"I'm going to choose a Hibble": Social and statistical reasoning in DEI contexts

Aarthi K. Popat (aarthi.popat@yale.edu)
Department of Psychology, 100 College Street
New Haven, CT 06510 USA

Yarrow Dunham (yarrow.dunham@yale.edu)
Department of Psychology, 100 College Street
New Haven, CT 06510 USA

Abstract

Diversity, equity, and inclusion (DEI) initiatives are widespread within progressive institutions. However, many such initiatives backfire because they diminish external perceptions of diverse hires' competence. Across four preregistered experiments, we develop a novel theory: that statistical reasoning about candidate competence is clouded in DEI cases by existing priors and stereotypes, as well as causal attributions about selector intention. Supporting this, we find that people do make logical statistical inferences with uniform populations and arbitrary selection processes. However, as people receive more information about the population and the selection process, their decisions move away from optimality. Our data suggest that this is a consequence of causal attribution, i.e., whether people attribute selections to competence or to group membership. Implications for DEI messaging and initiatives are discussed.

Keywords: statistical reasoning; causal reasoning; DEI; bias

Introduction

In 2020, President Biden vowed to select a woman as his running mate before choosing former Senator Kamala Harris. Upon taking office, Vice President Harris became simultaneously the first female, Black, and Asian person to occupy the position. Harris's selection was lauded by many as an important step toward representation in our government's highest branch (De Witte, 2021), but critics have since labeled her a "diversity hire," casting doubt on her merit and undermining her extensive qualifications (Herndon, 2023).

In general, formal and informal diversity, equity, and inclusion (DEI) efforts aim to increase the representation of historically marginalized groups in traditionally White and male fields (Sekaquaptewa et al., 2021; Settles et al., 2023). These efforts are well-intentioned: Visible gender representation in academic and workplace settings can reduce anti-woman bias (Fan et al., 2019), mitigate sex discrimination (Fine et al., 2020), and even increase young girls' motivation to pursue STEM fields (Shachnai et al., 2022; González-Pérez et al., 2020); conversely, under-representation can induce performance pressure and catalyze attrition for women and minority candidates (Cohen & Swim, 1995; Kanter, 1977).

However, many DEI efforts backfire because they call into question their targets' competence (e.g., Dover et al., 2020; Coate & Loury, 1993; Heilman & Welle, 2006); for example, DEI efforts may imply that diverse hires needed help to succeed, that they were not selected on the basis of merit, or that they had more opportunities than non-diverse counterparts (L. M. Leslie, 2019; Burnett & Aguinis, 2023). Existing research lacks a cognitive framework for understanding *how* these judgments come about. Here, we explore the cognitive and social factors that underlie negative judgments of successful candidates in DEI contexts.

First, we consider the impact of selection *process* on people's judgments of successful candidates. For example, President Biden explicitly subsetted the space of possible running mate candidates when he declared he would select a woman. We propose that in such DEI contexts, explicit population subsetting triggers basic statistical reasoning, which could affect judgments of candidate competence.

From early childhood, humans are statistical learners. Infants harness statistical reasoning to learn word meanings (Saffran & Kirkham, 2018), predict future events (Téglás et al., 2011; Denison & Xu, 2014), and understand object properties (Wu et al., 2011). Crucially, even eight-month-old infants can reason probabilistically (Xu & Garcia, 2008). Adults possess more sophisticated and complex statistical intuitions and harness them in everyday life. For example, adults can predict the emotions of others based on their prior emotional states (Thornton & Tamir, 2017). They can also invoke intuitive tenets of probabilistic reasoning to solve objective problems—for example, to determine lottery characteristics (Nisbett et al., 1983; Fong, 1983). Thus, people may correctly conclude that subsetting random populations has no mathematical effect on the probability of selecting a winning, or best, item or candidate. However, when asked to make judgments about people in non-random groupings, adults draw on a number of non-statistical heuristics (Nisbett et al., 1983; Fong, 1983). It is therefore plausible that statistical reasoning will not be the only process affecting people's judgments about selected candidates.

One such mechanism by which people scaffold statistical inferences about candidates' competence may be causal reasoning about *why* the selecting agent chose to subset the population. Reasoners readily and consistently ascribe intentions to other agents (Malle & Knobe, 1997; Dennet, 1987): "This person did that thing *because* ...". We constantly construct such causal hypotheses about the world around us (Kuhn, 1989; Nisbett & Ross, 1980; Klahr & Dunbar, 1988). When multiple possibilities for the outcome of an event exist (e.g., "selected for merit" or "selected for diversity"), success-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

ful learners evaluate evidence to discriminate a more probable hypothesis from less probable alternatives (Klahr et al., 1989; Klayman & Ha, 1989; Dougherty & Hunter, 2003). Expressed intent to select from a population subset (e.g., women) may constitute evidence for the possibility that a successful female candidate was selected for her gender, and not necessarily her merit. However, an alternative interpretation under these same conditions is that she was selected because women are equally or more competent than men. Why does general consensus favor the former inference, as suggested by Cauterucci, 2021?

