Causality, Fairness, and Information in Peer Review

Sheridan Grant

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Elena Erosheva, Chair

Marina Meilă, Chair

Thomas Richardson

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Causality, Fairness, and Information in Peer Review

Sheridan Grant

Co-Chairs of the Supervisory Committee:
Dr. Elena Erosheva
Statistics

Dr. Marina Meilă
Statistics

In this dissertation, I study peer review—the process by which scientists evaluate one another's work for publication or funding—through three distinct but related lenses. I focus on multi-step grant proposal peer review processes, in which reviewers score a research grant proposal on a set of criteria (via *criterion scores*) as well as overall. The National Institutes of Health (NIH) and American Institute of Biological Sciences (AIBS) both use this type of peer review.

We begin by analyzing racial disparities in NIH peer review scores, determining that the criterion scores explain racial disparities in overall scores. We also find—perhaps surprisingly—negligible racial differences in *commensuration*, the way in which criterion scores are weighed when determining the overall score and a potential vector for racial disparities. Our analysis uses hierarchical mixed-effects models to account for the intricate dependencies in the NIH's peer review structure and matching to nonparametrically adjust for covariates. Additionally, I discuss the conditions under which estimates from these models carry causal interpretations, and investigate the robustness of our estimates to deviations from these assumptions.

Outstanding questions from the NIH study motivate the subsequent chapters of the dis-

sertation. The unmeasurability of a grant proposal's underlying quality—a sure mediator of the relationship between demographics and peer review scores—leads us to explore a related question: how informative are peer review scores? We leverage the decimal AIBS scoring scale and the proven tendency of raters to round to define and study *refinement*, which characterizes the informativeness of a set of peer review scores. We find evidence that overall scores are more informative than criterion scores at AIBS.

Finally, the experimentally unverifiable causal structure underlying our NIH study models motivates us to adapt causal discovery techniques for use in peer review. We quantify uncertainty in discovery with a fully Bayesian approach—Bayesian Causal Discovery—that enables researchers to establish confidence in the causal structures that underpin future analyses.

TABLE OF CONTENTS

GLOSSARY

Italicized terms in the text are defined in the glossary, in addition to acronyms and common terms whose usage in this dissertation is specialized.

ADMINISTERING INSTITUTE: at NIH, one of twenty-four institutions that award grant funding in a particular research area.

AIBS: the American Institute of Biological Sciences.

APPLICATION: I use this term to refer to all components of an application for grant funding, including not just the research proposal but also, e.g., a biosketch.

BALANCE: the degree of similarity between the covariate distributions in the treatment and control groups.

BIBLIOMETRICS: statistics describing outcomes of scientific research, including publication counts, citation counts, and Journal Impact Factors (Borgman and Furner, 2002).

CAUSAL DIAGRAM: a graph relaying causal relationships among variables; see Pearl (1995).

CAUSAL DISCOVERY: the process of learning the causal relationships among a set of variables from observational data.

CEM: coarsened exact matching (Iacus et al., 2012). Exact matching on a coarsening of the observed covariates. A generalization of exact matching that achieves many nice properties (see Appendix A.2.1).

COMMENSURATION: the process of weighting, or combining, criterion scores into an overall score.

COMMENSURATION DIFFERENCES: differences—in Chapter 2, racial differences—in the average weighting of criterion scores into an overall score.

COMMENSURATION DISPARITIES: when commensuration differences, in aggregate, contribute to racial disparities in Overall Impact scores.

CONDITIONAL IGNORABILITY: a treatment T is ignorable for an outcome Y with potential outcomes $Y(0)$ and $Y(1)$ given covariates X if $Y(0), Y(1) \perp\!\!\!\perp T|X$.

CRITERION SCORES: in grant proposal peer review, scores that reflect specific aspects of a proposal. At NIH, the five criterion scores are Significance, Investigator, Innovation, Approach, and Environment. At AIBS, the four criterion scores are Significance, Investigator, Innovation, and Approach.

CROSSED RANDOM EFFECTS: non-nested random effects, which exist when one random factor level occurs for multiple levels of a different random factor in the experimental design.

CSR: at NIH, the Center for Scientific Review, which manages NIH's peer review processes.

EVIDENCE: in a Bayesian model M for data D parameterized by parameters $\theta \in \Theta$, the likelihood of the data under M: $\int_\Theta p(D|\theta)\, dP(\theta)$.

FINAL SCORE: at NIH, scores given by reviewers after SRG discussion of grant proposals—i.e., a revision of the preliminary score.

GAMING: while abiding the rules of a system, behaving in a way that serves one's own ends instead of the system-wide goal.

GOLD STANDARD: a widely-accepted measure of an ill-defined quality, such as "scientific importance."

HEAPING: the practice of rounding an estimate or report of a quantity, thereby introducing an error.

IRG: at NIH, Integrated Review Groups. These research area-specific units oversee many SRGs.

ITEMS: what is being score or ranked, e.g. in peer review.

MATCHING: a process often used to estimate a treatment effect whereby treated and control units are "matched" on the basis of other covariates and set aside for treatment effect estimation. Matching can be considered a form of semiparametric inference.

MULTISET: an extension of a set, in which multiple copies of an element are permitted.

NIH: the National Institutes of Health.

OVERALL SCORE: in grant proposal peer review, a score that reflects the overall quality of a proposal. At NIH, "Overall Impact Score"; at AIBS, "Merit Score."

PANEL: a group of peer reviewers that discuss a collection of items and score/rank them.

PERCENTILE SCORE: the percentile ranking of a proposal's average final Overall Impact score in the distribution of average final Overall Impact scores from the previous three rounds of review for a given SRG.

PI: Principal Investigator—in peer review, the lead scientist on a grant proposal.

PRELIMINARY SCORE: at NIH, scores given by reviewers prior to SRG discussion of grant proposals.

PROPOSAL: I use this term to refer specifically to the document that proposes the scientific research requiring funding, the most important of multiple components of a grant application.

PSEUDO-BIAS: occurs when an estimator is unbiased but has an extremely skewed distribution such that it almost certain to be an under- (or over-) estimate.

R01 GRANT PROPOSAL: the original and still-primary grant mechanism at NIH. "Provides support for health-related research and development based on the mission of the NIH" (NIH Staff, 2021d).

RELATIVE BIAS: if $\beta' = \beta + bias$ is estimated in place of β, the relative bias of β' for β is $bias/\beta'$.

SEP: at NIH, Special Emphasis Panels. These one-time units are convened to review grant proposals in a specific scientific area.

SRG: at NIH, Scientific Review Groups. These units review grant proposals and provide scores to administering institutes for use in funding decisions.

WEAK ORDERING: informally, a transitivity-respecting ordering in which some items are incomparable to others.

ACKNOWLEDGMENTS

This dissertation was made possible through collaboration with Drs. Elena Erosheva and Marina Meilă, who are both excellent advisers and researchers individually, but even better as a team. Thanks also to Drs. Carole Lee and Thomas Richardson, who made major contributions to subsets of these chapters, and who provided new perspectives and encouragement when they were needed most. Additionally, thanks to Dr. Kate Stovel for volunteering her time as my GSR, and for warmly welcoming me and many other graduate students to UW's interdisciplinary research culture through the Center for Statistics and the Social Sciences. Two chapters of this dissertation rely heavily on collaboration with or assistance from researchers at peer institutions—in particular, Richard Nakamura, Mei-Ching Chen, Mark Lindner (National Institutes of Health), and Steve Gallo (American Institute of Biological Sciences). Research in this dissertation was made possible by a contract from the National Institutes of Health and grant funding from the National Science Foundation and the National Security Agency. Finally, to my family, friends, and therapist: thank you for standing by me—whether literally, in Seattle, or from thousands of miles away—these last five years. Through dark Seattle winters, pandemics, and personal trials, I never felt alone because you reminded me that breakthroughs and a brighter world are always around the corner.

DEDICATION

to my family

Chapter 1

INTRODUCTION

This dissertation is motivated by practical problems in the statistical analysis of peer review. I will present new scientific findings about racial disparities and informativeness of scoring in peer review, as well as novel statistical methods for the analysis of peer review data and beyond. These range from the specific (quantifying information in structured rating scales, such as those used in peer review) to the general (evaluating uncertainty in causal discovery). This chapter expands upon these contributions and outlines the remaining chapters of the dissertation. It also introduces the concepts from peer review that will underlie much of the work.

1.1 Peer Review at the National Institutes of Health and American Institute of Biological Sciences

This dissertation uses data from two institutions that perform peer review for research grant proposals in the biomedical sciences: the National Institutes of Health (NIH) and the American Institute of Biological Sciences (AIBS). NIH funds scientific research on the basis of its peer review in addition to numerous other missions. AIBS focuses on three relatively distinct functions: science advocacy, publishing the journal *BioScience*, and peer review in the service of other organizations, such as NIH. Here, I describe the basic two-stage (pre- and post-panel) peer review process shared by NIH and AIBS, leaving the particularly complex structure of NIH R01 peer review to Section 2.2.1 of Chapter 2. R01 is the original and dominant NIH grant type, broadly reserved for "research and development based on the mission of the NIH" that "must be related to the stated program interests of one or more of the NIH Institutes and Centers based on their missions" (NIH Staff, 2021d).

The *panel* is a central component of peer review. A panel is typically composed of a set of peer reviewers, all of whom share expertise to some extent and are reviewing proposals in a particular scientific field. For NIH R01 grant reviews and some AIBS reviews, multiple panelists review a proposal independently before the panel discusses (possibly a subset of) the proposals. The panel may meet in-person or online, asynchronously, to discuss proposals.

At both NIH and AIBS, *preliminary* scores are given by reviewers after reading a proposal but before the panel meets; these may then be revised after panel discussion, yielding *final* scores. At NIH, only a subset of proposals are discussed by panels; these are chosen on the basis of the preliminary scores. Panel discussions induce dependencies between reviewers' scores, and since they are private, these dependencies cannot be modeled explicitly. In Chapter 2, we focus on NIH preliminary scores because we seek to explicitly model all dependencies between scores. In Chapter 3, we consider AIBS final scores; the metrics developed in this chapter assume the existence of latent dependencies between scores.

At NIH and AIBS, there are two types of scores: *criterion*, and *overall*. The criterion scores each concern a specific component of a proposal, while the overall score relays the overall quality of the proposal. At NIH, there are five criteria: Significance, Investigator(s), Innovation, Approach, and Environment (see Table A.2 for NIH's detailed descriptions of each criterion). AIBS uses just the first four of these criteria, with similar definitions. NIH refers to the overall score as the Overall Impact score, and tells reviewers that it should convey "the project's likelihood to have a sustained, powerful influence on the research field(s) involved" (NIH Staff, 2012). AIBS refers to the overall score as the Scientific Merit score.

It is important to note that reviewers are not required to determine the scores in a particular order. However, NIH notes that reviewers will "derive the Overall Impact score from the individual criterion scores" (NIH Staff, 2016b) and that "the impact score for an application is based on each individual reviewer's assessment of the scored criteria plus additional criteria..." (NIH Staff, 2013). Such language implies that criterion scores should be determined before overall scores. The causal ordering of scores will be relevant for our

NIH study in Chapter 2 and motivate the work in Chapter 4.

Final scores from panels are indirectly used to determine which grants to fund. NIH panels compute an average of the final Overall Impact scores for each application, multiplied by ten and rounded to the nearest integer, which I refer to as the panel Overall Impact score. Finally, panels compute *percentile scores* based on the ranking of panel Overall Impact scores for all applications reviewed in the three most recent meetings (NIH Staff, 2021e). These percentile scores correct for differences in how different panels use the NIH scoring scale, though they also erase information contained in the scores. Percentile scores then help NIH *administering institutes* determine which grants to fund. Since AIBS does not directly fund grants, it transmits final scores and other review content to the funding institutions that use its services, who then make funding decisions based on this information.

1.2 Contributions to the Science of Peer Review

Chapter 2 of this dissertation investigates racial disparities in NIH grant proposal peer review scores, using hierarchical mixed-effects models to rigorously interrogate NIH peer review. It relies on an analysis of confidential NIH data performed by myself, Drs. Elena Erosheva and Carole Lee, and a team of NIH researchers, published as Erosheva et al. (2020b). Specifically, we

- model Black-white[1] racial disparities—after adjusting for an array of mediating factors—in NIH preliminary scoring, the part of NIH's process that most drives racial disparities in grant funding (Hoppe et al., 2019);

- assess the extent of racial differences in commensuration (Lee, 2015), a subtle mechanism by which underrepresented applicants can be disadvantaged in complex review processes; and

[1] I follow the New York Times's capitalization policy for the racial terms "Black" and "white" (Coleman, 2020).

- quantify the contributions to variability in Overall Impact scores of various parts of NIH's multilevel peer review structure.

This study raises just as many questions as it answers, however. Two of these questions motivate Chapters 3 and 4 of the dissertation.

Our assessment of racial disparities in NIH grant review, in Chapter 2, is impeded by an unobservable correlate of race that strongly influences peer review scores: the underlying quality of a research proposal. This underlying quality cannot be measured, except by proxy, so in Chapter 3, we reframe the problem: how much do a set of scores tell us about the associated proposals? With Drs. Marina Meilă, Carole Lee, and Elena Erosheva, I develop a new measure—*refinement*—of the informativeness of a set of ratings in settings with no gold standard for the quality of the outcome. We apply our method to AIBS peer review data, finding significant differences in refinement for different types of scores.

1.3 Contributions to Statistical Methodology

Chapter 2 identifies a second thorny unknown in the NIH peer review process: the causal relationships among the criterion and overall scores. While reviewers are led to address criterion scores first, in a *weak ordering* (NIH Staff, 2016b), scoring happens in the mind, prohibiting experimental verification of the causal order of review scores. *Causal discovery* techniques that learn causal relationships from observational data (Spirtes et al., 2000) offer a potential solution—but most existing methods do not quantify the uncertainty in a learned causal structure, a scientific necessity when downstream analysis of data depends deeply on the assumed causal structure of the covariates. In Chapter 4, which covers joint work with Drs. Thomas Richardson, Marina Meilă, and Elena Erosheva, we fill this gap in the literature for the bivariate, additive regime with a novel, Bayes Factor-based approach to uncertainty in causal discovery. Further research is needed before this type of method can be applied to multivariate peer review scores; advances in Bayesian computation or numerical integration could permit such an extension.

Chapter 2

RACIAL BIASES AND DISPARITIES IN NIH PEER REVIEW SCORES

This chapter is based on work published in *Science Advances* (Erosheva et al., 2020b). It presents the same modeling results as the published paper, but further elaborates on the statistical methods and underlying theory. In particular, this chapter analyzes the assumptions that would lend our results causal interpretation, while Erosheva et al. (2020b) uses purely associative interpretations. Important differences from Erosheva et al. (2020b) are compiled for easy reference in Appendix A.6. No new data-based computations were performed for this chapter, except for those based solely on statistics published in Erosheva et al. (2020b) rather than the confidential data used in that paper.

The University of Washington's Institutional Review Board determined that the study did not involve human subjects. This work was supported by NIH contracts HHSN268201600310A and HHSN268201700300A, as well as NSF grant 1759825, all awarded to Drs. Elena Erosheva and Carole Lee.

Finally, because of the sensitive nature of the peer review data underlying the model estimates and other results in Erosheva et al. (2020b), which are repeated here, a public version of these confidential data—containing all records but only a subset of the study variables, in compliance with NIH policy—was made available for the purposes of reproducibility (Erosheva et al., 2020a). An R Markdown file accompanies the data, allowing anyone to reproduce model results comparable to—but not identical to, due to the restricted covariate set—those presented in this chapter (see Appendix A.5).

2.1 Introduction

Section 1.1 described peer review at NIH and AIBS, focusing on the process in which reviewers score a proposal on several criteria as well as an overall score which is then used to decide whether or not to fund the proposal. Racial disparities arising from such peer review processes have been the subject of intense research over the last decade. At NIH, the funding rate for R01 applications from Black Principal Investigators (PIs) has historically been substantially lower than that of applications from white PIs—45% lower in the years 2000–2006 (Ginther et al., 2011). A substantial portion of this disparity can be explained by racial differences in field-adjusted bibliometrics (publications, citations, and journal impact factors) (Ginther et al., 2018). Recent work from NIH, based on proposals from 2011–2015, found that racial disparities emerge when proposals are selected for discussion by a Scientific Review Group and when final Overall Impact scores are assigned; the tendency of Black PIs to propose research on topics with lower award rates also plays a role in funding disparities (Hoppe et al., 2019).

In this chapter, we consider a different step in the process that may impact racial disparities in NIH peer review: criterion scoring. Experts have proposed criterion scores as a means of increasing focus on factors related to merit and dampening social biases (Thorngate et al., 2009; Kahneman, 2013). However, studies have also found that ambiguity and uncertainty in *commensuration*—the process of weighting criterion scores into an overall score (Espeland and Stevens, 1998)—can amplify social biases (Hodson et al., 2002; Norton et al., 2006, 2004; Uhlmann and Cohen, 2005, 2007). Specifically, without instructions for how to commensurate the criterion scores into an overall score, the weightings of the criteria can differ based on applicants' characteristics, which we refer to as *commensuration differences*. *Commensuration disparities* occur when these differences disadvantage certain groups in aggregate (Lee, 2015). These conflicting possibilities motivate us to understand whether the 2009 introduction of criterion scores at NIH (NIH Staff, 2009) decreased funding disparities, to model racial commensuration differences at NIH, and to investigate the extent to which

criterion scores explain racial disparities in Overall Impact scores.

Using data from NIH's 2014–2016 funding years, we establish that the R01 funding rate gap between Black and white PIs is as large as it was in 2000–2006. We then narrow our focus to commensuration in preliminary scores, which are used to determine which proposals Scientific Review Groups (SRGs) will discuss and therefore substantially—though indirectly—impact funding decisions. We are the first researchers to analyze preliminary scores, enabling us to shed a unique light on the process by which proposals are selected for discussion, which Hoppe et al. (2019) called the "decision point that makes the largest single contribution to the funding gap." Our findings regarding commensuration and racial disparities in preliminary scores are twofold. First, we find evidence of modest racial differences in commensuration, but that the aggregate impact of these commensuration differences on preliminary Overall Impact scores (commensuration disparity) is negligible. Second, we demonstrate that preliminary criterion scores fully account for racial disparities—yet explain only part of the variability—in preliminary Overall Impact scores.

2.2 Study Data

Our data come from the IMPAC II (Information for Management, Planning, Analysis, and Coordination) grant data system, which stores information about each NIH application and self-reported demographics such as race and gender. Study variables include

- preliminary Overall Impact and preliminary criterion scores, defined in Table A.2;
- multilevel structural covariates, discussed in Section 2.2.1; and
- other applicant- and application-specific covariates that were previously shown to be associated with Overall Impact scores net of criterion scores (Eblen et al., 2016), detailed in Table A.1 and Appendix A.2.

