International Inequalities Institute

How the Reification of Merit Breeds Inequality: Theory and Experimental Evidence

Fabien Accominotti and Daniel Tadmon

Working paper 42

March 2020

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

LSE International Inequalities Institute

The International Inequalities Institute (III) based at the London School of Economics and Political Science (LSE) aims to be the world's leading centre for interdisciplinary research on inequalities and create real impact through policy solutions that tackle the issue. The Institute provides a genuinely interdisciplinary forum unlike any other, bringing together expertise from across the School and drawing on the thinking of experts from every continent across the globe to produce high quality research and innovation in the field of inequalities.

In addition to our working papers series all these publications are available to download free from our website: www.lse.ac.uk/III

For further information on the work of the Institute, please contact the Institute Manager, Liza Ryan at e.ryan@lse.ac.uk

International Inequalities Institute

The London School of Economics and Political Science

Houghton Street

London

WC2A 2AE

Email: Inequalities.institute@lse.ac.uk

Web site: www.lse.ac.uk/III

@LSEInequalities

Abstract

In a variety of social contexts, measuring merit or performance is a crucial step toward enforcing meritocratic ideals. At the same time, workable measures – such as ratings – are bound to obfuscate the intricacy inherent to any empirical occurrence of merit, thus reifying it into an artificially crisp and clear-cut thing. This article explores how the reification of merit breeds inequality in the rewards received by the winners and losers of the meritocratic race. It reports the findings of a large experiment (n = 2,844) asking participants to divide a year-end bonus among a set of employees based on the reading of their annual performance reviews. In the experiment's non-reified condition, reviews are narrative evaluations. In the reified condition, the same narrative evaluations are accompanied by a crisp rating of the employees' performance. We show that participants reward employees more unequally when performance is reified, even though employees' levels of performance do not vary across conditions: most notably, the bonus gap between top- and bottom-performing employees increases by 20% between our non-reified and reified conditions, and it rises by another 10% when performance is presented as a quantified score. Further analyses suggest that reification fuels inequality both by reinforcing the authoritativeness of evaluation and by making observers more accepting of the idea that individuals can be meaningfully sorted into a merit hierarchy. This has direct implications for understanding the rise of legitimate inequality in societies characterized by the proliferation of reifying forms of evaluation.

Keywords: Evaluation, inequality, reification, quantification, performance, meritocracy.

JEL Codes: J3, M1, M5, Z1

Introduction

In modern societies premised on ideals of equal opportunity, the legitimacy of inequality rests largely on its meritocratic character: disparities in rewards are acceptable as long as they reflect differences in effort, ability, or performance across individuals. In practice, this means that measuring merit or performance is a crucial step toward achieving a just allocation of deserts. This is partially why evaluation devices such as ratings and rankings, which we use to measure merit, have become ubiquitous: they help us to give individuals, organizations, or products their due.

There is, however, a basic trade-off at the heart of this meritocratic enterprise: to be effective, measures of merit often need to be simple; as a consequence, they tend to forsake the intricacy, ambiguity, and multidimensionality inherent to any actual occurrence of merit. In school already, grades reduce the complexity of academic achievement to make students commensurable, thereby ensuring that the meritocratic school system runs smoothly and efficiently. Likewise, workable measures of employee performance, borrowers' creditworthiness, or school quality have a tendency to portray them as crisp and clear-cut things – by presenting them as ratings or scores, for example. To make merit count, they erase its fuzziness and complexity and instead turn it into a seemingly straightforward property of those being evaluated. This is what we refer to as the reification of merit.

In this article we show that the tendency of evaluation systems to reify merit or performance widens the gap in the rewards received by the winners and losers of meritocratic races. Specifically, we use a large experiment to show that the reification of employee performance in organizations increases inequality in the rewards observers are willing to allocate to higher- and lower-performing employees.

A long sociological tradition has theorized the unintended consequences of institutionalizing merit as a basis for the distribution of rewards in society. Young's (1958) seminal critique of meritocracy thus revolved around the idea that it created unduly reified hierarchies – merit-based hierarchies, to be sure, but hierarchies still – out of the greater equality that could have been achieved by accepting the essential diversity of human worthiness. Bell (1973) likewise warns that the pursuit of greater meritocracy in educational or occupational attainment must not obscure how distributional inequality matters – even when individuals are sorted fairly into ranks and positions. And recent scholarship on evaluation and classification systems calls for an exploration of how these systems induce stratification and inequality (Bowker and Star 1999; Brandtner 2017; Espeland and Stevens 1998; Fourcade 2016; Fourcade and Healy 2017; Lamont, Beljean, and Clair 2014; Lamont and Pierson 2019; Mau 2019).

Nevertheless, there is a dearth of empirical research studying how the concrete tools used for the production of meritocracy may lead to greater or lesser inequality in the rewards

received by those these tools deem more or less deserving. This article takes one step toward filling this gap, by showing how the tendency of merit-based evaluation systems to reify merit polarizes rewards between individuals at the top and at the bottom of merit distributions.

We explore these dynamics through a study of performance evaluations in the workplace, where reifying measures of performance have become increasingly widespread in recent decades (Castilla 2008). Performance metrics in organizations are touted for reducing the cognitive load associated with processing complex information (e.g. Jin and Leslie 2003; Thaler and Sunstein 2008). They are also perceived to limit discretion and bias and therefore to increase fairness in the way individuals are evaluated (Espeland and Stevens 1998; Pager and Shepherd 2008; Reskin 2000). At the same time, research shows that these effects are strongly context-dependent (Bosk 2019; Christin 2018; Dobbin, Shrage, and Kalev 2015; Heilman 2012; Rivera and Tilcsik 2019; Timmermans and Epstein 2010) and scholars have argued that performance metrics' seeming neutrality may paradoxically work to whitewash bias against disadvantaged groups and minorities (Lamont 2012; Lamont and Pierson 2019).

We contribute to these debates by showing that reifying measures of performance polarize the distribution of rewards between higher- and lower-performing individuals. Our experiment asks participants to divide a year-end bonus between a set of employees based on the reading of their annual performance reviews. In the experiment's non-reified condition, reviews are narrative evaluations. In the reified condition, the same narrative evaluations are accompanied by a crisp rating of the employees' performance. We find that participants reward employees more unequally when evaluation reifies performance, even though employees' levels of performance and relative performance do not vary across experimental conditions. We further show that reification shapes inequality through two mechanisms: by reinforcing the perceived authority of evaluation, on the one hand; and by lowering observers' resistance to the idea that individuals can be sorted into a bona fide hierarchy of merit, on the other. The first mechanism acts on the perceived legitimacy of the ratings assigned to individuals, the second on the perceived legitimacy of the act of sorting itself.

voting for democrats/lesser evils

By highlighting their tendency to reify merit as a consequential but often overlooked property of meritocratic evaluation systems, this article adds to traditional critiques of meritocracy that portray it as an unfulfilled promise – showing how, even in supposedly meritocratic systems, race-, class-, and gender-related barriers and prejudices continue to disadvantage underprivileged groups and minorities (e.g. Bourdieu and Passeron 1990; Castilla 2008; Rivera 2015; Rivera and Tilcsik 2016; for reviews, see Domina, Penner and Penner 2017; Pager and Shepherd 2008). Unlike these critiques, here we do not focus on how evaluation systems fail to be meritocratic – that is, on how they fall short of accurately identifying who is worthy and deserving. Rather, we show that these systems also shape how much difference will emerge between the outcomes of those deemed deserving and of

others. We thus demonstrate that the mundane implementation of meritocracy itself, to the extent that it entails a reification of merit, widens the gap between the winners and losers of the meritocratic evaluation process. Beside their greater or lesser meritocratic character, we therefore emphasize a previously underexplored dimension of evaluation systems: their propensity to create greater or lesser distributional inequality in the rewards they are designed to allocate.

We proceed by first outlining the concept of reification in greater detail and by elaborating how evaluation systems that reify merit or performance can polarize rewards between individuals at the top and at the bottom of merit hierarchies. We then present our experimental design and population of experimental participants. The second half of the article describes our findings. It shows that, when asked to reward a set of employees for their performance, third parties do so more unequally if performance is reified by the evaluation system. Further analyses suggest that reification fuels inequality both by reinforcing the authoritativeness of evaluation and by making third parties more accepting of the idea that evaluated individuals can be meaningfully sorted into a merit hierarchy. Finally, we find that reification shapes inequality more powerfully for a set of men than for set of women employees, paving the way for an exploration of how reification interacts with broader cultural understandings of those being evaluated.

Evaluation, Reification, and Inequality

The Unintended Consequences of Evaluation

Evaluation systems are designed to sort individuals, objects, and organizations according to their merit or quality, thereby enabling third parties to make informed choices about these entities (Espeland and Vannebo 2007; LeGrand and Enthoven 2007). Evaluation therefore has a profound impact on how people perceive and behave toward social reality (Espeland and Stevens 1998; Lamont 2012). The concrete workings of evaluation systems are rarely benign, however, and the process of evaluating tends to shape social worlds in ways that go beyond the mere recognition of greater or lesser quality. These unintended consequences have been emphasized by scholars focusing on how actors and organizations alter their behavior in order to game evaluation and reward systems (Espeland and Sauder 2016; Miller 2001; Power 1997; Sauder and Espeland 2007). Likewise, research has stressed how different evaluation criteria, the adoption of which is often politically negotiated, can result in vastly different outcomes when it comes to deciding who is more or less deserving (Boltanski and Thévenot 2006; Fourcade 2009; Guetzkow, Lamont, and Mallard 2004; Lamont 2010). As a consequence, evaluation frequently reproduces and legitimizes extant inequalities between social groups by relying on criteria that favor the already advantaged (Bourdieu and Passeron 1990; Karabel 2005; Lamont 2012).

Here we show that evaluation does not just unintendedly shape who gets recognized as worthy and deserving; it also bears on how much difference will emerge between the outcomes of those deemed deserving and of others. This means that while evaluation systems stratify by design, certain characteristics of these systems further polarize the rewards received by those at the top and at the bottom of the hierarchies that they create. Specifically, we demonstrate the role of one major feature of many evaluation systems: by presenting merit and performance as artificially crisp and clear-cut things, they tend to reify them – that is, to make them seem like tangible characteristics of the evaluated.

Reification as a By-Product of Evaluation

One of the virtues of evaluation systems is that they can reduce the complexity of merit and performance down to more tractable constructs, such as scores and categories of worth. This is appealing because it turns intrinsically different objects or people into commensurate and comparable entities (Espeland and Stevens 1998; Sauder and Espeland 2007). On the other hand, the simplicity of metrics and categories tends to erase the intricacy, ambiguity, and multidimensionality – in short, the messy reality – characteristic of any empirical instance of performance (Desrosières 2002; Muller 2018).