Critically, a commitment to diversity can co-exist with the desire to select a maximally qualified candidate. That identity-based and merit-based inferences are often thought to be at odds with each other may implicate additional assumptions that are, to say the least, controversial and inaccurate (e.g., that there are not maximally qualified candidates in the target diverse pool). In DEI-type contexts, such assumptions may obfuscate objective causal reasoning and instead spur motivated reasoning. Reasoners who hold prior beliefs or biases might discredit evidence that *disconfirms* their priors to preserve their main hypothesis (Kunda, 1990; Lodge & Taber, 2000). For example, those who endorse stereotypes about women and minoritized people as less competent—which emerge early and persist into adulthood (Bian et al., 2017; Baharloo et al., 2022; S.J. Leslie et al., 2015)—may dismiss evidence that Vice President Harris is qualified to maintain the hypothesis that she was selected because she was a woman. In this way, learners "think what they feel" (Lodge & Taber, 2013). We propose that in DEI-type selection scenarios, people rely on their priors and existing biases when choosing which causal hypothesis to privilege. In this case, pre-existing stereotypes about women being less competent than men may block the possibility that President Biden simply favors women on competence-related dimensions. Other gendered and racial stereotypes, which situate White men as more competent than women and minoritized people (Biernat & Kobrynowicz, 1997), may also scaffold people's inferences about selecting agents' intentions.

The current study elucidates the cognitive and social factors that underlie judgments about candidate competence in DEI contexts. In the first of four experiments, we begin with more objective cases in which people's existing priors and biases are unlikely to influence their reasoning. In the subsequent experiments, we layer on additional complexity. This allows us to quantify the contribution of "purely" statistical inferences (including the possibility of biased statistical reasoning; Tversky & Kahneman, 1974; Kahneman & Tversky, 1982) to people's reasoning about selected candidates in such contexts, before adding some of the additional complexity present in real-world DEI cases.

In each study, we present a vignette where an agent selects an item or candidate from a population. Across studies, we manipulate a number of different factors. First, we vary the selection process used by the agent to determine how an agent's *intention* to select from a subset of a population impacts people's inferences about the selected candidate. People may interpret intentional subsetting of a population by a trait or characteristic as indicative of an agent's goal (e.g., the agent prefers this subset). Second, we manipulate population diversity to scaffold inferences about *why* the agent prefers a given subset. In particular, we expect people to ascribe more meaning to the selection of candidates from *diverse* (versus homogeneous, or uniform) populations. Finally, across studies, we will progressively increase the animacy and social salience of the population. We expect participants to make richer inferences about why a candidate was chosen as these factors increase. The series of studies culminates with a vignette about two novel social groups, one dominant and one historically marginalized, to more closely parallel current DEI contexts.

Experiment 1

Exp. 1 (pre-registered) assessed people's pure statistical inferences about a homogeneous population with no natural diversity.

Method

295 participants completed Exp. 1 on Prolific. 4 participants were excluded due to failing attention checks, yielding a final sample of 291 participants (gender: 150 women, 134 men, 6 non-binary, 1 did not report; race: 206 White, 28 Asian or Pacific Islander, 25 Multiracial, 20 Latine/x, 9 Black, 1 Native American or Alaskan Native, 2 did not report).

We presented participants with a simple vignette: a cartoon agent must choose a plastic egg from two baskets, both with an equal number of visually identical eggs. Our paradigm was a 2x3 design crossing population diversity (uniform or diverse) with subset type (intentional-subset, no-subset, or random-subset). In the uniform condition, all eggs were identical. In the diverse condition, half of the eggs were orange, and half were purple. We told participants that there were 30 eggs in each basket, and that ten eggs had $10 hidden inside.

Participants then watched the agent's selection process. In the random-subset condition, the agent flipped a coin to decide from which subset to select an egg (uniform: the basket on the left; diverse: purple eggs). In the intentional-subset condition, the agent intentionally decided to choose an egg from one of the subsets (left basket; purple eggs). In the no-subset condition, the agent selected an egg from the full population (both baskets; both colors). Importantly, the *mathematical* probability of selecting the best candidate remains the same across subset procedures.