This study concerns a full data set of 54,740 R01 applications submitted by Black and white PIs and reviewed by NIH's Center for Scientific Review (CSR) during council years

2014-2016. CSR reviews about 90% of R01 applications; reviews of applications submitted to funding opportunity announcements with special review criteria are sometimes managed entirely by the administering institutes. Because our focus is on Black-white peer review disparities, we excluded from this study a total of 1,771 applications submitted by PIs whose race was American Indian or Alaskan, Asian, Native Hawaiian or Pacific Islander, or by PIs who indicated more than one race, as well as 8,648 applications for which PI's race was withheld or unknown. At the time of application, PI demographics are voluntarily reported by applicants; NIH requests but cannot compel PIs to provide this information.

Self-reported demographics, including race, do not appear with the application when it is handled by reviewers or by the NIH SRG review committee, staff, or council. However, reviewers may infer applicant race from some combination of personal knowledge, external information, and the application materials (e.g., name, receipt of a minority fellowship/grant, or other NIH biosketch content).

Approximately 15% of the applications from Black and white PIs were missing information on PI gender, ethnicity (Hispanic/Latino or not), or educational degree, and were excluded from the study. The remaining 46,226 applications—1,015 (2.2%) from Black PIs and 45,211 (97.8%) from white PIs—were evaluated by 19,197 unique reviewers who wrote 139,216 reviews (Table 2.1). CSR is unaware of patterns among PIs not reporting their demographic characteristics, though this issue deserves further study because such patterns could lead to biased conclusions (Little and Rubin, 2019)—even if the joint distribution of the non-missing variables for PIs with missing data is similar to that of the rest of the applicant pool. More details about the study data are available in Appendix A.2.

2.2.1 Multilevel Nature of NIH's Peer Review and Funding Process

At NIH, panels that evaluate the scientific merit of proposals are known as Scientific Review Groups (SRGs). These are organized within Integrated Review Groups (IRGs) by general scientific area (NIH Staff, 2021c). In addition to SRGs, Special Emphasis Panels (SEPs) are formed within IRGs to review other topics and proposals for which an SRG member has a

Table 2.1: Sampled data summary statistics by application subset.

Subset	Unique PIs	Reviewers	Reviews	Applications
All Black	500	2,310	2,926	1,015
Matched Black	456	2,084	2,578	890
Matched White	1,497	3,866	4,893	1,676
Random White	1,904	4,460	5,669	2,030
Total	3,679	7,901	13,140	4,596

conflict of interest (NIH Staff, 2021a). That is, SRGs and SEPs are nested within IRGs.

After an SRG finishes discussion of proposals, reviewers update their preliminary scores, yielding final scores, and other discussants (who did not review or preliminarily score the proposal) also provide final scores. The average of the entire SRG's final Overall Impact scores is then compared to the distribution of such averages over the previous three rounds of review, and its percentile in that distribution is called its *percentile score*. These percentile scores are passed on to NIH administering institutes, which carry out a second round of review and ultimately make funding decisions (NIH Staff, 2019b).

Individual PIs may submit multiple applications, which may be reviewed by multiple SRGs, which may fall under multiple IRGs. Reviewers typically review multiple applications within an SRG, but just under 3% review for more than one SRG. Figure 2.1 displays a diagram of the NIH review structure that features two applicants, three applications, and six reviewers.

2.2.2 Matching and Study Data Subsets Selection

Due to the sensitive nature of NIH peer review records, we were only given access to a fraction of the full data—but, we were also given discretion as to which records would make up the study data. This section describes the study data selection process; further details

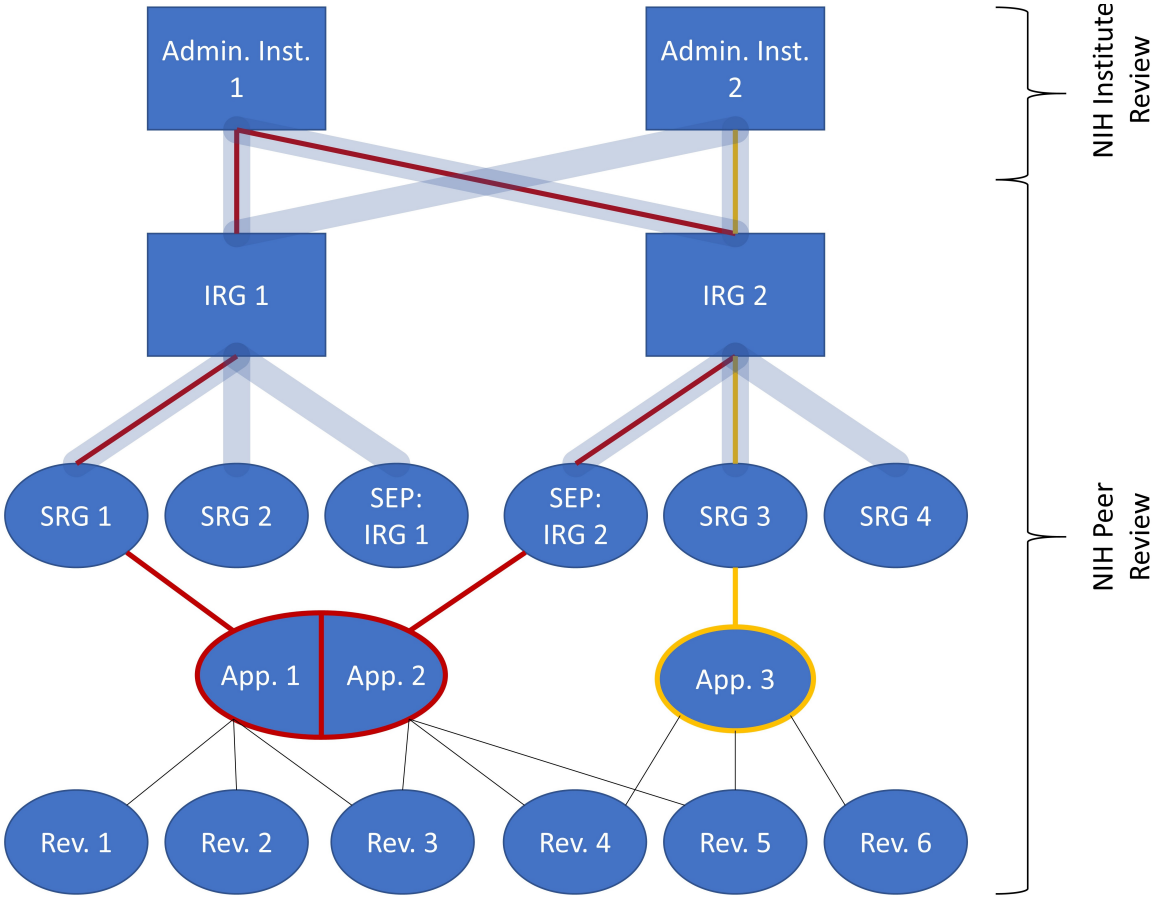


Figure 2.1: Multilevel NIH peer review structure for a hypothetical example of three applications (App. 1, 2, and 3) submitted by two PIs (yellow and red). Thick blue lines show structural connections. Thin lines show hypothetical reviewer assignments for the three applications. Rectangles are specified as fixed effects and ellipses as random effects in our mixed effects models in chapter 2.

can be found in Appendix A.2.1. Because our investigation primarily concerns covariate-adjusted Black-white differences in peer review scoring at NIH, we selected records that would yield high statistical power to detect such differences. Given the underrepresentation of Black investigators among NIH applicants, we obtained all applications from Black PIs. We combine these applications with a randomly selected subset of applications from white applicants to form the "random subset" of applications, which facilitates analyses requiring

representative samples of both the white and Black NIH R01 grant applicants.

Our analyses also rely on a *matched* subset of the data in which applications from Black applicants are matched to applications from white applicants ("matched Black" and "matched white" applications). Matching yielded a relatively large sample size, with 88% of Black applications retained. As a result, the matching—which can be viewed as a nonparametric regression adjustment technique (Ho et al., 2007)—yields enough statistical power to detect even substantively small effects. We used exact matching on eight key variables thought to be related to scores and award rates, summarized in Table 2.2.

Exact matching on a subset of covariates can be considered a version of Coarsened Exact Matching (CEM, Iacus et al. (2012)) with complete matching on selected variables and full coarsening on other variables (see Appendix A.2.1). We matched on only a subset of the covariates to balance the desire to nonparametrically adjust for those covariates thought likely to be strongly associated with the outcome with the fact that matching on more covariates yields smaller sample sizes. Relative to random sampling, matching improved balance—the similarity between covariate distributions for the white and Black subsets—on all the matching variables and on most other applicant- and application-specific covariates (Table A.5). This improved balance makes estimates based on the matched subset more robust to model misspecification and more efficient than analyses based on a random sample of the same size that use parametric controls (Ho et al., 2007; Erosheva et al., 2020b). Thus, when we wish to adjust for applicant- and application-specific covariates, we estimate models using the matched subset of the data.

Finally, the selection of records was subject to the constraint that no individual reviewer could have more than four reviews in the sample, in order to ensure the privacy and confidentiality of reviewers. Obeying this constraint while maintaining representativeness of the sampled data was nontrivial. Appendix A.2.1 presents the algorithms we used to accomplish this, as well as the matching algorithm.

Table 2.2: Matching Variables.

Name	Description
Applicant	
Gender	F/M, self-reported
Ethnicity	Hispanic/Latino or not, self-reported
Career Stage	Early Stage (ES), Experienced, or Non-ES New Investigator
Educational Degree	PhD, MD, MD/PhD, Other
NIH Funding Bin	FY 2014 total institution NIH funding; 5 bins
Application	
Application Type	New or Renewal
Amended Status	Amended or not
IRG	Integrated Review Group

Funding bins—with 20% of Black applications in each bin—were based on an ordering of the 1015 Black applications by total NIH funding received by the applicant's institution in fiscal year (FY) 2014 (see Table A.6).

2.3 Methods

While the available data are observational, the goal of this chapter is to estimate causal parameters that allow us to make counterfactual inferences. First, we would like to know what Black applicants' preliminary Overall Impacts scores would have been, on average, were their perceived race[1] (counter to fact) white, holding all other application factors perceived

[1]Race is a social construct, not an inherent trait (Coates, 2013). Furthermore, race—as it is constructed for the purposes of NIH peer review—is immutable, raising fundamental problems for defining any "causal effect of race" (Greiner and Rubin, 2011). Following these authors, we assume that any racial discrimination is based on race as perceived by reviewers. We also assume that applicants' perceived race aligns with the self-reported race. Recent work has criticized approaches such as ours, which treat (perceived) race as separate from other demographic characteristics such as educational background (Hu and Kohler-Hausmann, 2020). Future research on racial disparities should seek a more complete model for race and discrimination.

by reviewers fixed. We also wish to estimate this counterfactual with the criterion scores held fixed, so that we isolate any race effect in the production of overall impact scores from criterion scores. That is, we seek the total and direct effects of perceived race on preliminary Overall Impact scores; see Bertrand and Mullainathan (2003) for an experiment that identifies a similar total effect of perceived race on hiring outcomes.

Second, we seek to quantify the extent of commensuration differences: on average, how differently would Black applicants' preliminary criterion scores have been weighted when determining the preliminary Overall Impact score had their perceived race been (counterfactually) white but all other factors held equal? And third, do racial differences in commensuration disadvantage Black applicants on the whole?

Strong assumptions underlie any purportedly unbiased estimation of causal effects from observational data. In Section 2.3.3, we explicate these assumptions in the context of multilevel regression models for preliminary Overall Impact scores. When these assumptions hold, our results can be interpreted in terms of biases that stem from perceptions of race. We do not ask the reader to believe all these assumptions hold entirely, however, and in Section 2.4 we will interpret results in terms of covariate-adjusted racial disparities rather than biases to reflect that these identifying assumptions do not fully hold in practice. This chapter elucidates the conditions that would permit an unbiased estimate of a causal effect, how real-world conditions may differ, and how those differences may affect estimates.

2.3.1 Model Covariates and Random Effects Structure

Our multilevel model draws upon NIH grant review's hierarchical structure (Figure 2.1), modeling these dependencies through structural variables: IRG, SRG, and administering institute, as well as reviewer, PI, and application indicators. Reviews, the units of observation in this study, lie at the bottom of the hierarchy. Reviews are nested within applications, which are nested within PIs. But reviewers can review multiple PIs just as PIs are reviewed by multiple reviewers: reviewer and PI are *crossed*. Applications are nested within SRG, IRG, and administering institute, but PIs are not: over 200 PIs had applications reviewed

in more than one SRG, IRG, or administering institute. All SRGs are nested within IRG, while IRG is crossed with administering institute. All special emphasis panels (SEPs) within an IRG were modeled as a single study section/SRG.

The choice of whether to represent a structural covariate via fixed or random effects was based on our substantive knowledge of the NIH review process, as well as practical modeling considerations discussed in Section 2.3.3. Administering institutes and IRGs are fixed entities that rarely, if ever, increase or decrease in number, and are therefore modeled as fixed effects (rectangles in Figure 2.1). SRGs, on the other hand, are routinely created or disbanded, and thus—like reviewers and PIs—are considered to be a sample from a larger population and modeled with random effects (ellipses in Figure 2.1). Variability for application ID random intercepts is not reported because PI random intercepts were estimated to capture all variability in application ID. It is important to point out that individual differences between reviewers—reflected as reviewer random intercepts in our models—can be thought of as arising from individual differences in expertise, scientific interests, and value systems (Hargens and Herting, 1990; Lee, 2012). Likewise, individual differences between PIs are reflected by PI random intercepts, and average differences in preliminary Overall Impact scores between SRGs are captured by SRG random intercepts.

Other covariates include the applicant- and application-specific covariates from Table A.1. The five preliminary criterion scores can also be thought of as covariates that explain variability in preliminary Overall Impact scores. The causal relationship between these covariates and race is somewhat murky: some applicant- or application-specific covariates may influence perceptions of race (Hu and Kohler-Hausmann, 2020), though all model covariates are determined after a PI's birth, at which point we often consider race to be fixed. Regardless, estimating the effect of *perceived* race on preliminary Overall Impact scores (see Section 2.3.3) means that we must control for all covariates that are associated with both perceived race and Overall Impact scores.

2.3.2 Racial Disparity Models

Let Y_{ijklm} be the preliminary Overall Impact score for the ith review of the jth application from the kth PI (reviewed by the lth reviewer in the mth SRG), R_k a race indicator (1 indicates a Black PI), Z_{ij} the vector of criterion scores for the ith review of the jth application, and X_{jk} the vector of application- and applicant-specific control variables aside from PI ID, reviewer ID, and SRG. To estimate racial disparities, we consider the following mixed effects models:

$$Y_{ijklm} = \beta_0 + \beta_R R_k + \beta_C Z_{ij} + \beta X_{jk} + \gamma_k + \xi_l + \eta_m + \epsilon_{ij} \tag{2.1}$$

$$Y_{ijklm} = \beta_0 + \beta_R R_k + \beta_C Z_{ij} + \quad + \gamma_k + \xi_l + \eta_m + \epsilon_{ij} \tag{2.2}$$

$$Y_{ijklm} = \beta_0 + \beta_R R_k + \quad + \beta X_{jk} + \gamma_k + \xi_l + \eta_m + \epsilon_{ij} \tag{2.3}$$

$$Y_{ijklm} = \beta_0 + \beta_R R_k + \quad + \gamma_k + \xi_l + \eta_m + \epsilon_{ij} \tag{2.4}$$

where β_0 is the model intercept; β_R is the race coefficient; β_C is the vector of criterion score weights/coefficients; β is the vector of coefficients for control variables; γ_k, ξ_l, and η_m are crossed[2] random intercepts for PI, reviewer, and SRG; and the ϵ_{ij} are within-application independent Gaussian error terms. We refer to models (2.1) and (2.3) as the "direct" and "total" racial disparity models. Models (2.2) and (2.4) are the "unadjusted" versions of those models. Estimates from these four models are shown in Table 2.3; the red and orange colorings connect the models to Figure 2.2 and will be explained in the subsequent section.

Because the matching variables include seven applicant- and application-specific covariates and just one structural variable, we fit models including applicant- and application-specific control covariates to the matched subset of the data. The matching layers non-parametric adjustment for the matching subset of these covariates on top of the standard parametric regression controls. We fit models without applicant- and application-specific regression coefficients to the random subset to ensure that only structural covariates are adjusted for.

[2]There is no nesting among PIs, reviewers, and SRGs due to the mixed nature of the NIH peer review hierarchical structure.

2.3.3 Causal Assumptions

Denote $Y(1)$ the potential Overall Impact score for a review were the applicant perceived to be black and $Y(0)$ analogously were the applicant perceived to be white. Then the total racial bias in preliminary Overall Impact scores is the average effect of (perceived) race on the preliminary Overall Impact score:

$$\tau^{total} \equiv \mathrm{E}[Y(1) - Y(0)] \tag{2.5}$$

where the expectation is taken over the distribution of the review population. Under a collection of assumptions about the data-generating process, $\tau = \beta_R$ under the total racial disparity model (2.3). Fixing the criterion scores Z allows us to define

$$\tau^{direct} \equiv \mathrm{E}[Y(1) - Y(0)|Z], \tag{2.6}$$

the direct racial bias, which is estimated by the direct racial disparity model (2.1) under similar assumptions.

Before we delve into these assumptions, we address the possibility that not every reviewer perceives applicant race. Denote Q a review-specific binary covariate that is 1 if race was perceived and 0 otherwise. Then

$$\begin{aligned} \tau^{total} &= \mathrm{E}[Y(1) - Y(0)] \\ &= \mathrm{E}[Y(1) - Y(0)|Q = 1]\mathrm{P}(Q = 1) + \mathrm{E}[Y(1) - Y(0)|Q = 0]\mathrm{P}(Q = 0) \\ &= \mathrm{E}[Y(1) - Y(0)|Q = 1]\mathrm{P}(Q = 1) \\ \implies \tau^{total} &\leq \mathrm{E}[Y(1) - Y(0)|Q = 1]. \end{aligned} \tag{2.7}$$

This inequality is strict whenever $\mathrm{P}(Q = 1) < 1$, i.e. whenever some reviewers do not perceive applicant race. When $\mathrm{P}(Q = 1)$ is small, the total racial bias[3] that we estimate—which incorporates the nonexistent bias of reviewers who do not perceive race—may therefore be much smaller than the bias displayed by reviewers who do perceive race.

[3] An analogous derivation and result applies for direct racial bias as well.