The simplicity of performance metrics matters because, in the eyes of observers, it turns what is effectively a bundle of disparate attitudes and behaviors into a seemingly tangible attribute. This is what we refer to as the reification of performance. As an example, in his history of psychological testing in the United States, Carson (2007) demonstrates how the introduction of a single measure of IQ obscured the complexity of cognitive ability, thereby entrenching people's sense that general intelligence was an actual entity which individuals possessed to different extents. Here we argue that performance evaluations in the workplace likewise reify merit by obscuring its inherent complexity, thereby making it look like an objective attribute of those being evaluated.

The reification of performance is akin to the process of quantification that has been explored in recent research (for reviews, see Berman and Hirschman 2018; Espeland and Stevens 2008; Mennicken and Espeland 2019), yet the two are empirically different. First, reification can occur without quantification, for example when performance is summarized in the form of an evaluative category such as "proficient" or "outstanding." Second, not all instances of quantification reify: as an example, quantified point estimates presented as ranges to account for measurement error might work to thwart the reification of the constructs that they capture. That being said, and to the extent that they reduce performance to a neat and orderly score, quantified evaluations do entail an element of reification. In fact, the extreme crispness and seeming precision of simple numerical scores mean that they are particularly powerful reifiers. Thus, we suggest that the inequality-inducing effects of quantified evaluation measures can to a large extent be explained by their reifying character.

Mechanisms: How the Reification of Merit Breeds Inequality

We envision two key mechanisms whereby reification, as a byproduct of evaluation, may polarize rewards between those at the top and those at the bottom of merit hierarchies. The first is that reifying measures of merit reinforce the perceived authority of evaluation. Prior research has shown that observers trust numbers because they seem to emanate from objective and rational measurement procedures (Porter 1996; Espeland and Stevens 2008). Here we suggest that the firmness and seeming accuracy of reifying measures likewise lend a sense of greater robustness and objectivity to the evaluation process that generated them. From this it follows that reifying performance evaluations are powerful drivers of inequality, because they provide observers with seemingly objective and authoritative bases on which to differentiate between evaluated actors. As we explain at greater length in the empirical analysis below, we test this mechanism by examining whether reification shapes inequality differently for different kinds of observers: those typically skeptical of the authority of evaluation and reward systems, and those who tend to trust them.

The second reason why performance reification polarizes rewards between the winners and losers of evaluation processes, we hypothesize, is that it makes observers more accepting of the very idea of a hierarchy of merit among the evaluated. Scholars have stressed how evaluation and classification systems legitimize stratification – that is, the very act of stratifying – within populations of evaluated entities (Accominotti 2019; Espeland and Stevens 1998; Fischer et al. 1996; Fourcade 2016; Lamont 2012; Stevens 2007). This phenomenon is ubiquitous in data-rich societies, where scoring techniques increasingly sort individuals, products, and organizations by performance, risk, or creditworthiness, thereby entrenching the belief that there is such a thing as a meaningful hierarchy of performance, of risk, or of creditworthiness (e.g. Fourcade and Healy 2017; Kiviat 2019; Lauer 2018; Mau 2019).

Yet while prior work insists that evaluation legitimizes the idea of hierarchy, there is a dearth of empirical work exploring the direct effects of this legitimization on the polarization of outcomes among evaluated entities (see Sauder, Lynn, and Podolny 2012). To our knowledge, the only attempt in that direction is Sauder's (2006) work showing that, in the context of American legal education, the institutionalization of a formal ranking system increased inequality between top and bottom law schools by augmenting the number of distinctions actors perceived as relevant between them. We here build on this work to predict that when evaluating performance, reifying measures make observers more likely to see the evaluated as sortable in hierarchical terms, thereby leading to greater polarization in the rewards received by those at the top and at the bottom of the performance distribution.

To test this second mechanism, we ask whether reification shapes inequality differently when applied to groups of individuals who, culturally, might be more or less prone to be seen in terms of sharp hierarchies. Here we take advantage of a growing body of

scholarship showing how prevalent gender stereotypes make men more likely to be thought of as occupying extreme positions in hierarchies of competence. Research thus finds that in a variety of social contexts, high-performing men are more likely to be described as brilliant or exceptional than are their female counterparts (Bian, Leslie, and Cimpian 2017; Leslie et al. 2016; Musto 2019; Rivera and Tilcsik 2019). Conversely, work on performance evaluations at U.S. law and service firms suggests that similarly low levels of performance translate into lower numerical ratings and smaller salary growth for men than they do for women (Biernat, Tocci, and Williams 2012; Castilla 2008). If the reification of performance indeed shapes inequality by making observers more amenable to the idea of a hierarchy of merit – an outcome which, due to the aforementioned stereotypes, should be more easily achieved when evaluating a group of men than when evaluating a group of women – we expect it to have larger effects on a set of male than on a set of female actors.

In sum, we argue that reifying forms of evaluation compound inequality in the rewards accruing to individuals at the top and at the bottom of performance hierarchies through two main mechanisms: by lending evaluation greater authority, on the one hand, and by increasing observers' willingness to regard the evaluated in hierarchical terms, on the other.

performance reviews in office settings are usually promotion markers, you'd need ot really fuck up to not have a general upward trajectory in your review, although in service workers reviews are more punitive, not for increasing career trajectory. are you going to get fired or not? are you going to get punished or not?

tech and other office workers, for whom meritocracy begins in college admissions (or earlier, in G&T programs), and have it reinforced their whole lives, more likely to believe in it, more likely to "reward" with their social actions/solidarity those who they believe "deserve" it

Experimental Design

Overview

To test whether the reification of merit makes third-party observers more willing to reward individuals unequally, we created an artificial, small-scale meritocracy. Specifically, we devised an experiment in which we asked participants to divvy up a year-end bonus between three employees based on the reading of their annual performance reviews. We then manipulated the degree of reification of employee performance participants were exposed to across experimental conditions. In our baseline, non-reified condition, the performance of each employee was conveyed in a narrative report. It therefore retained some of the intricacy inherent to any occurrence of performance. In our two reified conditions, by contrast, the same narrative report was accompanied by a crisp rating of the employee's performance. Importantly, the introduction of these ratings did not alter the level of performance or relative performance of the employees, so that only the degree of reification of performance varied across experimental conditions. We come back to how this was achieved further down in this section.

We had written the three employee reports so that there would be a relatively salient top, medium, and bottom performer among the three employees. We therefore expected participants to reward employees unequally in all conditions. In line with our main

hypothesis, however, we also expected them to distribute the bonus more unequally when presented with crisp ratings in the reified conditions[1].

Procedure

We used Amazon's Mechanical Turk service (henceforth, MTurk) to recruit a total of 3,900 participants living in the United States. The service has been used widely in recent years for conducting experiments in various fields of social science (for reviews, see Buhrmester, Talaifar, and Gosling 2018; Hauser, Paolacci, and Chandler 2018). MTurk samples offer access to a diverse, if not perfectly representative, cross-section of the U.S. population, affording researchers more generalizability than traditional experimental subject pools (Horton, Rand, and Zeckhauser 2011; Weinberg, Freese, and McElhattan 2014; see also Coppock 2019). MTurk respondents, who perform tasks on the platform for compensation, have been shown to provide comparable or higher quality responses than those obtained from other online panels or undergraduate lab samples (Hauser and Schwarz 2016; Kees et al. 2017; Paolacci and Chandler 2014). This is particularly true for vignette-based experiments (Weinberg, Freese, and McElhattan 2014).

Our 3,900 participants were directed to an online survey platform, where they were invited to take part in a research study on "employee evaluations and how they influence the decisions managers make about employee compensation." The study was presented as a project by researchers at Columbia University. Participants were reminded that employee pay and compensation are important issues in today's economy – ones that have significant effects on the lives of almost everyone participating in the workforce. They also learned that if they agreed to be part of the study they would be presented with the performance reports of employees whose identities had been anonymized, and asked to answer a short survey about these reports.

Participants opting into the study were next invited to read the annual performance reviews of three employees before deciding how to divide a $10,000 year-end bonus between them. This step, which formed the main task of our experiment, was modeled on a two-stage performance-reward process commonly used in larger organizations (Castilla 2008; Castilla and Benard 2010). In this process, the performance of employees is first evaluated by a supervisor. In a second stage, and based on these performance evaluations, employees may then be awarded a bonus by a manager higher-up in the hierarchy. Our experiment placed participants in the position of this higher-up manager. We instructed them that while it might feel difficult to hand out bonuses based on the information they were provided, we simply expected them to do so to the best of their ability. This was meant to make respondents confident in the legitimacy of their bonus allocations. Finally, to ensure that participants would pay attention to the details of the study – and in particular to the text of

[1] The experiment was approved by Columbia University's Institutional Review Board (protocol number IRB-AAAR8920) and preregistered through the Open Science Framework (registration form available at https://osf.io/s7zu4/).

the performance reports – we indicated that we would later ask them questions about the reports and about their decisions. After reading these instructions, subjects proceeded to opening the reports and dividing the bonus.

The three employees were fictitious, as was the company they were employed at, although this was not known to participants. They were presented as occupying the same junior position in the same division at a medium-sized U.S. firm, and as having the same tenure in this position. This was intended to level the field of employees, so that the allocation of bonuses between them would rest solely on perceived performance, and not on seniority or job description. The employees' position was described as "business coordinator" – a middle-status occupation whose designation suggests that it is neither male- nor female-dominated, and which in the United States displays little evidence of a gender pay gap.[2] Our goal in selecting this occupation was to make the employees relatable to participants, thereby increasing the chances that they would feel competent to complete our task. We further redacted the names of the employees and of the supervisor who we claimed had written the performance reports, as well as the name and logo of the company they worked for. [3]

Finally, because we did not want gender bias to interfere with our main manipulation in shaping how participants handed out individual bonuses, we ran two separate versions of the experiment by manipulating the gendered pronouns appearing in the reports: in the first version the three employees were presented as men; in the second they appeared as women. In the analysis to follow we report the findings from these two versions in turn, and we use differences in outcomes across versions to test one of the mechanisms whereby we expect reification to polarize rewards among employees.

