Improving misrepresentations amid unwavering misrepresenters

Martin L. Jönsson1 · Jakob Bergman2

Received: 18 May 2021 / Accepted: 15 May 2022
© The Author(s) 2022

Abstract
In recruitment, promotion, admission, and other forms of wealth and power apportion, an evaluator typically ranks a set of candidates in terms of their competence. If the evaluator is prejudiced, the resulting ranking will misrepresent the candidates’ actual rankings. This constitutes not only a moral and a practical problem, but also an epistemological one, which begs the question of what we should do—epistemologically—to mitigate it. In a recent paper, Jönsson and Sjödahl in [Episteme 14(4):499–517, 2017], argue that the epistemic problem can be fruitfully addressed by way of a novel statistical method that changes the products of biased behaviour, i.e. the rankings themselves, rather than the biased persons. Jönsson and Sjödahl’s pioneering proposal is a both a welcome addition to the literature on implicit bias, due the problems with existing implicit bias interventions [see e.g. Lai et al. in J Exp Psychol Gen 143:1765–1785; J Exp Psychol Gen 145(8):1001–1016, 2014; 2016; Forscher et al. in J Person Soc Psychol 117(3):522–559, 2019] but also to the literature on prejudice more generally, where many proposed prejudice-reduction strategies enjoy less than adequate empirical support [Paluck and Green in Ann Rev Psychol 60(1):339–367, 2009]. Their proposal, however, needs supplementation in two ways: the circumstances that must hold in order for it to work needs to be refined, and their claim that it works as intended in these circumstances needs to be validated. We argue that four of Jönsson and Sjödahl’s method’s presumed presuppositions can be weakened, but needs to be supplemented by two additional assumptions, overlooked by Jönsson and Sjödahl. Moreover, we demonstrate that the method does work as intended by way of a statistical simulation.

✉ Martin L. Jönsson
martin.jonsson@fil.lu.se

Jakob Bergman
jakob.bergman@stat.lu.se

1 Department of Philosophy, Lund University, LUX, Helgonavägen 3, 221 00 Lund, Sweden
2 Department of Statistics, Lund University, Tycho Brahes väg 1, 223 63 Lund, Sweden

Published online: 30 June 2022
Keywords Prejudice · Epistemology · Prejudice interventions · Implicit bias · Ranking · Veracity

1 Introduction

In recruitment, promotion, admission, and other forms of wealth and power apportion, an evaluator typically ranks a set of candidates in terms of their competence. If the evaluator is prejudiced, the resulting ranking will misrepresent the candidates’ actual rankings. This constitutes not only a moral and a practical problem, but also an epistemological one, which begs the question of what we should do—epistemologically—to mitigate it.

In a recent paper, Jönsson and Sjödahl (2017) suggest that we can tackle this problem, at least in the context of implicit bias, by estimating an evaluator’s degree of bias from their history of evaluations, and then use this estimate to improve the veracity of the ranking itself, after it has been produced. This contrasts with the overwhelmingly most common, preventive, strategy in the literature on deprejudicing (see Madva, 2020 for an overview) inherent in both structural interventions that change the circumstances under which the ranking takes place (e.g. the introduction of anonymization, or the requirement that criteria–based decision making is used) in order to prevent prejudiced behavior (such as prejudiced ranking), and individual interventions that attempt to make the evaluator less prejudiced and thereby prevent further prejudiced behavior.

Jönsson and Sjödahl (2017)’s pioneering approach looks—as they suggest—attractive in the context of implicit bias, since existing implicit bias interventions tend not to have long term effects and seem to have limited impact on prejudiced behaviour (see e.g. Greenwald, et al., 2009; Oswald et al., 2013; Lai et al., 2014, 2016; Forscher et al., 2019), but can well be employed in the context of prejudice more generally, where many proposed prejudice–reduction strategies enjoy less than adequate empirical support (Paluck & Green, 2009). Their approach promises deprejudicing both beyond the scope of our existing prejudice interventions, and improved, complementary, deprejudicing within that scope (cf. Jönsson forthcoming). Such a promise warrants careful investigation.

In what follows we will argue that although Jönsson and Sjödahl’s proposal has much to recommend it, it requires supplementation in two main ways—the circumstances that must hold in order for it to work needs to be refined and spelled out, and that the method actually works as intended in these circumstances needs to be validated. We will proceed by first describing the ranking problem that Jönsson and Sjödahl set out to solve in a little more detail (Sect. 1), and then describe their general approach to solving it, as well as the specific method—dubbed GIRU—which they employ (Sect. 2). We will then (in Sects. 3 and 4) carefully investigate GIRU’s presuppositions—what needs to be true in order for GIRU to be able to solve the
epistemological problem. We will show that four out of five of GIRU’s presumed (by Jönsson and Sjödahl) presuppositions can be weakened, but also, that the method presupposes two additional assumptions, overlooked by Jönsson and Sjödahl. We will conclude the paper (Sect. 5) with a validation of GIRU by means of a statistical simulation which demonstrates that the method does work as Jönsson and Sjödahl intended, given that the aforementioned presuppositions obtain.

2 The problem

Consider the following situation. A prejudiced evaluator $E$ is responsible for assigning a score to jobseekers to indicate that person’s degree of competence (with respect to some form of specific competence, e.g. some form of technical competence, communication skill, problem solving ability, or any other form of measurable competence). Assume that $E$ produces the assignments on the left, but that the jobseekers’ actual competence scores correspond to the assignments on the right.

|   | $r_0$ |   | $r_{0^*}$ |
|---|-------|---|-----------|
| Mike | 8     |   | Mike | 8     |
| Mark | 7     |   | Mark | 7     |
| Felicia | 6 |   | Felicia | 9 |
| Gordon | 4     |   | Gordon | 4 |
| Sarah | 3     |   | Sarah | 6     |
| Latifah | 3 |   | Latifah | 6 |

The first list—$r_0$—clearly misrepresents the people being evaluated.

Since the misrepresentation is a ranking, the degree of misrepresentation—or conversely, veracity—can be easily measured precisely. One natural way to do this is in terms of the rank order correlation between $r_0$ and $r_{0^*}$, i.e. the degree to which the two rankings order the people being ranked in the same way. This measure can be captured by Spearman’s rank correlation coefficient, a number between 1 (signifying a perfect positive correlation) and –1 (signifying a perfect negative correlation). According to this measure, $r_0$’s veracity amounts to 0.58, which is thus some ways off from a perfect representation.

1 Here and in what follows, a ranking will be construed broadly in the usual way as a total preorder over a set of people i.e. an ordering which relates all members of the set, but allows for two or more members to have the same rank. It is thus a very widely applicable concept, applying as it does, for instance, to hiring situations featuring only two people (who might either be evaluated as different or equivalent).