In the following developments, we rely on Figure 2.2, the *causal diagram* (Pearl, 1995; Pearl and Verma, 1987) corresponding to our model of NIH peer review. The nodes in the diagram represent variables, observed and unobserved. The directed edges represent directions of causality (with some caveats) while the single bi-directed edge means that causality can run both directions between the two sets of variables U and $\tilde{X}$. Thus, in this diagram, R is a cause of both the criterion scores Z and the overall impact score Y, and Z is a cause of Y. Actual race A is unobserved, and is determined prior to all other variables in the model; we do not refer to causal effects of A, as it is immutable (Greiner and Rubin, 2011). $\tilde{X}$ and U are observed and unobserved confounders of the relationship between Z and Y, which also induce dependence between R and Y via A. The portion of this dependence that is mediated via $\tilde{X}$—which represent the application- and applicant-specific covariates in X, as well as PI ID, reviewer ID, and SRG—is colored orange, and can be adjusted for by controlling for $\tilde{X}$ in a regression. The blue edge represents the direct effect of R on Y, τ^{direct}, while the sum of the blue and red edges represents the total effect τ^{total}.

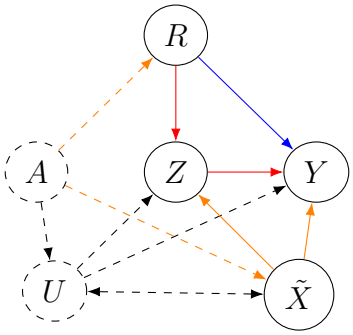


Figure 2.2: Causal structure of NIH's preliminary scoring process. A represents race, R race as perceived by reviewers, $\tilde{X}$ the observed application- and applicant-specific characteristics (as well as reviewer ID) Z the criterion scores, and Y the Overall Impact score. U represents unobserved information that reviewers use to determine Y. Blue arrow is the direct effect of R on Y for fixed Z; red arrow is the effect of R on Y as mediated through Z; orange path represents d-connection of R and Y through $\tilde{X}$.

The first assumption required for unbiased estimation of τ^{direct} is that the causal ordering for the individual-level preliminary reviewing process aligns with NIH instructions to "derive

the Overall Impact score from the individual criterion scores" (NIH Staff, 2016b) and ensure that "the impact score for an application is based on each individual reviewer's assessment of the scored criteria plus additional criteria..." (NIH Staff, 2013). That is, we assume

Assumption 1. *All reviewers assign preliminary criterion scores prior to Overall Impact scores.*[4]

If some criterion scores are in fact determined after the overall impact score (Z_2 in Figure 2.3), $\beta_R \neq \tau^{direct}$ under the direct racial disparity model (2.1) due to collider bias induced by conditioning on Z_2 erroneously.

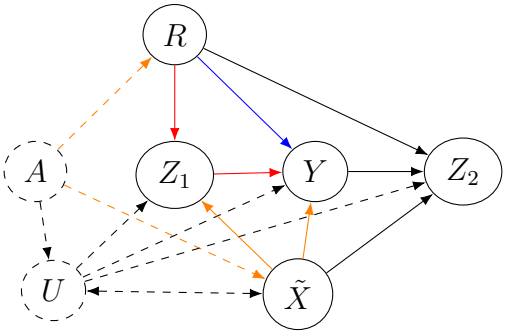


Figure 2.3: Causal structure of NIH's preliminary scoring process with some criterion scores Z_1 determined before the overall impact score Y and some determined after (Z_2). See caption of Figure 2.2 for other definitions.

Second, one assumes that perceived race R is independent of the potential outcomes $Y(0)$ and $Y(1)$ given $\tilde{X}$. Formally, we assume

Assumption 2. *Conditional ignorability:* $Y(1), Y(0) \perp\!\!\!\perp R | \tilde{X}$.

Equivalently, we assume that U in Figure 2.2 is either empty or that all variation in U is explained by $\tilde{X}$.

In practice, U—which represents factors whose relationships with race and the preliminary Overall Impact score are not completely explained by the observed covariates $\tilde{X}$—is

[4]This assumption is difficult to verify using observation data, a challenge discussed further in Chapter 4.

nonempty. For example, PI's educational history—which is in the NIH biosketch that reviewers read, but not our data set—is associated with race A and may affect Y, e.g. by affecting writing style in a way that is not captured by the criterion scores. The dashed bidirectional arrow in Figure 2.2 illustrates that $\tilde{X}$ can reduce the bias induced by U if measured variables in $\tilde{X}$ are good proxies for unmeasured variables in U. Institution NIH funding bin, educational degree, and ethnicity may be in aggregate a reasonably good proxy for one's larger educational history. Additionally, the criterion scores may mediate much of the effect of an applicant's educational pedigree on the Overall Impact score.

One important unobserved factor in U that cannot be fully explained by a proxy in $\tilde{X}$ is the unique content of the research proposal. It is difficult to speculate about the extent to which preliminary criterion scores mediate the proposal content's effect on the preliminary Overall Impact score, making the degree of omitted variable bias difficult to ascertain. Because we cannot adjust for differences in proposal content and some components of the NIH biosketch, we shall maintain the associative language of racial "disparities" from Erosheva et al. (2020b) rather than using the causal language of "biases." The reader should keep in mind that we wish to estimate causal effects but are unable to do so with zero bias. We instead aim to mitigate that bias to the greatest extent possible.

The final requirement for unbiased estimation of τ is:

Assumption 3. *The direct and total disparity models* (2.1) *and* (2.3) *are correctly specified in the following senses:*

- *The errors* ϵ_{ij} *are independent.*
- *The conditional expectation of* Y*, averaging over the random effects and errors, is written correctly.*
- *The random effects are independent of the errors:* $\gamma, \xi, \eta \perp\!\!\!\perp \epsilon$ *(exogeneity of random effects).*

Our estimates may be inefficient if the errors display heterogeneity, and coefficient standard error estimates may be biased to the extent that the errors are non-normal, but these are minor concerns given the large sample size. The linear model is justified for multiple reasons. We assume linearity in the criterion scores since reviewers are led to conceptualize the Overall Impact score as a weighted average of the criterion scores by NIH scoring guidance (NIH Staff, 2013). All other covariates are categorical—except for log requested budget—so the only constraints imposed by linearity are additivity (no interactions) and linearity in log requested budget. Finally, the matching procedure is a nonparametric control that additionally adjusts for potential interactions between the eight matching variables (Table 2.2). We therefore argue that our model accurately approximates the conditional expectation function to first order.

Finally, we verify that the scientific justification for our random effect specifications also guards against endogeneity. If random effects are indeed exogenous, then fixed-effect estimates from a correctly-specified mixed-effects model are unbiased and more efficient than if fixed effects were used in place of random effects. However, random effect endogeneity leads to biased fixed-effect estimates even if the model is otherwise correct (Hausman, 1978)—a non-issue if the random effects are replaced with fixed effects. Given our large sample size and efficiency gains from matching, loss of efficiency is much less concerning than bias, so we use random effects only when they are strongly justifiable. We are constrained by degrees of freedom to use random intercepts for reviewers and PIs due to the large number of unique reviewers and PIs in the data (Table 2.1). A Hausman test (Hausman, 1978) for endogeneity of the SRG random intercepts finds no evidence for endogeneity (see Appendix A.3.1 for a discussion of this test), further justifying the random effects specification that reflects our substantive knowledge about SRGs. All other effects in the model are fixed.

2.3.4 Commensuration Practices Models

To study commensuration practices, we estimate interaction effects between race and preliminary criterion scores. The linear commensuration model for the preliminary Overall Impact

score Y_{ijklm} of the ith review of the jth application from the kth PI (reviewed by the lth reviewer in the mth SRG) is

$$Y_{ijklm} = \beta_0 + \beta_R R_k + \beta_C Z_{ij} + \beta_I R_k Z_{ij} + \beta X_{jk} + \gamma_k + \xi_l + \eta_m + \epsilon_{ij} \tag{2.8}$$

where β_0 is the model intercept; β_R is the race coefficient; β_C is a vector of preliminary criterion score coefficients; β_I is the vector of commensuration coefficients for the interactions between race and the preliminary criterion scores; β is the vector of coefficients for control variables X_{jk}; γ_k, ξ_l, and η_m are random intercepts for PI, reviewer, and SRG; and the ϵ_{ij} are within-application independent Gaussian error terms. For commensuration models, the control variables X include applicant- and application-level characteristics as well as IRG and NIH administering institute. Estimates from model (2.8) carry causal interpretations when assumptions 1, 2, and 3 are met.

There are six parameters of interest in this model: β_R and the five components of β_I. Any component of β_I being nonzero indicates racial differences in commensuration. Under assumptions 1, 2, and 3, commensuration differences can also be conceptualized in terms of a heterogeneous causal effect:

$$\tau(z) \equiv \mathrm{E}[Y(1) - Y(0) | Z = z] \tag{2.9}$$

which is not constant in z. The heterogeneity allows us to estimate a black PI-specific marginal effect of perceived race, analogous to an Average Treatment Effect on the Treated (ATT):

$$\tau' \equiv \mathrm{E}_{R=1}[\tau(Z)] \tag{2.10}$$

which we refer to as commensuration bias. (2.10) is the change in preliminary Overall Impact score Black PIs would expect to see—on average—were they perceived as white and their criterion scores weighted as such.

2.4 Results

First, we compare award rates for Black and white applicants to determine whether the funding gap between Black and white applicants found for the years 2000–2006 by Ginther et al. (2011) persisted in 2014–2016. For all CSR-reviewed R01 applications in council years 2014–2016, the award rate for Black applications was 45% lower than that of white applications (10.2% versus 18.5%). In the random subset—the representative sample to which we had access—the award rate for Black applications was 44% lower than that of random white applications (11.03% versus 19.66%), reassuringly similar to the gap in the full data set. In the matched subset, however, this gap was only 25% (11.57% vs 15.39%). Thus, matching on variables that include area of science represented by the Integrated Review Group (IRG) reduces the award disparity between Black and white applications by 56%.

Funding disparities for Black applications are driven by disparities in peer review scores (Ginther et al., 2011, 2012, 2016), with the largest driver of the disparity being the selection of applications for SRG discussion based on preliminary Overall Impact scores (Hoppe et al., 2019). We now turn to estimates of racial disparities in these scores based on models 2.1–2.4.

2.4.1 Racial Disparities in Preliminary Overall Impact Scores

Histograms of preliminary Overall Impact scores (ranging from 1 to 9) illustrate that white applications tend to receive better (lower) scores than Black applications (Figure 2.4, bottom right), and that this difference is tempered—though not eliminated—for the comparison between matched applications (Figure 2.4, top right). We use the linear mixed-effects regression models (Raudenbush and Bryk, 2002; Goldstein, 2011) described in Section 2.3.2 to evaluate whether these marginal racial disparities in preliminary Overall Impact scores can be explained by the NIH review structure (Figure 2.1), applicant- and application-specific covariates, and criterion scoring.

Table 2.3 provides estimates of racial disparities in preliminary Overall Impact scores, controlling for structural and various other combinations of covariates. To indicate statistical

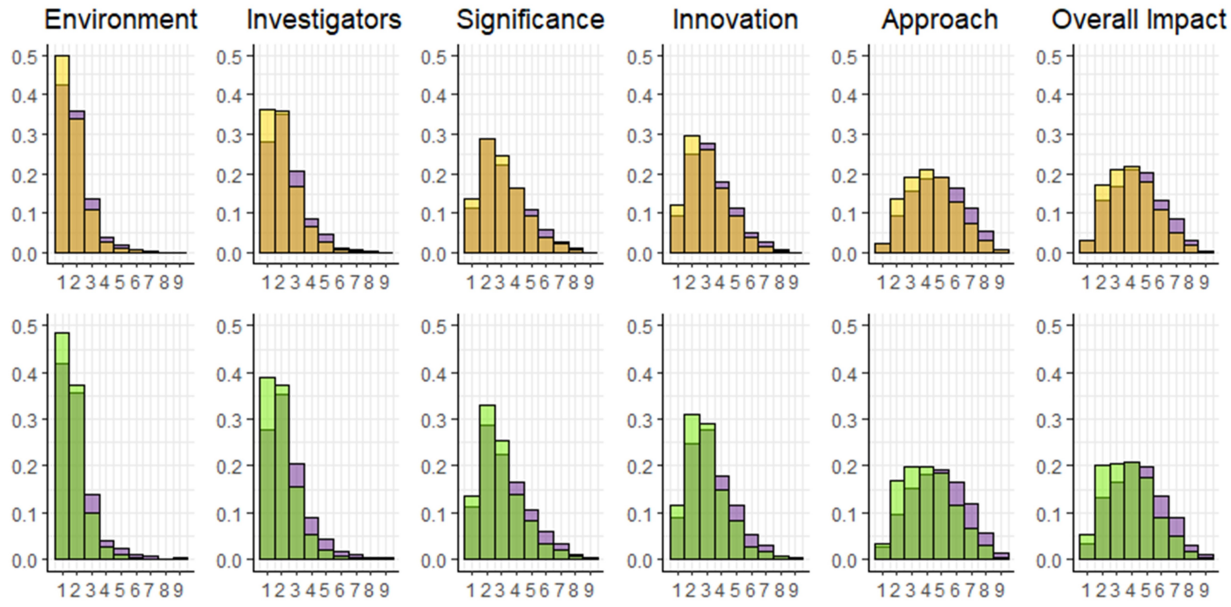


Figure 2.4: Frequency histograms for the 5 preliminary criterion scores and the preliminary Overall Impact score. Top row: matched Black (purple) and matched white (yellow) applications comparison, with overlap in orange; bottom row: all Black (purple) and random white (light green) applications comparison, with overlap in dark green.

significance, we use the recommended 0.005 p-value cutoff for "new discoveries" (Benjamin et al., 2018). p-values are frequently (but naturally) misinterpreted as the posterior probability of a hypothesis given the data (Diamond and Forrester, 1983). We follow Berger and Sellke (1987) by analyzing the posterior distribution $\mathrm{P}(\beta_R|p)$ under a reasonable class of prior distributions for β_R, which put 50% probability mass on $H_0 : \beta_R = 0$ and 50% on distributions that are unimodal and symmetric about 0. Using the stricter $p < 0.005$ standard for significance, we find that

$$\mathrm{P}(\beta_R = 0|p = 0.005) > 0.07 \tag{2.11}$$

That is, from a Bayesian standpoint, a p-value of 0.005 counter-intuitively admits at least a 0.07 (≈ 0.05) chance of the null hypothesis holding (smaller p-values than 0.005 yield smaller lower bounds than 0.07). Since the matched subset sample size is $7,471$ and the random

subset sample size is $8,595$, our analyses are amply powered, suggesting that (2.11) is a reasonably tight lower bound. In summary, the combination of strict significance standard ($p < 0.005$) and high power in this study mean that claims of significance in our study provide guarantees that approximately align with common, intuitive (mis)understandings of "$p < 0.05$."

For practical significance, we argue that a difference of 0.3 points or more in preliminary Overall Impact score for applications near the funding cutoff is substantial. For example, at the 15th percentile of sampled preliminary Overall Impact scores, increasing (decreasing) an application's final Overall Impact score by 0.3 points moves that application up to the 20th (down to the 12th) percentile. Because NIH award rates are typically between 10 and 20% for an SRG, a difference of 0.3 points in the Overall Impact score can tangibly affect funding decisions. Since standard error estimates for $\hat{\beta}_R$ are quite small—ranging from 0.017 to 0.064—our high standard for statistical significance is still much less strict than that of practical significance.

Accounting only for structural dependencies, including area of science (Model 2.4), we estimate that Black applicants' preliminary Overall Impact scores are on average 0.7 points worse than those of white applicants', a statistically and practically significant difference. This difference halves to a still-significant 0.350 points after additionally controlling for applicant- and application-level characteristics via matching and regression adjustment (Model 2.3). However, the difference becomes practically and statistically negligible when preliminary criterion scores are included as control variables in addition to the applicant- and application-level characteristics (Model 2.1).

Criterion scores alone explain much, though not all, of the variability in preliminary Overall Impact scores: their inclusion in the model reduces the residual standard deviation from about 1.3 (Models 1 and 2) to approximately 0.6 (Models 3 and 4). In contrast, controlling for application- and applicant-specific covariates—including via matching—barely changes residual error variances: the difference between Models 1 and 2 (as well as 3 vs. 4) is negligible. Controlling for criterion scores (Models 3 and 4) also substantially reduces

Table 2.3: Selected parameter estimates from models 2.1–2.4.

	Model 2.4	Model 2.3	Model 2.2	Model 2.1
Non-Structural Covariates	$\emptyset$	X	Z	X, Z
Matching?	No	Yes	No	Yes
Race Fixed Effect				
Coefficient	0.700*	0.350*	0.031	0.014
(Std. Err.)	(0.064)	(0.051)	(0.017)	(0.018)
p-value	< 0.005	< 0.005	0.071	0.431
Random Effects				
Reviewer Std. Dev.	0.490	0.500	0.274	0.286
PI Std. Dev.	0.836	0.578	0.093	0.082
SRG Std. Dev.	0.306	0.271	0.084	0.075
Residual Std. Dev.	1.312	1.284	0.567	0.562

Race coefficient estimates, their standard errors, and variance components estimates from four hierarchical linear models for preliminary Overall Impact scores. Control variables are listed in Table A.1. Coefficient estimates for control variables are not shown. Significance * is reported for $p < 0.005$.

the estimated variability explained by PI, reviewer, and SRG indicators (the standard deviation for PI random effects sees a massive 10-fold reduction). The 0.286 reviewer random intercept standard deviation estimated for Model 3 (and 4) means that—conditional on other structural covariates, matching variables, and criterion scores—a randomly chosen reviewer's assigned preliminary Overall Impact score will typically differ from average by 0.286, a number roughly equal to our threshold for practical significance. Thus, reviewer-specific scoring characteristics play an important role in determining the preliminary Overall Impact score, even once the criterion scores are fixed.

Substantial variability in preliminary Overall Impact scores still remains after accounting

for criterion scores. But, crucially, racial disparities are nullified after adjusting for criterion scores. On this basis, we argue that perceived race's effect on preliminary Overall Impact scores is largely mediated through the criterion scores. Race's relationship with the criterion scores is the topic of future work, and is being investigated experimentally via the NIH's anonymization studies (NIH Staff, 2020). Systematic racial differences also hold for preliminary criterion scores (Figure 2.4). The disparity is largest for Approach score, with a mean of 4.75 for Black applications and 4.12 for white applications ($p < 0.005$). Approach is the criterion weighed most heavily in determining the preliminary Overall Impact score in our analyses, as well as in prior research (Eblen et al., 2016).

2.4.2 Commensuration Practices in Preliminary Scores

In models that admit racial differences in commensuration practices (2.8), we control for all structural, application- and applicant-specific characteristics and estimate the key first-order commensuration coefficients—the interactions between the race indicator and the preliminary criterion scores—for the matched subset of the data. Table 2.4 contains relevant parameter estimates from the linear commensuration model (2.8); estimates for other control variables are not shown.