Manipulating the Reification of Performance

Participants were randomly allocated to one of three experimental conditions. In the *non-reified* condition, the performance report of each employee was made up of three paragraphs of narrative evaluation, totaling around five hundred words. The paragraphs appeared under the headings "Evaluate and discuss the employee's job performance," "Are there areas of particular performance that should be noted?" and "Are there areas of performance needing more attention or improvement?" This format was modeled on a typical performance review form used as teaching material at a major U.S. business school. Besides writing the reports in a way that conveyed a difference in performance between the three employees, we also strived to eschew language that might have been interpreted

[2] According to job search engine Glassdoor, the base pay of business coordinators in the U.S. was $42,836 in 2019, and women business coordinators earned on average 0.5% more than their men counterparts.
[3] The reports were photocopied and manually redacted to provide them a genuine look. Responses to an open-ended question soliciting participants' feedback indicated that the reports were largely perceived as authentic.

more or less positively in light of employee gender.[4] This was meant to facilitate the comparison of reification's impact on inequality between the experiment's male- and female-gendered versions.

In the experiment's *reified* condition, participants were presented with the same narrative reports as in the baseline, non-reified condition, yet these reports were followed by a crisp rating of each employee's performance. This rating appeared under the heading "Overall performance assessment" and took the form of a tick on a horizontal bar, the nine graduations of which corresponded to verbal descriptions ranging from "unacceptable" to "exceptional."

A key feature of our manipulation was that, unlike the degree of reification of performance, the level of performance or relative performance of the three employees would not vary across experimental conditions. We therefore needed to ensure that the individual rating each employee received in the reified condition did not convey a different level of performance from the one communicated by their narrative report. Here we proceeded in two steps. After we had drafted our narrative reports, we first asked an independent group of respondents, recruited from the same pool of Mechanical Turk workers as experimental participants, to rate each report on the same nine-point scale as the one that would appear in the experiment's reified condition. This independent set of respondents was exposed to the same instructions as experimental participants: they were asked to read the three reports before rating the employees; the order of appearance of reports was randomly rotated from one respondent to the next (as it was in the experiment); respondents had to answer the same attention checks; and we applied the same quality criteria to decide which responses to include in our analysis. The only difference was that they were asked to rate employees instead of dividing a bonus between them.[5] In a second step, we inserted the average rating each employee had received from this set of independent respondents as the crisp rating in this employee's reified performance report. Thus, ratings in the reified condition conveyed the same level of performance to experimental participants on average as did the attached narrative reports.[6]

It is important to note that participants were not aware that the ratings they were looking at had been crowd-sourced. As far as they knew, the whole performance report had been

[4] In fact, when we asked an independent set of online respondents to rate the three reports, for some in their male and for others in their female version, we did not find the average ratings of employees to vary significantly by gender condition.

[5] We used the data of 79 independent raters for the male version of the reports, and 101 for the female version. Respondents who had served as raters in this first stage were not eligible to participate in the experiment. There were no noticeable differences in reported income, political orientation, gender makeup or race and ethnicity makeup between raters and experimental participants. Raters' mean age, however, was slightly lower (34.6 vs. 37.9, $p < .01$).

[6] To be sure, some individuals assigned to the reified condition might have experienced a discrepancy between the narrative portion of each report and its attached rating. On average, however, we can expect these discrepancies to have been experienced to the same extent in one and in the other direction.

composed by the employees' direct supervisor. It is also important to stress that by introducing ratings under the heading "Overall performance assessment," we aimed to suggest that they were a summary, by the supervisor, of their foregoing narrative assessment, and therefore that they do not channel fresh information that would have been available to the supervisor but not to participants.

Finally, participants could be assigned to a third experimental condition, where the crowd-sourced rating of each employee's performance was presented as a quantified score. This score appeared on a horizontal scale of 1 to 9, the units of which also bore the verbal descriptions "unacceptable" to "exceptional." The performance reports in this third condition, which we refer to as the *reified-quantified* condition, were otherwise identical to the ones featured in the reified condition. By dissociating quantification from reification, the introduction of this condition enables us to ask whether the latter shapes inequality even when it lacks the authority observers typically associate with numbers (Porter 1996; Espeland and Stevens 2008). On the other hand, and to the extent that numerical scores convey greater precision than mere ticks on a verbal scale, quantification also acts as a particularly forceful agent of reification. As a consequence, we expect bonus allocations to be more unequal in the reified-quantified condition than in the reified one.

After they had completed our main task, participants were asked to answer a series of comprehension questions and to fill out a short survey of their demographics. They were then prompted to comment on how they had dealt with our main task in two open-ended questions. Finally, they were debriefed and compensated.[7]

Exclusion Criteria

In the analyses to follow, we exclude the responses of participants whose behavior on the survey platform indicated that they had not engaged with all three reports, who did not spend a meaningful amount of time one the task, or who failed to answer two out of three comprehension questions about the reports' narrative sections.[8] The same exclusion criteria were applied to the responses we used when crowd-sourcing the ratings of the three employees. Excluding respondents who had not gone through the reports carefully means that participants in the reified and reified-quantified conditions were exposed to at least some of the intricacy inherent to each employee's performance. This, we stress, is a conservative choice as far as identifying a causal effect of reification is concerned. In fact, we find that in another version of the analysis, where we relax our inclusion criteria to

[7] Respondents were compensated for completing the survey regardless of the quality of their responses to comprehension questions. We paid them $1 for an approximately six-minute survey.

[8] We provide a detailed account of our exclusion criteria in appendix A. We signaled the presence of later comprehension questions ahead of our main task to secure our participants' attention and to suggest they might be held accountable for their choices. We did not explicitly suggest that respondents might not be compensated if they answered these attention checks incorrectly, however, and in fact we compensated everyone who completed the study.

consider responses by everyone who had opened the three reports, reification and reification-quantification act more powerfully on inequality.[9]

After applying our exclusion criteria, we analyze the responses of 2,844 unique participants.[10] These participants' demographics are balanced across experimental conditions: we find no significant difference in the average age, gender, racial identification, household income, or political leaning reported by participants in the three conditions. Among those who passed our exclusion criteria, we also find no significant difference by condition in the time participants spent on our main task, nor in the average number of attention checks they answered correctly.

Reification and the Polarization of Rewards

We first focus on inequality in the average bonuses received by employees across experimental conditions: we therefore adopt the perspective of employees and look at what each could expect if we were to average the decisions of multiple bonus allocators under more or less reifying evaluation systems. To better understand how reification shapes inequality in average outcomes, we then turn to how it affected the bonuses awarded by individual experimental participants. We report the findings from the male-gendered version of our experiment, before comparing them to those obtained in the female-gendered version.

In line with the hierarchy of performance we had built into the three reports, participants in the non-reified condition rewarded the three employees unequally: our top, medium, and bottom performers received an average of $4,264, $3,359, and $2,377 in this condition (all pairwise differences were significant at the $p < .001$ level). The difference in average compensation between the top and bottom performers – that is, the bonus gap between them – therefore amounted to $1,887. In this section we use this intuitive notion of the "top-to bottom-performer bonus gap" as our main measure of inequality. In a scenario such as ours, featuring three employees and a set bonus to divide among them, it captures the same information as a Gini coefficient.

Even though employees' level of performance or relative performance did not vary across conditions, participants allocated rewards more unequally when performance was reified. Compared to our baseline condition, the bonus gap between top and bottom performers rose by a substantial $368, or 20%, to $2,255 in the reified condition where performance reports were accompanied by a simple tick on a verbal scale ($p < .001$). Figure 1 illustrates how this rise was almost equally accounted for by a $191 increase in the top performer's

[9] The Gini coefficient measuring inequality in the bonuses of the three employees is 16% higher when we look at the responses passing these laxer inclusion criteria.

[10] We report summary statistics on our sample's demographics in appendix B, where we also compare these demographics to those of representative samples of the U.S. population.

average bonus and a $177 decrease in the bottom performer's one. The bonus of the medium performer, on the other hand, was left virtually unchanged by this first layer of reification.

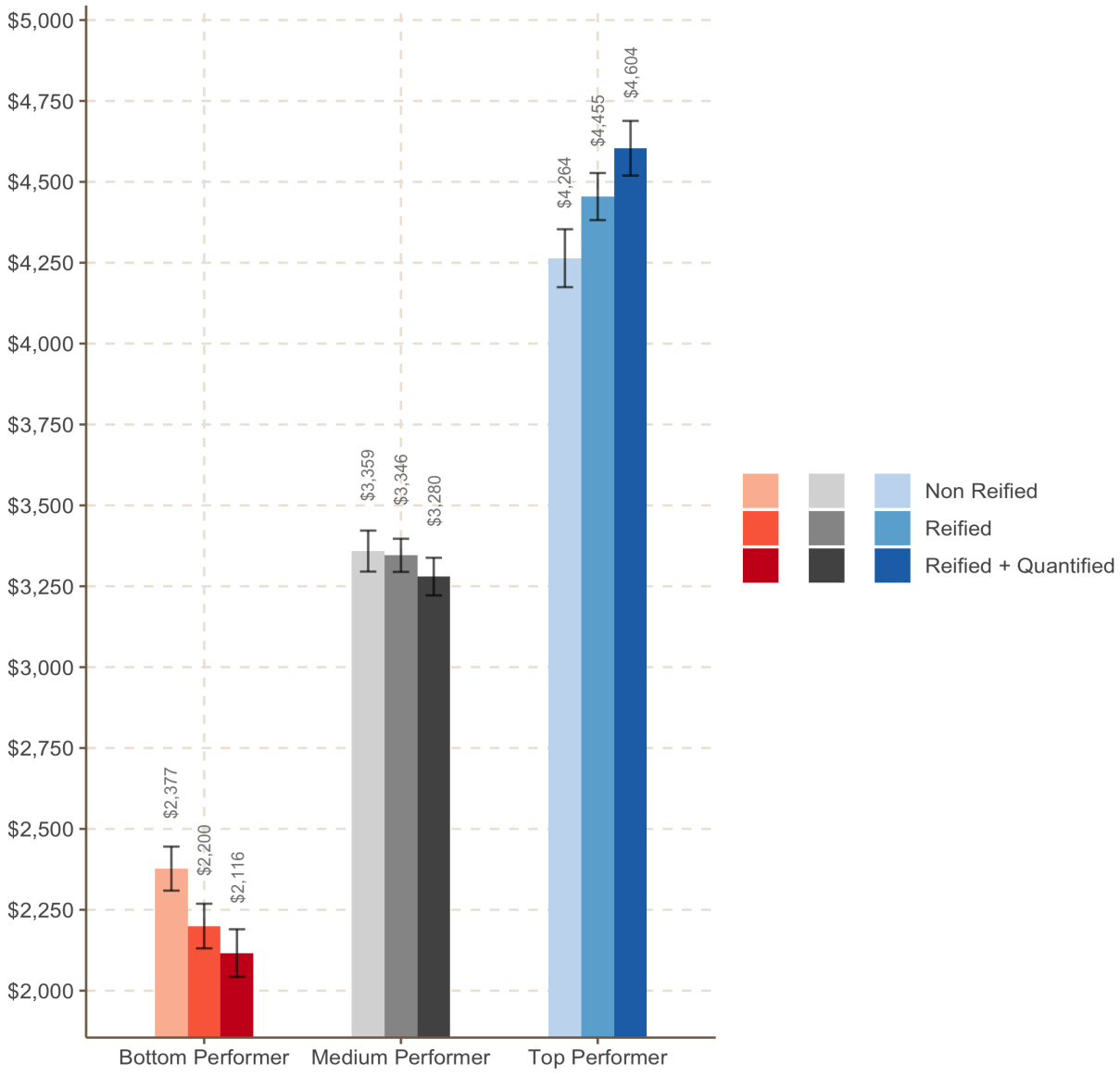

Figure 1. Average bonuses allocated to top-, medium-, and bottom-performing employees in the non-reified, reified, and reified-quantified conditions (throughout figures, brackets report 95% confidence intervals).

