4

Blinding to Remove Biases in Science and Society

Robert J. MacCoun

Abstract

This chapter examines the use of blinding methods to potentially bias information to improve the validity and/or fairness of judgments in scientific data analysis, scientific peer review, and the screening of job applicants. Some of the major findings in empirical tests of these procedures are reviewed, addressing potential concerns with blinding, and identifying directions for new theory and research.

Introduction

In this chapter, I examine the promise, and the limitations, of the use of methods of blinding as one way to achieve deliberate ignorance (see Hertwig and Engel, this volume, 2016) in situations where a decision maker's knowledge of some variables might bias judgments or create unfairness in the decision process. Readers will be familiar with the notion of blinding in at least two ways. First, everyone has seen depictions of the Roman goddess Iustitia (Justice), whose scales and blindfold depict the aspiration for unbiased judgment in legal systems around the world. Second, double-blinding (of patients and physicians) in medical trials is one of a handful of methodological principles (with placebos and sample size) familiar to most lay people. A recent edited volume (Robertson and Kesselheim 2016) offers a thorough treatment of blinding in medical science, forensic science, and legal procedures, and so I will only make brief mentions of those literatures here.

In this essay I will examine blinding in three domains, deployed in pursuit of two different normative goals (see Table 4.1).

According to Gosseries and Parr (2005), the fact "that transparency and accountability are social goods is taken as self-evident in contemporary democracies." As Louis Brandeis famously put it: "Sunlight is said to be the best of disinfectants, electric light the most efficient policeman." Transparency refers

From "Deliberate Ignorance: Choosing Not to Know," edited by Ralph Hertwig and Christoph Engel.
Strüngmann Forum Reports, vol. 29, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262045599

Table 4.1 Domains and goals.

Domains of blinding	Goal of blinding
Data analysis	Validity
Scientific peer review	Validity, fairness
The job market	Fairness

to openness and visibility, while accountability implies that the actor must be able to explain his or her choices, and that there are consequences for those choices.

In a 2006 essay, I argued that, whatever its abstract merits might be, there are psychological constraints that make true transparency and accountability difficult to achieve, and that can lead to unintended and undesirable effects, and I reviewed theory and evidence for four propositions:

1. Introspective access to our cognitions is very limited.
2. Accountability can have perverse effects.
3. Group processes can actually amplify individual biases.
4. Being explicit can distort goals and the willingness to make tradeoffs.

The Rawlsian tradition in philosophy offers a rich debate on the merits of a "veil of ignorance" as a guarantor of unbiased assessments of social distribution and welfare. Although the details may differ, the underlying logic seems basically the same as that used to motivate blinding in job screening, data collection, data analysis, and other situations.

To make the logic more concrete, I will use Egon Brunswik's "lens-model" approach to investigating the quality and determinants of human judgment (Cooksey 1996; Dhami et al. 2004; Hammond and Stewart 2001; Karelaia and Hogarth 2008). Figure 4.1 shows a typical lens-model diagram. The right side of the "lens" depicts the true relationships among a set of "cues" or predictor variables and some outcome of interest. The left side of the lens shows the relationships among these cues and a judgment (prediction, decision) made by some judge (referee, editor, scientist, selection committee)—their implicit "judgment policy." I vary the thickness of the arrows to show the strength of the relationships on each side. By comparing the judgment to the outcome, we can assess the validity of the judgment. But a lens-model analysis tells us more by allowing us to compare the signs and magnitudes of the arrows on each side of the lens. It can show where judges are using a "bad cue" or missing a "good cue," in which case we might intervene with training, blinding, or simply replacing the judge with the algorithmic model on the right side of the lens.