Using $p < 0.005$ as the standard of statistical significance (Benjamin et al., 2018), we find that the contribution of the preliminary Approach score to the preliminary Overall Impact score is higher for reviews of Black applications (the interaction coefficient is 0.041; $p < 0.005$ even after a 5-fold Bonferroni multiple testing correction) than those of white applications. This estimated weighting difference is almost as large for the Significance score, but it falls short of the $p < 0.005$ significance standard. The other three criterion scores' interaction coefficients are small, less than half that of Approach.

Differences in commensuration practices for white vs. Black applications are small overall. It is no surprise, then, that there is little evidence for commensuration disparities, since there are on average no major differences in commensuration practices that might contribute to large expected differences in preliminary Overall Impact scores between white and Black

Table 2.4: Selected parameter estimates, commensuration model (2.8).

Variable	Estimate (Std. Err.)	p-value
Fixed Effects		
Significance	0.258* (0.008)	< 0.005
Investigator	0.057* (0.011)	< 0.005
Innovation	0.129* (0.008)	< 0.005
Approach	0.598* (0.007)	< 0.005
Environment	0.022 (0.011)	0.057
PI Race = Black	-0.024 (0.047)	0.610
Significance * PI Black	-0.034 (0.013)	0.010
Investigator * PI Black	0.018 (0.017)	0.298
Innovation * PI Black	-0.020 (0.014)	0.144
Approach * PI Black	0.041* (0.012)	< 0.005
Environment * PI Black	-0.010 (0.018)	0.596
Random Effects		
Reviewer Intercepts Std. Dev.	0.286	
PI Intercepts Std. Dev.	0.079	
SRG Intercepts Std. Dev.	0.076	
Residual Variability Std. Dev.	0.562	

Preliminary criterion score, race, and commensuration coefficient estimates, and variance components estimates, for preliminary Overall Impact scores. Control variables (coefficient estimates not shown) include structural and applicant/application-specific covariates from Table A.1. Unadjusted significance * is reported for $p < 0.005$.

applications. This can be seen clearly in Figure 2.5, which shows the distribution of expected changes in preliminary Overall Impact score for all matched Black applications if those applications' applicants' perceived race were white, but all other observed variables—structural,

application- and applicant-specific, and criterion scores—were held constant, under the commensuration model (2.8). Under the conditional ignorability assumption (Section 2.3.3), this histogram represents the distribution of estimates of the effect of perceived race on preliminary Overall Impact scores, an effect which is heterogeneous because it depends on the criterion scores.

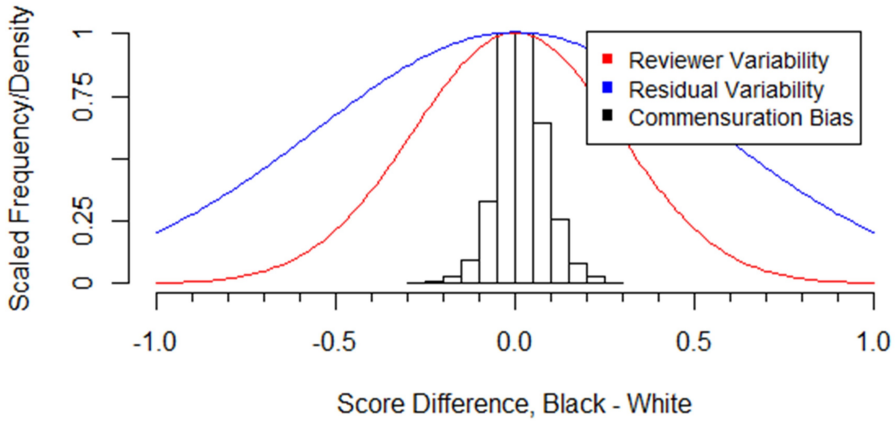


Figure 2.5: Distribution of estimated expected preliminary Overall Impact score differences due to commensuration (histogram) and distributions of reviewer intercepts and model residuals (colored lines) from the commensuration model (2.8) fit to the matched subset (Table 2.4). Histogram and densities have been scaled to have a common maximum for ease of visualizing differences in variability.

Examining this empirical distribution further, we find that for just 15% of Black applications would we expect an otherwise identical (on the observed covariates) white application to score differently by at least 0.1 points (better, 11%; worse, 4%) in the preliminary Overall Impact score. A difference of 0.1 points is still small relative to the 0.3 point standard of practical significance, which is approximately the standard deviation of the reviewer-specific random intercepts and about half the estimated residual standard deviation from the commensuration model (2.8). Under this model, 29% of all applications can expect the preliminary Overall Impact score to differ by more than 0.3 points from expectation due to reviewer variability alone, and 59% due to residual variability that is unexplained by the model. *Zero*

Black applicants in our data set would expect to see a score difference of greater than 0.3 due to commensuration practices. These remarkably small differences result partially from the small magnitude of the interaction coefficients in Table 2.4. In addition, differences in the signs of these coefficients combined with positive correlations between criterion scores result in a cancellation effect when computing expected differences due to race.

Finally, to illustrate the largest potential impacts of commensuration differences on the preliminary Overall Impact score, we consider two situational pairs, each of a hypothetical Black applicant and a hypothetical white applicant (Table 2.5). In these scenarios, the discrepancies between preliminary Approach and Significance/Innovation scores are extreme but still plausible: each combination of criterion scores does occur in our data set both for white and Black applicants.[5] In each hypothetical scenario, we assume the two applicants and applications are identical on all the observed covariates except for race and the criterion scores.

Under the "Innovative" scenario, the application review scores indicate that the proposed research is innovative and significant, but that the approach is sub-par. Based on our matched subset commensuration analysis, in such a scenario, the matched white researcher's score would be 0.12 points better than the Black researcher's score ($p < 0.005$). This difference occurs because, on average, reviewers weigh the preliminary Approach score more heavily for Black applicants than for matched whites and the preliminary Innovation and Significance scores less heavily. Conversely, under the "Thorough" scenario, in which the research proposals are scored as rigorous but not significant or innovative, our model predicts that the Black applicant will, on average, receive an impact score 0.20 points better than the white researcher ($p < 0.005$). As noted earlier, these differences are small in magnitude as compared to reviewer random effect variability or residual variability.

[5]The "Innovative" preliminary criterion score combination occurs twice in the set of reviews of Black applications and once in the set of reviews of white applications; the "Thorough" preliminary criterion score combination occurs twice in the set of white applicants and twice in the set of Black applicants.

Table 2.5: Hypothetical Commensuration Scenarios

	Innovative		Thorough	
Applicant Race	White	Black	White	Black
Significance	1	1	5	5
Investigator	2	2	2	2
Innovation	2	2	5	5
Approach	5	5	2	2
Environment	2	2	2	2

Hypothetical preliminary criterion score scenarios; "Innovative" scenario has relatively high Innovation and Significance scores and a low Approach score, and vice-versa for the "Thorough" scenario.

2.4.3 Robustness of Estimates to Unobserved Covariates

A crucial component of most observational studies that target causal effects is a strategy for plausibly satisfying Assumption 2 (conditional ignorability), such as identification of an instrumental variable (Angrist and Pischke, 2008). In this study, no such strategy is available because there is no way to account for the crucial information contained by the unobserved research proposals themselves. I instead use the "Robustness Value" approach of Cinelli and Hazlett (2019) to assess the robustness of our effect estimates to unmeasured variables U (see Figure 2.2). This section briefly introduces the Robustness Value and applies it to effect estimates of interest from Sections 2.4.

First, recall the definition of omitted variable bias. Simplifying the regression models used in this chapter somewhat for clarity, suppose the regression of interest is

$$Y = \beta_0 + \beta X + \gamma U + \epsilon$$

but we instead fit

$$Y = \beta_0' + \beta' X + \epsilon$$

since U is unobserved. Then we can derive β' as a biased version of the true β:

$$\begin{aligned}
\beta' &= \text{Cov}(X, Y)/\text{Var}(X) \\
&= \text{Cov}(X, \beta_0 + \beta X + \gamma U + \epsilon)/\text{Var}(X) \\
&= \text{Cov}(X, \beta X)/\text{Var}(X) + \text{Cov}(X, \gamma U)/\text{Var}(X) \\
&= \beta + \gamma \text{Cov}(X, U)/\text{Var}(X) \\
&= \beta + \gamma\delta.
\end{aligned}$$

The relative bias of β' for β is then

$$\frac{\gamma\delta}{\beta'}.$$

Cinelli and Hazlett (2019) derive the relative bias in terms of regression parameters. Denoting $R^2(A \sim B|C)$ the explained variance of the residuals of $A \sim C$ on B—i.e., the amount of variance in A that B explains in addition to that already explained by C—and recalling that Cohen's f is a simple transformation of R^2:

$$f = \sqrt{R^2/(1 - R^2)},$$

the first key result is

$$\left|\frac{\gamma\delta}{\beta'}\right| = \frac{|R(Y \sim U|X) f(X \sim U)|}{|f(Y \sim X)|}.$$

We now seek to understand how important U must be for the relative bias to be $100q\%$, where q parameterizes the magnitude of the relative bias. Moving from population-level regressions to regression estimates, let t be the t statistic for the coefficient of interest and df be the regression's degrees of freedom. Define $f_q = q \times df \times t$. Then the Robustness Value is

$$\text{RV}_q \equiv \frac{1}{2}\left[\sqrt{f_q^4 + 4f_q^2} - f_q^2\right].$$

RV_q has a simple but powerful interpretation: if U explains a RV_q proportion of the variability in both $Y|X$ and X—i.e. $R^2(Y \sim U|X) = R^2(X \sim U) = \text{RV}_q$—then the magnitude of the relative bias due to U may be as high as $100q\%$. Analogous results and interpretations also pertain in settings with additional regression control covariates. Note also that the Robustness Value is not explicitly a causal quantity; it is derived entirely in terms of regression statistics. Thus, while Cinelli and Hazlett (2019) focus on the common case of omitted confounders, it applies equally well when U is any other sort of covariate we must adjust for in order to estimate a causal effect unbiasedly.

Using the Robustness Value, we find that the size of the Approach-race interaction coefficient is very tenuous in light of unmeasured variables U. We find that if U explains just $RV_q = 1\%$ of the variability—that is not already explained by $\tilde{X}$—in both preliminary Overall Impact scores and the product of the Approach score and the race indicator, it could move the Approach-race interaction coefficient below the $p < 0.005$ significance threshold. We also $RV_1 = 4\%$, meaning that U explaining just 4% of the residual variability in preliminary Overall Impact scores and the Approach-race interaction would be enough to completely nullify this interaction coefficient. It is easy to imagine proposal contents being this influential, since they likely affect both the Approach and Overall Impact scores significantly even after accounting for the observed covariates.

Considering the racial disparity model estimates in Table 2.3, we find that for Model 1, $RV_1 = 11.1\%$: U explaining 11% of the residual variability (after conditioning on $\tilde{X}$) in R and Y could change the race coefficient from 0.7 to 0—or from 0.7 to 1.4, since the relative bias is unsigned. Thus, we cannot draw a strong causal conclusion from Model 1, which does not adjust for $\tilde{X}$ or Z.

Finally, for model (2.1)—in which the race coefficient estimate is very small—the question is: how strong would U need to be to mask a practically significant race effect of 0.3? Such U would need to increase $\hat{\beta}_R$ by 2043%. Still, $RV_{20.43} = 16.7\%$ for model (2.1). While it is not certain that proposal contents could explain nearly 17% of residual variability in perceived race and preliminary Overall Impact scores, it is plausible. And, as Cinelli and

Hazlett (2019), the Robustness Value provides a sufficient but not necessary condition for the size of relative bias; U explaining more than 17% of variability in preliminary Overall Impact scores, but less than 17% of variability in perceived race, could also inflate the race effect estimate to the level of practical significance.

2.4.4 Final (Post-Discussion) Overall Impact Scores

Our conclusions regarding racial disparities in final Overall Impact scores are largely the same as for preliminary Overall Impact scores: final criterion scores fully explain Black-white disparities in final Overall Impact scores (see Table A.7). However, estimated variability in reviewer random intercepts and residual variability were both considerably lower for final than for preliminary scores. This is consistent with the idea that panel discussions lead reviewers toward consensus (Fleurence et al., 2014). One cannot rigorously study commensuration differences in final scores, which also reflect SRG discussions, because commensuration is conceptualized as happening at the individual reviewer level (Lee, 2015). Of the assigned reviewers who change their Overall Impact scores after discussion, only 43% recorded respective changes in their criterion scores (see Table A.6); it is unknown why some reviewers change their criterion scores and others do not.

2.5 Discussion

The work this chapter builds on, Erosheva et al. (2020b), was broadly motivated by the persistence of a large racial funding gap: for 2014–2016 R01 applications, the overall award rate for Black applications was 55% of that for white applications (10.2% versus 18.5%). This substantial 45% funding gap is practically identical to the 45% gap found for applications from 2000–2006 (Ginther et al., 2011, 2012, 2016), though the comparison is complicated by the 2009 introduction of scored criteria and methodological differences such as use of self-reported race in this study as opposed to self-reported race in addition to supplemental information from the Association of American Medical Colleges Faculty Roster in Ginther et al. (2011, 2012, 2016). The funding gap remains despite psychological research suggesting

that using scored criteria can focus attention on merit-related factors and decrease bias in expert judgment under complex evaluative conditions (Thorngate et al., 2009; Kahneman, 2013; Kahneman and Klein, 2009). Our work is the first to examine disparities in individual reviewers' preliminary Overall Impact scores, which determine whether or not an application is discussed by an SRG and are therefore a major driver of the funding gap (Hoppe et al., 2019).

Our study finds striking racial disparities in preliminary Overall Impact scores, even after adjusting for a wide array of applicant- and application-specific variables—but not criterion scores. Adjusting solely for structural factors (Figure 2.1), we find that Black PIs on average score 0.700 points worse than white PIs on NIH's 1–9 point scale (Model 1, Table 2.3), an enormous difference given the low funding rates and competitive application process for R01 grants. This difference remains large after controlling for applicant- and application-specific covariates (0.350 points; Model 2, Table 2.3). Yet further adjustment for criterion scores reduces this large gap to a negligible 0.014 points (Model 4, Table 2.3).

Allowing criterion score weights to vary by PI's race also leads to a finding of near-zero racial disparities after controlling for criterion scores, because the weights used for Black applications are on average insubstantially different from those of white applications. While the Frequentist hypothesis testing approach does not allow confirmation of a null hypothesis, the data are much more consistent with a model in which racial disparities are null than models in which average preliminary Overall Impact scores differ by a material 0.3 points or more. These results hold despite our study efficiently and robustly adjusting for important adjustment covariates via matching, as well as being highly powered (with sample sizes of roughly $8,000$).

The observational nature of the data, combined with the complexity of NIH peer review, prevent us from couching these results in the causal language of racial bias and commensuration bias (Lee, 2015). Unobserved factors that we cannot adjust for—such as applicant bibliometrics, academic network information, and the contents of proposals—make causal interpretations inappropriate. Bibliometrics have been found to explain a substantial por-

tion of the Black/white R01 funding gap (Ginther et al., 2018). Likewise, underrepresented researchers were found to have smaller intra-institutional coauthor networks, which was associated with lower publication and citation counts (Warner et al., 2016). It is even possible that estimated racial disparities—small after adjusting for criterion scores—could be estimated as larger in magnitude if we were able to adjust for such unobserved covariates. This inherent weakness of any observational study highlights the importance of experimental approaches that can unbiasedly estimate certain types of disparities or biases—e.g. the famous "audit study" of Bertrand and Mullainathan (2003), in which white- or Black-sounding names are placed on fake resumes and submitted in response to job postings. NIH has run a small pilot study on the peer review of anonymized grant proposals (NIH Staff, 2021b), finding that anonymization decreases the Black-white score gap by a statistically significant but practically small amount. NIH is also currently undertaking a larger anonymization study (Lauer, 2020).

Additionally, missing data on demographic characteristics deserve further attention. We only considered applications from PIs who self-identified as Black or white; in addition, 15% of the applications from Black and white PIs were missing information on PI gender, ethnicity (Hispanic/Latino or not), or educational degree, and were excluded from the study (see Appendix A.2). Finally, our study focused on the individual reviewer-level preliminary scoring process and did not scrutinize other steps in NIH review, such as the assignment of final scores by SRG discussants who did not preliminarily review an application. Detailed analyses of other components of NIH peer review are required for a complete picture of the drivers of the funding gap.

In particular, more research is necessary to understand the reasons behind Black-white differences in preliminary criterion scores for NIH R01 applications. Black investigators on average receive worse preliminary scores for all five criteria—Significance, Investigator(s), Innovation, Approach, and Environment—even after matching (Figure 2.4). This finding is consistent with multiple compatible explanations: implicit racial preferences (Greenwald et al., 1998) which may be expressed more strongly when evaluators have more discretion

to interpret, apply, and prioritize criteria (Hodson et al., 2002; Norton et al., 2006, 2004; Uhlmann and Cohen, 2005, 2007; Dovidio and Gaertner, 2000); Black PIs disproportionately pursuing research in areas on which reviewers may not place a high priority (Hoppe et al., 2019); Black-white differences in research productivity or impact (Ginther et al., 2018); and/or the cumulative effect of disparities experienced over a PI's academic career that result in lower-quality grant applications, including differences in mentorship and social networks that can increase a researcher's productivity, impact, and grantsmanship (Ginther et al., 2011, 2018; Warner et al., 2016; Blau et al., 2010). Future research should evaluate the extent to which these possibilities account for racial disparities in preliminary criterion scores.

Chapter 3

REFINEMENT: MEASURING INFORMATIVENESS OF RATINGS IN THE ABSENCE OF A GOLD STANDARD

This joint work with Drs. Marina Meilă, Elena Erosheva, and Carole Lee is under revision at the *British Journal of Mathematical and Statistical Psychology* as Grant et al. (2020) and won a Joint Statistical Meetings (JSM) student paper award from the Social Statistics Section (SSS), Government Statistics Section (GSS), and Survey Research Methods Section (SRMS). The data that support the findings of this study in Section 3.4 are publicly available at https://doi.org/10.6084/m9.figshare.12728087 (Gallo, 2021). This work was partly supported by NSF grant #1759825, awarded to Drs. Elena Erosheva and Carole Lee.

3.1 Introduction

In this chapter, we are interested in how experts—peer reviewers in particular—communicate complex judgments via numerical ratings. Institutions often make decisions based on such ratings when the objectives of the decision process cannot be—or are too complex to be—mathematically formalized, when only human experts have the requisite decision-making knowledge, or when we want human value judgments to be expressed.[1] Humans typically make collective decisions via a formal system of rankings, ratings, or comparisons.[2]

[1]For example, Drury and Sinclair (1983) found that humans outperformed a machine in an industrial inspection task even though the machine was excellent at finding faults, because the machine was worse at determining the severity of the faults. The decision objective—fault severity—was difficult to formalize and the technology limited enough that human judgment was subtler and superior. Algorithms that aid judges in felony sentencing assess the risk of a defendant reoffending as well or better than humans can, yet judges routinely give younger defendants shorter sentences than recommended by algorithms "in line with a long-standing practice of treating youth as a mitigator in sentencing, due to lower perceived culpability" (Stevenson and Doleac, 2019).