Presenting performance as a quantified score further polarized the distribution of rewards. The bonus gap rose by an additional $233, or 10%, between the reified and reified-quantified conditions ($p < .05$). As shown in figure 1, this rise again reflected both an increase in the compensation going to the top performer and, to a lesser extent, a decrease in the bonus of the bottom achiever. Figure 2 summarizes how reification and quantification magnify inequality between employees by polarizing the rewards allocated to top and bottom performers.

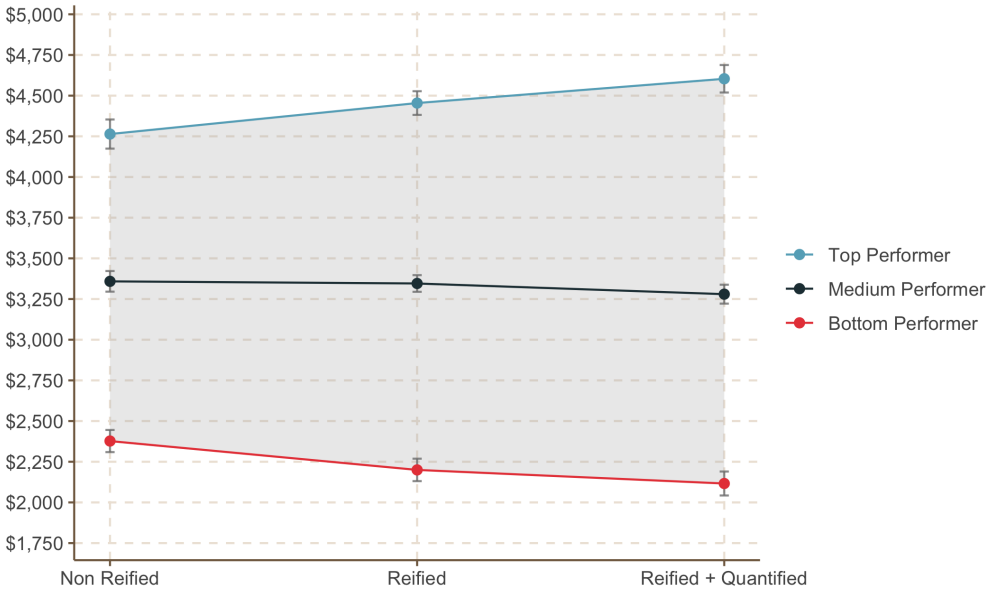

Figure 2. Reification polarizes the distribution of rewards between top and bottom performers: Average bonuses awarded to top- and bottom-performing employees, by experimental condition.

Because the ratings in the two reified conditions had been crowd-sourced, they did not alter the mean levels of performance or relative performance participants perceived in our set of employees. As such, reification did not shape inequality through the introduction of additional information about each employee's performance. On the other hand, the reifying ratings were stronger signals of employees' performance: they presented the same levels of performance more saliently. We therefore expected lesser dispersion around each employee's mean bonus in the two reified conditions. Here we find that the variance of employee bonuses indeed decreased significantly from the non-reified to the reified conditions ($p < .01$ in either case, Fligner-Killeen tests for the homogeneity of variances). We find no significant difference in the dispersion of each employee's bonus between the reified and reified-quantified conditions, suggesting that the crispness of the reifying signal, and not its numerical character, was doing the heavy-lifting of reducing the dispersion observed in the non-reified scenario.

Our first series of findings therefore demonstrates that reifying measures of workplace performance do not just vacuum up the complexity and ambiguity characteristic of any instance of performance: in doing so they also increase the level of inequality one will typically observe between high and low achievers. Reification thus works as a diffraction prism polarizing rewards between the deserving and others. Our next section examines the evidence in support of two mechanisms that may explain this outcome: it asks whether reification shapes inequality by reinforcing the perceived authority of evaluation, on the one hand; and by lowering observers' resistance to the idea that individuals can be sorted into a bona fide hierarchy of merit, on the other.

Pathways to Inequality

To understand how the reification of merit polarizes rewards, we need to shift from focusing on the average outcome each employee could expect and to analyze how reification shaped the bonuses awarded by each individual participant. From this point on we therefore look at the level of inequality found in each individual respondent's allocation of bonuses. While this information could equivalently be expressed as a top- to bottom-performer bonus gap, to clearly distinguish these findings from earlier ones we present all respondent-level analyses using the Gini coefficient associated with each participant's bonus distribution.[11] Before proceeding to mechanisms, we model this respondent-level Gini as an outcome of respondents' experimental condition and individual characteristics. This enables us to compare the effect of reification to that of other typical drivers of individuals' willingness to accept inequality in economic outcomes, such as respondents' income or political leaning. Model 1 in table 1 shows that participants' Gini in the non-reified condition was .147 on average, and that the reification and quantification of performance increased this Gini by respectively .018 (or about 12%; $p < .01$) and .033 (22%; $p < .01$). The magnitude of these effects remains stable when we adjust for participants' characteristics in model 2. A stepwise regression, the outcome of which we report in model 3, shows that modeling participants' income, political orientation, age, as well as whether they had received a performance review at one point in their career slightly improves our model fit. Here we find that leaning Republican, reporting a higher income, being older, and having previously received a performance review were all associated with a greater willingness to reward employees unequally. Specifically, the difference in respondents' Gini between our non-reified and reified conditions (.017) was of similar magnitude to that between participants whose household income was under $10,000 and over $150,000, respectively the bottom and top end of our 12-point income scale (.002 x 11 = .022). Participants' increased propensity to inequality from the non-reified to the reified scenario also topped the difference between a strong Democrat and a strong Republican (.002 x 6 = .012) or between a subject in their 20s and one in their 60s (.0004 x 40 = .16). Finally, we find no

[11] Our findings are robust to using other measures of inequality, such as the standard deviation of bonuses awarded to the three employees.

statistically significant interaction effects between participants' characteristics and our two treatment conditions.

Table 1. OLS Regressions Predicting Inequality in Participants' Bonus Allocations

	Dependent variable: Participant's Gini		
	Model 1	Model 2	Model 3
Reified	0.0178**	0.0172**	0.0174**
	(0.0061)	(0.0061)	(0.0061)
Reified + Quantified	0.0325**	0.0321**	0.0321**
	(0.0062)	(0.0062)	(0.0061)
Income [1-12 scale]		0.0023**	0.0022**
		(0.0009)	(0.0008)
Political Leaning (D-R) [1-7 scale]		0.0021	0.0021
		(0.0013)	(0.0013)
Age		0.0004+	0.0004+
		(0.0002)	(0.0002)
Gender		0.0003	
		(0.0051)	
Received Performance Report		0.0131	0.0118
		(0.0081)	(0.0078)
Gave Performance Report		-0.0031	
		(0.0055)	
Constant	0.1468**	0.0996**	0.1007**
	(0.0043)	(0.0123)	(0.0119)
Observations	1,334	1,334	1,334
F	13.99***	5.90***	7.82***

Note: ** $p<.01$; * $p<.05$; + $p<.1$ (two-tailed tests). Standard errors in parentheses.

The Authority of Reifying Information

We next test our first mechanism and ask whether reification increases inequality by reinforcing the perceived authority of performance evaluations: even though the crisp ratings in our two treatment conditions were presented as mere summaries of the foregoing narrative reports, their reifying character may have come with a sense of greater robustness and objectivity that might have strengthened participants' feeling that they were basing their decisions on trustworthy information (Espeland and Stevens 2008; Porter 1996; Springer 2019). By contrast, respondents in the non-reified condition may have erred on the side of equality because they did not feel they had authoritative enough information to discriminate strongly between profiles.

To test this mechanism, we observed whether reification acted differently on different kinds of respondents: if reification increases observers' willingness to be unequal by augmenting the authority of evaluation, we reasoned, it should act more powerfully on individuals who are more skeptical of this authority in the first place. We therefore measured how respondents scored on a scale measuring their general trust in evaluation and reward systems. To this end, three weeks after we ran our experiment, we reached out again to participants who had been assigned to the male-gendered version of it and asked them to fill out an ostensibly unrelated survey. The delay between experiment and survey, together with the lack of any apparent relationship between the two (we used different institutional affiliations and graphical templates when introducing each), made it unlikely that respondents' answers to the survey would have been primed by their experience of the experiment.