After the agent made a selection, we asked participants the following questions to assess their inferences about the chosen object's objective and comparative quality: (1) "How likely do you think it is that [the agent] will be happy with the chosen egg?" (quality likelihood estimate) and (2) "How likely do you think it is that the egg [the agent] chose had $10 inside?" (superiority likelihood estimate). Here and across studies we focus on the superiority likelihood measurement,

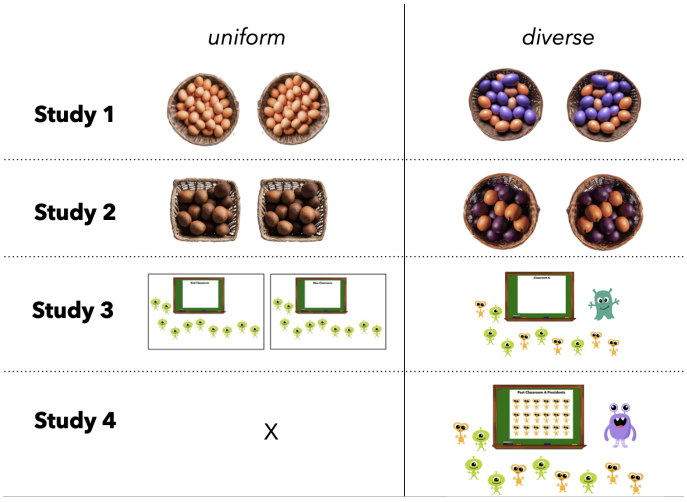

Figure 1: Populations were either uniform or diverse in Studies 1-4.

as results for both measures were similar and judgments of superiority are more directly linked to statistical judgments.

We hypothesized that population diversity (diverse vs. uniform) would impact judgments of item superiority in the intentional-subset condition. We also predicted that when participants in the diverse condition saw an agent intentionally choose an egg from one of the subsets (intentional-subset condition), they would rate object superiority likelihood higher.

Results and Discussion

All results were preregistered unless stated as exploratory. Models were linear regressions predicting superiority likelihood estimates alongside posthoc pairwise comparisons.

Diversity. As predicted, we observed an effect of diversity on participants' superiority likelihood estimates in the intentional-subset condition ($t(96) = 3.95, p < 0.001$). Conversely, participants in the uniform condition made equivalent superiority likelihood estimates across subset types ($p > 0.05$). Further exploratory analyses showed that, compared with uniformity and across all subset types, diversity increased participants' ratings of the likelihood that the selected egg had \$10 in it ($t(289) = 7.64, p < 0.001$). Because the probability of success is equivalent regardless of diversity or subset type, some other inference must have driven participants' skewed ratings in the diverse condition.

Subset type. As predicted, there was no effect of subset type in the uniform condition ($ps > 0.05$). However, contrary to our hypothesis, there was also no effect of subset type on superiority likelihood estimates in the diverse condition, notably between the no-subset and intentional-subset selection types ($t(143) = 0.67, p = 0.78$). One possible explanation for this effect is that when population diversity increases, both expressing intent to select from a subset and simply selecting from that subset convey a level of intentionality.

Exp. 1 demonstrates that when there are otherwise no differences in a population or context, participants make accurate likelihood estimates about the superiority of a chosen item. However, perceptual diversity can inflate these estimates—regardless of selection process. Thus, these results point towards *non-statistical* influences on inferences about selection. In Exp. 2, we increase the *natural* diversity of the population to elucidate the non-statistical factors contributing to judgments of selected candidates.

Experiment 2

Exp. 2 (pre-registered) tested how population subsetting and *greater* population diversity would impact inferences about a selected item from a population of salak fruit.

Method

306 participants completed Exp. 2 on Prolific. One participant was excluded due to attention check failure, resulting in a final sample of 305 participants (gender: 138 women, 156 men, 8 non-binary, 3 did not report; race: 199 White, 33 Black, 24 Asian or Pacific Islander, 23 Multiracial, 20 Latine/x, 2 Middle Eastern or North African, 1 Native American or Alaskan Native, and 3 did not report).

Exp. 2 was identical to Exp. 1 except that the agent chose from two baskets of salak fruit instead of plastic eggs. We used this unusual fruit because it is less known than more common fruit (e.g., apples, oranges) and displays natural variation in a way that the identical plastic eggs did not. This opened the possibility of richer causal inferences about agent selection rationale.

We anticipated that here, like in the diverse condition in Exp. 1, participants' ratings would diverge from statistical optimality. We predicted a positive effect of intention on judgments of superiority ("How likely do you think it is that the chosen salak fruit was the best salak fruit?"), particularly within the diverse condition. We also expected a difference in superiority likelihood estimates across diversity levels between the intentional-subset and no-subset conditions.