2 Jönsson and Sjödahl are not explicit about how they measure veracity, but doing so in terms of a rank order correlation is very much in the spirit of their proposal, since they ultimately care only about the rank order of the evaluations (cf. Jönsson and Sjödahl: 513). Other natural measures that could have been used include the mean absolute difference (between the scores in $r_1$ and $r_{1^*}$) or the mean square error.
The challenging epistemological problem then, is how to reliably improve the veracity of $r_0$ (and rankings like it) without having recourse to $r_0^\ast$.

Although there are many strategies for coming to terms with different forms of prejudice (see e.g. Paluck & Green, 2009, and Madva, 2020 for good overviews), most of them—both individual and structural strategies—attempt to prevent the prejudice from manifesting, either by making people less prejudiced, or preventing their prejudice (via e.g. anonymization or criteria based decision making) from triggering. Jönsson and Sjödahl’s (2017) approach, which is the focus of this paper, is interestingly different. It attempts to use an evaluator’s history of evaluations to estimate their degree of prejudice, and then use this estimate to improve the veracity of rankings such as $r_0$ after the fact, by way of what we might call a post hoc intervention. Since the method does not attempt to make people any less prejudiced, it can be even be used amid misrepresenter that remain unwavering in their prejudice in the face of other prejudice interventions.³

Although the evaluation situations where Jönsson and Sjödahl’s method applies by no means exhaust the situations where prejudice is manifested, it applies to a large range of situations which are tremendously important from the perspective of the distribution of wealth and power in society and the overall integration of its members. So even though a focus on evaluation situations involves a considerable restriction (excluding as it does, for instance, manifestations of hostility and violence), it still targets tremendously important situations.⁴ It also bears emphasis, that these situations often feature prejudice, as is demonstrated particularly vividly by so-called CV–studies (See e.g. Steinpreis et al., 1999; Bertrand & Mullainathan, 2004; Correll et al., 2007; Moss–Racusin et al. 2012; Döbrich et al., 2014; Agerström & Rooth, 2011; Rooth, 2010) which demonstrate that identical, or comparable, CVs give rise to different evaluations when they signal membership in different irrelevant social groups (e.g. male/female or Black/White).

### 3 GIRU

Jönsson and Sjödahl (2017) describe and examine three different post hoc interventions—Solipsistic Ordinal Scale Update (SOU), Multiplicative, Informed Ratio Scale Update (MIRU) and Generalized Informed Ratio Scale Update (GIRU), all of which are attempts to improve rankings produced by prejudiced evaluators. But since they conclude that only the third one, GIRU, reliably improves the veracity of the rankings it is applied to, it will be the main focus of this paper.

³ This paper will—for space reasons—not compare Jönsson and Sjödahl’s approach with other approaches that aim to come to terms with bias and prejudice (e.g. anonymization, increased intergroup contact, implicit bias interventions, quota—based affirmative action or various forms of structural interventions) beyond these brief remarks. See Jönsson (forthcoming) for such a comparison where the presuppositions of GIRU are compared with the corresponding presuppositions of individual prejudice interventions. See Jönsson (submitted) for a critical discussion of implicit bias interventions.

⁴ It might be worth mentioning that the generality of many other interventions might also be restricted, but less obviously so, considering the diverse ways in which prejudice manifests itself, and the complex etiology of prejudiced behavior.
GIRU is meant to be applied to, and thus improve the veracity of, a set of competence assignments such as \( r_0 \)—the target evaluation—that an evaluator \( E \) has produced for the purpose of ranking people in terms of their suitability with respect to some end (such as hiring, or promoting, or for some other purpose).

In order to determine whether GIRU needs to update \( r_0 \), GIRU first consults \( E \)’s history of evaluations, a set of sets just like \( r_0 \) consisting of evaluations that \( E \) has previously made. In \( E \)’s history of evaluations, GIRU checks the mean scores of members of a number of target groups—groups that \( E \) might be prejudiced against (e.g. groups such as men and women, or different ethnic groups). GIRU then compares \( E \)’s group means with the means one would expect to find if the populations corresponding to the target groups are assumed to be identical in terms of their distribution of competence (or whatever other property one is interested in). GIRU then proceeds to assume that \( E \) is prejudiced if there is a statistically significant difference between the mean competence scores of the target groups in the history of evaluations (since it is presupposed that there is none in the corresponding populations). It then identifies a set of veridical corrective functions \( v \) such that that each of its members is such that if it is (individually) applied to the scores in the history of evaluations, the target groups obtain the same means. GIRU then suggests that \( r_0 \) should be updated along the lines of what the corrective functions in \( v \) jointly agree on.

To make this more vivid, consider the following illustration from Jönsson and Sjödahl’s article, where the history of evaluations is assumed to contain three sets of competence evaluations, \( h_1 \), \( h_2 \) and \( h_3 \), and we are considering whether to update the target evaluation \( r_0 \)

\[
\begin{array}{cccc}
1 & John & 8 & 1 \\
2 & Luke & 7 & 1 \\
3 & Sarah & 4 & 1 \\
4 & Amber & 4 & 1 \\
5 & Isa & 3 & 1 \\
\end{array}
\quad
\begin{array}{cccc}
1 & Mike & 7 & 1 \\
2 & Catherine & 6 & 1 \\
3 & Billy & 6 & 1 \\
4 & Richard & 5 & 1 \\
6 & Molly & 2 & 1 \\
\end{array}
\quad
\begin{array}{cccc}
1 & Brittney & 6 & 1 \\
2 & Jamal & 5 & 1 \\
3 & Richard & 3 & 1 \\
4 & Susan & 3 & 1 \\
5 & Jennifer & 4 & 1 \\
\end{array}
\quad
\begin{array}{cccc}
1 & Mike & 8 \\
2 & Mark & 7 \\
3 & Felicia & 6 \\
4 & Gordon & 4 \\
5 & Sarah & 3 \\
\end{array}
\]

If we assume that the target groups are men and women, we can note that the mean competence scores in the history of evaluation for these groups are 5.9 and 3.8 respectively. Given that this is a statistically significant difference, and that there is no difference between men’s and women’s population means, GIRU concludes that \( E \) is prejudiced and that \( r_0 \) needs to be updated. One way to do this (which would remove the mean difference between men and women in the history of evaluations) is by means of adding 2.1 to the score of each woman in \( r_0 \) (Felicia, Sarah and Latifah)—and one corrective function \( f_1 \) thus corresponds to ‘\( f_1(x) = x + 2.1 \)’—another is to multiply each score by 1.55—and ‘\( f_2(x) = x \times 1.55 \)’ thus corresponds to another
corrective function \( f_2 \). If either of these functions is applied to the scores in the history of evaluations, the difference in men’s and women’s means would disappear. Other corrective functions might have the same result. GIRU then suggests that each such function is individually applied to the values in the target evaluation \( r_0 \), resulting in \( n \) different rankings \( f^1(r_0), f^2(r_0), \) etc., one for each corrective function \( f_i \). To the extent that these rankings converge on the same results (e.g. Felicia being ranked higher than Mark) GIRU recommends that \( r_0 \) be updated accordingly. If we assume that \( f^1 \) and \( f^2 \) are the only relevant corrective functions in this example, GIRU thus recommends that \( r_0 \) be updated in accordance with what the following two rankings converge on (i.e. by ranking Felicia first, and ranking Sarah and Latifah before Gordon.