Figure 4.1 is of course an oversimplification. Typical lens-model applications depict a multiple regression or path coefficient for each link, along with additional links showing cue intercorrelations. A more ambitious extension might be to depict each side of the lens as a directional acyclic graph (Pearl 2000) which could show that the causal structure of the judgment process (left side) misrepresents the

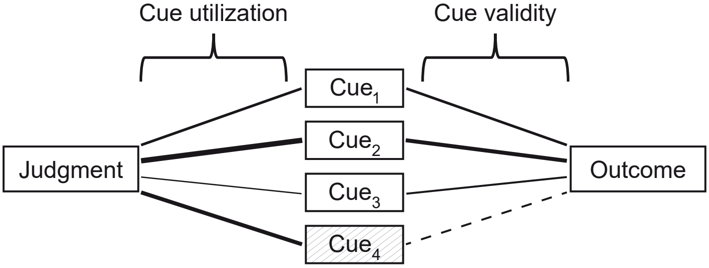

Figure 4.1 A simplified example of the lens-model approach to assessing the validity of judgments. The left side of the diagram depicts the human judgment process, where the cues are predictor variables, and the thickness of the lines represents the weight placed on cues (which here are shown as positive, for simplicity). The right side of the diagram depicts the objective relationship between various cues and a later observable outcome that corresponds to the judgment (e.g., job performance if hired). A comparison of the cue utilization weights and the cue validity weights reveals cues that are being underutilized (here, Cue$_3$) or overweighted (here, Cue$_4$).

causal structure that produces the outcomes (right side); for example, a judgment might overutilize a cue that is actually a spurious correlate (no causation) or even a consequence (reverse causation) of the outcome.

Although I have not seen it used in this way, the lens-model framework provides an explicit framework for thinking about how and when to blind effectively. Blinding is appropriate when current judgments give undue weight to a particular cue or use a cue that is actually spurious, as seen in Figure 4.2. Blinding may be unnecessary when a valid cue is used appropriately, or when an invalid cue is being ignored. But a lens-model analysis might also show that blinding (whether of humans or of algorithms) might have unintended consequences when good and bad cues are intercorrelated, a point I return to later.

The lens model is most useful for questions of validity: What are the true predictors of an outcome and does the judge have a valid mental model? It does not readily depict cue utilization with respect to other normative criteria.

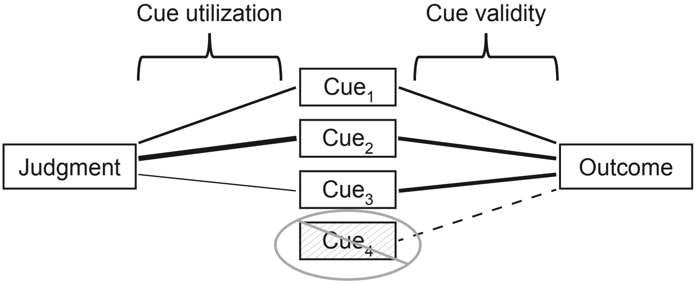

Figure 4.2 Blinding the judge by blocking or obscuring a cue that would bias the judgment.

In particular, as noted in Table 4.1, some applications of blinding are motivated by concerns about fairness rather than (or in addition to) validity. Even then, the lens model can clarify our discussions of fairness. Is a cue "unfair" because it has low validity, or are some cues unfair even when they are valid predictors? Are there normative reasons to retain some cues even when they are low in validity?

Blinding in Data Analysis

In the course of analyzing data, the analyst must make a host of judgments about what variables to include, how to handle outliers and other data anomalies, what statistical tests and estimators to use, and so on. It is well established (see MacCoun 1998) that such decisions are often biased by examinations of the data, which can reveal whether a particular approach will produce a test result that is favorable to a preferred (or abhorred) hypothesis. Although this problem plagues all empirical disciplines, its effects on the replicability of psychological research are now well known.

I had the pleasure of teaching an undergraduate course for several years with Nobel Laureate physicist Saul Perlmutter, and when he heard me lecturing about psychology's problems with confirmation bias and replicability, he asked: "Don't you perturb your data before analyzing them?" I had no idea how to interpret this kinky-sounding question, but then he explained that many lab groups, in particle physics and cosmology, routinely add noise or bias to their data before analyzing it, so that any preconceptions or careerist motivations can't bias their inferences. A blinding method is selected to facilitate intermediate analytic decisions while precluding choices that would favor one hypothesis over others. The blind is then lifted once all analytic decisions are made.