The survey consisted of the seven items proposed by Lipkus (1991) to measure the classic social psychological construct of belief in a just world – that is, respondents' perception that in social life, individuals tend to get the rewards that they deserve (Rubin and Peplau 1973, 1975; for a review, see Furnham 2003). We used Lipkus's belief in a just world scale because it measures respondents' beliefs that the world is just with others, as opposed to self (Lipkus, Dalbert, and Siegler 1996). Respondents who score high on this scale – just world believers – perceive interindividual differences in outcomes to be just because they believe deserts are generally allocated through fair procedures. Respondents who score low – just world skeptics – challenge the idea that rewards go to the deserving.[12]

We then looked back at how just world believers and just world skeptics had compensated the three employees, and at how their bonus allocations had responded to the reification of performance in our experiment. Figure 3 shows that just world believers were more likely to reward the employees unequally than were just world skeptics ($p < .14$). Crucially, believers

[12] Of the 1,414 experimental participants we reached out to, 594, or about 42%, completed our follow-up survey. There were no significant differences in household income, political leaning, gender makeup, or race and ethnicity makeup between participants in the experiment and respondents to the survey. Survey respondents' mean age, however, was slightly higher (39.7 versus 37.3, $p < .01$). Cronbach's alpha between the seven items of Lipkus's belief in a just world scale was .90.

and skeptics responded differently to our experimental manipulation. In figure 4 we divide participants who completed our follow-up survey into three equally sized groups: those with respectively the lowest, middle, and highest scores on our belief in a just world scale. The figure shows that the willingness of strong believers to reward employees unequally was virtually unchanged by the reification of performance. Strong skeptics' willingness to be unequal, on the other hand, significantly increased in the reified and reified-quantified conditions ($p < .05$, one-way ANOVA), while the magnitude of medium believers' response fell somewhere in the middle. This means that the more skeptical participants were of the authority of evaluation and reward systems in the first place, the more traction reification had in making them see the employees as unequally deserving. This lends support to the idea that reification shapes inequality by increasing the perceived authoritativeness of the evaluation process.[13]

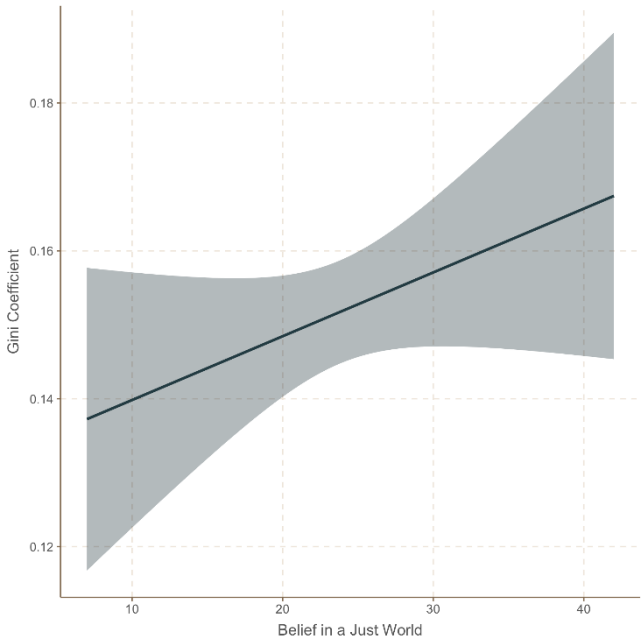

Figure 3. Gini coefficient of participants' bonus allocations, by participant belief in a just world.

[13] We found a negative interaction between respondents' exposure to reification and their beliefs in a just world when modeling participants' bonus allocations and adjusting for other predictors of inequality. This interaction was not statistically significant, however, possibly due to our limited statistical power.

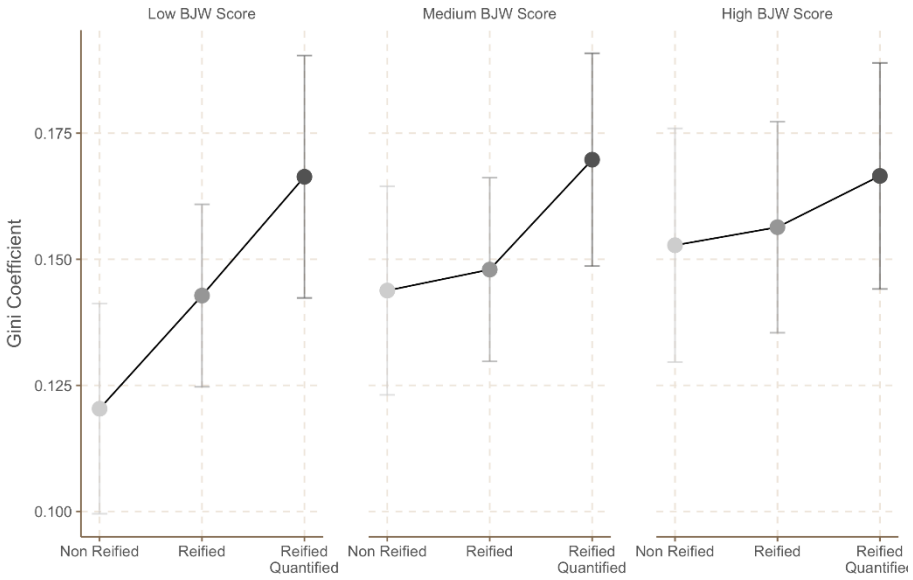

Figure 4. Gini coefficient of participants' bonus allocations, by experimental condition, for participants with respectively high, medium, and low levels of belief in a just world.

Lowering Resistance to Hierarchy

Does the reification of merit fuel inequality by making observers more amenable to the very idea of a merit hierarchy? In the context of our experiment, this would mean that reification acted by increasing participants' likelihood to regard different employees as unequal in terms of their merit and deservingness. To substantiate this mechanism, we take two further analytical steps.

We first look more closely at how participants in each condition divided the overall bonus among the employees. We focus in particular on the proportion of respondents who split the bonus perfectly equally across the three profiles, thereby making a statement that they did not want to see them as unequally deserving. We define a "perfect equalizer" as a participant whose bonus allocation allowed for a gap of no more than one dollar between the bonuses of any two employees. This might mean that two employees got $3,333 each while the third received $3,334, for example – or that two employees got $3,333.33 and the third $3,333.34.[14] There were 4.4% perfect equalizers among the participants assigned to our non-reified condition, more than double the 2.0% and 1.9% we observed in our reified and reified-quantified conditions ($p < .001$ in either case, chi-squared tests). This is a first

[14] It is important to note that perfect equalizers did not fail to engage with our task: to be included in our analysis they had to pass our exclusion criteria, and in fact we find that they did not significantly depart from other participants in the time they spent on the task nor in the number of attention checks they answered correctly.

piece of evidence that reification makes observers less reluctant to the idea that individuals can be sorted into a merit hierarchy.

Reification, Gender, and the Acceptability of Hierarchy

In a second step, we finally examine how the reification of performance acts differently on different kinds of evaluated employees. Specifically, we test whether reification shapes inequality differently when applied to groups of individuals who, culturally, might be more or less prone to be seen in terms of sharp hierarchies. Here we take advantage of a growing body of gender research showing how, in a variety of social settings, men are more likely to be thought of as occupying extreme positions in hierarchies of competence (Bian, Leslie, and Cimpian 2017; Biernat, Tocci, and Williams 2012; Leslie et al. 2016; Musto 2019; Rivera and Tilcsik 2019). If the reification of performance indeed works by making observers more accepting of the idea of a hierarchy of merit – an outcome which due to these gender stereotypes should be more easily achieved when evaluating a group of men than when evaluating a group of women – we reason that it should shape inequality more forcefully for a set of men than for a set of women employees.

To test this expectation, we ran a second version of our experiment, in which we switched all gendered pronouns in our narrative performance reports from male to female pronouns.[15] Figure 5 reports the average bonus gaps we observed between top and bottom performers in this version, alongside those we had observed when the same employees were portrayed as men. It shows that performance reification shaped inequality more forcefully in the group of male employees than it did in the group of female employees: while the bonus gap grew with reification in both gender conditions, it did so more sharply between top- and bottom-performing men than it did between top- and bottom-performing women. A two-way ANOVA shows a significant interaction between experimental and gender conditions ($p < .05$). In figure 6 we show that this finding holds true when we focus only on men or only on women respondents. Figure 7 finally disaggregates the finding further, by showing that it arises from participants' tendency to both over-reward the top male profile and over-penalize the bottom male profile, relative to their female counterparts, when performance is reified.[16] This means that the comparatively small growth of the bonus gap for women in the reified and reified-quantified conditions does not emerge solely from a reluctance to extend high rewards to top-achieving women. Instead, we find that

[15] In a first step, we asked an independent set of online participants (n = 101) to rate the three female reports on a scale of "unacceptable" to "exceptional," as we had with the male reports. We did not find the average ratings of our top, medium, or bottom performers to differ significantly between the male and female versions of the reports, possibly because we had avoided language whose perceived value might have depended on employee gender when drafting them in the first place. The female version of the experiment was completed by 1,430 participants.

[16] Importantly, participants were not systematically more unequal with employees of either gender: in the non-reified condition the top- to bottom-performer bonus gap of women employees was slightly larger than that of their men counterparts; there was no difference in the reified condition; and the gap was slightly smaller for women in the reified-quantified condition

performance reification “sticks more” with a set of men employees than it does with a set of women ones. This lends further support to the idea that reification acts by lowering respondents’ skepticism of the existence of a meaningful hierarchy of merit – something that is more easily achieved when evaluating men than when evaluating women.

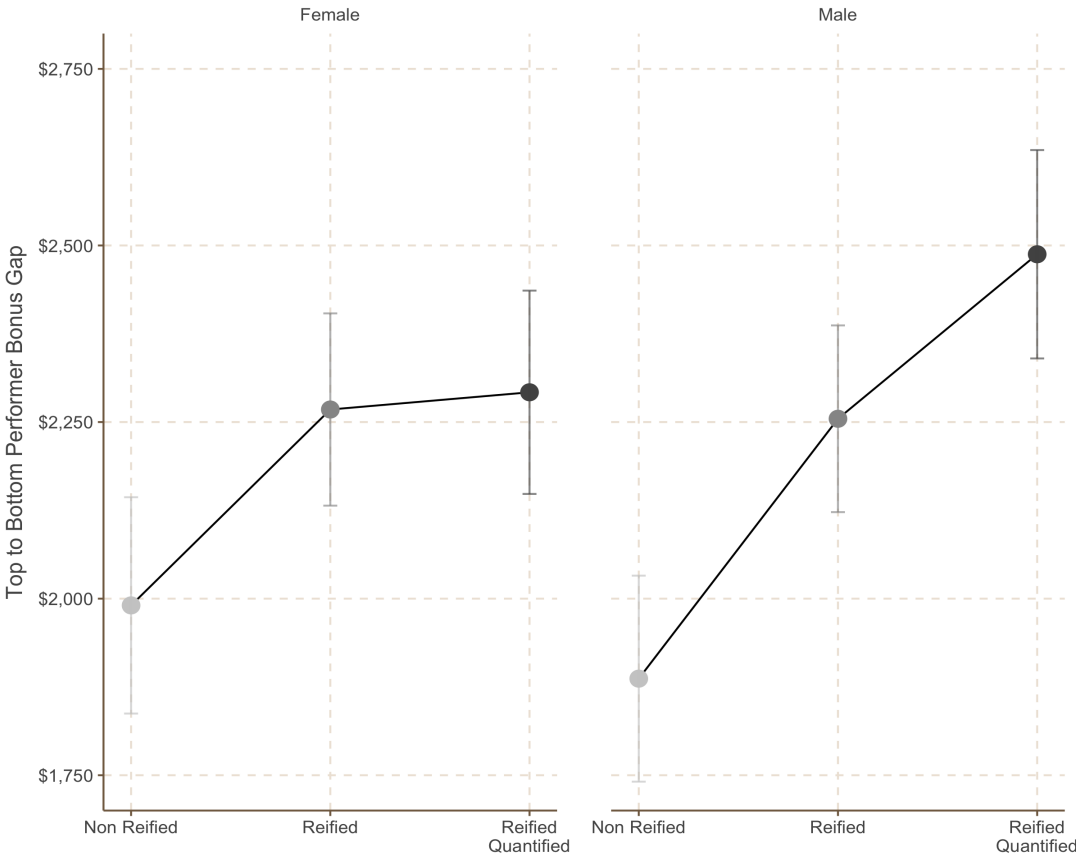

Figure 5. Average bonus gap between top- and bottom-performing employees in the experiment’s female and male versions, by experimental condition.