\[
\begin{align*}
  f^1(r_0) & \\
  \text{Mike} & 8 & \text{Mike} & 8 \\
  \text{Mark} & 7 & \text{Mark} & 7 \\
  \text{Felicia} & 8.1 & \text{Felicia} & 9.3 \\
  \text{Gordon} & 4 & \text{Gordon} & 4 \\
  \text{Sarah} & 5.1 & \text{Sarah} & 4.65 \\
  \text{Latifah} & 5.1 & \text{Latifah} & 4.65
\end{align*}
\]

In this toy example, this will clearly improve the veracity of \( r_0 \) (since it is now perfectly rank-order correlated with \( r_{0^*} \)).

What is required in general for this method to work? According to Jönsson and Sjödahl, GIRU will tend to increase the veracity of prejudiced rankings if the following five presuppositions are true:

1. The estimated competence of a candidate can be meaningfully captured by a number on a ratio scale.
2. There is a (sufficiently large) history of evaluations.
3. Candidates from all relevant social groups are drawn from populations that have the same overall distribution of the competence that is being evaluated (the equal qualifications assumption).\(^5\)
4. The evaluator persists unchanged in her prejudice between the history of evaluations and the target evaluation.
5. GIRU groups people into the same social groups as the evaluator.

However—as we will see in the next section—these presuppositions are individually too strong as stated, but they also require supplementation by other assumptions to be jointly sufficient for GIRU to work as intended.

\(^5\) It should be mentioned that Jönsson and Sjödahl express some optimism concerning the possibility to weaken this assumption. See Subsection 3.3 for support for this optimism.
4 The presumed presuppositions of GIRU

In this section, we will first discuss each of the aforementioned presuppositions in detail, and then, in the next section, turn to consider three additional _prima facie_ presuppositions. We will argue that four of Jönsson and Sjödahl’s five presumed presuppositions can be weakened, but that they have to be supplemented by two of the additional presuppositions, which were overlooked by Jönsson and Sjödahl (one will turn out not to be a substantial additional requirement after all).

Moreover, even though we find Jönsson and Sjödahl’s arguments for the efficacy of GIRU convincing, a more straightforward validation of the method would be welcome. We thus conclude the paper (in Sect. 5) by providing such a validation by means of a statistical simulation.

4.1 The estimated competence of a candidate can be meaningfully captured by a number on a ratio scale

Jönsson and Sjödahl (2017, pp.~505–507) realize early on in their review of different post hoc interventions, that post hoc interventions in terms of ordinal numbers cannot solve the problem they are concerned with, since ordinal numbers cannot be arithmetically manipulated in a meaningful way. They therefore switch to a method using numbers on a ratio scale.

This is an overreaction. It is sufficient for Jönsson and Sjödahl’s purposes that competence values (or other values of interest) are numbers on an interval scale—i.e. numbers where the difference between two numbers correspond to a proportional difference in that which is being measured—in order for them to be employed in the relevant statistical tests (and in order for them to be meaningfully manipulated by the suggested corrective functions). The additional assumption of a ratio scale—that there is a meaningful zero–value—is not needed for the statistical tests discussed by Jönsson and Sjödahl. It would only be needed if, for instance, we were to calculate the geometric mean or another statistic where multiplying measurements with each other is required.

What is required of an evaluator to satisfy the interval scale assumption? The requirement presupposes that the competence values assigned to people by the evaluator are roughly proportional to the perceived degree of difference between the competence of these people. If this is the case, one can, for instance, infer from the fact that the difference between the competence values of Anna and Boris is the same as the difference between the competence values of Cecilia and David, that the perceived difference in competence between Anna and Boris is the same as that between Cecilia and David. This distinguishes values on an interval scale from values on an ordinal scale, where no such inferences can be made.

It is not obvious to what extent evaluators always conform to this requirement. But it is a commonplace assumption in recruitment, promotion, grading, and other forms of assessment such as grant proposal approval, that they do (cf. Gatewood et al., 2015, pp. 210–212). Adherence to the requirement is sometimes non–obvious though. Consider for instance the grades awarded in Swedish schools, which although marked
by letters of the alphabet—which suggest that a nominal scale is used—underlie calculations of a child’s sum of merits (‘Meritvärde’), a calculation that presuppose an underlying interval scale (cf. https://utbildningsguiden.skolverket.se/gymnasieskolan/om-gymnasieskolan/behorighetsregler-och-meritvarde). Or consider when the U.S. Department of State recommends using detailed rating scales to help quantify subjective data when conducting structured interviews (cf. Kim, 2016), or the fact that the Swedish Research Council asks evaluators to rate the merits of applicants for research grants using such scales. As soon as measurements are manipulated arithmetically, an interval scale is presupposed.

A ratio scale on the other hand presupposes a ‘true zero point’, i.e. that a score of zero corresponds to a complete absence of that which is being measured. A measure of height is a typical example of a measure on a ratio scale, since being (exactly) 0 cm tall indicates not having a height, while a measure of temperature in terms of Celsius is not an example of a measure on a ratio scale since it being 0 degrees Celsius is not a matter of there not being any temperature. The difference is important in the present context since not all measures underlying social selection have true zero points. Most scales do not feature zero points at all, and even if such values are inferred from the scales, it often seems unreasonable to think that they correspond to an absence of that being measured. Consider for instance, a scale of competence from 1 to 5 (with no explicit zero point), and the difference in competence between persons B and C, where B receives a 1 on the scale and C a 2. Now assume that the difference between A and B is the same as that between B and C. If we assume that the scale is a ratio scale, this amounts to thinking that A has no competence whatsoever, which for many scales would be an unreasonable thing to maintain. The fact that Jönsson and Sjödahl’s method only presupposes an interval scale thus significantly widens its scope of application.

4.2 There is a (sufficiently large) history of evaluations

Even though GIRU essentially presupposes the existence of a history of evaluations of a certain size (in order for a prejudice to be statistically detectable), Jönsson and Sjödahl does not spell out this presupposition in any detail.

To elucidate the presupposition, we can begin by noting that the required size of a history of evaluations depends on what one is willing to accept in terms of the probability of thinking that there is a prejudice where there is none (making a Type 1 error) and not finding a prejudice where there is one (making a Type 2 error). Since calculations of these probabilities are test dependent, size cannot be discussed independently of one’s choice of statistical test, which in turn depends on 1) which other assumptions one is prepared to make and 2) how many target groups and factors (where factors correspond to ‘dimensions of prejudice’, such as gender or ethnicity) one is interested in testing.

