, & Hodson, 2017). The zip code proxies can also be biased if gym-goers systematically deviate from other people within the general population. Together, this suggests that any inferences from the zip code analyses should be interpreted with caution. On the other hand, our participant gender

[5] These results remain consistent when we use logit or probit regressions rather than an OLS regression with robust standard errors ($b = -0.103$, $SE = 0.086$, $p = 0.231$ and $b = -0.058$, $SE = 0.042$, $p = 0.172$, respectively).

[6] The lack of interaction between treatment and gender remains consistent when we use logit or probit regressions ($b = -0.149$, $SE = 0.188$, $p = 0.426$ and $b = -0.064$, $SE = 0.091$, $p = 0.486$, respectively).

[7] The lack of moderation by racial diversity of the participant's zip code remains consistent when we use logit or probit regressions ($b = 0.394$, $SE = 0.359$, $p = 0.272$ and $b = 0.185$, $SE = 0.176$, $p = 0.293$, respectively).

[8] The lack of moderation by political leaning of the participant's zip code remains consistent when we use logit or probit regressions ($b = 0.519$, $SE = 0.701$, $p = 0.459$ and $b = 0.310$, $SE = 0.343$, $p = 0.365$, respectively).

analyses do not face these same issues and provide convergent results.

We also feel it is important to explicitly state the constraints on the generality of our study findings. First, we present the results of only a single field experiment. While our field experiment is a large-scale, preregistered field experiment measuring real-world behaviors in a policy-relevant context (financial support for demographically diverse scientific teams), the sums of money at stake are small, and it is possible that our findings do not generalize to non-science contexts or to other settings where diversity (or lack thereof) may be more salient to people. Indeed, the lack of demographic diversity in STEM fields has been well-documented (Guterl, 2014; Handelsman et al., 2005; Moss-Racusin et al., 2012), so people may be satisfied with lower levels of demographic diversity here than in other contexts.

Second, our participants come from zip codes that are less White and less Republican than the United States at large. Our results may have differed if we had a more representative sample, given that past research has shown that White people in the U.S. react differently to diversity than do racial minorities (Craig & Richeson, 2014; Danbold & Huo, 2015; Plaut, Garnett, Buffardi, & Sanchez-Burks, 2011), and surveys have shown that Democrats are significantly more likely than Republicans to have positive views of diversity in the United States (Fingerhut, 2018).

Finally, the United States has seen large increases in calls for racial justice in 2020 in the wake of the police killings of George Floyd, Breonna Taylor, and countless other Black people. Societal support for Black Lives Matter has increased drastically, and many organizations have made explicit commitments to being more anti-racist (Cohn & Quealy, 2020). Our field experiment was conducted in 2018, prior to these events. It is possible that we would find different results if we ran this field experiment today, given increased attention to and salience of racial justice.

How people respond to demographic diversity in teams is a question of theoretical importance as we seek to understand how diversity affects team performance, and it is a question of practical importance as the United States becomes increasingly diverse. These results help contribute to our understanding of this question by showing that in a real-world context measuring real behaviors, we find null effects. Although there are many audit experiments showing how people respond to individuals of different demographic identities, we present the first–to our knowledge–audit experiment testing how people respond to teams with different levels of demographic diversity. While there are many reasons to be cautious in interpreting null effects, our paper suggests that more field experimental research is needed to understand how diversity in teams affects outsiders' behaviors towards those teams.

5. Open practices

Data, code, and survey materials are available at https://osf.io/g5x7a/?view_only=65d79a3e57b248dcb20f57bfb59179e0. The Study 1 preregistration can be found at https://aspredicted.org/blind.php?x=y55xk6, and the field experiment preregistration can be found at http://aspredicted.org/blind.php?x=zu28g5.

Declaration of Competing Interest

None.