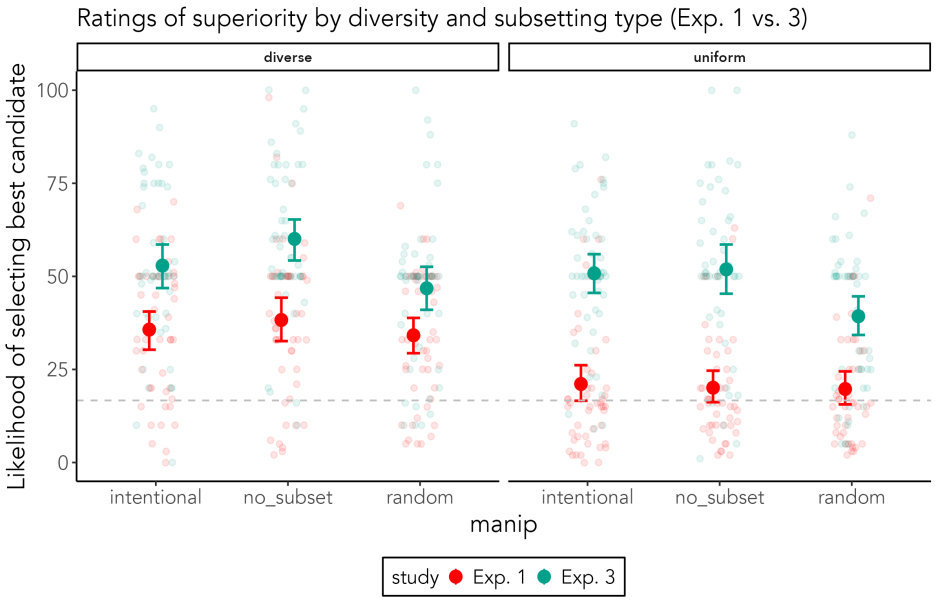

Figure 2: Participants in Exp. 1, who rated the likelihood of an egg being one that has $10 in it, computed and adhered to accurate probabilities for success in the uniform condition (see dashed line at the probability of selection, p($10) = 1/6). Given more information—in the diverse condition of Exp. 1 and in all conditions of Exp. 3—participants' ratings rise inaccurately.

Results and Discussion

Diversity. Contrary to our prediction, diversity alone did not impact participants' estimates of salak fruit superiority likelihood ($t(299)$ = -0.71, p = 0.48). It is possible that the natural heterogeneity of the salak fruit population (which, again, is greater than that of Exp. 1's plastic eggs) rendered artificial diversity a less powerful manipulation.

Exploratory analyses comparing superiority likelihood estimates between Exp. 1 and Exp. 2 support this possibility: Superiority likelihood estimates in Exp. 2 increased from the uniform condition in Exp. 1, whether a brown (uniform; $F(592) = 19.12, p < 0.001$) or purple (diverse; $F(594) = 21.32, p < 0.001$) fruit was selected. Given that the actual superiority likelihood estimate in Exp. 2 is *lower* than it was in Exp. 1, it is reasonable to conclude that participants in Exp. 2 deviate from pure statistical reasoning in this context.

Subset type. Population subset type (intentional, random, or none) predicted superiority likelihood ratings across levels of population diversity ($F(3,301) = 3.23, p = 0.02$). Crucially, consistent with our hypothesis, selector intentionality predicted higher superiority likelihood estimates in the diverse condition($t(303)$ = 2.04, p = 0.04).

These data suggest that natural *and* intra-population heterogeneity affect statistical inferences about selected item traits. Within the salak fruit population, the potential for added diversity (e.g., bruised fruit; sweet fruit) may have expanded the available possibilities for *why* the fruit was selected. This would allow people to make richer inferences about agent intentionality—*why* an agent selected the specific fruit—and therefore, exactly *how good* that fruit might be.

It remains unclear, however, how these mechanisms might scaffold inferences about agentic populations. Thus, Exp. 3 introduces a population of two novel groups to add a more meaningful layer of natural heterogeneity.

Experiment 3

Exp. 3 (pre-registered) assessed whether more meaningful natural variation—in this case, membership in a novel social group—would induce richer causal inferences about the quality of a chosen candidate.

Method

We recruited 301 adult participants on Prolific. Seven participants were excluded due to attention check failure, resulting in a final sample of 294 participants (gender: 144 women, 146 men, 2 non-binary, 2 did not report; race: 207 White, 27 Multiracial, 25 Black, 19 Asian or Pacific Islander, 11 Latine/x, 1 Native American or Alaskan Native, 4 did not report).

In Exp. 3, the selecting agent is trying to select a class (diverse condition) or school (uniform condition) president. In the diverse condition, two groups, Hibbles and Glerks (see Figure 1), are in the same school classroom. In the uniform condition, there are only Hibbles at the school, but they are split into the Red Classroom and the Blue Classroom.