In order to make some headway, consider the following—non-exhaustive—list of statistical tests (each of which is a candidate test to discover a prejudice in an evaluator’s history of evaluations) together with their assumptions. As indicated by Table 1, which test one should use depends on what distributional assumptions one
Table 1 Tests for detecting differences in means

|                                      | t-test | One-way anova | Two-way anova | z-test | Mann–Whitney test | Kruskal–Wallis test | Welch’s t-test |
|--------------------------------------|--------|---------------|---------------|--------|-------------------|---------------------|---------------|
| Populations have the same variance   | x      | x             | x             | -      | x                 | x                   | -             |
| Populations are normally distributed | x      | x             | x             | -      | -                 | -                   | x             |
| Populations have the same distributational shape | x | x | x | - | x | x | x |
| Central limit theorem applies        | -      | -             | -             | x      | -                 | -                   | -             |
| Only one factor (e.g. gender, ethnicity) | x | x | - | x | x | x | x |
| Only two groups (e.g. men and women)  | x      | -             | -             | x      | x                 | -                   | x             |
is prepared to make, whether the history is large enough for the central limit theorem to be applicable, and how many comparisons one wants to make (e.g. if one is only concerned with one factor, such as either gender or ethnicity, or several, and how many groups one wants to compare within each factor, e.g. White people, Black people and Latinos).\(^6\)

For instance, if one only wants to compare men and women, one can assume that the competence being evaluated is normally distributed among men and women and these two populations have the same variances, a \(t\)-test may be used. If, on the other hand, one wants to compare White people, Black people and Latinos, with only a small number of evaluations in each group (so the central limit theorem is not applicable), and the evaluations do not seem to be normally distributed, a Kruskal–Wallis test may be used instead.

We can calculate the required size of the history of evaluations (or more precisely, the needed size of the target groups in the history of evaluations) exactly given that we have chosen a particular test. To illustrate this, let’s assume that we are concerned with two groups (e.g. men and women) and that the corresponding populations that we sample from are normally distributed with respect to the property of interest (e.g. competence). In this situation it is fairly straightforward to calculate the power of a test, i.e. the probability of detecting a mean difference between the groups of a certain size, and given this calculation we can derive an equation to determine the number of evaluations from each group \(n_1\) and \(n_2\) needed for the test to have power \(P\) of detecting a difference \(\delta\), given a level of significance \(\alpha\) (i.e. probability of committing a Type 1 error) and variances \(\sigma_1^2\) and \(\sigma_2^2\). If we, for simplicity, are only concerned with detecting a higher mean value in group 1, we solve the following equation, derived from a \(z\)-test,

\[
\delta \left( \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)^{-\frac{1}{2}} = z_\alpha - \Phi^{-1}(1 - P),
\]

where \(z_\alpha\) is the \(\alpha\) quantile of the standard normal distribution and \(\Phi^{-1}\) is the inverse of the standard normal cumulative distribution function.\(^7\) In the special case where we assume the same number of evaluations \(n\) and the same variance \(\sigma^2\) in both groups, we find the solution

\[
n = \frac{2\sigma^2(z_\alpha - \Phi^{-1}(1 - P))^2}{\delta^2}.
\]

In addition to the variance of the two groups, the size required thus depends on both a level of significance, the desired power and the size of the difference we want to detect. Fortunately, in the present context we are just interested in getting the relative positions of people in the target evaluation right (so that e.g. the person ranked first

---

\(^6\) Whether the Central Limit Theorem is applicable depends on the distribution of the measurements: if the measurements come from a symmetric and ‘well-behaved’ distribution, a dozen evaluations might be enough, but if the distribution is skew and very far from normal, hundreds of evaluations might be required.

\(^7\) There is no analytical expression for \(\Phi^{-1}\), but it is widely available in computer packages, e.g. as NORM.S.INV() in Excel and qnorm() in R.
is really the best person, and so on). This means that we are not interested in finding prejudices that are smaller than the least difference between a man and a woman in the target evaluation, for even if we find and correct for such a prejudice, it won’t influence the order of the target evaluation. Let us assume that this is 1 (i.e. that \( \delta = 1 \)). We can then conclude that if \( \sigma^2 = 1 \) and \( \alpha = 0.05 \), i.e. \( z_\alpha = 1.6449 \), then to obtain a power of at least 75% we would need 11 evaluations from each group in the history (i.e. only slightly more than in the toy example in Sect. 2) and that 18 evaluations from each group in the history would be needed to obtain a power of 90%. If we want to detect a difference of 2 units, then only 3 evaluations from each group in the history would be sufficient to obtain a power of 75% and 5 evaluations from each group for a power of 90%. So the size of the history of evaluation need not be very big, even if we want quite powerful tests. 

### 4.3 Candidates from all relevant social groups are drawn from populations that have the same overall distribution of competence (the Equal Qualifications Assumption).

The equal qualifications assumption (i.e. the assumption that candidates from all relevant social groups are drawn from populations that have the same overall distribution of competence) is clearly a problematic presupposition of GIRU. For one thing, it is somewhat vague as it stands, and more importantly, if interpreted in a strong sense—according to which the relevant populations must be identical with respect to competence distribution—it is clearly implausible. 

In order to discuss the assumption in more precise terms it can be unpacked into three separate assumptions about the relevant competence distributions (or other property distributions):

a) The relevant populations have the same distributional shape.
b) The relevant populations have the same variance.
c) The relevant populations have the same means.

That neither (a) nor (b) is necessary was shown in the previous subsection since these assumption are not presupposed by all of the statistical tests presented there (the \( z \)-test, for instance, makes neither assumption).

Assumption (c) is not necessary either as long as it is substituted with another, weaker, assumption (c*) according to which the actual mean difference between the relevant populations is known to be of a certain value. If this is assumed, we can use the aforementioned tests to check whether the difference in the corresponding means in the history of evaluations significantly deviate from the difference assumed to exist in the populations.

---

8 See Sect. 5 for exactly how well GIRU performs in terms of veracity improvement, with histories of evaluation with a median value of 14 prejudiced evaluations. Even though this is less than what one would have expected, it can still be problematic in the cases of intersectional prejudice, e.g. when a prejudice is directed towards Black men specifically rather than Black people or men in general. See Jönsson (forthcoming) for a discussion of this problem.

9 In their defense, Jönsson and Sjödahl (2017, p. 513) admit as much, and express—as it turns out—well-founded hope that it can be weakened.

10 Knowing that the mean difference is zero, i.e. that c) obtains, is now just a limiting case.
We cannot reasonably use GIRU without c*) being true though, since without it we have no way of distinguishing between actual competence and the evaluator’s prejudice. There is a noteworthy trade–off here between how much we need to assume about the relevant populations, and how much we need to assume about how the prejudice operates. And in order to do without c*) we have to make specific assumptions about how the prejudice operates (e.g. ‘it subtracts 1 from the scores of all women’) which makes correcting for the prejudice trivial, and estimating the prejudice from a history of evaluations redundant. So in order to retain the central feature of GIRU—that it does not need to make any specific assumptions about how the prejudice operates it must instead assume c*). But even so, c*) is a significantly weaker assumption than the equal qualifications assumption suggested by Jönsson and Sjödahl.