Acknowledgments

We thank Katherine Milkman and Jennifer Dannals for their insightful feedback on this work; the Behavior Change for Good team for their invaluable research and organizational support; Joseph Kay, Timothy Lee, Yeji Park, and Alex Rohe for their research support; and the Wharton Behavioral Lab, the Behavior Change for Good Initiative, and the Baker Retailing Center for providing financial support. This material is based upon work supported by an NSF Graduate Research Fellowship under Grant DGE-1845298. The views expressed here do not necessarily reflect the views of these entities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2020.104099.

References

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters, 80*(1), 123–129.

Apfelbaum, E. P., Sommers, S. R., & Norton, M. I. (2008). Seeing race and seeming racist? Evaluating strategic colorblindness in social interaction. *Journal of Personality and Social Psychology, 95*(4), 918–932.

Avery, D. R. (2003). Reactions to diversity in recruitment advertising—Are differences black and white? *Journal of Applied Psychology, 88*(4), 672–679.

Bell, J. M., & Hartmann, D. (2007). Diversity in everyday discourse: The cultural ambiguities and consequences of "happy talk". *American Sociological Review, 72*(6), 895–914.

Berdahl, J. L., & Min, J.-A. (2012). Prescriptive stereotypes and workplace consequences for east Asians in North America. *Cultural Diversity and Ethnic Minority Psychology, 18*(2), 141–152.

Bertrand, M., & Duflo, E. (2017). Field experiments on discrimination. In , *Vol. 1. Handbook of economic field experiments* (pp. 309–393). Elsevier.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review, 94*(4), 991–1013.

Bonam, C. M., Bergsieker, H. B., & Eberhardt, J. L. (2016). Polluting black space. *Journal of Experimental Psychology: General, 145*(11), 1561–1582.

U.S. Census Bureau. (2015). *New census bureau report analyzes U.S. population projections* (press release). U.S. Census Bureau https://www.census.gov/newsroom/press-releases/2015/cb15-tps16.html.

Chang, E. H., Milkman, K. L., Chugh, D., & Akinola, M. (2019). Diversity thresholds: How social norms, visibility, and scrutiny relate to group composition. *Academy of Management Journal, 62*(1), 144–171.

Chatman, J. A., & O'Reilly, C. A. (2004). Asymmetric reactions to work group sex diversity among men and women. *Academy of Management Journal, 47*(2), 193–208.

Cohn, N., & Quealy, K. (2020, June 10). *How public opinion has moved on black lives matter*. The New York Times. https://www.nytimes.com/interactive/2020/06/10/upshot/black-lives-matter-attitudes.html.

Craig, M. A., & Richeson, J. A. (2014). On the precipice of a "majority-minority" America: Perceived status threat from the racial demographic shift affects white Americans' political ideology. *Psychological Science, 25*(6), 1189–1197.

Danbold, F., & Huo, Y. J. (2015). No longer "all-American"? Whites' defensive reactions to their numerical decline. *Social Psychological and Personality Science, 6*(2), 210–218.

Danbold, F., & Unzueta, M. M. (2020). Drawing the diversity line: Numerical thresholds of diversity vary by group status. *Journal of Personality and Social Psychology, 118*(2), 283–306. https://doi.org/10.1037/pspi0000182.

Dezső, C. L., Ross, D. G., & Uribe, J. (2016). Is there an implicit quota on women in top management? A large-sample statistical analysis. *Strategic Management Journal, 37*(1), 98–115.

Ely, R. J., & Thomas, D. A. (2001). Cultural diversity at work: The effects of diversity perspectives on work group processes and outcomes. *Administrative Science Quarterly, 46*(2), 229–273.

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*(2), 175–191.

Fingerhut, H. (2018, June 14). *Most in US say growing racial and ethnic diversity makes country better*. Pew Research Center. https://www.pewresearch.org/fact-tank/2018/06/14/most-americans-express-positive-views-of-countrys-growing-racial-and-ethnic-diversity/.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878–902.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology, 117*(3), 522–559. https://doi.org/10.1037/pspa0000160.