In the random-subset condition, the selecting agent (classroom teacher or school principal) flips a coin, and the result reveals which subset to choose from: Hibbles (vs. Glerks; diverse) or the Red Classroom (vs. Blue Classroom; uniform). The agent then chooses a student from that subset. Participants in the intentional-subset condition watched the agent decide to choose a student from one of those subsets. In the

no-subset condition, the agent chose a student from the full population (Hibbles and Glerks; both classrooms). Participants then rated the likelihood that the chosen student (1) would be a good class president and (2) was the best student for the job (which we focus on here). Finally, to assess the extent to which priors about social group stratification impacted participants' judgments of the selected candidate, we administered a social dominance orientation (SDO; Pratto et al., 1994) questionnaire.

We predicted effects of intentionality and diversity, such that participants in diverse condition who witnessed intentional subsetting would rate the selected candidate higher than participants in the uniform condition and participants who witnessed random subsetting. We also predicted that participants who saw no subsetting would rate the chosen candidate higher in the diverse (vs. uniform) condition.

Results and Discussion

Diversity. Contrary to our predictions, we did not find an effect of diversity on ratings of candidate superiority in the intentional-subset condition ($t(101) = 0.52, p = 0.61$). However, exploratory analyses showed that diversity positively impacted ratings of candidate superiority across manipulations ($t(290) = 2.450, p = 0.01$).

Subset type. As predicted, in the diverse condition, participants who saw the no-subset manipulation rated candidate superiority higher than participants who saw the random-subset manipulation ($t(145) = 3.14, p = 0.006$). However, there was no difference in the diverse condition between intentional-subset and random-subset participants' ratings of candidate superiority ($t(145) = 1.44, p = 0.32$).

One possible explanation for this finding is that participants have made opposing inferences: Some participants may have concluded that intentional subsetting would increase candidate quality and others that it would decrease candidate quality. This possibility would implicate participant's prior social beliefs as a driving mechanism for their superiority likelihood estimates. If so, we might observe reliable individual differences in judgments.

SDO. Exploratory analyses revealed a null effect of SDO score on participants' overall judgments of candidate superiority ($t(290) = -0.71, p = 0.48$). However, an interaction between SDO score and intentional subsetting predicted higher ratings of candidate superiority ($t(288) = 2.02, p = 0.04$). In other words, people with higher social dominance scores—those who endorse more rigid social strata—rated the selected candidate's likelihood of superiority higher when the candidate was selected via an intentional subsetting process.

The interaction between intentional subsetting and high SDO corroborates the possibility that the null subset effects are due to opposing inferences: It suggests an influence of individual beliefs about the impact of population subsetting. Also supporting this theory is that competence ratings between the intentional- and random-selection subset types within the uniform condition were essentially equivalent: It is plausible that the number of possible causal inferences about selection rationale compounded in Exp. 3 and produced a noisier distribution of superiority ratings across subset types. Because of this, it is uncertain whether participants were merely failing to make statistically viable competence evaluations at all or whether our procedure induced different effects in different participants (DEI-based, merit-based, or some combination), supported by different real-world priors. We thus propose that causal reasoning governs these effects. Exp. 4 explores the role of causal reasoning in scaffolding judgments of selected candidates by introducing a more explicit paradigm that mirrors a real-world DEI context.

Experiment 4

In Exp. 4 (pre-registered), we added additional context about the historical status of Hibbles and Glerks to more closely parallel the real world. We tested the effect of this context and examine participants' causal attributions for selection.

Method

We recruited 300 adult participants on Prolific. Two participants were excluded due to attention check failure, resulting in a final sample of 298 participants (gender: 138 women, 149 men, 6 non-binary, 5 did not report; race: 201 White, 36 Black, 26 Asian or Pacific Islander, 20 Multiracial, 13 Latine/x, 2 did not report).

In this paradigm, we described more substantial differences between Hibbles and Glerks. Critically, because we are now working with more meaningful social groups, we removed the artificial diversity component. We told participants that in the past, only Glerks were able to attend school. A few years ago, Hibbles joined, and now there are both Hibbles and Glerks at school and in Classroom A (see Figure 1). However, all the past classroom presidents have been Glerks, even since Hibbles have been allowed to attend school. There is a new teacher (Teacher Ro) in Classroom A. The teacher is supposed to select a new class president. In the intentional-subset condition, participants heard that Teacher Ro will select a Hibble for class president. In the no-subset condition, participants heard that Teacher Ro will select a student. In the random condition, the result of a coin flip tells Teacher Ro to select a Hibble. Moreover, participants also reported *why* they thought Teacher Ro made their selection and how much they liked Teacher Ro.

We predicted a negative effect of intentional-subsetting and no-subsetting on judgments of candidate superiority and candidate quality, anticipating that participants would view any selection of a marginalized candidate as a DEI-type initiative. We also predicted that superiority likelihood estimates would be lower under Exp. 4's intentional-subsetting and no-subsetting manipulations than under the same manipulations in Exp. 3.