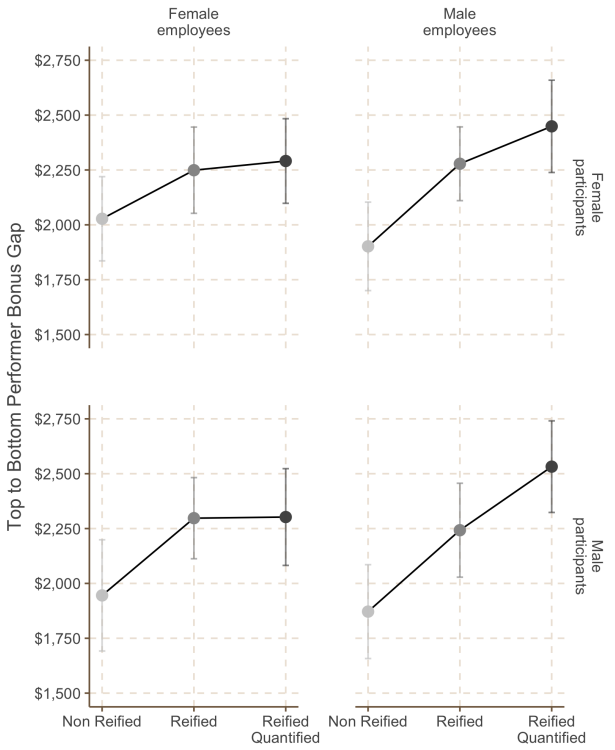

Figure 6. Average bonus gap between top- and bottom-performing employees in the experiment's female and male versions, by experimental condition and participant gender.

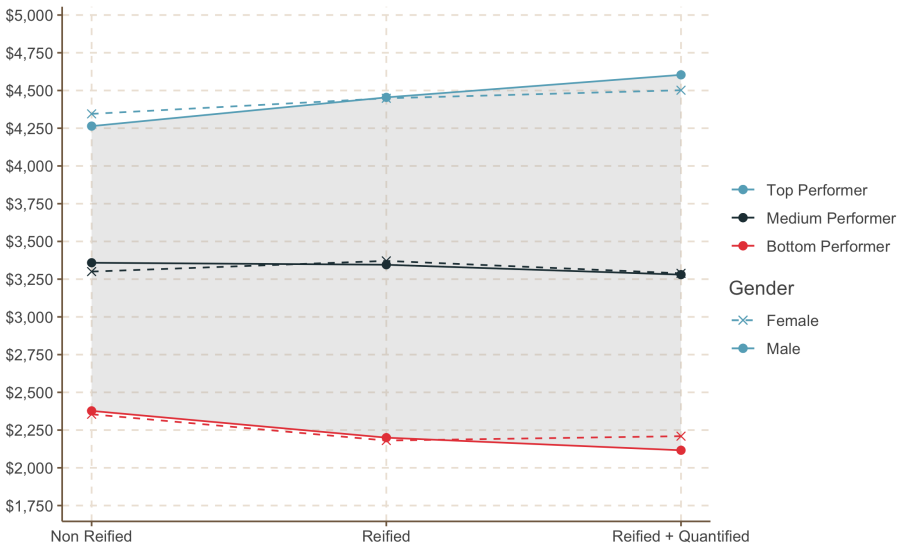

Figure 7. Reification polarizes the distribution of rewards more forcefully among men than among women: Average bonuses awarded to top- and bottom-performing employees in the experiment's female and male versions, by experimental condition.

Conclusion

In this article we created a small-scale meritocracy and we used this setting to explore how a key feature of merit-based evaluation systems shapes inequality in the rewards received by the winners and losers of the meritocratic race. Through a controlled experiment, we showed that evaluation systems that reify merit or performance make third-party observers willing to be more unequal in how they reward those these systems deem to be more or less deserving. Our analysis also provides support for two mechanisms whereby reification polarizes the distribution of rewards between the deserving and others: we find, first, that reifying measures of performance increase the authoritativeness of evaluation; and, second, that they make observers less resistant to sorting individuals into a hierarchy of performance.

There are several notable limitations to our study. First, our group of participants, though diverse, does not constitute a nationally representative sample of the U.S. population, nor is it representative of a professional group – such as managers – whose job it is to hand out bonuses to individuals in the workforce. Additional steps in our analysis would be warranted to test whether our results generalize to these populations of interest. Second, while we find that the reification of performance shapes inequality more powerfully for a group of men than for a group of women employees, we do not conduct a systematic examination of how reification interacts with broader cultural understandings of evaluated entities. Further research could explore how cultural expectations attached to different kinds of evaluated entities – whether they are people, objects, or organizations, for example – shape the effect of reification on observers' willingness to reward them unequally. Third, this article studies how reification affects inequality in performance pay in the case of the U.S. workplace. This begs the question of how the dynamics of reification and inequality might play out in other national contexts, for example those characterized by lesser income inequality or lesser reliance on performance pay.

These limitations notwithstanding, we wish to stress three broader implications of our findings. First, our argument offers a way of thinking about the production of distributional inequality in societies characterized by the increasing ubiquity of merit- or performance-based classification systems. Scholars have long argued that these systems induce inequality (e.g. Bowker and Star 1999; Espeland and Stevens 1998; Mau 2019), yet we know surprisingly little about how their concrete workings bear on the distribution of rewards between the individuals they deem more or less deserving. Here we argue that, in employee performance evaluations but likely in other domains too, evaluation systems that reify merit lead to greater levels of distributional inequality in the deserts these systems are designed to allocate. We further suggest that reification shapes inequality by entrenching the belief that populations of evaluated entities can be meaningfully sorted into merit hierarchies. This means that the proliferation of reifying classification systems may have a tendency to undermine the pursuit of greater equality by thwarting the very idea of an inherent diversity and multidimensionality of merit.

By showing how merit-based classification systems breed distributional inequality by reifying merit, our experiment also highlights a tension at the core of the meritocratic enterprise. To recognize merit and reward it fairly, modern organizations are drawn to measuring it in a standardized way. Formalized measures indeed come with the promise of reducing discretion and bias in the way different individuals are evaluated, and therefore of helping to enforce ideals of equity (Grodsky and Pager 2001; Pager and Shepherd 2008; Reskin 2000). At the same time, though, standardized measures have a tendency to reify merit or performance. This, we have shown, means that their use moves organizations farther away from another potential ideal – that of greater levels of outcome equality – than would less reifying forms of evaluation.

As the terms of this tension make clear, it might be a value question whether an organization, or any other evaluative body, should want to adopt classification tools that entail a greater or lesser reification of merit. More generally, a number of reasons might explain why organizations rely on reifying devices when it comes to evaluating merit or performance: these devices are typically perceived to enable more effective decision-making (e.g. Thaler and Sunstein 2008), to project authority and trustworthiness (Porter 1996), and to more effectively persuade colleagues and superiors (Springer 2019). Against this backdrop, our work highlights one more dimension worth considering when adopting them: reifying evaluation systems come with the unintended consequence of inflating disparities in rewards. Whether this is desirable is for their users to decide. Suffice it to say, however, that by choosing less reifying forms of evaluation, one may be able to curb the undue inequality it generates, while retaining what we praise about it: that it helps us give merit its due.

Appendix A – Exclusion Criteria

To make sure that the participants we included in our analyses had engaged seriously with our main task, we applied a series of exclusion criteria to the initial set of responses collected when fielding the experiment. Participants' responses were excluded if they had not (1) spent a minimum of six seconds on the instructions page; (2) opened all three performance reports; (3) spent a minimum of fifty seconds on the bonus allocation task; and (4) correctly answered two out of three attention checks appearing at the end of our survey and asking about the content of the reports' narrative sections. The same exclusion criteria had been applied to the responses we used in crowd-sourcing the ratings of the three employees.

We devised criteria (1) and (3) after piloting our experiment with a small set of trial respondents. Plotting the time spent by our overall set of participants on respectively the instructions page (figure A1) and the main experimental task (figure A2) shows a bimodal distribution: in both cases a sizeable group of participants spent very little time on the task, apparently breezing through without engaging with it carefully; a second group, in contrast, went seriously about the task. In both cases the time thresholds set by criteria (1) and (3) exclude the bulk of the first group.

In total, 32% of participants who had completed our study failed at least one of the four aforementioned criteria and were therefore excluded from our analysis. An open-ended question at the end of our experiment did not suggest that any participant had uncovered the true aim of our study and would therefore have needed to be excluded further. The results reported in this article are based on the responses of 2,844 participants who passed all of our exclusion criteria.

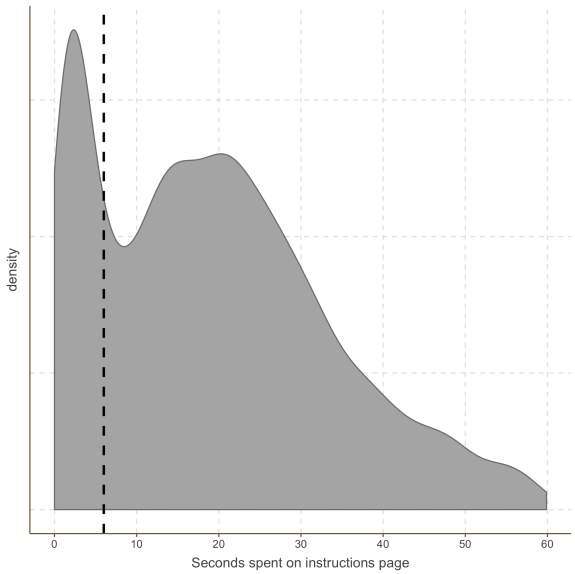

Figure A1. Frequency distribution of the time spent by participants on the instructions page; the dashed line indicates our exclusion threshold.