4.4 The evaluator persists unchanged in her prejudice between the history of evaluations and the target evaluation.

Jönsson and Sjödahl assume that in order for GIRU to work, the evaluator needs to persist unchanged in her prejudice between the history of evaluations and the target evaluation. Even though this is a natural assumption to make—since one wants to avoid GIRU correcting a non–prejudiced person’s evaluation due to her now gone, but previously existing prejudice—it is also an assumption that can be somewhat relaxed. As we will see in Sect. 5, GIRU can still be applied to cases where an evaluator’s prejudice randomly fluctuates considerably. In fact, the conservative corrective function recommended by Jönsson and Sjödahl lead to improvements in two thirds of the cases where it recommends updates for evaluations using a 10–point scale and a fluctuating prejudice with a fluctuation with a standard deviation of 4. However, it bears emphasis that the improvements in these cases are very small. If GIRU is to have a likely noticeable positive effect the prejudice can still fluctuate, but the fluctuations have to be quite small with respect to the size of the prejudice. Still, the fact that GIRU doesn’t presuppose that an evaluator’s prejudice is constant significantly widens its scope of application.11

If the prejudice does not fluctuate randomly but instead changes by increasing or decreasing slightly over time, GIRU will still work but not work optimally. If the prejudice is increasing, the method will undercompensate, but still generate an improvement. If the prejudice is decreasing however, the method will overcompensate, which is more problematic since it might lead GIRU to intervene in a way that lowers veracity. This problem can be mitigated however, by only making use of the most recent portion of the history of evaluation corresponding to the portion required in order to apply GIRU with a desired level of power.

4.5 GIRU groups people into the same social categories as the evaluator

In order for GIRU to be able to correct for a prejudice towards a particular group, it must correctly test that group, and it is thus natural that Jönsson and Sjödahl suggests

---

11 For instance, measures of implicit prejudice are known to fluctuate considerably. See, Gawronski and Bodenhausen (2017) and Gawronoski (2019) for a good discussion of the variability of implicit bias.}
that GIRU must group people into the same social categories as the evaluator. But, similarly to most of the previous assumptions, this assumption can also be relaxed somewhat.

It is sufficient to assume that GIRU groups people into—possibly proper—subsets of the groups that the evaluator makes us of. If the subsets are proper (e.g. if the intervention makes use of “Black” instead of “Person with a dark complexion”), the method will still improve the veracity of scores from the subgroup (but only those that belong to the subgroup). It might be worth noticing however, that the likelihood of statistically identifying a prejudice might be reduced in this situation as the means of the relevant groups will be muddled. Still, as long as this assumption is in place, GIRU will not distort evaluations it is applied to.

The situation is different if GIRU categorizes using supersets of the groups that the evaluator makes us of, since it runs risk of compensating for prejudice where there is none (e.g. if GIRU corrects the scores of people with a dark complexion even though the evaluator is only prejudiced towards Black people). It also runs the risk of not finding any prejudice even if there is one in a subgroup of the relevant group (since the average in the relevant group will look less affected by prejudice since only a part of it really is). The same problems arise if the group of the evaluator and the group of the post hoc intervention overlap but neither is a subset of the other.

5 Other presuppositions of GIRU

In the previous section we showed that all but one of the five presuppositions that Jönsson and Sjödahl discussed in their paper could be relaxed, on occasion quite significantly. However, in addition to the presuppositions that Jönsson and Sjödahl highlighted themselves, there are three additional prima facie presuppositions that bear discussion. The first of these will turn out not to be a substantial additional assumption, but bears discussion nonetheless.

5.1 The correct populations have to be identified

The parametric tests listed in Subsection 3.2 presuppose various things about the relationship between the populations that the history of evaluations sample. In order for GIRU to only correct for prejudices if there is prejudice, it is central that it employs information about the right populations. Otherwise, what is thought to be a prejudice might really be due to the fact that what is sampled in the history of evaluations are other populations. If, for instance, all the men who apply to a certain position are well educated but the women who apply for this position are not, then this might generate a big difference between the mean competence scores for men and women in a history of evaluations. But this difference does not correspond to a prejudice, and GIRU should thus not correct for it (even though the difference might of course reflect other, structural, societal problems that need addressing in other ways).

However, even if an example like this underlines the importance of identifying the right populations in an application of GIRU, it does not undermine GIRU or amount
to any substantial additional assumptions. If the history of evaluations is made up of samples of well-educated men and less educated women, the relevant populations are not men and women simpliciter, but well-educated men and less educated women. And given that the parameters for these populations are known, GIRU will not try to update a non–prejudiced evaluation but will proceed (in the way described above in Sect. 2) without update. So even if it is important—and far from trivial—that the history of evaluations really sample from the populations that we think they do (i.e. that we have identified the right populations) this does not correspond to any separate substantial assumption since it is built into the assumption that the (right) populations means are known (cf. Subsection 3.3).12

Another kind of example raises a related point. In real life situations, people will sometime reapply for jobs they didn’t get the first time around, and they might continue to do so even in the light of repeated rejection. If one or more persons that belong to a group that an evaluator is prejudiced against are both persistent in this way and genuinely underqualified, they might contribute to a skewed estimate of prejudice. For if the history of evaluation features them several times (and there are no similarly persistent and underqualified persons in the other target groups), the calculated group average might be lower, than if the history of evaluations contained only unique individuals. GIRU will thus overestimate the existing prejudice, and ‘correct’ for prejudice even in cases where this is not called for.

This is a similar problem to the aforementioned one, but one which is more easily avoided in practice, since we can just weed out multiple instances of a certain person from a history of evaluations, favouring later occurrences if there are any differences in their estimated qualifications (and thus allowing for people to become more competent).

5.2 The prejudice operates on discrete properties

The discussion above (and the simulation below) has been carried out on the assumption that people are prejudiced categorically (plus or minus some random variation) towards certain groups of people (e.g. women or Black people), but prejudice could also operate continuously as a function of the degree to which someone has a certain property (e.g. degree of femininity or skin tone). For instance, in the literature following Maddox (2004)’s seminal paper on racial phenotypicality bias, it has been demonstrated, for instance, that self–reported skin tone is a significant predictor of the frequency of perceived discrimination and perceived stress among older African Americans (Monk et al., 2021).

Although GIRU—as it was presented above—clearly presupposes that people are prejudiced categorically, the intervention is easily modified to operate on a continuous form of prejudice instead, e.g. by using regression analysis instead of t–tests or ANOVAs. However, this requires valid measurements of the property in question (e.g.