Results and Discussion

Subset type. Contrary to our predictions, the no-subset and intentional-subset manipulations both predicted higher rather than lower superiority ratings (no-subset: $t(295) = 6.94, p <$

0.001; intentional-subset: $t(295) = 5.65, p < 0.001$), corroborated by preregistered planned contrasts.

Comparison with Exp. 3. Against our predictions, participants in the intentional-subset condition in Exp. 4 also rated candidate superiority higher compared to the diverse condition in Exp. 3 ($t(196) = 4.14, p < 0.001$). Participants in the no-subset condition in Exp. 4 also yielded higher candidate superiority ratings in the diverse condition ($t(196) = 3.96, p < 0.001$).

SDO. Crucially, controlling for condition, exploratory analyses revealed that participants' SDO scores predicted ratings of superiority ($t(294) = 2.27, p = 0.02$). Specifically, an interaction between no subsetting and SDO score predicted lower ratings of candidate superiority ($t(292) = -2.38, p = 0.02$). In this case, when participants with high social dominance scores watch the selection of a marginalized candidate, they believe it is less likely that the candidate was the best one.

Causal attribution. We qualitatively coded participants' causal attributions for Teacher Ro's selection as merit-based (*merit*), group membership-based (*DEI*), or merit- and group membership-based (*both*). Exploratory analyses showed that participants who believed Teacher Ro's selection was based on merit estimated a higher likelihood of selected candidate superiority than participants who Teacher Ro selected based on group membership ($t(232) = 5.94, p < 0.001$) or even both attributes ($t(232) = 3.76, p < 0.001$).

Taken together, the results of Exp. 4 suggest that judgments of candidate competence in this DEI-adjacent scenario are scaffolded by participants' prior beliefs and their causal attributions. As people learn context about the novel groups that more closely parallels real-world social groups (e.g., men and women), their judgments about DEI interventions vary as a function of their broader views about social hierarchy as well as their causal reasoning about the successful candidate's selection. With this in mind, it is possible that participants are purposely skewing their competence ratings to express broader support for (or disagreement with) DEI initiatives. Providing even stronger context information (e.g., cues to the presence of an ability-related stereotype), could elucidate these judgments and their relationship with SDO.

General Discussion

Certain diversity, equity, and inclusion efforts involve explicitly signaling the intention to select candidates from a demographic subset of the candidate population (e.g., selecting from group of only women). While well-intentioned, these initiatives have been shown to backfire by calling into question their targets' competence. We elucidate the cognitive and social factors that scaffold these inferences: In particular, we find that statistical reasoning about population subsetting is clouded by social priors and causal reasoning about *selector intent*. Across four preregistered studies, we assessed beliefs about selected candidate superiority under several increasingly complex vignettes. We found that reasoners are capable of making basic statistical inferences about superiority likelihood in purely random, objective cases. However, this capacity is affected by social and causal inputs as people receive more social and contextual information.

Populations with low heterogeneity—like the plastic eggs in Exp. 1—constrain the number of inferences people may make about them. Consistent with literature that suggests reasoners struggle to employ statistical reasoning in more subjective contexts, complete population uniformity allowed for accurate statistical inferences about candidate selection. However, when even *artificial* variation was introduced in this random context (i.e., the eggs became orange and purple), people's ratings of candidate superiority displayed bias.

This trend held under different subset types and even with objects with more natural diversity, like fruit. Crucially, people's inferences became noisier and more complex when they were reasoning about animate populations (i.e., Hibbles and Glerks) in meritorious contexts (i.e., in a classroom). Ratings were affected by social dominance orientation and causal reasoning about *why* a candidate was selected, indicating that individual differences in social cognition (e.g., prior beliefs, causal inference) may be responsible for inconsistent results in more convoluted contexts. It is also possible that people used the opportunity to rate the selected candidate to express their attitudes toward the DEI initiative itself. In other words, someone who supports DEI initiatives may have overcorrected their ratings of candidate superiority to express their support, while someone against DEI initiatives may have rated the chosen candidate lower. Ongoing work aims to validate this possibility by manipulating the status of the selected candidate (i.e., whether a Hibble or Glerk is selected).

Across studies, we found evidence for inaccurate statistical inferences driven by social cognitive mechanisms. As our experimental paradigm got closer to mirroring a real-world DEI initiative, people's inferences diversified. Crucially, the novel groups in our paradigm—even with context about their past marginalization—cannot replicate the implicit and nuanced inferences people make about gender and race in the real world. Future work will investigate whether more deeply held automatic priors—e.g., gender stereotypes—more universally skew statistical reasoning in DEI contexts.