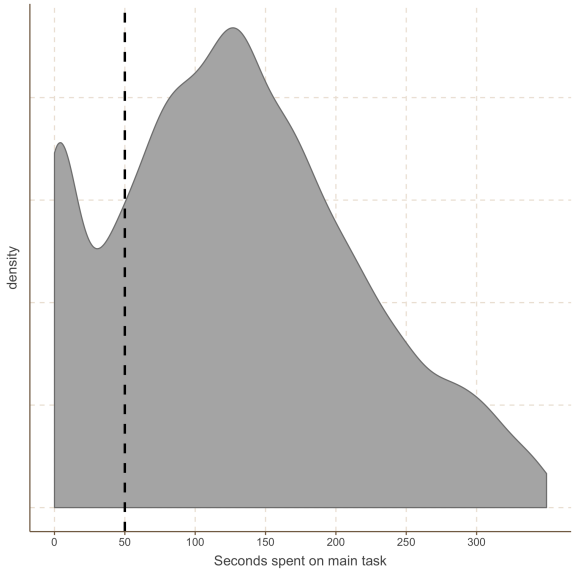

Figure A2. Frequency distribution of the time spent by participants on the main experimental task; the dashed line indicates our exclusion threshold.

Appendix B – Participant Demographics

Here we survey the demographics reported by the 2,844 participants who passed our exclusion criteria and compares these demographics with those of two nationally representative samples: the General Social Survey 2018 (GSS) and the Census Bureau's Current Population Survey 2018 (CPS). We found no significant between the gender composition of our participant pool (56% women) and that of the U.S. population according to the GSS (55% women). On the other hand, participants tended to report a lower age, lower household income, and more Democratic political leaning than are found in the U.S. population (figures B1 to B3). Our group of participants also under-represents individuals identifying as Black or African-American and over-represents those identifying as White or as Asian (figure B4).

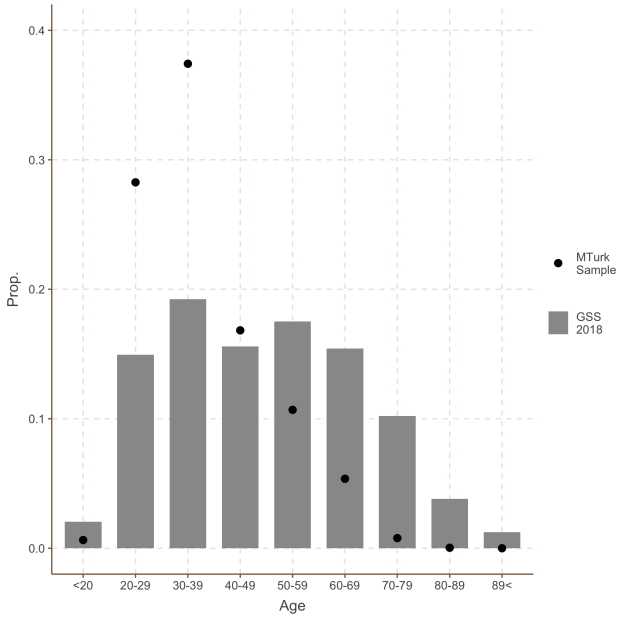

Figure B1. Age of experimental participants vs. age of the U.S. population based on the 2018 General Social Survey.

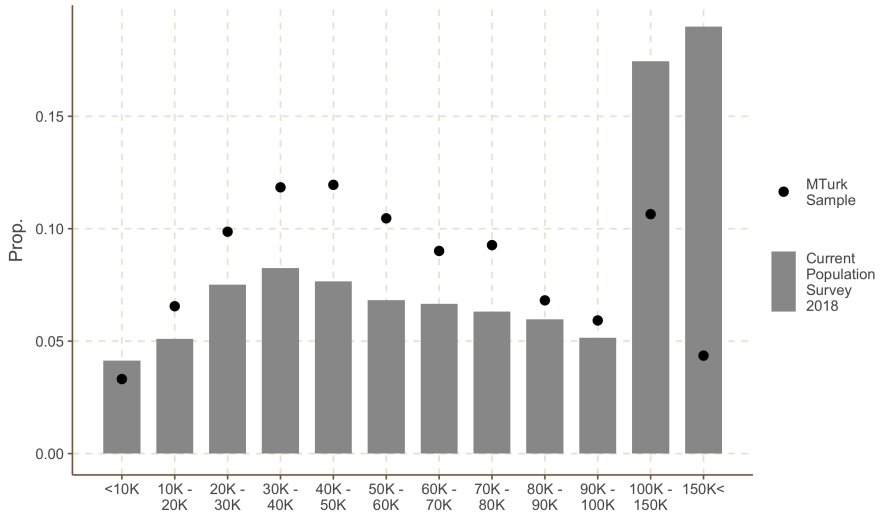

Figure B2. Household income of experimental participants vs. household income in the U.S. population based on the 2018 Current Population Survey.

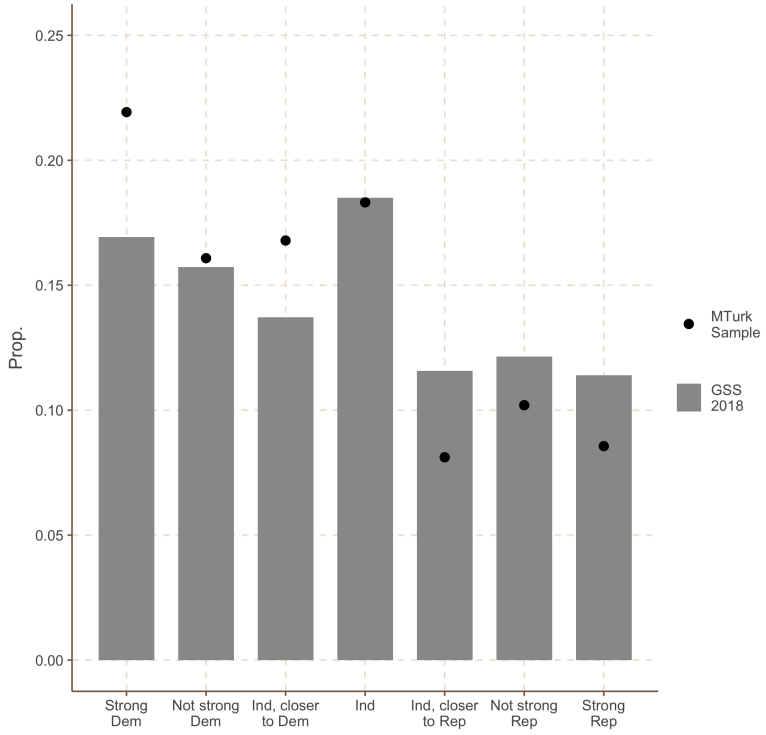

Figure B3. Political affiliation of experimental participants vs. political affiliation of the U.S. population based on the 2018 General Social Survey.

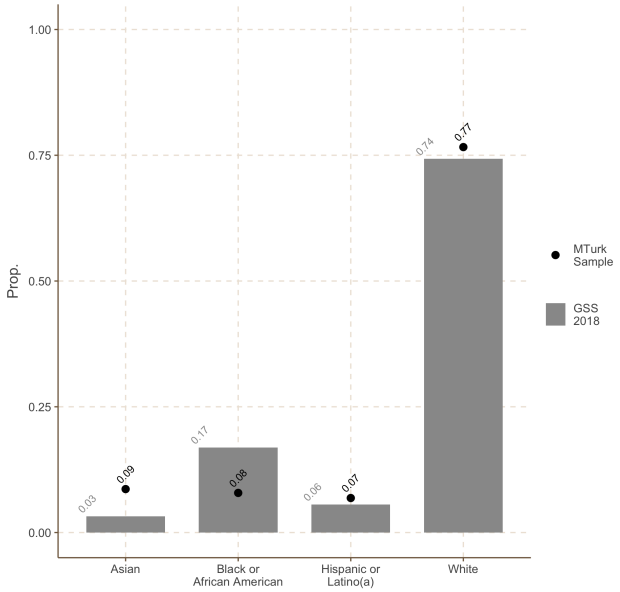

Figure B4. Race and ethnicity of experimental participants vs. race and ethnicity of the U.S. population based on the 2018 General Social Survey.

Appendix C – Dealing with the Recent Rise in Low-Quality Responses on MTurk

In the summer of 2018 Amazon Mechanical Turk requesters noticed an increase in low quality responses on the platform (Dennis, Goodson, and Pearson 2019; Moss and Litman 2018). While these responses – which by some estimates comprised up to 9% of all responses filed on the platform (Ryan 2018) – apparently originated from a small number of geolocations in the U.S. producing multiple responses, they seem to have been generated by overseas workers routing their traffic through U.S.-based servers. On the other hand, only a few of geolocations with multiple responses are problematic and many do in fact yield good quality responses (Gautam et al. 2018).

In our data, we did notice a correlation between geolocations yielding multiple responses and the likelihood that these responses would be of low quality as signaled by their failure to pass our exclusion criteria. By applying these criteria, however, we make sure we eliminate problematic responses originating from multiply-appearing geolocations while retaining high quality ones. As a robustness check, we also ran our analyses after excluding all responses from multiply-appearing locations. This yielded similar findings to the ones we report in the paper.

References

Accominotti, Fabien. 2019. “Consecration as a Population-Level Phenomenon.” *American Behavioral Scientist*, forthcoming.

Bell, Daniel. 1973. *The Coming of Post-Industrial Society: A Venture in Social Forecasting*. New York: Basic Books.

Berman, Elizabeth Popp, and Daniel Hirschman. 2018. “The Sociology of Quantification: Where Are We Now?” *Contemporary Sociology* 47: 257-266.

Bian, Lin, Sarah-Jane Leslie, and Andrei Cimpian. 2017. “Gender Stereotypes about Intellectual Ability Emerge Early and Influence Children's Interests.” *Science* 355: 389-391.

Biernat, Monica, M.J. Tocci, and Joan C. Williams. 2012. “The Language of Performance Evaluations: Gender-Based Shifts in Content and Consistency of Judgment.” *Social Psychological and Personality Science* 3: 186-192.

Boltanski, Luc, and Laurent Thévenot. 2006. *On Justification: Economies of Worth*. Princeton, NJ: Princeton University Press.

Bosk, Emily A. 2019. “Iron Cage or Paper Cage: The Interplay of Worker Characteristics and Organizational Policy in Shaping Unequal Responses to a Standardized Decision-Making Tool.” *Social Problems*, forthcoming.

Bourdieu, Pierre, and Jean-Claude Passeron. 1990. *Reproduction in Education, Society and Culture*. London: Sage.

Bowker, Geoffrey C., and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Brandtner, Christof. 2017. “Putting the World in Orders: Plurality in Organizational Evaluation.” *Sociological Theory* 35: 200-227.

Buhrmester, Michael D., Sanaz Talaifar, and Samuel D. Gosling. 2018. “An Evaluation of Amazon’s Mechanical Turk, Its Rapid Rise, and Its Effective Use.” *Perspectives on Psychological Science* 13: 149-154.

Carson, John. 2007. *The Measure of Merit: Talents, Intelligence, and Inequality in the French and American Republics, 1750-1940.* Princeton, NJ: Princeton University Press.