[2]NIH grant proposal reviewers provide integer-scale ratings that are, after some discussion and possible revision, averaged (NIH Staff, 2012). Maine began voting by ranked choice in 2018 (Maine State Legislature

Although researchers proposed statistical prediction as a replacement for clinical assessment decades ago (Meehl, 1954; Morera and Dawes, 2006), the more recent development of black-box machine learning algorithms has dramatically accelerated the switch from human to machine decision systems (Shortliffe and Sepúlveda, 2018; Athey, 2018). Because machine decisions can be formalized mathematically, they are analytically tractable. Specifically, the objective of the decision process can often be framed as an optimization problem in which the machine attempts to minimize predictive risk, a measure of how far from the true or optimal outcome a machine prediction/decision is. In contrast, humans often make decisions in contexts without a well-defined true outcome, which we will refer to going forward as a *gold standard*.

Current popular methods for analyzing human decision-making in the absence of a gold standard make comparisons to some other point of reference. For example, inter-rater reliability evaluates the extent to which one rater's ratings are replicated by a different rater. We introduce the concept of *refinement*, an information-theoretic measure of the informativeness of a set of ratings from a single rater that makes no comparisons to a gold standard or other point of reference. Our exposition takes the context of peer review, in which human decision making is critical due to the lack of a gold standard by which to judge the predictive or external validity of peer review scoring practices (Bailar and Patterson, 1985; Feurer et al., 1994; Jayasinghe et al., 2001, 2003; Lauer and Nakamura, 2015; Lee and Moher, 2017; van Rooyen et al., 1999). Note that in peer review, prior research has shown that reliability may be a poor proxy for the normative credibility of review scores and review content (Bornmann et al., 2010; Lee et al., 2013; Hargens and Herting, 1990).

Different formal systems of human decision-making can lead to substantially different decisions (Langfeldt, 2001). The outcome of the popular vs. electoral college votes in the 2000 and 2016 U.S. presidential elections is a prominent example. Comparisons between formal systems have typically focused on differences between rating scales. Schwarz et al.

Staff, 2019). Traditional first-past-the-post voting simply aggregates comparisons (Curtice, 2009).

(1991) studied how changing global scale parameters without changing the internal structure of a scale affects raters' usage of the scale. In 1988, the National Institutes of Health (NIH) tested a move from a 1–5 decimal scale to a 1–5 scale with increments of 0.5 (Green et al., 1989); more recently, NIH tested whether adding multiples of 0.5 to a 1–9 integer scoring system changed the distribution of average scores derived later in the process (NIH Staff, 2019c). Neither study directly measured the utility of the decisions produced.

Attempts to circumvent the problem via a proxy gold standard have not found strong signals. In grant proposal peer review, a textbook example of a cardinal rating system, Li and Agha (2015) found statistically significant gains in bibliometrics/productivity accruing from better NIH grant proposal scores. But Fang et al. (2016) and Lauer et al. (2015) find that on the whole these gains are practically modest, or even negligible. Note that the use of bibliometrics as a proxy for quality of scientific research is debated (Higginson and Munafò, 2016; Wang et al., 2017; Smaldino and McElreath, 2016; Lindner et al., 2018; Lindner and Nakamura, 2015).

Before introducing refinement in the peer review context, we briefly review rating systems, which are often used in contexts like peer review. In a rating system, raters score each item on a scale—possibly multiple scales, each representing a different aspect of the item. A cardinal scale's levels have intrinsic numerical meaning via ratios or differencing (such as 0-100 essay grades), whereas Likert-type scales have a neutral central value, and the differences between adjacent levels of the scale are qualitatively identical. Shah et al. (2014) found that pairwise comparisons are faster and, when aggregated, yield a more accurate ranking of the items than ratings. However, we restrict our attention here to cardinal scales, which yield fine-grained detail about the rated items in addition to an overall ranking. For example, in grant proposal peer review, ratings allow us to determine which applications meet a standard of quality rather than simply identifying the best ones. They also facilitate providing applicants feedback that is more informative than simply their rank in a pool of anonymized applications. It is our goal to quantify the information produced by these complex systems.

3.1.1 Refinement

Refinement describes how finely a rater distinguishes between items of similar quality: do they give them all the same round score, or do they use small scale denominations to differentiate them? To what extent do the ratings imply an unambiguous ordering of the items? Refinement thus characterizes a set of scores from a single rater over multiple items, in contrast to reliability, which is a characteristic of ratings from multiple reviewers. As such, the rater is presumed to be interested in making distinctions among the items, meaning that the items must be comparable, such as grant proposals falling under the same round of review.

Refinement meets the immediate, practical need for a measure of the utility of a set of ratings in the absence of a gold standard. At NIH, "there have been concerns that [the current 1-9 integer scale], which is functionally cut in half for the 50% of applications that are considered competitive, is not sufficient to express a study section's judgment of relative merit" (Nakamura, 2019). The NIH Staff (2019c) study directly addressed this concern, but those analyses used aggregate ratings from multiple reviewers and did not consider individual reviewers' use of the scale. This study followed NIH's 2009 switch from a richer 1.0-to-5.0 single-decimal scale to the current 1-9 integer scale, a change motivated by the "compress[ed] score range" observed under the 1.0-to-5.0 scale which "effectively reduc[ed] the usefulness of scores for NIH funding decisions," as well as the difficulty of "[making] 41 reliable discriminations of application merit" (NIH Staff, 2019a).

The NIH Staff (2019c) study, which focused on panel Overall Impact scores, noted that "score compression and ties indicate that the review panel did not distinguish among the applications for impact and the lack of clear distinction among applications makes funding decisions more difficult, particularly when several applications receive identical scores and/or percentile ranks within the same study section." Thus, ambiguity of the ranking induced by the scores was a primary concern. The metrics used to measure this "score compression" were the frequency of ties at scores that were multiples of 10 $(10, 20, \ldots, 90)$, and the percentiles of various common scores in the funding cutoff range—neither of which directly measure

the extent to which the scores imply an unambiguous ranking of the applications or the quantity of information conveyed by the ratings. Refinement enables us to directly assess the usefulness of a scale via the informativeness of ratings made on it.

We will adopt the language of the American Institute of Biological Sciences (AIBS) grant proposal peer review system, in which reviewers review applications/proposals—in general, *items*—providing scores/ratings on a set of criteria as well as an overall "merit" score. More precisely, we adopt the scenario in which a reviewer has several proposals to rate on a given scale. We shall measure the refinement of a set of ratings in a way that is sensitive to the fact that some reviewers can be assigned sets of proposals more similar in quality than others. Our measure will also account for the natural tendency of raters to prefer round ratings, which is explored at greater length in the next section.

Measuring refinement, or the degree of ranking disambiguity on a scale, will depend on the scale's allowing sufficiently fine comparisons between proposals close in value. We design a refinement measure for decimal-based scales that admit multiples of 0.1 as ratings. The exposition employs the AIBS scale, which runs from 1.0 to 5.0, 1.0 being best, and admits a single decimal (Gallo et al., 2016). We denote the set of allowable scores $\mathbb{S}$, so that in this case $\mathbb{S} = \{1.0, 1.1, \ldots, 5.0\}$. Note that the refinement measures may depend on the scale used; here they will be tailored to the AIBS rating scale. We stress that refinement applies to rating systems generally, not just peer review at AIBS.

The next section lays out the refinement framework and our proposed primary measure of refinement—Entropic Refinement—with two additional metrics briefly discussed for contrast. Section 3.3 compares mathematical and statistical properties of the metrics. Section 3.4 analyzes the refinement of the scores in a data set comprised of reviews of AIBS grant applications. The final section explains how refinement fits into the study of peer review, ratings, and decision-making more generally, and suggests directions for future work.

3.2 Measuring Refinement

In this section, we focus on a set of n univariate review scores from a single reviewer, denoted $[Y_1, \ldots, Y_n] = \mathbf{Y} \in \mathbb{S}^n$ (bold uppercase denotes random vectors, while standard uppercase denotes random variables). The n scores need not be unique. Hence, technically, $\mathbf{Y}$ is a *multiset*: a collection in which the elements need not be unique and order does not matter. For simplicity, however, we will continue to call it a set, with the understanding that the multiplicities are to be considered.

When $\mathbf{Y}$ contains multiple scores with the same value, then the ordering of the respective items is not fully determined. A set of ratings $\mathbf{Y}$ is more refined when $\mathbf{Y}$ conveys more information about the relative ranking of the scored items. More specifically, the set of scores will distinguish finely between applications of similar quality, meaning there will be relatively fewer ties and the ratings will imply a near-total ordering.

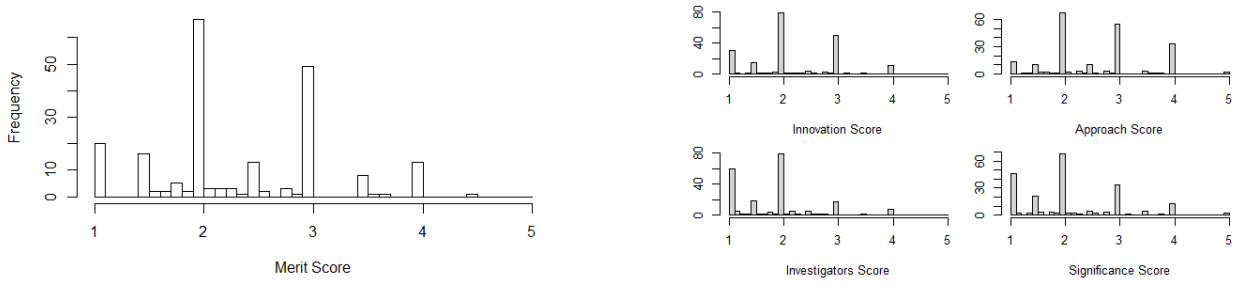
As we discuss in the next section, raters tend to provide round scores, inflating the likelihood of ties and decreasing refinement as we shall measure it. On the whole, finding ways to increase rating refinement should be useful to grant funding agencies and other institutions that use ratings to make decisions. However, we emphasize that refinement measures information, not the quality or accuracy of reviews. It may be that a peer reviewer, after careful consideration of a set of proposals, scores many of them equally. We therefore do not advocate blindly maximizing refinement; rather, refinement complements other techniques in the toolbox of ratings analysis.

3.2.1 Score Rounding

A rounding tendency has been shown to occur in multiple arenas, such as pricing (Lynn et al., 2013), price estimation (Simonsohn, 2013), age reporting (Gráda, 2006), height reporting (Bopp and Faeh, 2008), cigarette-smoking reporting (Klesges et al., 1995), and length-ratio estimation (Plug, 1977). Relatedly, *heaping* describes how survey responses are often reported with an error that rounds the response to an integer number of units, e.g. "years

married" or "income in thousands of dollars" (Bar and Lillard, 2012). Response set biases (Cunningham et al., 1977), such as extreme response bias (Erosheva et al., 2007), may also explain striking patterns in scoring such as the tendency to provide integer scores. For clarity, and since we do not wish to imply that providing a round score must involve an error, we use the term "rounding" to refer to the tendency to provide scores that are multiples of 1 or 0.5 going forward.

The aggregated AIBS review data also provide strong evidence of rounding: AIBS reviewers use integer scores much more often than scores that are multiples of 0.5, which in turn are more frequent than other scores. This pattern is evident for merit scores and especially true for criterion scores; see Figure 3.1 (Section 3.4.3 also lends support to this claim). There is thus ample evidence that AIBS reviewers gravitate towards rounder scores. Yet providing a less ambiguous comparison of the proposals requires resistance to this pull. Our refinement metric shall measure the extent to which reviewers do so.



(a) Merit score frequencies.

(b) Criterion score frequencies.

Figure 3.1: Histograms of scores from the AIBS data, all 216 reviews from all 26 reviewers. Rounder score levels (multiples of 0.5, particularly multiples of 1) are clearly preferred.

3.2.2 Entropic Refinement

We now introduce an entropy-based refinement metric that explicitly measures the extent to which reviewers resist the rounding tendency, via the decrease in entropy induced by rounding the scores. Entropy is an information-theoretic measure of how unpredictable the

data generated from a probability distribution is. For a probability mass function p on k elements defined by $p_1, \ldots, p_k$, the Shannon entropy of p is

$$H(p) = -\sum_{i=1}^{k} p_k \log p_k$$

(Cover and Thomas, 2012).[3] In our application, p is the empirical distribution on $\mathbb{S}$ induced by $\mathbf{Y}$, and we will use the shorthand $H(\mathbf{Y})$ instead for clarity. Let $R_t(y)$, $t \in \{0.5, 1\}$ be a rounding function that rounds y to the nearest multiple of 0.5 or 1, and extend it to vectors and matrices in the natural way. Define $\mathbb{S}_t \equiv R_t(\mathbb{S})$ the set of possible scores after rounding to level t; here, for example, $\mathbb{S}_1 = \{1, 2, 3, 4, 5\}$.

Entropic Refinement is our proposed refinement metric and is defined as the decrease in entropy induced by rounding:

$$r_E(\mathbf{Y}; t) \equiv H(\mathbf{Y}) - H(R_t(\mathbf{Y})). \tag{3.1}$$

Moving forward, we will assume integer rounding ($t = 1$) unless otherwise specified and will drop t from the notation unless it is needed for clarity.

Note that, because rounding is a form of quantization and quantization can only reduce entropy (Cover and Thomas, 2012), Entropic Refinement is non-negative. Entropic Refinement will tend to be higher when scores are not disproportionately round, as is observed in Figure 3.1, but rather spread evenly around round score levels (see Section 3.3.1).

Entropic Refinement aligns with the following behavioral reviewing example. Suppose that reviewers first choose a plausible score from a coarser subset of the available scores, such as the integers or multiples of 0.5. After this initial scoring, there are likely to be numerous ties between proposals' scores, which could prompt reviewers to then adjust the scores by small amounts based on more nuanced evaluation or by comparison to previously-rated proposals. The greater the extent of these adjustments, the more information the scores provide, and the larger Entropic Refinement grows. In the context of this example,

[3] We take the log base e; this is merely an arbitrary choice of scale, and base e makes certain mathematical manipulations of H simpler.

Entropic Refinement (3.1) can be interpreted as the increase in entropy induced by adjusting the initially rounded scores.

We formally analyze the properties of Entropic Refinement in Section 3.3. First, however, we briefly discuss two alternate approaches to measuring refinement which, while intuitive and appealing at first glance, will be shown to be inadequate. They are useful as bases of comparison when considering the properties of Entropic Refinement.

3.2.3 Alternative Refinement Metrics

The following alternate metrics, Fractional and Tiebreak Refinement, were each constructed to target a specific aspect of refinement: the tendency to avoid rounding and the tendency to break ties. We do not advocate for these measures because, as we illustrate in what follows, each ignores an important facet of refinement.

Fractional Refinement

Given a decimal-valued scale like that used by AIBS, we may assert that utilizing decimal values beyond just multiples of 1.0—or even 0.5—conveys more refinement.

Taking Y_i to be the ith score in $\mathbf{Y}$, let $n_{0.5}$ be the number of scores that are multiples of 0.5 but not 1 and $n_{0.1}$ be the number of scores that are not a multiple of 0.5 (or 1). That is, $n_{0.5} \equiv \sum_{i=1}^{n} \mathbf{1}[Y_i \pmod 1 \neq 0]\mathbf{1}[Y_i \pmod{0.5} = 0]$ and $n_{0.1} \equiv \sum_{i=1}^{n} \mathbf{1}[Y_i \pmod{0.5} \neq 0]$. Let $w \in (0, 1)$ be a weight parameter. Then we define the *Fractional Refinement* of $\mathbf{Y}$ to be

$$r_F(\mathbf{Y}; w) \equiv \frac{1}{n}(n_{0.1} + w n_{0.5}). \tag{3.2}$$

Thus, Fractional Refinement is a linear combination of the frequencies of the different types of scores, where rounder ratings receive less weight (integer scores receive zero weight). Fractional Refinement is a straightforward way of determining whether reviewers are utilizing all the types of levels the scale provides.

However, Fractional Refinement does not directly measure informativeness. For example, a set of scores $\mathbf{Y} = [3.2, \ldots, 3.2]$ has maximal Fractional Refinement, but tells us little about

the relative quality of the applications.

In contrast, our next secondary refinement metric—Tiebreak Refinement—directly measures the extent to which applications are ranked unambiguously.

Tiebreak Refinement

A refined set of ratings conveys small differences between applications' perceived quality and will thus contain relatively more small differences between scores than ties.[4] We think of these small differences as potential evidence that reviewers recognize when applications are of similar quality but then break rating ties in order to indicate the applications' relative ranking. This motivates the Tiebreak Refinement metric.

Let $Y_{(i)}$ be the ith order statistic of $\mathbf{Y}$, with ties broken arbitrarily and $\mathbf{D}(\mathbf{Y}) \equiv \left\{Y_{(i+1)} - Y_{(i)} : i \in [n-1]\right\}$ be the multiset of distances between consecutive scores. Then

$$z(\mathbf{Y}) = |\{x \in \mathbf{D}(\mathbf{Y}) : x = 0\}|$$
$$l(\mathbf{Y}, c) = |\{x \in \mathbf{D}(\mathbf{Y}) : 0 < x \leq c\}|$$

define the "zero" sorted distances (ties) and the "little" sorted distances for some $c < 1$ (just z and l when context is clear). We then define *Tiebreak Refinement* as the fraction of sorted distances less than c that are nonzero:

$$r_T(\mathbf{Y}; c) \equiv \frac{l}{z + l}. \tag{3.3}$$

If $z = 0$ and $l \neq 0$, then $r_T = 1$; if $l = 0$ and $z \neq 0$, then $r_T = 0$. If both l and z are zero, then we set $r_T = 1$ because all sorted distances are large and there are no ties. For $n = 1$, because there are no sorted distances, r_T is undefined.

The choice of c is application-dependent; for the AIBS scale, we recommend $c < 0.5$, and in our Section 3.4 application we choose $c = 0.2$ so that every score is at most a "little" distance from exactly one multiple of 0.5.

[4] Ties may not reflect true evaluative equality when n is not sufficiently smaller than the number of levels on the rating scale, an issue that arises for a small subset of the reviewers in our AIBS application and that we also address in Section 3.3.2.