We subsequently coauthored two papers describing a variety of data-blinding methods and advocating their use in other empirical disciplines (MacCoun and Perlmutter 2015, 2017). These are the basic approaches and terminology:

- *Noising*: Add a random deviate to each data point.
- *Biasing*: Add a systematic offset to each data point.
- *Cell scrambling*: Swap the labels of different cells (arms) of the experimental design.
- *Row scrambling*: Swap the labels on each row of the data matrix, so that observations from the same cell are no longer grouped together.

Two others we did not discuss are:

- *Salting*: Adding fake data points to a real data set
- *Masking*: Simply hiding or anonymizing the identity of a data unit

Masking, of course, is the kind of blinding that is used in peer review and in job screening procedures, but this list shows that there are many other possibilities worth considering. In simulations, we found that these blinding methods had different effects on what was and was not obscured in the data, suggesting that they might be suitable for different situations or purposes. The empirical literature on data blinding and its consequences is still very sparse, and we argued that like any other intervention, data blinding should be assessed to establish its benefits, costs, and any boundary conditions on its effectiveness.

Blinding in Scientific Peer Review

Two decades ago (MacCoun 1998), I reviewed evidence on the myriad forms of bias that occur when people use and interpret research data, suggesting that traditional remedies like peer review are only partial solutions. Evidence since then (especially in my own discipline of psychology) suggests that if anything, I probably understated the problem. Carroll (2018), adapting a famous quip by Winston Churchill, recently argued that peer review is "the worst way to judge research, except all the others." There are hundreds of papers critiquing the peer review system, dozens of empirical papers on inter-referee reliabilities, and a handful examining the question of double-blind reviewing (i.e., blinding of author identity, since blinding of reviewer identity has long been the norm).

Journals that use double-blind reviewing are still the exception, not the rule. In an interdisciplinary sample of journals, Walker and Rocha da Silva (2015) found that 70% used single-blinding (author names visible to reviewers) and 20% used double-blinding. At least one major journal (*American Economic Review*) has abandoned the practice and there is growing support in the "open science" movement for the use of fully unblinded peer review, in which referee reports are signed and publicly archived. The paucity of evidence on these procedures explains how two opposite strategies are each being endorsed as solutions to the same set of problems.

Like blinded-data analysis, blinding in peer review has been primarily motivated by the goal of increasing decision validity, but it is also seen as a mechanism for promoting fairness. The most well-known studies focus on blinding to improve the quality of published research—a validity criterion. McNutt et al. (1990) reported what appears to be the first double-blinded experiment on double-blinded review, an experiment in which the *Journal of General Internal Medicine* sent 137 manuscripts to pairs of reviewers, one of whom was randomly selected, to receive an anonymized version of the submission. Editors—themselves blinded to the selection—rated the quality of reviews as significantly greater for double-blinded reviews, although the effect was small (3.5 vs. 3.1 on a five-point scale). Blinding did not affect the rate at which reviewers signed their reviews, and signing was unrelated to quality ratings.

Around the same time, Blank (1991) reported an experiment in the *American Economic Review*, in which 1,498 manuscripts were randomly assigned to receive either single- or double-blind peer review. Double-blinded manuscripts had a higher referee response rate, were accepted at a lower rate, and received significantly more critical ratings. Blinding had less effect on manuscripts from top- and low-ranked institutions than on those in the middle of the pack. One caveat is that an editorial assistant "automatically assigned any paper that she felt could not be handled as a blind paper to the nonblind category" (Blank 1991:1050).

Both of these studies have a criterion problem, and a self-referential one at that: If reviewing processes are flawed, can we really infer whether blinding improves matters by using acceptance rates and subjective quality ratings? Okike et al. (2016) addressed this problem by randomizing whether a decoy manuscript contained five "subtle errors." Like the earlier studies, they found lower acceptance rates and quality ratings for double-blind manuscripts. However, they were not able to detect a difference in the frequency with which the planted errors were detected.