12 It should be conceded though that populations means are seldom available directly. Instead, they have to be estimated by either using objective measures of competence or from subjective evaluations from blinded evaluators, expert evaluators, evaluators evaluating under particularly good conditions (e.g. under less stressful circumstances) or other evaluators which we have reason to think are well equipped to estimate population means.
the degree of femininity or the skin tone), which might in practice be hard to obtain. GIRU can also be applied as is if the continuous property in question is discretised into a number of different categories (“very light,” “light,” “medium,” “dark,” and “very dark.”) as is done in some of the work on phenotypicality bias (e.g. Monk et al., 2021). So the existence of this kind of prejudice doesn’t seem to undermine the general approach that GIRU embodies.

5.3 The prejudice operates in an approximately linear way.

A noteworthy omission in Jönsson and Sjödahl’s discussion is the possibility of prejudice that operates non-linearly (i.e. which is such that the prejudiced evaluation $y$ cannot, even approximately be described in terms of the function $y = a + bx$ where $x$ is the actual perceived competence of the candidate and $a$ and $b$ are constants).

Consider for instance situations where $E$ assigns competence scores to members of a certain group with disregard, at least to some extent, for the member’s actual competence scores. An extreme example of this is an evaluator who fails all students of a certain ethnicity regardless of their actual competence. A less extreme case is when $E$—in the grip of the stereotype that there are no female geniuses—makes use of a lower maximum score for women than for men. If the top competence score is, say, 5 this would be the case if $E$ assigns all women with real competence score of 4 or 5, a 4.

Both of these cases are problematic for GIRU due to the non-linearity of the exemplified prejudices. In each case, GIRU cannot restore the non-prejudiced scores since information is lost in the wake of the prejudice. Consider for instance the second case and say that $E$ displays no prejudice at all in the evaluations of women with lower scores than 4. In this case, GIRU is likely to distort competences scores instead of improving them, since a discovery that the mean value of women in the history of evaluations have too low scores (due to the absence of 5’s) will be corrected by increasing the scores of all women, many of which haven’t received prejudiced scores. Even if GIRU could detect specifically that there are too many fours and too few fives, all information about which candidate that deserves which of these two scores have been lost. So in order for GIRU to operate correctly, we have to assume that the prejudice operates in an approximately linear way.14

13 Although this is not impossible, since e.g. reflectance spectrophotometry can be used to measure skin tone.

14 This is slightly stronger than what is really motivated by the aforementioned examples. What is strictly speaking required by these is just that the prejudiced score is some function of the unprejudiced score, not that the function is linear. However, the linearity assumption is operative in statistically testing GIRU in the next section, so the version of GIRU being validated there is one where the linearity assumption is presupposed. One possible instance where linearity is not satisfied but GIRU might be possible to apply anyway is where an evaluator’s degree of prejudice is greater for a certain range of the scale being used than for other parts of the scale, e.g. where the evaluator is more prejudiced towards people receiving non-extreme scores. There is some data that suggests that this is the case for people in the grip of aversive racism, who might be prone to discriminatory behavior only “when the guidelines for appropriate behavior are vague… or when one can justify or rationalize a negative response on the basis of some factor other than race.” (cf. Dovidio et al., 2017, p. 271). For instance, Dovidio and Gaertner (2000) demonstrated that
5.4 Interim Conclusion

Even though we could conclude from the previous section that all but one of the five presuppositions that Jönsson and Sjödahl discussed in their paper could be relaxed, we can now see that GIRU’s presuppositions are more demanding than Jönsson and Sjödahl assumed in another sense. In particular, GIRU presupposes that prejudice operates 1) in an approximately linear way, 2) on discrete properties.

6 A Statistical simulation of GIRU’s performance

Jönsson and Sjödahl’s original claim that GIRU will in fact improve the veracity of rankings produced by prejudiced evaluators was only backed up by informal (albeit convincing) arguments—no actual test was carried out to support their claim. To remedy this, and to explore the relative merits of various different corrective functions, we ran a number of statistical simulations of prejudices operating in a linear way.

We simulated 5 000 evaluators. For each evaluator we generated a history of random length of 30–50 evaluations consisting of uniformly distributed evaluations between 1 and 10. A random sample between a third and half of the evaluations was reserved for prejudiced treatment. We then generated a new set of 6–10 evaluations in the same way as the history of evaluations and again reserved between a third and half of the evaluations for prejudiced treatments. For each evaluator we thus have a history of evaluations and a target evaluation, both of which are non–prejudiced except for a designated subset of each.

We then simulated 19 different prejudices and applied them to the designated subsets. Each of these can be described as the function $f(x) = x + a \times b$ with the addends $a$, and factors $b$ described in Table 2.

The first of these corresponded to a baseline of no prejudice, the next six to purely additive prejudices, the following five to purely multiplicative prejudices, and the last seven to hybrid prejudices. Apart from the constant prejudice, we also simulated prejudices with fluctuations, where the additive prejudice was a random number from a normal distribution with the prescribed value as expected value and the multiplicative prejudice was a random number from the gamma distribution (to ensure strict positive values) with the prescribed value as expected value. We changed the amount of fluctuation by using different standard deviations when generating the random numbers. For the additive components, we used standard deviations of 0.5, 1, 2, 3, and 4. (As a comparison, the standard deviation from the uniform distribution used for generating the evaluations is approximately 2.6.) For the multiplicative components, the additive standard deviation was divided by four, i.e. 0.125, 0.25, 0.5, 0.75, and 1, for a reasonable variation. We have in total $19 \times 6 = 114$ different patterns of prejudice (including the baseline of no prejudice, and no fluctuation).

For every combination of prejudice (including no prejudice) and amount of fluctuation (including no fluctuation), we tested, for each evaluator, if there was a difference

Footnote 14 continued

White college students only recommended Black candidates significantly less often than White candidates (with exactly the same credentials) for a campus position, when the candidates had moderate qualifications.
in mean score in the history between the non–prejudiced and prejudiced evaluations, using a two sample t–test. If the test was significant at a five per cent level, three different basic prejudice correction functions were tried: an additive correction, a multiplicative correction, and standardisation. For the additive correction, we added the difference in mean between the non–prejudiced and prejudiced evaluations in the history to the prejudiced evaluations in the new set. For the multiplicative correction, we multiplied the prejudiced evaluations in the new set with the ratio of the means of the non–prejudiced and prejudiced evaluations in the history. For the standardisation correction, the labelled and unlabelled evaluations in the new set were standardised with the group’s mean and standard deviation in the history, respectively. The standardised values were then multiplied with the original new set overall standard deviation and the overall mean was added to transform the scores back to the original scale.

In addition to the three basic prejudice corrections, we also tested a conservative corrective function corresponding to the one originally suggested by Jönsson and Sjödahl. This function only modified the prejudiced evaluation to the degree that the basic corrective functions converged on the same relative order (cf. the example described in Sect. 2).