Tiebreak Refinement is the foil of Fractional Refinement: it directly measures ties, but in no way accounts for rounding and the structure of the scale. Consider two sets of scores $\mathbf{Y} = [1, 1.5, 2, 2, 2.5]$ and $\mathbf{Y}' \equiv \mathbf{Y} + 0.1 = [1.1, 1.6, 2.1, 2.1, 2.6]$. Then $r_T(\mathbf{Y}; 0.5) = r_T(\mathbf{Y}'; 0.5) = 3/4$, but rounding may have taken place for the $\mathbf{Y}$ scores, while it certainly has not for $\mathbf{Y}'$. What *is* clear for both sets of scores is the rank order of the proposals.

3.3 Properties of Refinement Metrics

This section derives mathematical properties of Entropic Refinement and compares them to those of Fractional and Tiebreak Refinement. We also highlight how these properties should inform applications and interpretations of refinement. Properties are summarized in Table 3.1.

Table 3.1: Properties of refinement metrics.

Property (Section)	Entropic Refinement	Fractional Refinement	Tiebreak Refinement
Decomposition (3.3.1)	Basins of Attraction	Trivial	None
Range (3.3.2)	$[0, \log \max_{s \in \mathbb{S}}\{\lvert B(s)\rvert\}]$	$[0, 1]$	$\left[0, \frac{\lvert\mathbb{S}\rvert - 1}{n-1}\right]$
Large-n (3.3.3)	Normalized	Normalized	$\lim_{n\to\infty} r_T(\mathbf{Y}; c) = 0$

3.3.1 Decomposition

Rounding is a common operation on scores that also naturally induces a partitioning of the scale $\mathbb{S}$. We now demonstrate how Entropic Refinement can be decomposed in terms of this partitioning.

For every $s \in \mathbb{S}_t$, define its *basin of attraction* $B_t(s) \equiv \left\{R_t^{-1}(s)\right\}$ to be the set of scores that yield s when rounded to level t. Clearly, the sets $B_t(s)$ for $s \in \mathbb{S}_t$ partition $\mathbb{S}$. These

basins need not all be the same size: for AIBS, $|B(1)| = |\{1.0, 1.1, 1.2, 1.3, 1.4\}| = 5$, $|B(5)| = |\{4.5, 4.6, 4.7, 4.8, 4.9, 5.0\}| = 6$, and $|B(s)| = 10$ for $s \in \{2, 3, 4\}$.

Again for every $s \in \mathbb{S}_t$, let $\mathbf{Y}_{s,t} \equiv \{y \in \mathbf{Y} \cap B_t(s)\}$. Also recall that p is the empirical probability mass function on $\mathbf{Y}$, so that $p(B(s))$ is the fraction of the observed scores that lie in $B(s)$. We then have the following decomposition result, whose proof can be found in Appendix B.1:

Proposition 1. *For any score vector* $\mathbf{Y}$ *and rounding level* t*, Entropic Refinement is a weighted average of entropies over rounding basins:*

$$r_E(\mathbf{Y}; t) = \sum_{s \in \mathbb{S}_t} p(B_t(s)) H(\mathbf{Y}_{s,t}). \tag{3.4}$$

The weights $p(B_t(s))$ are the fractions of the observed scores in each basin, and the entropies $H(\mathbf{Y}_{s,t})$ represent the quantity of information conveyed by the scores within each basin. No disambiguation between scores that are in the same basin leads to zero refinement, whereas breaking ties within a basin (disambiguation) leads to increased within-basin entropy and increased Entropic Refinement. Scores that are close in that they are in the same basin interact with one another in determining Entropic Refinement, but not with scores lying in other basins.

We now briefly analyze decomposition properties for Fractional and Tiebreak Refinement. For Fractional Refinement

$$r_F(\mathbf{Y}; w) = \frac{1}{n} \sum_{i=1}^{n} r_F(Y_i; w)$$

by linearity, since r_F is simply a weighted average. Such trivial decomposability is an undesirable property, because it means that r_F fails to take into account relationships between the scores.

Tiebreak Refinement cannot be decomposed at all: hiding the value of a single Y_i makes $Y_{(j)}$ indeterminate for all j, so Tiebreak Refinement of proper subsets of the observed scores $\mathbf{Y}$ cannot fully inform us of the Tiebreak Refinement of the full set. While Tiebreak Refinement

captures dependence among the scores, it ignores rounding and does not respect the local structure of the scale, as Entropic Refinement does via basins of attraction.

3.3.2 Extrema and Range

We now turn to upper and lower bounds for each refinement metric. These results show how different scales may not be easily comparable in terms of refinement. We recommend only comparing refinement metrics derived from the same scale.

Proposition 2. *The Entropic Refinement r_E takes values between* 0 *and* $\log \max_{s \in \mathbb{S}_t} |B(s)|$*, attaining its minimum when there is only one unique observed score value in any basin of attraction and its maximum when scores are only located in the maximally sized basins of attraction, and are uniformly distributed within each such basin.*

See Appendix B.2 for proof. To see how this maximum is attained, and that it depends on the rounding level t, consider the following 3 sets of scores: $\mathbf{Y}_A = \{1.0, 1.1, \ldots, 5.0\}$; $\mathbf{Y}_B = \{1.5, 1.6, \ldots, 4.4\}$; and $\mathbf{Y}_C = \{1.3, 1.4, \ldots, 4.7\}$. $\mathbf{Y}_A$ is uniform over the entire scale, $\mathbf{Y}_B$ over the maximum-size basins for $t = 1$, and $\mathbf{Y}_C$ over the maximum-size basins for $t = 0.5$. Hence,

$$\begin{aligned} r_E(\mathbf{Y}_A; t = 1) &= \frac{5}{41}\log(5) + 3 \times \frac{10}{41}\log(10) + \frac{6}{41}\log(6) \\ &\approx 2.14 \\ &< \log(10) = r_E(\mathbf{Y}_B; t = 1). \end{aligned}$$

While $\mathbf{Y}_B$ maximizes $r_E(\mathbf{Y})$ for $t = 1$, $\mathbf{Y}_C$ does for $t = 0.5$:

$$\begin{aligned} r_E(\mathbf{Y}_C; t = 0.5)) &= 7 \times \frac{1}{7}\log(5) \\ &\approx 1.61 \\ &> 1.50 = r_E(\mathbf{Y}_B; t = 0.5). \end{aligned}$$

This property holds not only for different levels of rounding but also for different scales. Consider $2\mathbb{S}$, a 2–10 scale that admits only multiples of 0.2. The only difference between this

scale and the AIBS scale is a factor of 2, but refinement metrics calculated from each will be incomparable because, for example, $2\mathbb{S}$ contains nine integers instead of five.

Fractional Refinement achieves its minimum of zero when $\mathbf{Y}$ is integral and its maximum of 1 when $\mathbf{Y}$ does not contain multiples of 0.5.

Tiebreak Refinement achieves its minimum of zero if and only if $l = 0$ and $z > 0$, for example when only integer scores are present and there is at least one tie. It achieves its maximum of 1 whenever there are no ties, i.e., $z = 0$. However, when $n > |\mathbb{S}|$, $z > 0$ necessarily. In this case, the maximum Tiebreak Refinement for a given $n > |\mathbb{S}|$ is $\frac{|\mathbb{S}|-1}{n-1}$, which occurs whenever every level of the scale is utilized, i.e. when $\mathbb{S} \subseteq \mathbf{Y}$.

3.3.3 Large-n Behavior

Here, we consider the dependence on n of the Entropic Refinement r_E. For $n = 1$ and any $\mathbf{Y}$, $r_E(\mathbf{Y}) = 0$. For $n = 2$, r_E is zero when the two scores are identical or in different basins of attractions, and $\log(2)$ if the two scores are different but in the same basin of attraction; but $\log(2)$ is still much smaller than the maximum r_E on the AIBS scale for arbitrary n, which is $\log(10)$. It is clear that, for small values of n, the dependence of r_E on n is strong. Therefore, when sample sizes are modest, comparisons of refinement statistics must be stratified by n.

However, the behavior of r_E as $n \to \infty$ tells another story. The following analysis draws a contrast between Entropic and Fractional Refinement on one side and Tiebreak Refinement on the other. For both Fractional and Entropic Refinement, given any set of scores $\mathbf{Y} \in \mathbb{S}^n$, we can construct an infinite sequence of sets of scores $\mathbf{Y} \in \mathbb{S}^n, [\mathbf{Y}, \mathbf{Y}] \in \mathbb{S}^{2n}, [\mathbf{Y}, \mathbf{Y}, \mathbf{Y}] \in \mathbb{S}^{3n}, \ldots$ (repeat each score in $\mathbf{Y}$ once, twice, etc.) such that refinement is constant within the sequence. We call a refinement metric for which such a sequence exists for any $\mathbf{Y} \in \mathbb{S}^n$ "sample size-normalized" for large n.

This is not the case, however, for Tiebreak Refinement. It follows directly from the fact that $r_T \leq \frac{|\mathbb{S}|-1}{n-1}$ (Section 3.3.2) that $\lim_{n\to\infty} r_T = 0$, and hence such an infinite sequence with constant refinement does not exist. The differentiating factor is that Fractional and Entropic Refinement are functions only of the empirical distribution on $\mathbf{Y}$—given the empirical distri-

bution, they are independent of the sample size n and the observed scores themselves—while Tiebreak Refinement depends on the scores themselves. We argue that the informativeness of a set of ratings should only depend on its distribution and not tend to zero as more ratings are given.

3.3.4 Multivariate Extensions

In some peer review systems, such as those at AIBS and NIH, reviewers provide C criterion scores $X^1, \ldots, X^C$ in addition to the merit score Y. The criterion scores are intended to be preliminary to the merit score, as well as provide more detailed feedback to applicants on various aspects of their application. They are rated on the same scale $\mathbb{S}$ as the merit score Y, which is used to determine proposal funding. Here we present methods for assessing the refinement of multivariate scores such as criterion scores.

Let C be the number of criteria in the reviewing system at hand and assume (as is the case for AIBS) that each criterion is measured on the same scale as the merit score Y. Let $X^1, \ldots, X^C$ denote the individual criterion scores with $\mathbf{X} \in \mathbb{S}^C$ the vector of criterion scores. The most immediate way of measuring multivariate refinement is to average over the C dimensions: abusing notation slightly, we define

$$r^{avg}(\mathbf{X}) \equiv \frac{1}{C} \sum_{k \in [C]} r\left(\mathbf{X}^k\right)$$

for a given metric r. For Fractional and Tiebreak Refinement, this average is the only clear choice. However, for Entropic Refinement, we can also consider the entropy of the empirical joint p.m.f. of the criterion scores,

$$r_E^{joint}(\mathbf{X}) \equiv H(\mathbf{X}) - H(R(\mathbf{X})).$$

This extension is fundamentally different than the average over individual criteria. The next proposition clarifies the relationship between the two.

Proposition 3. *For any* $\mathbf{X} \in \mathbb{S}^{C\times n}$ *and a given* t*,*

$$\frac{1}{C}\sum_{k\in[C]} r\left(\mathbf{X}^k\right) = r_E^{avg} \leq r_E^{joint} \leq C r_E^{avg} = \sum_{k\in[C]} r\left(\mathbf{X}^k\right) \tag{3.5}$$

See Appendix B.3 for proof. The left-hand inequality becomes equality when, for example, all criterion scores are identical for each proposal. The right-hand inequality becomes equality when $X^1, \ldots, X^C$ are mutually independent with respect to their joint empirical distribution. In practice, the criterion scores are likely to be correlated to some extent, and both inequalities will be strict. Thus scaling r_E^{joint} so as to be comparable between vectors of scores with different values of C is impractical. For this reason, we use average multivariate refinement in Section 3.4's application to AIBS peer review data.

3.4 Refinement in AIBS Grant Proposal Peer Review Scores

Using the refinement metrics introduced above, we analyze the scoring behavior of reviewers at the biomedical science grant agency AIBS (American Institute of Biological Sciences). The University of Washington's Institutional Review Board confirmed that this study did not directly involve human subjects.

3.4.1 AIBS Review Data

The AIBS data set consists of review scores of 72 grant applications, all from the same round of review, reviewed by AIBS through an intramural collaborative biomedical research funding program for the biomedical sciences (Gallo, 2021). For each application, exactly three reviewers provide four criterion scores—Innovation, Approach, Investigator, and Significance—on a 1-5 scale in single-decimal (0.1) increments, where 1 is best and 5 is worst. AIBS reviewers also supply a merit score that attempts to capture the quality of the entire application, rather than just an aspect of it as the criterion scores do. Funding decisions are made largely on the basis of these merit scores.[5]

[5]The criterion scores, in addition to other factors such as the topic of the proposed research, can also play roles in these decisions.

Let $N = 26$ refer to the number of reviewers and n_i be the number of reviews performed by the ith reviewer. Table 3.2 displays summary statistics for the data set, and Figure 3.2 displays a histogram of the number of scores given by each reviewer, i.e. a histogram of $\{n_i : i \in [N]\}$.

Table 3.2: AIBS data set summary statistics.

Statistic	Value
Number of reviewers N	26
Number of applications	72
Total number of reviews $\sum_{i=1}^{26} n_i$	216 $(= 3 \times 72)$

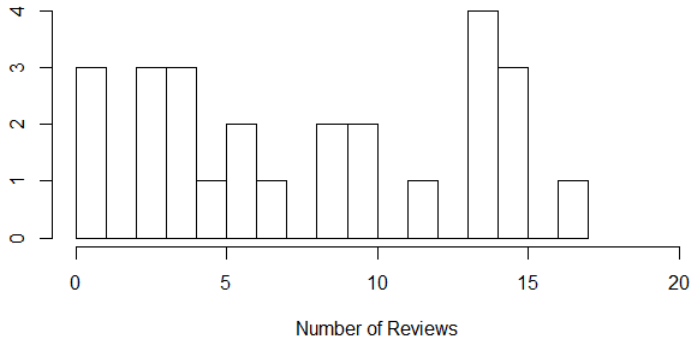


Figure 3.2: Histogram of the number of proposals reviewed by each reviewer.

3.4.2 Testing Refinement Hypotheses

We exemplify the use of refinement for the study of reviewer behavior by testing a hypothesis regarding refinement that we believed *a priori* to hold for AIBS and more broadly across peer review systems.

Hypothesis 1. ***Merit score refinement is higher than criterion score refinement:*** For AIBS as well as other funding agencies, the merit score (or equivalent) is the primary score used in funding decisions. Reviewers may therefore attempt to finely distinguish between similar-quality applications via merit scores, while allowing for more ties in criterion scores, which may instead be considered a mechanism for providing detailed feedback to applicants. If this were so, we would expect merit scores to display more refinement than criterion scores. The null hypothesis we test is that merit score refinement is less than or equal to criterion score refinement.

We utilize average criterion score refinement, r_E^{avg}, when computing Entropic Refinement as the criterion scores and merit score are of different dimensions (see Section 3.3.4). We use the paired t-test to test the null hypothesis that merit score refinement is no greater than the average criterion score refinement: $r(\mathbf{Y}) \leq r^{avg}(\mathbf{X})$. We use the Wilcoxon signed-rank test (Wilcoxon, 1946) to test the null hypothesis that $\mathrm{P}\left(r(\mathbf{Y}) > r^{avg}(\mathbf{X})\right) \leq 0.5$,[6] with the alternative hypothesis being that $\mathrm{P}\left(r(\mathbf{Y}) > r^{avg}(\mathbf{X})\right) > 0.5$. The sample size for both tests is the number of reviewers $N = 26$.

Both tests are paired, so that merit and criterion score refinement are always compared on an individual level. However, we do not stratify the tests by the number of reviews completed. While this does not harm the Type I error rates of the tests, it means that reviewers with larger n_i are weighted more heavily in the tests. We believe this is appropriate, given that these reviewers completed more reviews, but we do not claim that our testing strategy is fully efficient—developing maximally efficient tests for refinement is a new problem entirely.

Both of these tests operate under the assumption of independent observations, i.e. that the refinement of the scores from one reviewer is independent of the refinement of the scores from a different reviewer.[7] We assume that two sets of scores are independent when the

[6] If one randomly samples a reviewer and associated review scores from their respective hypothetical populations, then there is at most a 50% probability that the merit score refinement will be greater than the average criterion score refinement.

[7] We make no assumptions regarding the dependencies among the scores from a single reviewer. Since the

underlying proposals reviewed do not overlap, so that refinement of one reviewer's scores does not depend on other reviewers' scores except possibly when reviewers review the same proposal. Per AIBS, "online discussion [among reviewers] was limited and most scoring did not change [after online discussion]," so the only plausible reason for dependencies between refinement statistics is overlap in the underlying proposals reviewed. In our data, the rate of overlapping reviewer assignment is small (see Figure 3.3), so we believe the assumption of independent observations made by these tests is reasonable.

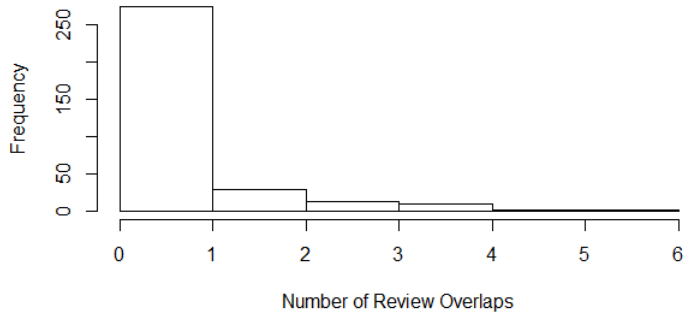


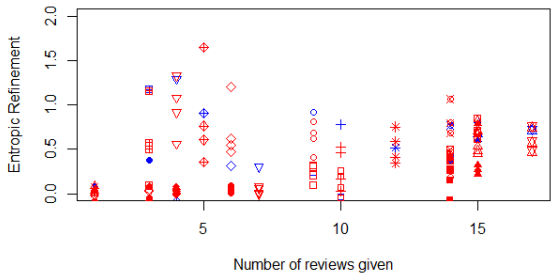
Figure 3.3: Histogram of the number of overlapping proposal assignments for each pair of reviewers. There were two pairs for which this quantity was 6, the maximum, both of which involved reviewers of 14 or 15 proposals.

3.4.3 Results

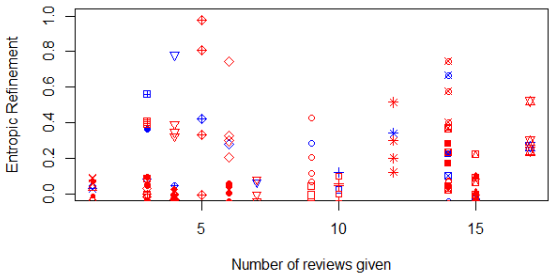
We compute Entropic Refinement for the merit and each of the four criterion scores individually, for each of the 26 reviewers, with rounding either to the nearest integer or the nearest multiple of 0.5. First, we plot r_E in Figure 3.4 for merit and the criteria.