Goff, P. A., Steele, C. M., & Davies, P. G. (2008). The space between us: Stereotype threat and distance in interracial contexts. *Journal of Personality and Social Psychology, 94*(1), 91–107.

Gomila, R. (2020). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General. Advance online publication*. https://doi.org/10.1037/xge0000920.

Guterl, F. (2014, October 1). Diversity in science: Where are the data? *Scientific American*. https://doi.org/10.1038/scientificamerican1014-40.

Handelsman, J., Cantor, N., Carnes, M., Denton, D., Fine, E., Grosz, B., Hinshaw, V., Marrett, C., Rosser, S., & Shalala, D. (2005). More women in science. *Science, 309*(5738), 1190–1191.

Hewlett, S. A., Marshall, M., & Sherbin, L. (2013). How diversity can drive innovation. *Harvard Business Review, 91*(12), 30.

Holger, D. (2019, October 26). *The business case for more diversity*. Wall Street Journal. https://www.wsj.com/articles/the-business-case-for-more-diversity-11572091200.

Hoogendoorn, S., Oosterbeek, H., & Van Praag, M. (2013). The impact of gender diversity on the performance of business teams: Evidence from a field experiment. *Management Science, 59*(7), 1514–1528.

Jones, S., & Donnelly, G. (2017, June 16). *Why we logged every fortune 500 company's diversity data, or lack thereof*. Fortune. http://fortune.com/2017/06/16/why-we-logged-every-fortune-500-companys-diversity-data-or-lack-thereof/.

Kawakami, K., Dunn, E., Karmali, F., & Dovidio, J. F. (2009). Mispredicting affective and behavioral responses to racism. *Science, 323*(5911), 276–278.

Kozlowski, S. W., & Bell, B. S. (2003). Work groups and teams in organizations. In *Handbook of psychology* (pp. 333–375).

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259–269.

Lount, R. B., Jr., Sheldon, O. J., Rink, F., & Phillips, K. W. (2015). Biased perceptions of racially diverse teams and their consequences for resource support. *Organization Science, 26*(5), 1351–1364.

MacInnis, C. C., Page-Gould, E., & Hodson, G. (2017). Multilevel intergroup contact and antigay prejudice (explicit and implicit) evidence of contextual contact benefits in a less visible group domain. *Social Psychological and Personality Science, 8*(3), 243–251.

Milkman, K. L., Akinola, M., & Chugh, D. (2015). What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *Journal of Applied Psychology, 100*(6), 1678–1712.

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology, 81*(1), 33–43.

Morton, F. S., Zettelmeyer, F., & Silva-Risso, J. (2003). Consumer information and discrimination: Does the internet affect the pricing of new cars to women and minorities? *Quantitative Marketing and Economics, 1*(1), 65–92.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109*(41), 16474–16479.

Plaut, V. C., Garnett, F. G., Buffardi, L. E., & Sanchez-Burks, J. (2011). "What about me?" perceptions of exclusion and Whites' reactions to multiculturalism. *Journal of Personality and Social Psychology, 101*(2), 337–353.

Rosette, A. S., Leonardelli, G. J., & Phillips, K. W. (2008). The White standard: Racial bias in leader categorization. *Journal of Applied Psychology, 93*(4), 758–777.

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues, 57*(4), 743–762.

Van Dijk, H., Van Engen, M. L., & Van Knippenberg, D. (2012). Defying conventional wisdom: A meta-analytical examination of the differences between demographic and job-related diversity relationships with performance. *Organizational Behavior and Human Decision Processes, 119*(1), 38–53.

Wilton, L. S., Sanchez, D. T., Unzueta, M. M., Kaiser, C., & Caluori, N. (2019). In good company: When gender diversity boosts a company's reputation. *Psychology of Women Quarterly, 43*(1), 59–72.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science, 330*(6004), 686–688.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036–1039.