When there is no prejudice, the test was significant for 276 evaluators (5.5%), in line with the five per cent level of significance. As expected, the proportion of significant tests increased with the size of the prejudice and decreased with the fluctuation, ranging from 5 to 99%. With a constant additive prejudice of two (prejudice pattern 4 in Table 2), the prejudice was detected for 60% of the evaluators, but with a random prejudice with an expected value of 2 and a standard deviation of 2 the prejudice was only detected for 29% of the evaluators. For more than half of evaluators to be significant with a fluctuation of 0.5 or less, the prejudice needed to be rather large: an additive prejudice of 2 or more, or a multiplicative prejudice of 1.5 or more. This is of course an effect of the variability in the evaluations (a standard deviation of 2.6) and relatively small samples (the median history is 40 and median number of prejudiced evaluations is 14).

For every corrected set of evaluations, we calculated the Spearman’s rank correlation coefficient (i.e. the correlation coefficient between ranks) between the non–prejudiced evaluation scores and the corrected scores. We compared this to the rank correlation between the non–prejudiced evaluation scores and the prejudiced

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| 1 | 0.0 | 1.00 | 8 | 0.0 | 1.05 | 13 | 0.5 | 1.05 |
| 2 | 0.5 | 1.00 | 9 | 0.0 | 1.10 | 14 | 0.5 | 1.10 |
| 3 | 1.5 | 1.00 | 10 | 0.0 | 1.50 | 15 | 0.5 | 1.50 |
| 4 | 2.0 | 1.00 | 11 | 0.0 | 2.00 | 16 | 0.5 | 2.00 |
| 5 | 2.5 | 1.00 | 12 | 0.0 | 2.50 | 17 | 1.0 | 1.05 |
| 6 | 3.0 | 1.00 |   |   |   | 18 | 1.0 | 1.10 |
| 7 | 3.5 | 1.00 |   |   |   | 19 | 1.0 | 1.50 |

| Table 2 Addends and factors of tested prejudices | a | b | a | b | a | b |
|---|---|---|---|---|---|---|
| 1 | 0.0 | 1.00 | 8 | 0.0 | 1.05 | 13 | 0.5 | 1.05 |
| 2 | 0.5 | 1.00 | 9 | 0.0 | 1.10 | 14 | 0.5 | 1.10 |
| 3 | 1.5 | 1.00 | 10 | 0.0 | 1.50 | 15 | 0.5 | 1.50 |
| 4 | 2.0 | 1.00 | 11 | 0.0 | 2.00 | 16 | 0.5 | 2.00 |
| 5 | 2.5 | 1.00 | 12 | 0.0 | 2.50 | 17 | 1.0 | 1.05 |
| 6 | 3.0 | 1.00 |   |   |   | 18 | 1.0 | 1.10 |
| 7 | 3.5 | 1.00 |   |   |   | 19 | 1.0 | 1.50 |
scores, calculating the improvement as the difference between the corrected correlation and the prejudiced correlation. Figure 1 displays the boxplots contrasts the degrees of veracity of the four corrective functions with the veracity of the uncorrected rankings.

All four corrective functions led to improvements in the majority of cases when the prejudice was constant or had a fluctuation of 0.5, but only standardisation and the corrective function would work consistently across the different kinds of prejudice for higher degrees of fluctuation. In order to get a good estimate of how well the four corrective functions performed we noted how many of the 4000 cases where the corrective functions caused a difference in rank order correlation, that were improved. Table 3 summarizes the corresponding proportion for each corrective function and degree of fluctuation.

We can see that all four functions perform worse with an increased degree of fluctuation but that improvement remains more likely than distortion even if high degrees of fluctuation is included. The conservative function does slightly better than the other functions (and much better than the additive function) in all intervals and has an impressive 93% improvement rate for constant prejudices.

It should be noted though that the mean degree of improvement in rank order correlation across prejudices for fluctuating prejudices with a standard deviation greater than 1 is very close to zero. For constant prejudices, all corrective function led to noticeable improvements for most prejudices. However, constant prejudices corresponding to prejudice patterns 1, 2, 8, 9 and 13, i.e. those corresponding to no prejudice or very small prejudices did not lead to improvements for any corrective functions. It should be noted though that these prejudices are so small that they can only serve as ‘prejudiced tiebreakers’, i.e. such that they can misrepresent one out of two equally competent persons as better than the other. They are not big enough to reverse the order of two candidates of unequal qualifications. If the prejudiced tiebreakers are ignored, the four corrective functions gave rise to a mean median improvement score of about 0.1 across the different constant prejudice patterns, typically making sure that corrected rank order correlation was over 0.9, i.e. fairly close to a perfect correlation. The mean median improvement dropped to around 0.05 for fluctuating prejudices with a standard deviation of 0.5 and dropped even further for greater fluctuation. See Table 4 for a summary.

It can be seen that the additive corrective function performs worse than the other functions but that the latter three functions yield comparable results. Jönsson and Sjödahl’s claim that the conservative function would tend to improve the veracity of prejudiced rankings is thus borne out both for constant prejudices as well as for somewhat fluctuating prejudices, even though the mean median score improvement for the latter cases is quite modest.

In addition to the test described above, we also tested the effects of doubling the size of the learning history. As expected, this led to detecting more prejudice. It did not, however, have any dramatic changes in average improvement.
Fig. 1 Boxplots of Spearman’s rank correlation coefficients between the unbiased and the corrected scores for different standard deviations (panels), bias patterns, and corrective functions (coloured boxes)
Table 3  Proportion of improvements in the cases where the rank order was changed

| Degree of fluctuation | 0       | ≤ 0.5   | ≤ 1     | ≤ 2     | ≤ 3     | ≤ 4     |
|-----------------------|---------|---------|---------|---------|---------|---------|
| Additive              | 85%     | 74%     | 66%     | 60%     | 57%     | 56%     |
| Multiplicative        | 87%     | 80%     | 74%     | 70%     | 67%     | 65%     |
| Standardization       | 87%     | 78%     | 71%     | 68%     | 67%     | 67%     |
| Conservative          | 93%     | 83%     | 76%     | 72%     | 70%     | 69%     |

Table 4  Mean median improvement ignoring prejudice patterns 1, 2, 8, 9, and 13

| Degree of fluctuation | 0       | ≤ 0.5   | ≤ 1     | ≤ 2     | ≤ 3     | ≤ 4     |
|-----------------------|---------|---------|---------|---------|---------|---------|
| Additive              | .09     | .03     | –.01    | –.02    | –.02    | –.01    |
| Multiplicative        | .10     | .05     | .02     | 0       | 0       | 0       |
| Standardization       | .10     | .04     | .02     | 0       | 0       | 0       |
| Conservative          | .10     | .05     | .01     | 0       | 0       | 0       |

7 Summary and Conclusion

This concludes our review of GIRU. We have clarified its presuppositions by explaining how four out of the five presumed presuppositions can be relaxed, and why these presuppositions need to be supplemented by two additional assumptions, overlooked by Jönsson and Sjödahl.15

In summary, in order for GIRU to be able to improve the veracity of a certain evaluator $E$’s evaluations of a set of candidates in terms of some form of competence, the following seven assumptions have to be satisfied:

1. $E$’s evaluations correspond to numbers on an interval scale.
2. $E$ has a history of evaluations which is large enough to reliably find prejudices of the desired size with a suitable statistical test.
3. The difference in population means between the relevant populations is known to be of a certain value.
4. GIRU groups people into subsets of the groups $E$ is prejudiced towards.
5. Any fluctuations in $E$’s prejudice are small compared to the size of the prejudice.
6. $E$’s prejudice operates in an approximately linear way.
7. $E$’s prejudice operates on discrete groups.