We also compare the entirety of the empirical distributions of merit and criteria Entropic Refinement for the $N = 26$ reviewers, which are displayed in Figure 3.5. The empirical distribution of merit refinement stochastically dominates that of criteria refinement in the

unit of observation for testing this hypothesis is a refinement statistic for *all* scores from a given reviewer, these dependencies are not material to this hypothesis test.



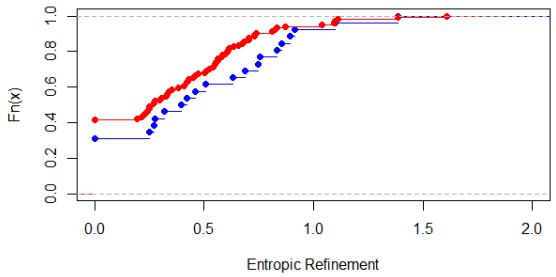
(a) Entropic Refinement with rounding to the nearest $t = 1$.



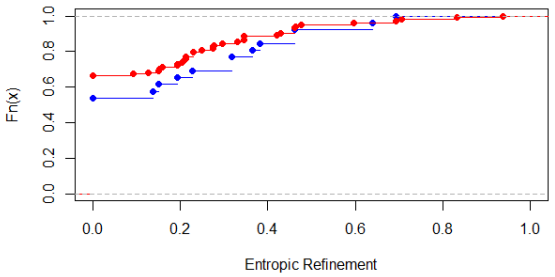
(b) Entropic Refinement with rounding to the nearest $t = 0.5$

Figure 3.4: Entropic Refinement for the AIBS reviewers, with merit score refinement in blue and criterion score refinement in red. Each symbol corresponds to a single reviewer. The x axis represents the number of reviews given by the reviewer, which one must condition on since r_E is not sample-size normalized for small n. The scores are jittered for clarity of visualization.

$t = 1$ case, and nearly does so in the $t = 0.5$ case.



(a) Entropic Refinement with rounding to the nearest $t = 1$.



(b) Entropic Refinement with rounding to the nearest $t = 0.5$.

Figure 3.5: Entropic Refinement empirical CDFs for the AIBS reviewers, with merit score refinement in blue and average criterion score refinement in red.

Table 3.3, below, illustrates the results of both hypothesis tests—paired t-test and Wilcoxon signed-rank—applied to the AIBS merit and criterion refinement. Each test was specified to be one-sided, with the alternative that merit refinement is greater than criterion refinement. For reference, tests using the Fractional and Tiebreak metrics were included as

well.

Table 3.3: Test statistics and p-values for the three types of refinement.

Refinement Metric	Parameter	t-test mean dif. (p-val)	Wilcoxon p-val
Entropic	$t = 1$	0.12 (0.004)	0.003
	$t = 0.5$	0.049 (0.04)	0.05
Fractional	$w = 1$	0.10 (0.001)	0.001
Tiebreak	$c = 0.2$	0.11 (0.01)	0.02

For all tests, the evidence suggests that merit scores display greater refinement. p-values for Fractional Refinement are significant at the 0.005 level, meaning that rounder scores are used significantly less often for the merit score than for the criterion scores on the whole. p-values for Tiebreak Refinement are only significant at the 0.05 level, suggesting that merit score ranking ambiguity is higher for the criterion scores than for the merit scores. Finally, for Entropic Refinement, p-values are significant at the 0.005 level when rounding to the nearest integer but barely significant at the 0.05 level for rounding to the nearest multiple of 0.5. There is thus strong evidence for merit scores being more informative within integer rounding basins, but weaker evidence for a difference in informativeness within half-integer rounding basins. All in all, the evidence is moderately strong that AIBS merit scores, which help determine proposal funding decisions, are more refined than criterion scores, even in our fairly small sample of reviews.

3.5 Conclusion

In this chapter, we articulated the concept of refinement as a novel way of quantifying the informativeness of human ratings that is particularly useful in the absence of a gold standard, as is the case in peer review. We introduced an information-theoretic metric for measuring

it—Entropic Refinement—and examined refinement for a set of reviews of grant proposals submitted to AIBS.

Entropic Refinement captures disambiguation through the difference in entropy before and after rounding the scores. As Proposition 1 demonstrates, this is equivalent to measuring the entropy—information content—of the scores within basins of attraction, and taking a probability-weighted average. While Entropic Refinement is more complex than the two simpler metrics we introduced, its decomposability property, sensible asymptotic behavior, and behavioral motivation make Entropic Refinement our recommended metric.

Because refinement measures the degree of disambiguation and informativeness of a reviewer's scores without comparison to some external baseline, such as a different reviewer's scores, it is distinct from inter-rater reliability metrics and particularly well suited to the no-gold-standard paradigm, as typically holds in peer review (Bailar and Patterson, 1985; Feurer et al., 1994; Jayasinghe et al., 2001, 2003; Lauer and Nakamura, 2015; Lee and Moher, 2017; van Rooyen et al., 1999).

Refinement may also provide important context in conjunction with inter-rater reliability. For example, when reliability is low, low refinement across reviewers means that there is neither consensus nor abundant information about the relative merits of the proposals. High refinement paired with low reliability, however, suggests that while they may not agree, reviewers are effectively disambiguating the proposals they rate—potential evidence of the use of a variety of evaluative perspectives. High reliability with low refinement implies the opposite, and may indicate that the scale at hand is insufficiently fine-grained for reviewers to assert their unique perspectives (this may or may not be desirable, depending on the setting). Finally, high reliability and high refinement—likely a rare outcome[8]—would imply that consensus is not merely the product of a coarse scale or heavily rounded ratings.

In psychology, ratings are often modeled as being a combination of a latent, unobserved "true response" and a measurement error (Schmidt and Hunter, 1996). Applying this con-

[8]Rounding behavior will tend to decrease refinement but increase reliability (as rounding may turn disagreements into agreements, but not the other way around).

cept to peer review, Johnson (2008) proposed a model for ratings in which NIH reviewers' errors are defined by the extent to which their ratings tend to be higher or lower than other reviewers' on average. Johnson (2008) then analyzes how NIH's funding decisions would differ if they were to adjust for these measurement errors. Other approaches use multilevel regression modeling to account for differences in reviewers' average scores but do not explicitly characterize these differences as arising from measurement error (Erosheva et al., 2020b; Jayasinghe et al., 2003).

In this vein, we can assess the refinement of estimated latent ("true") scores rather than the observed scores. We can even do so without explicitly estimating the latent scores: if we instead have an error distribution p_ϵ, we can solve the deconvolution $p_L \oplus p_\epsilon = p$ for p_L, the distribution of the latent scores (per Efron and Hastie (2016), this may be difficult). Entropic Refinement can then be computed for the distribution p_L. While the addition of independent noise increases entropy, Entropic Refinement is a difference of entropies, so latent score refinement may be higher or lower than that of the observed scores.

In our exposition of refinement and our application in Section 3.4, we use the observed scores and do not adjust for measurement error. Our approach aligns with the standard current practice of using unadjusted peer review scores.

With a small sample size of $N = 26$, we found moderate support for the hypothesis that AIBS merit scores—which are the only score used to make final funding decisions—are on average more refined than criterion scores. With a larger data set, more complex hypotheses about peer review informativeness could be tested with sufficient power. Consider the following hypothesis that could not be tested with currently available data:

Hypothesis 2. ***Reviewers who review applications they perceive as competitive display more refinement:*** When a reviewer believes an application's quality puts it near the funding cutoff, they may elect to expend the extra effort to distinguish that application from potential competitors by fine-tuning its scores.[9] This would manifest itself in higher

[9]If this fine-tuning does not accurately reflect a reviewer's evaluation but rather stems from a desire to

entropy in rounding basins near a (perceived) funding cutoff.

Threshold-based incentives spur improved performance in other arenas, e.g. ultramarathoning (Grant, 2016); we hypothesize that increased perceived likelihood of determining an application's funding similarly incentivizes reviewers to provide more refined ratings. According to an AIBS representative, there is no formal or informal "funding cutoff" known to reviewers, so data from a different funding institution and a survey of reviewers regarding their beliefs about a funding cutoff would be needed to test this hypothesis.

One limitation of Entropic Refinement is that it is specialized to decimal scales $\mathbb{S}$ such as the one used by AIBS. Our analysis reveals that such scales—in contrast to, for example, integer scales—provide raters with the ability to first conceptualize a round rating and then further refine. Nevertheless, extensions of refinement to other popular scale types are needed. The authors are currently investigating refinement for integer scales, such as the $\{1, \ldots, 9\}$ scale used by NIH. With additional data, future analyses could assess whether refinement is greater for competitive-seeming applications, or could track reviewers over time to assess whether or not their scores' refinement increases as they gain experience. These types of studies will help illuminate the intricacies of ratings and human decision-making.

---

influence the proposal's funding outcome, it can be considered *gaming* (Coveney et al., 2017). Gaming is considered by some panelists to be unacceptable (Lamont, 2009).

Chapter 4

BAYESIAN CAUSAL DISCOVERY WITH BIVARIATE ADDITIVE MODELS

This work with collaborators Elena Erosheva, Thomas Richardson, and Marina Meilă, was partly supported by NSF grant #1759825, awarded to Drs. Elena Erosheva and Carole Lee.

4.1 Introduction

Causal discovery—the identification of causal relationships among the variables of interest from observational data—is a critical prerequisite to the estimation of causal parameters. While experiments/interventions obviate the need for causal discovery in some contexts, in others they can be impossible or impractical. For example, it is particularly difficult to identify causality in fMRI data (Smith et al., 2011) as interventions in the brain are impractical, expensive, and potentially unethical. In climate science, causal discovery has been used to determine causal relationships between atmospheric variability in different regions (Ebert-Uphoff and Deng, 2012); computational challenges in spaciotemporal dynamical systems can make such large-scale causal relationships difficult to infer from known, smaller-scale causal structure.

Yet causal discovery is not in wide practical use. One roadblock to its adoption is that many discovery algorithms do not reveal all the causal relationships among the observed variables, but rather a set of causal structures that are all compatible with the data (Spirtes et al., 1993). These collections of plausible structures can be large, though, and scientists may require more precision in order to proceed with model specification and estimation. A second roadblock is statistical uncertainty: while some approaches quantify uncertainty in

the causal structure—e.g. Cooper and Herskovits (1992); Friedman et al. (1999); Komatsu et al. (2010)—these are the exception to the rule. Confidence in causal inference-based scientific results is based on the degree of certainty that the causal structure underlying the inference is correct.

The bivariate case is fundamental to causal discovery, and is particularly important to those interested in learning exact causal structures since classical techniques that rely solely on conditional independence and the faithfulness assumption to produce a Markov Equivalence Class (Andersson et al., 1997) of possible structures cannot be used in the bivariate setting. In this chapter, we develop bivariate Bayesian Causal Discovery, which naturally quantifies uncertainty about the causal relationship between the two variables under identifying assumptions. We formulate Bayesian versions of two foundational discovery techniques, Additive Noise Models (Hoyer et al., 2009) and Linear non-Gaussian Additive Models (Shimizu et al., 2006). Finally, we illustrate Bayesian Causal Discovery's efficacy in terms of prediction and uncertainty quantification on simulated and real data.

4.1.1 Bivariate Causal Discovery

Let $X \in \mathbb{R}$ and $Y \in \mathbb{R}$ be jointly distributed according to $P_{X,Y}$. We observe bivariate, non-interventional data $\{X, Y\}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, assumed i.i.d. A set of *a priori* conditions must be met to identify the causal relationship between X and Y. First, we assume that X and Y are not independent, which can be tested in great generality using, e.g., the Hilbert Space Independence Criterion of Gretton et al. (2008). Second, we assume that it is not the case that both X causes Y and Y causes X (acyclicity). The scientific context will often justify this assumption, as causal discovery in the cyclic setting often requires interventional data (Hyttinen et al., 2012). These two assumptions are common to bivariate causal discovery approaches. Finally, one often assumes that X and Y are not confounded, as in Hyvärinen and Smith (2013) and Hoyer et al. (2009)—though in some cases this assumption can be relaxed, complicating the causal discovery problem (Hoyer et al., 2006; Shimizu and Bollen, 2014; Chen and Chan, 2013). We assume unconfoundedness via

the independence of the cause variable and error term. Modeling unobserved latent variables by allowing for dependence between the cause and error is a challenge beyond our scope; doing so would introduce additional parameters, degrading the reliability of the computations described in Section 4.2.2 (see also Appendix C.3).

We are left with two possible causal models: M_X—in which X causes Y—and, conversely, M_Y. Under M_X, we assume that $X \sim P_X \in \mathbb{P}_X$, $\epsilon \sim P_\epsilon \in \mathbb{P}_\epsilon$ (with $\mathrm{E}[\epsilon] = 0$) and that Y is generated as

$$Y = f(X) + \epsilon.$$

We refer to X as the "cause," $f \in \mathbb{F}$ as the "mechanism," and ϵ as the "noise." Similarly, under M_Y, we assume that $Y \sim P_Y \in \mathbb{P}_Y$, $g \in \mathbb{G}$, $\eta \sim P_\eta \in \mathbb{P}_\eta$ with $\mathrm{E}[\eta] = 0$, and $X = g(Y) + \eta$. As the mathematical developments for M_Y will mirror those of M_X exactly, we will henceforth assume X causes Y and leave M_Y aside unless it is strictly needed.

4.1.2 Additive Noise Models and Linear Non-Gaussian Additive Models

Additive Noise Models (ANMs) and Linear Non-Gaussian Additive Models (LiNGAMs), the two causal discovery methods we consider, leverage nonlinearity of f and non-Gaussianity of ϵ, respectively, to distinguish between M_X and M_Y.[1] They adapt straightforwardly to the Bayesian framework because the nonlinearity and non-Gaussianity assumptions can be expressed via prior distributions on the mechanism f and noise ϵ, as we shall see in Section 4.2. While we focus on ANMs and LiNGAMs in the main text of this chapter, Bayesian Causal Discovery can be applied more widely: see Appendix C.1 for an application to the Information-Geometric Causal Inference technique of Janzing et al. (2012), and Spirtes and Zhang (2016) for a broader overview of causal discovery.

Hoyer et al. (2009) introduce ANMs, showing that when M_X holds with nonlinear f and $\epsilon \perp\!\!\!\perp X$, no model M_Y can also generate data from $P_{X,Y}$ (except in highly contrived scenarios; see Zhang and Hyvärinen (2009) for an accounting of these). Thus, assuming

[1] Rothenhäusler et al. (2016) unite these two identification strategies under a single framework.

that such an ANM obtains in the true causal direction, one can estimate ANMs in both directions and test for the independence of the residuals from the independent variable in both models. Hoyer et al. (2009) compare the p-values from these two independence tests, concluding that the model with the larger p-value (more plausible independence) is the true causal model. However, these p-values do not have any direct interpretation in terms of the relative plausibility of M_X and M_Y, and Bayesian interpretation of such p-values is not straightforward (Casella and Berger, 1987).

LiNGAMs (Shimizu et al., 2006) identify bivariate causal structure using non-Gaussianity instead of nonlinearity. LiNGAMs guarantee consistent discovery of the causal relationship between X and Y provided that f is linear and ϵ is non-Gaussian. LiNGAM was originally estimated using Independent Components Analysis (ICA) (Hyvarinen, 1999), but faster and more practical methods have since been developed (Shimizu et al., 2011; Hyvärinen and Smith, 2013). A related approach (Wang and Drton, 2018) extends LiNGAM to the case where there are more variables than samples (in which the ICA problem cannot be solved). There have also been investigations of LiNGAMs in the presence of latent variables, relaxing the unconfoundedness assumption (Hoyer et al., 2006; Shimizu and Bollen, 2014; Chen and Chan, 2013).

4.1.3 Contributions of this Chapter

Bivariate causal discovery algorithms have been derived under a variety of frameworks: Frequentist likelihood-based (Hyvärinen and Smith, 2013), black-box prediction (Lopez-Paz et al., 2015), and ad-hoc comparisons of relevant statistics (Janzing et al., 2012; Hoyer et al., 2009). Yet most do not quantify uncertainty in the causal discovery process. This multiplicity of frameworks and lack of uncertainty estimation are major barriers to the use of causal discovery by applied scientists. Our proposed framework, Bayesian Causal Discovery, is a fully Bayesian technique that identifies a causal direction from observational data via a Bayes Decision Rule. Bayesian model selection is a perfect fit to the causal discovery problem: Gelman et al. (2013) note that "Bayes Factors can work well when the underlying

model [class] is truly discrete and [when] it makes sense to consider one or the other model as being a good description of the data."

In the next section, we lay out the general framework of additive bivariate Bayesian Causal Discovery. In Section 4.2.3 we formulate Bayesian ANMs using a Gaussian Process model for the mechanism and show how to efficiently compute them, while Section 4.2.4 briefly discusses Bayesian LiNGAMs. Section 4.3 shows how Bayesian Causal Discovery quantifies the strength of the evidence for a causal model, and contrasts Bayesian ANMs and LiNGAMs with classical ANMs and LiNGAMs via an application to simulated data that present various difficulties for causal discovery algorithms. Section 4.4 tests the predictive performance of Bayesian ANMs and LiNGAMs on the collection of real-world bivariate data sets. We finish with a discussion of outstanding problems and future directions.

4.2 Bivariate Bayesian Causal Discovery

Our fully Bayesian approach treats the causal model itself as random: nature chooses M_X with probability p_{M_X} and M_Y with $p_{M_Y} = 1 - p_{M_X}$. Under M_X, (P_X, f, P_ϵ) are randomly generated from a "mother distribution" $\mathcal{P}_{P_X,f,P_\epsilon}$ which is a joint distribution over two distributions and a function (Lopez-Paz et al., 2015). We denote the marginal prior on P_X as $\mathcal{P}_{P_X}$, the marginal prior on f as $\mathcal{P}_f$, and the marginal prior on P_ϵ as $\mathcal{P}_{P_\epsilon}$; P_X, f, and P_ϵ need not be independent. Figure 4.1 illustrates how the randomness of the model (M_X or M_Y) and the model priors $\mathcal{P}_{P_X,f,P_\epsilon}$ and $\mathcal{P}_{P_Y,g,P_\eta}$ hierarchically define a distribution on (X, Y).