Our findings build a framework of social and cognitive factors that shape judgments about diverse candidates selected under DEI messaging. While statistical reasoning offers a foundation for assessing candidate selection processes, social cognition—shaped by stereotypes, priors, and causal inferences about selectors—complicates this assessment in predictable ways. Thus, it is not enough to simply implement policies aimed at increasing diversity. Rather, it is critical that DEI initiatives are designed with an acute awareness of the cognitive underpinnings that shape perceptions of merit. Our future work will investigate exactly which DEI messages genuinely enhance inclusivity without inadvertently reinforcing the very biases they seek to eliminate.

Acknowledgments

We thank members of the Social Cognitive Development Lab for input and feedback.

References

Baharloo, R., Fei, X., & Bian, L. (2022). The development of racial stereotypes about warmth and competence.

Bian, L., Leslie, S.-J., & Cimpian, A. (2017, January). Gender stereotypes about intellectual ability emerge early and influence childrenÄôs interests. *Science*, *355*(6323), 389–391. Retrieved 2024-02-01, from https://www.science.org/doi/10.1126/science.aah6524 (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.aah6524

Biernat, M., & Kobrynowicz, D. (1997). Gender-and race-based standards of competence: lower minimum standards but higher ability standards for devalued groups. *Journal of personality and social psychology*, *72*(3), 544.

Burnett, L., & Aguinis, H. (2023, November). How to prevent and minimize DEI backfire. *Business Horizons*. Retrieved 2023-12-20, from https://www.sciencedirect.com/science/article/pii/S0007681323001179 doi: 10.1016/j.bushor.2023.11.001

Cauterucci, C. (2021, Jun 25). *Kamala harris has been set up to fail.* https://slate.com/news-and-politics/2021/06/kamala-harris-set-up-to-fail.html. (Accessed: 2024-02-01)

Coate, S., & Loury, G. C. (1993). Will Affirmative-Action Policies Eliminate Negative Stereotypes? *The American Economic Review*, *83*(5), 1220–1240. Retrieved 2023-12-20, from https://www.jstor.org/stable/2117558 (Publisher: American Economic Association)

Cohen, L. L., & Swim, J. K. (1995, September). The Differential Impact of Gender Ratios on Women and Men: Tokenism, Self-Confidence, and Expectations. *Personality and Social Psychology Bulletin*, *21*(9), 876–884. Retrieved 2024-02-01, from https://doi.org/10.1177/0146167295219001 (Publisher: SAGE Publications Inc) doi: 10.1177/0146167295219001

Denison, S., & Xu, F. (2014, March). The origins of probabilistic inference in human infants. *Cognition*, *130*(3), 335–347. Retrieved 2024-01-27, from https://www.sciencedirect.com/science/article/pii/S0010027713002370 doi: 10.1016/j.cognition.2013.12.001

Dennet, D. (1987). True believers. *Intentional Stance*, *13*, 36.

De Witte, M. (2021, Feb 12). *Breaking barriers: Madame vice president kamala harris.* Stanford News. Retrieved from https://news.stanford.edu/2020/12/11/breaking-barriers-madame-vice-president-kamala-harris/ (Accessed: yyyy-mm-dd)

Dougherty, M. R. P., & Hunter, J. E. (2003, July). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica*, *113*(3), 263–282. Retrieved 2023-12-20, from https://www.sciencedirect.com/science/article/pii/S0001691803000337 doi: 10.1016/S0001-6918(03)00033-7

Dover, T. L., Kaiser, C. R., & Major, B. (2020). Mixed Signals: The Unintended Effects of Diversity Initiatives. *Social Issues and Policy Review*, *14*(1), 152–181. Retrieved 2023-12-20, from https://onlinelibrary.wiley.com/doi/abs/10.1111/sipr.12059 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/sipr.12059) doi: 10.1111/sipr.12059

Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019, February). Gender and cultural bias in student evaluations: Why representation matters. *PLOS ONE*, *14*(2), e0209749. Retrieved 2023-12-22, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0209749 (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0209749

Fong, G. T. (1983). The Effects of Statistical Training on Thinking Everyday Problems.

González-Pérez, S., Mateos de Cabo, R., & Sáinz, M. (2020). Girls in STEM: Is It a Female Role-Model Thing? *Frontiers in Psychology*, *11*. Retrieved 2023-12-22, from https://www.frontiersin.org/articles/10.3389/fpsyg.2020.02204

Heilman, M. E., & Welle, B. (2006). Disadvantaged by Diversity? The Effects of Diversity Goals on Competence Perceptions1. *Journal of Applied Social Psychology*, *36*(5), 1291–1319. Retrieved 2023-12-20, from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0021-9029.2006.00043.x (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0021-9029.2006.00043.x) doi: 10.1111/j.0021-9029.2006.00043.x

Herndon, A. W. (2023, October 10). *In search of kamala harris.* https://www.nytimes.com/2023/10/10/magazine/kamala-harris.html. (Accessed: 2024-02-01)

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, *11*(2), 123–141.