In addition, we reported the results from a statistical simulation that validates Jönsson and Sjödahl’s claim that GIRU can in fact reliably solve the epistemological problem it started out with.

As was noted above, the seventh assumption is just a presupposition of GIRU as described and tested here. GIRU could easily be revised to operate on continuous forms of prejudice as well.
The foregoing also has a terminological consequence. GIRU is more aptly named GIIU (Generalized Informed Interval scale Update) due to the first of these presuppositions, and we thus suggest that the method should henceforth be referred to by this acronym.

The results reported here pave the way for investigating in which particular real–life contexts GIIU can be employed. One such investigation is already underway (Bergman and Jönsson in preparation) where we study assessments of a large number of research grant applications. We investigate how gender affects the assessments while controlling for a number of relevant factors, e.g. subject, seniority, and affiliation, assuming that on average the grades of females and males should be the same after removing the differences due to the available relevant factors. We then explore the effects of applying GIIU to this data as a way to counter gender bias.

Although this will have to be settled conclusively in another article (see Jönsson forthcoming), it can be noted that none of the seven presuppositions are in obvious conflict with other kinds of interventions, such as individual interventions, and GIIU can thus, at least, prima facie be used to complement these interventions. Some care has to be taken though, so that, for instance, an individual intervention is not administered between the rankings in the history of evaluations takes place, and the target evaluation takes place. For if that intervention turns out to be successful, GIIU might overestimate an evaluator’s degree of prejudice when they carry out the target evaluation and overcompensate, which might lead to a decrease in veracity. However, the interventions can complement each other without the risk of overcompensation, if the individual intervention is administered before the history of evaluation has taken place. Then, that intervention might lower E’s prejudice to some degree, and GIIU can monitor and correct for any residual prejudice. And if the individual intervention proves to be ineffective, GIIU enables us to improve the misrepresentations amid the unwavering misrepresenters.

Acknowledgements This paper was made possible partly by a research grant from the Swedish Research Council (dnr. 2017–02193) awarded to the first author. The authors would like to thank Åsa Burman, Julia Sjödahl, the participants of the Higher Seminar in Theoretical Philosophy at Lund University, as well as three anonymous reviewers reviewing for this journal for fruitful discussion and helpful comments.

Funding Open access funding provided by Lund University. The research was funded by a research grant from the Swedish Research Council (Dnr. 2017-02193).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. 
References

Agerström, J., & Rooth, D. O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology, 96*, 790–805.

Bergman, J., & Jönsson, M. (in preparation) Gender Bias in Grant Applications. Inquiry and the Potential for a Post Hoc Remedy

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? *American Economic Review, 94*, 994–1013.

Correll, S. J., Benard, S., & Paik, I. (2007). Getting a job: Is there a motherhood penalty? *American Journal of Sociology, 112*, 1297–1339.

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science, 11*, 319–323.

Dovidio, J. F., Gaertner, S. L., & Pearson, A. R. (2017). Aversive racism and contemporary bias. In C. G. Sibley & F. K. Barlow (Eds.), *The Cambridge handbook of the psychology of prejudice* (pp. 267–294). Cambridge University Press.

Döbrich, C., Wollersheim, J., Welpe, I. M., & Spörrle, M. (2014). Debiasing age discrimination in HR decisions. *International Journal of Human Resources Development and Management, 14*(4), 219–241.

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

Gatewood, R., Field, H., & Barrick, M. (2015). *Human resource selection* (8th ed.). Cengage Learning.

Hodson, G., Dovidio, J. F., & Gaertner, S. L. (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin, 28*, 460–471.

Jönsson, M. (forthcoming). On the Prerequisites for Improving Prejudiced Ranking(s) with Individual and Post Hoc Interventions. *Erkenntnis*.

Jönsson, M. L., & Sjödahl, J. (2017). Increasing the veracity of implicitly biased rankings. *Episteme, 14*(4), 499–517.

Gawronski, B. (2019). Six Lessons for a Cogent Science of Implicit Bias and Its Criticism. Perspectives on Psychological Science.

Gawronski, B., & Bodenhausen, G. (2017). Beyond Persons and Situations: An Interactionist Approach to Understanding Implicit Bias. *Psychological Inquiry, 28*(4), 268–272.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*, 17–41.

Kim, M. (2016) Conducting Effective Structured Interviews Resource Guide for Hiring Managers and Supervisors. Retrieved August 19, 2020, from https://2009-2017.state.gov/documents/organization/107843.pdf.

Lai, C., Marini, M., Lehr, S., Cerruti, C., Shin, J., Joy-Gaba, J., Ho, A., Teachman, B., Wojcik, S., Koleva, S., Frazier, R., Heiphetz, L., Chen, E., Turner, R., Haidt, J., Kesebir, S., Hawkins, C., Schaefer, H., Rubichi, S., Nosek, B. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*, 1765–1785.

Lai, C., Skinner, A., Cooley, E., Murrat, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y., Pedram, C., Marshburn, C., Simon, S., Blanchar, J., Joy-Gaba, J., Conway, J., Redford, L., Klein, R., Roussos, G., Schellhaas, F., Burns, M., … Nosek, B. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General, 145*(8), 1001–1016.

Madva, A. (2020). Individual and structural interventions. In E. Beeghly & A. Madva (Eds.), *An introduction to implicit bias: Knowledge, justice, and the social mind*. New York: Routledge.

Maddox, K. (2004). Perspectives on racial phenotypicality bias. *Personality and Social Psychology Review, 8*(4), 383–401.

Monk, E., Kaufman, J., & Montoya, Y. (2021). Skin tone and perceived discrimination: Health and aging beyond the binary in NSHAP 2015. *The Journal of Gerontology: Series B, 76*(3), S313–S321.

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 USA, 109*, 16474–16479.
Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology, 105*, 171–192.

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology, 60*(1), 339–367.

Rooth, D. O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics, 17*, 523–534.

Steinpreis, R. E., Anders, K. A., & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A National Empirical Study. *Sex Roles, 41*, 509–528.

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.