Castilla, Emilio. 2008. “Gender, Race, and Meritocracy in Organizational Careers.” *American Journal of Sociology* 113: 1479-1526.

Castilla, Emilio, and Stephen Benard. 2010. “The Paradox of Meritocracy in Organizations.” *Administrative Science Quarterly* 55: 543-576.

Christin, Angèle. 2018. “Counting Clicks: Quantification and Variation in Web Journalism in the United States and France.” *American Journal of Sociology* 123: 1382-1415.

Coppock, Alexander. 2019. “Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach.” *Political Science Research and Methods* 7: 613-628.

Dennis, Sean A., Brian Matthew Goodson, and Chris Pearson. 2019. “Online Worker Fraud and Evolving Threats to the Integrity of MTurk Data: A Discussion of Virtual Private Servers and the Limitations of IP-Based Screening Procedures.” Working paper, available at SSRN: http://dx.doi.org/10.2139/ssrn.3233954

Desrosières, Alain. 2002. *The Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge, MA: Harvard University Press.

Dobbin, Frank, Daniel Schrage, and Alexandra Kalev. 2015. "Rage against the Iron Cage: The Varied Effects of Bureaucratic Personnel Reforms on Diversity." *American Sociological Review* 80: 1014-1044.

Domina, Thurston, Andrew M. Penner, and Emily Penner. 2017 "Categorical Inequality: Schools as Sorting Machines." *Annual Review of Sociology* 43: 311-330.

Espeland, Wendy N., and Michael Sauder. 2016. *Engines of Anxiety: Academic Rankings, Reputation, and Accountability*. New York: Russell Sage.

Espeland, Wendy N., and Mitchell L. Stevens. 1998. "Commensuration as a Social Process." *Annual Review of Sociology* 24: 313-343.

Espeland, Wendy N, and Mitchell L. Stevens. 2008. "A Sociology of Quantification." *European Journal of Sociology* 49: 401-436.

Espeland, Wendy N., and Berit I. Vannebo. 2007. "Accountability, Quantification, and Law." *Annual Review of Law and Social Science* 3: 21-43.

Fischer, Claude S., Michael Hout, Martin Sanchez Jankowski, Samuel Lucas, Ann Swidler, and Kim Voss. 1996. *Inequality by Design: Cracking the Bell Curve Myth*. Princeton, NJ: Princeton University Press.

Fourcade, Marion. 2009. *Economists and Societies: Discipline and Profession in the United States, Britain, and France, 1890s to 1990s.* Princeton, NJ: Princeton University Press.

Fourcade, Marion. 2016. "Ordinalization: Lewis A. Coser Memorial Award Lecture for Theoretical Agenda Setting 2014." *Sociological Theory* 34: 175-195.

Fourcade, Marion, and Kieran Healy. 2017. "Seeing Like a Market." *Socio-Economic Review* 15: 9-29.

Furnham, Adrian. 2003. "Belief in a Just World: Research Progress over the Past Decade." *Personality and Individual Differences* 34: 795-817.

Gautam, R., M. Kerstein, Aaron. J. Moss, and Leib Litman. 2018. "Understanding Geolocations and Their Connection to Data Quality." Available at: https://blog.turkprime.com/understanding-geolocations-and-their-connection-to-data-quality.

Grodsky, Eric, and Devah Pager. 2001. "The Structure of Disadvantage: Individual and Occupational Determinants of the Black-White Wage Gap." *American Sociological Review* 66: 542-567.

Guetzkow, Joshua, Michèle Lamont, and Grégroire Mallard. 2004. "What Is Originality in the Humanities and the Social Sciences?" *American Sociological Review* 69: 190-212.

Hauser, David J., Gabriele Paolacci, and Jesse J. Chandler. 2018. "Common Concerns with MTurk as a Participant Pool: Evidence and Solutions." Available at: https://psyarxiv.com/uq45c.

Hauser, David J., and Norbert Schwarz. 2016. "Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks Than Do Subject Pool Participants." *Behavior Research Methods* 48: 400-407.

Heilman, Madeline E. 2012. "Gender Stereotypes and Workplace Bias." *Research in Organizational Behavior* 32: 113-135.

Horton, John J., David G. Rand, and Richard J. Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14: 399-425.

Jin, Ginger Zhe, and Phillip Leslie. 2003. "The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards." *Quarterly Journal of Economics* 118: 409-451.

Karabel, Jerome. 2005. *The Chosen: The Hidden History of Admission and Exclusion at Harvard, Yale, and Princeton*. Boston: Houghton Mifflin Harcourt.

Kees, Jeremy, Christopher Berry, Scot Burton, and Kim Sheehan. 2017. "An Analysis of Data Quality: Professional Panels, Student Subject Pools, and Amazon's Mechanical Turk." *Journal of Advertising* 46: 141-155.

Kiviat, Barbara. 2019. "The Moral Limits of Predictive Practices: The Case of Credit-Based Insurance Scores." *American Sociological Review*, forthcoming.

Lamont, Michèle. 2010. *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge, MA: Harvard University Press.

Lamont, Michèle. 2012. "Toward a Comparative Sociology of Valuation and Evaluation." *Annual Review of Sociology* 38: 201–221.

Lamont, Michèle, Stefan Beljean, and Matthew Clair. 2014. "What Is Missing? Cultural Processes and Causal Pathways to Inequality." *Socio-Economic Review* 12: 573-708.

Lamont, Michèle, and Paul Pierson. 2019. "Inequality Generation and Persistence as Multidimensional Processes: An Interdisciplinary Agenda." *Daedalus* 148(3): 5-18.

Lauer, Josh. 2018. *Creditworthy: A History of Consumer Surveillance and Financial Identity in America*. New York: Columbia University Press.

LeGrand, Julian, and Alain Enthoven. 2007. *The Other Invisible Hand: Delivering Public Services Through Choice and Competition*. Princeton, NJ: Princeton University Press.

Leslie, Sarah-Jane, Andrei Cimpian, Meredith Meyer, and Edward F. Freeland. 2015. "Expectations of Brilliance Underlie Gender Distributions Across Academic Disciplines." *Science* 347: 262-265.

Lipkus, Isaac M. 1991. "The Construction and Preliminary Validation of a Global Belief in a Just World Scale and the Exploratory Analysis of the Multidimensional Belief in a Just World Scale." *Personality and Individual Differences* 12: 1171-1178.

Lipkus, Isaac M., Claudia Dalbert, and Ilene C. Siegler. 1996. "The Importance of Distinguishing the Belief in a Just World for Self versus Others: Implications for Psychological Well-Being." *Personality and Social Psychology Bulletin* 22: 666-677.

Mau, Steffen. 2019. *The Metric Society: On the Quantification of the Social*. London: Polity Press.

Mennicken, Andrea, and Wendy N. Espeland. 2019. "What's New with Numbers? Sociological Approaches to the Study of Quantification." *Annual Review of Sociology* 45: 223-245.

Miller, Peter. 2001. "Governing by Numbers: Why Calculative Practices Matter." *Social Research* 68: 379-396.

Moss, Aaron J., and Leib Litman. 2018. "After the Bot Scare: Understanding What's Been Happening with Data Collection on MTurk and How to Stop It." Available at:

https://blog.turkprime.com/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it.

Muller, Jerry Z. 2018. *The Tyranny of Metrics*. Princeton, NJ: Princeton University Press.

Musto, Michela. 2019. "Brilliant or Bad: The Gendered Social Construction of Exceptionalism in Early Adolescence." *American Sociological Review* 84: 369-393.

Pager, Devah, and Hana Shepherd. 2008. "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets." *Annual Review of Sociology* 34: 181-209.

Paolacci, Gabriele, and Jesse J. Chandler. 2014. "Inside the Turk: Understanding Mechanical Turk as a Participant Pool." *Current Directions in Psychological Science* 23: 184-188.

Porter, Theodore M. 1996. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Power, Michael. 1997. *The Audit Society: Rituals of Verification*. Oxford: Oxford University Press.

Quadlin, Natasha. 2018. "The Mark of a Woman's Record: Gender and Academic Performance in Hiring." *American Sociological Review* 83: 331-360.

Reskin, Barbara F. 2000. "The Proximate Causes of Employment Discrimination." *Contemporary Sociology* 29: 319-328.

Rivera, Lauren. 2015. *Pedigree: How Elite Students Get Elite Jobs*. Princeton, NJ: Princeton University Press.

Rivera, Lauren, and András Tilcsik. 2016. "Class Advantage, Commitment Penalty: The Gendered Effects of Social Class Signals in an Elite Labor Market." *American Sociological Review* 81: 1097-1131.

Rivera, Lauren, and András Tilcsik. 2019. "Scaling Down Inequality: Rating Scales, Gender Bias, and the Architecture of Evaluation." *American Sociological Review* 84: 248-274.

Rubin, Zick, and Letitia Anne Peplau. 1973. "Belief in a Just World and Reactions to Another's Lot: A Study of participants in the National Draft Lottery." *Journal of Social Issues* 29: 73-93.

Rubin, Zick, and Letitia Anne Peplau. 1975. "Who Believes in a Just World?" *Journal of Social Issues* 31: 65-89.

Ryan, Timothy J. 2018. "Data Contamination on MTurk." Available at: http://timryan.web.unc.edu/2018/08/12/data-contamination-on-mturk/.

Sauder, Michael. 2006. "Third Parties and Status Systems: How the Structures of Status Systems Matter." *Theory & Society* 35: 299-321.

Sauder, Michael, and Wendy N. Espeland. 2007. "Rankings and Reactivity: How Public Measures Recreate Social Worlds." *American Journal of Sociology* 113: 1-40.

Sauder, Michael, Freda Lynn, and Joel M. Podolny. 2012. "Status: Insights from Organizational Sociology." *Annual Review of Sociology* 37: 267-283.

Springer, Emily. 2019. "Bureaucratic Tools in (Gendered) Organizations: Performance Metrics and Gender Advisors in International Development." *Gender & Society*, forthcoming.

Stevens, Mitchell L. 2007. *Creating a Class: College Admissions and the Education of Elites*. Cambridge, MA: Harvard University Press.

Thaler, Richard H., and Cass R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

Timmermans, Stefan, and Steven Epstein. 2010. "A World of Standards but not a Standard World: Toward a Sociology of Standards and Standardization." *Annual Review of Sociology* 36: 69-89.

Weinberg, Jill D., Jeremy Freese, and David McElhattan. 2014. "Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and a Crowdsource-Recruited Sample." *Sociological Science* 1: 212-310.

Young, Michael. 1958. *The Rise of the Meritocracy, 1870-2033: An Essay on Education and Equality*. London: Thames & Hudson.