To the extent that reviewer biases favor certain categories of authors—white males, elite universities, Americans—then efforts to improve the validity of peer review also serve to make it a more fair system. Still, fewer empirical studies have directly addressed this criterion. Blank's 1991 experiment was unable to detect an effect of acceptance rates for female authors, but cautions that only 8% of the papers had a primary author who was female. Budden et al. (2008) argue that double-blinding at the journal *Behavioral Ecology* led to an increase in accepted papers with female first authors, though a number of subsequent critiques (reviewed by Lee et al. 2013) indicate that the apparent finding was probably artifactual. Tomkins et al. (2017) report that double-blinding of submissions to a computing conference reduced the influence of author fame and institutional prestige on acceptance rates.

Blinding in the Job Market

Discrimination on the basis of economically irrelevant or legally protected categories (by gender, race, ethnicity, sexual orientation, religion, or ideology) is the subject of vast empirical literatures in economics, sociology, psychology, and other disciplines. Many of these studies are "observational" in the econometric sense, meaning that they involve multivariate analysis of correlational data. Two methods have been helpful in overcoming the myriad problems with causal identification in such studies.

Correspondence studies are controlled experiments (usually "in the field") in which an assessor is randomly assigned a job or school application in which potentially biasing demographic or other characteristics are varied (through deception) while holding other (more probative) information constant (see Pager and Shepherd 2008). In a meta-analysis of 738 different tests from 43 separate studies, Zschirnt and Ruedin (2016:1128) find that "[e]quivalent minority candidates need to send around 50% more applications to be invited for an interview than majority candidates." Audit studies are field experiments in which matched pairs of actors differ in some visual demographic characteristic but are otherwise given identical fake credentials and trained to behave similarly. Pager and Shepherd's (2008:187) review cites audit estimates of white advantage ranging from 50%–240%.

There are a variety of proposed solutions to these forms of discrimination, including legal sanctions against discriminators, legal remedies for the discriminated, various affirmative action policies, and training and education, including "implicit-bias" training. My focus in this chapter is exclusively on the use of various methods of blinding or anonymity designed to make it difficult or impossible for the decision maker to react to potentially biasing factors.

In 2000, Claudia Goldin and Cecilia Rouse published what is probably the most famous study of blinding in the marketplace, a paper that has been cited almost 1,400 times (as of 2/1/19) according to *Google Scholar*. After documenting robust strong biases against women in the classical music industry, Goldin and Rouse note that major orchestras have gradually adopted a blind audition procedure, in which the auditioning musician performs behind a screen so that the selection committee can hear but not see them (Goldin and Rouse 2000:721):

> In blind auditions (or audition rounds) a screen is used to hide the identity of the player from the committee. The screens we have seen are either large pieces of heavy (but sound-porous) cloth, sometimes suspended from the ceiling of the symphony hall, or what look like large room dividers. Some orchestras also roll out a carpet to muffle footsteps that could betray the sex of the candidate. Each candidate for a blind audition is given a number, and the jury rates the candidate's performance next to their number on a sheet of paper. Only the personnel manager knows the mapping from number to name and from name to other personal information.

Goldin and Rouse (2000:716) assembled roster data and audition data for eleven different orchestras:

> Among the major orchestras, one still does not have any blind round to their audition procedure (Cleveland) and one adopted the screen in 1952 for the preliminary round (Boston Symphony Orchestra), decades before the others. Most other orchestras shifted to blind preliminaries from the early 1970s to the late 1980s. The variation in screen adoption at various rounds in the audition process allows us to assess its use as a treatment.

Using difference-in-difference and fixed effects methods, the authors argue that blind auditions have had a profound effect on orchestral hiring. For example, the audition data set suggests that "the screen increases—by 50%—the probability that a woman will be advanced from certain preliminary rounds and increases by severalfold the likelihood that a woman will be selected in the final round." Similar analyses of the orchestra roster data suggest that up to 30% of the increase in female representation in orchestras in the 1970–1996 period is attributable to blind auditioning.