Kanter, R. M. (1977). *Men and women of the corporation*. New York: Basic Books. Retrieved 2024-02-01, from https://catalog.hathitrust.org/Record/000169677

Klahr, D., & Dunbar, K. (1988). Dual Space Search During Scientific Reasoning. *Cognitive Science*, *12*(1), 1–48. Retrieved 2024-02-01, from https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1201_1 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1201_1) doi: 10.1207/s15516709cog1201_1

Klahr, D., Dunbar, K., & Fay, A. L. (1989). DESIGNING GOOD EXPERIMENTS TO TEST BAD HYPOTHESES.

Klayman, J., & Ha, Y.-W. (1989, July). Hypothesis Testing in

Rule Discovery: Strategy, Structure, and Content. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 596–604. doi: 10.1037/0278-7393.15.4.596

Kuhn, D. (1989). Children and Adults as Intuitive Scientists.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. (Place: US Publisher: American Psychological Association) doi: 10.1037/0033-2909.108.3.480

Leslie, L. M. (2019, July). Diversity Initiative Effectiveness: A Typological Theory of Unintended Consequences. *Academy of Management Review*, *44*(3), 538–563. Retrieved 2023-12-20, from https://journals.aom.org/doi/abs/10.5465/amr.2017.0087 (Publisher: Academy of Management) doi: 10.5465/amr.2017.0087

Leslie, S. J., Cimpian, A., Meyer, M., & Freeland, E. (2015, January). *Expectations of brilliance underlie gender distributions across academic disciplines.* Retrieved 2024-02-01, from https://www.science.org/doi/full/10.1126/science.1261375?casa_token=rPxAEXsxPYIAAAAA:MRUYaj5zNtFANTtQm7jc301sDV2KMaiBq4gsXmAcSJRn549ybqJqCBCiYQXsQNh7UirM6ZD2sFNxw88

Lodge, M., & Taber, C. (2000). Three steps toward a theory of motivated political reasoning. *Elements of reason: Cognition, choice, and the bounds of rationality*, *183*.

Lodge, M., & Taber, C. S. (2013). *The rationalizing voter*. Cambridge University Press.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of experimental social psychology*, *33*(2), 101–121.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological review*, *90*(4), 339.

Nisbett, R. E., & Ross, L. (1980). Human inference: Strategies and shortcomings of social judgment.

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of personality and social psychology*, *67*(4), 741.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant Statistical Learning. *Annual Review of Psychology*, *69*(1), 181–203. Retrieved 2024-01-27, from https://doi.org/10.1146/annurev-psych-122216-011805 (_eprint: https://doi.org/10.1146/annurev-psych-122216-011805) doi: 10.1146/annurev-psych-122216-011805

Sekaquaptewa, D., Takahashi, K., Malley, J., Herzog, K., & Bliss, S. (2019, February). An evidence-based faculty recruitment workshop influences departmental hiring practice perceptions among university faculty. *Equality, Diversity and Inclusion: An International Journal*, *38*. doi: 10.1108/EDI-11-2018-0215

Settles, I. H., Linderman, J. J., Rivas-Drake, D., Saville, J., & Conner, S. (2023, September). Three strategies for engaging campus leaders in transformative initiatives to retain faculty of color. *Journal of Diversity in Higher Education*. Retrieved 2024-02-01, from http://doi.apa.org/getdoi.cfm?doi=10.1037/dhe0000511 doi: 10.1037/dhe0000511

Shachnai, R., Kushnir, T., & Bian, L. (2022, November). Walking in Her Shoes: Pretending to Be a Female Role Model Increases Young Girls' Persistence in Science. *Psychological Science*, *33*(11), 1818–1827. Retrieved 2024-02-01, from https://doi.org/10.1177/09567976221119393 (Publisher: SAGE Publications Inc) doi: 10.1177/09567976221119393

Thornton, M. A., & Tamir, D. I. (2017). Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences*, *114*(23), 5982–5987.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, *185*(4157), 1124–1131.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011, May). Pure Reasoning in 12-Month-Old Infants as Probabilistic Inference. *Science*, *332*(6033), 1054–1059. Retrieved 2024-01-07, from https://www.science.org/doi/full/10.1126/science.1196404 (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.1196404

Wu, R., Gopnik, A., Richardson, D. C., & Kirkham, N. Z. (2011). Infants learn about objects from statistics and people. *Developmental Psychology*, *47*(5), 1220–1229. Retrieved 2024-02-01, from http://doi.apa.org/getdoi.cfm?doi=10.1037/a0024023 doi: 10.1037/a0024023

Xu, F., & Garcia, V. (2008, April). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012–5015. Retrieved 2024-01-07, from https://www.pnas.org/doi/abs/10.1073/pnas.0704450105 (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.0704450105