The logic of blinding in orchestra auditions is premised on the compelling intuition that musical excellence should be judged by auditory and not visual criteria. Surprisingly, Tsay (2013) found that participants "reliably select the actual winners of live music competitions based on silent video recordings, but neither musical novices nor professional musicians were able to identify the winners based on sound recordings or recordings with both video and sound."

Blind auditions have not eliminated gender imbalance. According to an analysis in *The Washington Post* (Edgers 2018), "although women make up nearly 40% of the country's top orchestras, when it comes to the principal, or titled, slots, 240 of 305—or 79%—are men. The gap is even greater in the "big five"—the orchestras in Boston, Chicago, Cleveland, Philadelphia, and New York. Women occupy just 12 of 73 principal positions in those orchestras." In 2000, Goldin and Rouse noted that most orchestras unblinded the late rounds of auditions, and *The Washington Post* analysis suggests that this was still true in 2018.

Despite the well-deserved attention that the Goldin and Rouse analysis has received, the use of physical screens is not very representative of actual blinding practices in the marketplace. More typically, blinding is done by redacting information on a document or a computer screen. Most of these studies use the term "anonymity" rather than blinding, but I prefer the latter term, both for continuity, and because "anonymity" can also refer to issues of privacy, confidentiality, or secrecy, where the goals and the context often differ.

Aslund and Skans (2012) report on a nonexperimental study of anonymous job applications in Gothenburg, Sweden from 2004–2006. Using a differences-in-differences model, they found that anonymity increased the rate of interview callbacks for both women and ethnic minorities, but that women, not minorities, received an increase in job offers.

In a 2011 unpublished paper, Bøg and Kranendonk describe two experiments in a Dutch city from 2006–2007. Participation by municipal departments was voluntary. In the first experiment, seven departments were randomly assigned to use either standard or anonymous screening procedures for job applications during the test period. In the second experiment, these assignments were reversed. Note that because the logic of random assignment is based on the law of large numbers, this is a far weaker design than random assignment at the level of the individual application. The authors found that the majority-minority gap in interview callbacks was reduced by the experiment. But in fact, there were similar rates of interview invitations and job offers for minority candidates in the treatment and control conditions, and the reduced gap was produced by fewer callbacks for majority applicants in the anonymous condition.

Behaghel et al. (2015) report a study of anonymous application procedures in a French public employment service from 2010–2011. Private-sector firms who agreed to participate received either anonymous or standard applications from the employment service. The unit of randomization was the job vacancy rather than the firm (though not the job applicant), so this design clearly improves on Bøg and Kranendonk (unpublished). Unexpectedly, the authors found that "participating firms become less likely to interview and hire minority candidates when receiving anonymous résumés." The authors attribute this result to two factors. First, the decision to participate in the experiment may have screened out those firms most likely to discriminate. Second, among the participating firms, anonymization prevented them from providing more favorable treatment to minorities. Anonymization did help women applicants but only to a limited extent because for many vacancies, applicants were either all male or all female.

Krause et al. (2012) studied five private and three public-sector German organizations. Like Behaghel et al. (2015), they found that anonymity had unintended consequences. Female applicants actually fared better than males under standard applications, and blinding removed this advantage. Applicants in a nonrandomized blinded sample were compared to two different comparison groups—standard applications from the previous year, or applications from the study cohort that were not blinded. Results were mixed; under some circumstances minorities fared better with blinded applications, but under other circumstances they fared somewhat worse. The authors conclude that "the introduction of anonymous job applications can lead to a reduction of discrimination—if discrimination is present in the initial situation. Anonymous job application can also have no effects if no discrimination is present initially, and they can stop measures such as affirmative action that may have been present before. In any case, the effects of anonymous job applications depend on the initial situation" (Krause et al. 2012:12).

Between 1993 and 2010, the U.S. military adopted a personnel policy that is clearly a form of deliberate ignorance, and can be viewed as a form of blinding (with the onus of concealment placed on the employee rather than the employer). Under this "Don't ask, don't tell" policy, gay and lesbian service personnel were permitted to serve in the military provided that they concealed their sexual orientation from their peers. This kind of mandated self-concealment has serious limitations. Whereas other blinding approaches are temporary, "Don't ask, don't tell" required ongoing blinding for the course of service—something which proved very difficult to maintain for sexual orientation and impossible to achieve for visible attributes like gender or race. And, of course, the motivation for blinding was very different; whereas blinding in the application process is usually designed to protect the applicant, "Don't ask, don't tell" was essentially designed to protect the unit from the applicant. As I documented elsewhere as part of an assessment that contributed to the repeal of this policy (MacCoun 1993; MacCoun and Hix 2010), this logic was based on a false premise: the idea that knowledge of a unit member's gay or lesbian orientation would somehow impair the unit's ability to work together to accomplish its mission.

Concerns about Blinding

In this review, I have highlighted the potential benefits of blinding as a way of achieving deliberate ignorance when some kinds of knowledge would jeopardize the validity and/or fairness of judgments in science and the marketplace. But blinding also has some potential drawbacks and limitations.

The following five issues constitute a research agenda for a comprehensive assessment of blinding.

Does Blinding Actually Blind?

In their study of orchestras, Goldin and Rouse (2000:722) note, but dismiss, the possibility that listeners can still infer gender from auditory cues. They suggest that because "the candidates play only predetermined and brief excerpts from the orchestral repertoire," there is "little or no room for individuality to be expressed and not much time for it to be detected." In the peer review literature (see Largent and Snodgrass 2016), a sizeable fraction of reviewers (25%–50%) believe they can identify the masked author; in some cases, they misidentify the author, which is arguably worse than no blinding at all. In our simulations of data blinding, we found that in some situations, adding noise or bias to data failed to obscure the true experimental outcome. I once participated in a professional meeting in which we tried and failed to identify a foolproof placebo for trials of LSD psychotherapy.

Does Blinding Do More Harm than Good?

As we have seen, blinded job screening can't eliminate discrimination that isn't there, and it can block the application of normatively prescribed biases like affirmative action.

In the clinical trial context, Meinert (1998) argues that "[m]asking should not be imposed if it entails avoidable risks for patients. Masked monitoring denies the monitors the key information they need to perform in a competent fashion, and incompetent monitoring poses a risk to research subjects." This concern is hardly groundless, but it is surely an argument for smart blinding rather than no blinding. MacCoun and Perlmutter (2015:188) argue that "when safety is at stake, such as in some clinical trials, it often makes sense to set up an unblinded safety monitor while the rest of the analytical team is in the dark." According to the 2013 statement of the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) group,[1] an international consortium of clinical trial experts:

> To maintain the overall quality and legitimacy of the clinical trial, code breaks should occur only in exceptional circumstances when knowledge of the

[1] https://www.spirit-statement.org/emergency-unblinding/, accessed on October 18, 2019

From "Deliberate Ignorance: Choosing Not to Know," edited by Ralph Hertwig and Christoph Engel.
Strüngmann Forum Reports, vol. 29, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262045599

> actual treatment is absolutely essential for further management of the patient.… Unblinding should not necessarily be a reason for study drug discontinuation.

Other dangers seem more remote. Cain et al. (2005) have found that disclosures of a conflict of interest can "morally license" agents to act in a more biased fashion, but blinding seems less likely to have this effect because it mechanically blocks the agent from acting on their biases. Various lines of research indicate that being anonymous can promote (or reveal) antisocial impulses and actions (Postmes and Spears 1998), but this seems unlikely in the domains examined here because anonymity is not being offered to active participants in the process where blinding occurs (data analysis, review of applications, etc.).

Will Biases "Find a Way"?

To adapt a line from Jurassic Park, another concern is that even if blinding works, somehow biases will find a way—that is, blocking a bias through blinding will just open a path to a different manifestation of bias. This could happen in several different ways. When a target attribute is inaccessible or difficult to cognitively process, individuals often substitute one cue for another, a process Brunswick called "vicarious functioning" (see Gigerenzer and Kurz 2001; Kahneman and Frederick 2002). For example, if reviewers do not know a performer's gender, they may put more weight on presumed proxy variables like the volume or dynamics of the music. If the substitute variable is actually a good cue (relative to some normative system), so much the better. But there are ways in which substitution could make things as bad, or worse than the original situation. In a strong case of taste discrimination, the judge may reject all candidates rather than risk the possibility of selecting a member of the disliked class. Or the judge may reject all candidates who have a proxy cue that the judge associates with the disliked class.

There's a troubling real-world example. Based on evidence that prison records were making it difficult for many African American men to find jobs, many jurisdictions adopted "ban the box" policies that prohibited employers from including a "criminal history" checkbox on job application forms. Unfortunately, there is convincing observational (Holzer et al. 2006) and experimental (Agan and Starr 2018) evidence that this policy has the opposite effect—it significantly reduces the hiring of members of groups that employers associate with criminality. In essence, when blinding blocks employers from considering criminal justice information, they will often use race or ethnicity as a proxy, potentially replacing a smaller category (men with criminal records) with a larger one (men of color).

Given a set of available variables, what correlational and causal structures are most conducive to effective, ineffective, or even pernicious applications of blinding? This is a topic that merits further theory and research.

Does Blinding Crowd Out Better Solutions?

In psychology and sociology there has been a lively debate about the relative merits of "color blindness" versus "multiculturalism" as remedies for racial and ethnic discrimination. For example, Boddie (2018) complains that

> The problem is that no one is colorblind, and acting as if we are makes us worse off, not better.…While whites may be conscious of others' race, they often are not conscious of their own because they do not have to be. Colorblindness, therefore, forces race underground. It turns people of color into tokens and entrenches whiteness as the default.

Plaut et al. (2018:204) offer a nuanced empirical review of the tradeoffs inherent in the choice between color blindness and multiculturalism, concluding:

> Color blindness, while often heralded as a remedy for racism, can foster negative outcomes for people of color (e.g., interpersonal discrimination). Moreover, color blindness serves to reify the social order, as it allows Whites to see themselves as nonprejudiced, can be used to defend current racial hierarchies, and diminishes sensitivity to racism. Multiculturalism can provoke threat and prejudice in Whites, but multicultural practices can positively affect outcomes and participation of people of color in different institutional arenas. Yet it also has the potential to caricature and demotivate them and mask discrimination.

Does blinding crowd out other data collection and analysis strategies in science? Possibly. Many have argued that the conditions required to implement a proper double-blind randomized trial create unrepresentative—and hence misleading—circumstances. And I suppose blinded data collection could crowd out preregistration and other bias-control policies if we let it. In data analysis, a bigger concern is that blinding might blunt the possibility of making unanticipated discoveries in the data. Blind methods allow the analyst to supplement preregistered analyses with more exploratory analyses, while still minimizing the effects of wishful thinking on interpretation.

Blinding When Normative Systems Collide

The logic of blinding is relatively straightforward when there is a single normative system (e.g., "find the truth") for defining bad cues. In most domains, there are multiple normative systems making claims on our conduct—truthfulness, fairness, collegiality, and the like. I don't think anything in my review points to stark differences in how blinding can or should work for these different normative systems. But certainly, things get more tricky when there are conflicting normative demands—for example, validity versus fairness. Then, a cue might be "good" with respect to one system but "bad" with respect to another. These issues have been explored in depth in the professional literature on the psychometrics of ability testing and assessment, but they have not been solved or resolved, and I suspect similar issues will arise in applications of blinding in other domains.

From "Deliberate Ignorance: Choosing Not to Know," edited by Ralph Hertwig and Christoph Engel.
Strüngmann Forum Reports, vol. 29, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262045599