To make posterior inferences about whether M_X or M_Y generated the data, we begin by placing prior probabilities on each: $p_{M_X} = \mathrm{P}(M_X)$ and $p_{M_Y} = \mathrm{P}(M_Y)$ with $p_{M_X} + p_{M_Y} = 1$. We choose a causal model by comparing the posterior probabilities of M_X and M_Y given observed data $\{X, Y\}_n$ via the ratio

$$\frac{\mathrm{P}(M_X|\{X,Y\}_n)}{\mathrm{P}(M_Y|\{X,Y\}_n)}. \tag{4.1}$$

(4.1) can be rewritten in terms of the Bayes Factor K_n, which is the ratio of the marginal

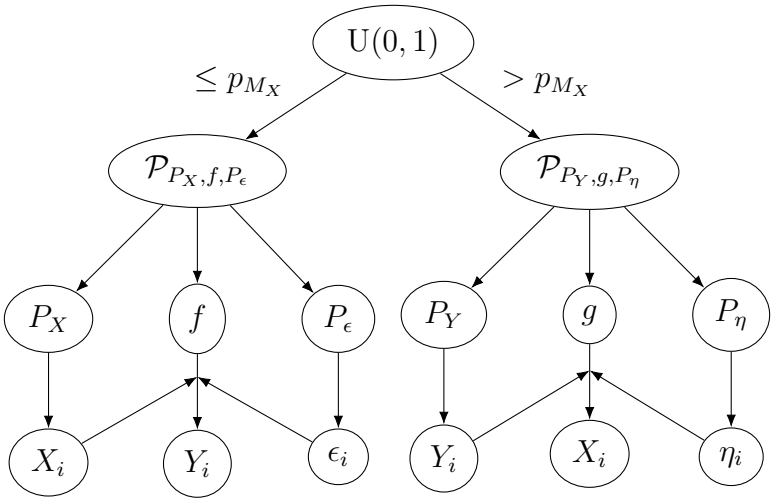


Figure 4.1: Hierarchical model for bivariate Bayesian Causal Discovery.

likelihoods of the data under each model:

$$\frac{\mathrm{P}(M_X|\{X,Y\}_n)}{\mathrm{P}(M_Y|\{X,Y\}_n)} = \frac{\mathrm{P}(\{X,Y\}_n|M_X)}{\mathrm{P}(\{X,Y\}_n|M_Y)}\frac{p_{M_X}}{p_{M_Y}} \equiv K_n\frac{p_{M_X}}{p_{M_Y}}. \tag{4.2}$$

See Kass and Raftery (1995), Section 3.2, for a brief overview of the interpretation of Bayes Factors.

In the common scenario that we have no *a priori* reason to favor one causal direction over the other, $p_{M_X} = p_{M_Y} = 0.5$ and the Bayes Factor is equivalent to the posterior probability ratio. We shall take $p_{M_X} = p_{M_Y}$ going forward, which allows us to choose a causal model on the basis of K_n (just K when n is fixed and known). M_X (M_Y) is more likely when $\log(K)$ is positive (negative) with $|\log(K)|$ capturing the relative evidence for the favored model over the other. Note that $|\log(K_n)|$ will tend to grow with n, which is appropriate since more data ought to lend greater support for or against a model.

We make a decision d regarding the causal direction via

$$\begin{aligned} d = M_X \text{ if } \log(K) > 0 \\ d = M_Y \text{ if } \log(K) < 0. \end{aligned} \tag{4.3}$$

Because our Bayesian model assigns probability to each causal direction, we can formulate causal model selection in terms of Bayesian decision theory (Cyert and DeGroot, 1987; Harsanyi, 1978; Minton et al., 1961), in which the optimal model choice depends on the posterior model probabilities and the cost of choosing $d = M_X$ when M_Y holds and vice-versa. Given $p_{M_X} = p_{M_Y}$, (4.3) is the optimal Bayes Decision Rule when these costs are equal (see Proposition 5 of Appendix C.3). In our Section 4.4 real-data application, we extend this decision framework to admit a third option, M_0: indecision, which is useful when the evidence for M_X or M_Y is not decisive enough to be scientifically useful. Appendix C.2 analyzes decisions under this three-model setup.

We compute K via its numerator and denominator separately. The numerator of K is $\mathrm{P}(\{X,Y\}_n|M_X)$, known generally as the marginal likelihood of $\{X,Y\}_n$ under model M_X; in the context of Bayes Factors, it is known as the *evidence* for model M_X because it conveys the relative amount of support, or evidence, the data provide for M_X vs. M_Y. Let us assume that under each of M_X and M_Y the model prior distribution admits a probability density with respect to some dominating measure μ, which we will suppress in the notation unless it is needed. For example, if the model prior is parameterized in terms of some $\theta \in \mathbb{R}^d$, then the joint distribution on θ, P_θ, is dominated by the Lebesgue measure. We compute the evidence by marginalizing over the model parameters P_X, f, ϵ:

$$\begin{aligned}\mathrm{P}(\{X,Y\}_n|M_X) &= \int_{\mathbb{P}_X,\mathbb{F},\mathbb{P}_\epsilon} p(\{X,Y\}_n|P_X,f,P_\epsilon)\, d\mathcal{P}_{P_X,f,P_\epsilon} \\ &= \int_{\mathbb{P}_X,\mathbb{F},\mathbb{P}_\epsilon} \prod_{i=1}^{n} \left[p(Y_i|X_i,f,P_\epsilon)p(X_i|P_X)\right]\, d\mathcal{P}_{P_X,f,P_\epsilon} \end{aligned} \tag{4.4}$$

In all but the simplest cases, analytical calculation of the evidence (4.4) is impractical, and it must be approximated computationally. Algorithm 1 details a Monte Carlo algorithm for approximating the evidence.

If we could correctly specify p_{M_X}, $\mathcal{P}_{P_X,f,P_\epsilon}$, and $\mathcal{P}_{P_Y,g,P_\eta}$ (see Figure 4.1) and were able to calculate the evidence under those mother distributions exactly, we could compute the true posterior probability ratio (4.1) exactly. In reality, our model priors must be specified

Algorithm 1 `MonteCarloEvidence`

procedure MonteCarloEvidence($\{X, Y\}_n, \mathcal{P}_{P_X,f,P_\epsilon}, \mathcal{P}_{P_Y,g,P_\eta}, B$)
  for $i \in [B]$ **do**
    Draw $P_X \sim \mathcal{P}_{P_X}$
    Draw $f \sim \mathcal{P}_{f|P_X}$
    Draw $P_\epsilon \sim \mathcal{P}_{P_\epsilon|f,P_X}$
    Compute $\mathrm{P}(\{Y\}_n|\{X\}_n, f, P_\epsilon)$
    Compute $\mathrm{P}(\{X\}_n|P_X)$
    $L_i \leftarrow \mathrm{P}(\{Y\}_n|\{X\}_n, f, P_\epsilon) \times \mathrm{P}(\{X\}_n|P_X)$
  end for
  return $\frac{1}{B}\sum_{i=1}^{B} L_i$
end procedure

without knowing the true data-generating mechanism and we can usually only approximate the evidence. Both of these steps introduce the potential for error, the subject of the next two sections.

4.2.1 Model Prior Specification

In selecting model priors $\mathcal{P}_{P_X,f,P_\epsilon}$ and $\mathcal{P}_{P_Y,g,P_\eta}$, we incorporate *causal identification assumptions*, which facilitate choosing a causal model.

Definition 1 (Causal Identification Assumption)**.** *A causal identification assumption is one that, when expressed via the model priors $\mathcal{P}_{P_X,f,P_\epsilon}$ and $\mathcal{P}_{P_Y,g,P_\eta}$, yields*

$$E_{M_X}\left[\log\left(\frac{P(X,Y|M_X)}{P(X,Y|M_Y)}\right)\right] > 0 \tag{4.5}$$

and

$$E_{M_Y}\left[\log\left(\frac{P(X,Y|M_X)}{P(X,Y|M_Y)}\right)\right] < 0. \tag{4.6}$$

In general, the inequalities (4.5) and (4.6) hold weakly/inclusively, since Kullback-Leibler divergences are non-negative. By employing a causal identification assumption, we exclude model priors that lead to equal evidence for M_X and M_Y, on average over a chosen model.

Specification of model priors is challenging because, as Kass and Raftery (1995) note, prior selection impacts Bayes Factors substantially more than standard Bayesian posterior inferences. For posterior inference, sufficiently large n will typically render a reasonable prior irrelevant, but for Bayes Factors this is not so. The influence of the model priors motivates two more principles for prior specification that prevent factors unrelated to causality from impacting causal discovery: *marginal agnosticism* and *mechanism symmetry*.

First, we assume that the marginal locations and scales of the variables are not informative of the causal direction in the sense of the following definition.

Definition 2 (Marginal Agnosticism)**.** *$\mathcal{P}_{P_X,f,P_\epsilon}$ and $\mathcal{P}_{P_Y,g,P_\eta}$ are jointly marginal cause-agnostic if*

$$P(\{X\}_n|M_X) = \int_{\mathbb{P}_X} P(\{X\}_n|P_X)\, d\mathcal{P}_{P_X} = \int_{\mathbb{P}_Y} P(\{Y\}_n|P_Y)\, d\mathcal{P}_{P_Y} = P(\{Y\}_n|M_Y)$$

and marginal effect-agnostic if

$$E[Y|M_X] = E[X|M_Y] = 0.$$

In this chapter, we only use model priors that satisfy marginal cause- and effect-agnosticism.[2] Identification strategies that exploit differences in variables' marginal distributions do exist, but require burdensome assumptions such as equality of error variances—e.g. Chen et al. (2019); Peters and Bühlmann (2014)—that may be true in specialized settings but are not true in general as the measured variables may be on different scales.

When both model priors are marginal cause-agnostic, the terms involving P_X and P_Y in the Bayes Factor (4.2) cancel, so that the conditional distribution of the effect given the cause and the joint prior distribution of the mechanism and error term determine the Bayes

[2]Other authors (Hoyer and Hyttinen, 2009; Shimizu and Bollen, 2014) also model all exogenous variables with the same distribution, though they do not require the marginal likelihoods of the exogenous variables to be equivalent. Shimizu and Bollen (2014) take an empirical Bayes approach and estimate differing hyperparameters for the distributions of exogenous variables.

Factor. Assuming marginal cause-agnosticism and using the simplifications from (4.4) yields

$$K = \frac{\int_{\mathbb{P}_X,\mathbb{F},\mathbb{P}_\epsilon} p(\{Y\}_n|\{X\}_n, f, P_\epsilon)p(\{X\}_n|P_X)\, d\mathcal{P}_{P_X,f,P_\epsilon}}{\int_{\mathbb{P}_Y,\mathbb{G},\mathbb{P}_\eta} p(\{X\}_n|\{Y\}_n, g, P_\eta)p(\{Y\}_n|P_Y)\, d\mathcal{P}_{P_X,f,P_\epsilon}} \tag{4.7}$$

$$= \frac{\int_{\mathbb{P}_X,\mathbb{F},\mathbb{P}_\epsilon} p(\{Y\}_n|\{X\}_n, f, P_\epsilon)\, d\mathcal{P}_{f,P_\epsilon|P_X}}{\int_{\mathbb{P}_Y,\mathbb{G},\mathbb{P}_\eta} p(\{X\}_n|\{Y\}_n, g, P_\eta)\, d\mathcal{P}_{g,P_\eta|P_Y}}. \tag{4.8}$$

One straightforward way to achieve marginal cause-agnosticism is to center and scale each variable by subtracting empirical means and dividing by empirical standard deviations, and then set $\mathcal{P}_{P_X}$ and $\mathcal{P}_{P_Y}$ to be $\mathrm{N}(0,1)$ with probability 1. The same approach works with other distributions and scalings. For example, centering by $\bar{X}$, scaling by $\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \bar{X}\right|$, and specifying $P_X = P_Y = \mathrm{Laplace}(0,1)$ also yields marginal cause-agnosticism (we use this approach with LiNGAM in Section 4.3.2). These strategies also help facilitate marginal effect-agnosticism: it is easy to verify that $\mathcal{P}_{P_X,f,P_\epsilon}$ is marginal effect-agnostic when $\mathcal{P}_{P_X}$ ensures that P_X is symmetric about zero with probability 1, when $P_X \perp\!\!\!\perp f$,[3] and when $\mathcal{P}_f$ is specified such that $\mathcal{P}_f(f) = \mathcal{P}_f(-f)$.

Next, we require that $\mathcal{P}_f$ and $\mathcal{P}_g$ be identical and symmetric about the identity function for any invertible functions in their support, a property we refer to as *mechanism symmetry*.

Definition 3 (Mechanism Symmetry)**.** *$\mathcal{P}_{P_X,f,P_\epsilon}$ and $\mathcal{P}_{P_Y,g,P_\eta}$ obey mechanism symmetry if, for any invertible function $h \in \mathbb{F}$,*

$$\mathcal{P}_f(h) = \mathcal{P}_g(h) = \mathcal{P}_f(h^{-1}) = \mathcal{P}_g(h^{-1}).$$

For example, asymmetric model priors over f and g that only admit strictly convex functions would be a poor choice, because causal discovery would then proceed on the basis of how well convex mechanisms fit the data in each direction, a consideration which is unrelated to causality. All model priors we use satisfy mechanism symmetry.

Finally, note that for identical model priors under M_X and M_Y that satisfy marginal agnosticism and mechanism symmetry, joint Gaussianity is not a causal identification assumption.

---

[3] See Appendix C.1 for an example of a model prior in which f and P_X are dependent.

Proposition 4. *Suppose that* $\mathcal{P}_{P_X,f,P_\epsilon} = \mathcal{P}_{P_Y,g,P_\eta}$*, that* $\{X,Y\}_n$ *is centered and scaled to have sample mean zero and sample variance one, that* $P_X, P_Y, P_\epsilon, P_\eta$ *are Gaussian with probability one, that* f, g *are linear with probability one, and that* $\mathcal{P}_F$ *and* $\mathcal{P}_g$ *satisfy mechanism symmetry. Then* $E_{M_X}[\log(K)] = E_{M_Y}[\log(K)] = 0$.

$\{X,Y\}_n$ are bivariate Normal, so the likelihood is determined solely by five sufficient statistics: the empirical means, standard deviations, and correlation. Due to centering and scaling, the means and standard deviations are fixed under both models, and the sample correlation is invariant to swapping X and Y so its distribution is identical under M_X and M_Y. Proposition 4 then follows because the model priors are identical and therefore assign the same probability to any given sample correlation.

4.2.2 Evidence Approximation

We now turn to computation of the evidence (4.4). Monte Carlo approximation (Algorithm 1) has numerous advantages: it is straightforward to implement, unbiased, consistent, and converges at a $\sqrt{n}$ rate.[4] However, its accuracy can be suspect: Kass and Raftery (1995) note that "when sample sizes are moderate or large, the integrand becomes highly peaked around its maximum," leading to a phenomenon known as *pseudo-bias* (Lenk, 2009). In Appendix C.3, we analyze the accuracy of Monte Carlo approximation of the evidence (Algorithm 1) under a Gaussian likelihood with a Gaussian prior for the mean, a setting in which the marginal likelihood can be computed analytically. We find that, in fact, the complexity of the model prior determines the difficulty of approximating the evidence accurately to a greater extent than the sample size does. Pseudo-bias most often occurs because all Monte Carlo samples (P_X, f, P_ϵ) satisfy

$$\mathrm{P}(\{X,Y\}_n | P_X, f, P_\epsilon) < \mathrm{P}(\{X,Y\}_n | M_X),$$

[4]The CLT applies when, for example, the covariance of the likelihood (with the data $\{X,Y\}_n$ fixed and the stochasticity coming from a parametric model prior) is finite. The LLN applies under slightly weaker conditions. See Giné and Nickl (2016) for analogous results on infinite-dimensional model priors.

leading to an underestimate of the evidence. In our analysis, increasing the number of parameters that must be marginalized beyond three virtually guarantees pseudo-bias, while for a fixed number of parameters the likelihood of pseudo-bias grows relatively slowly in n.

Numerous alternatives to Algorithm 1 for approximating marginal likelihoods exist—see the surveys of Llorente et al. (2021); Evans and Swartz (1995)—but many have drawbacks of implementation complexity. For example, the MCMC approach of Carlin and Chib (1995) requires intensive hyperparameter tuning even for relatively simple models. Others have poor convergence properties: see Wolpert and Schmidler (2012)'s analysis of the importance sampling-based harmonic mean estimator. We tested the Carlin and Chib (1995) approach but found that results were highly sensitive to the MCMC specification. Importance sampling approaches did not appear to converge as quickly as Algorithm 1. The most modern algorithm we tried—the `bridgesampling` R package (Gronau et al., 2020), which is designed for approximating marginal likelihoods—performed worse in both run time and approximation accuracy than Monte Carlo approximation (Algorithm 1) on bivariate Gaussian test data that were known to have Bayes Factor approximately 1 (see Appendix C.4).

In this chapter, we aim for accurate and consistent evidence approximations first by ensuring through analytical means that the parameter space has dimension at most 3 so as to avoid pseudo-bias. Second, we use numerical integration or the basic Monte Carlo approach of Algorithm 1 to perform the marginalization (4.4) as these standard approaches are less error-prone in implementation than more complex techniques, and appeared most stable in testing.

4.2.3 Bayesian Additive Noise Models

We now develop Bayesian Additive Noise Models (ANMs, Hoyer et al. (2009)), detailing the model prior formulation and evidence approximation. Our model specification is similar to that of Stegle et al. (2010), with two important differences. First, Stegle et al. (2010) explicitly model the cause variable with a Gaussian mixture model, which does not satisfy marginal cause-agnosticism (Definition 2), in contrast to our approach: centering-and-scaling

with a $\mathrm{N}(0, 1)$ prior. Second, these authors use an approximate maximum a posteriori (MAP) for the mechanism parameters instead of integrating over them to compute the evidence. Our fully Bayesian procedure directly approximates the evidence, capturing all uncertainty expressed in the model prior.

Model Specification and Computation for Bayesian ANMs

The key assumptions employed in Frequentist ANMs are nonlinearity of the mechanism f and independence of the error term ϵ and the cause variable X. Regarding error independence, our Bayesian framework immediately yields

$$X \perp\!\!\!\perp \epsilon | P_X, P_\epsilon. \tag{4.9}$$

Note that this conditional independence would not necessarily hold if we generalized the error distribution to $P_{\epsilon|X}$. Equation (4.9) implies that only the location of $\mathrm{P}(Y|X)$ depends on X, an identification assumption also employed by Mitrovic et al. (2018). For Bayesian ANMs, we additionally model the distribution of the cause as independent of the mechanism and noise distribution; that is, $\mathcal{P}_{P_X,f,P_\epsilon} = \mathcal{P}_{P_X}\mathcal{P}_{f,\epsilon}$.[5] This ensures that $X \perp\!\!\!\perp \epsilon$ marginally in addition to (4.9).

Like Hoyer et al. (2009), we employ Gaussian Processes (Rasmussen and Williams, 2006) as the nonlinear mechanism for the ANM: under M_X, the mechanism f will be a realization of a Gaussian Process (GP):

$$\mathrm{Cov}(f(s), f(t)) = k_{h,l}(s, t) = l^2 \exp(-h(s - t)^2).$$

That is, the process f has a Gaussian covariance kernel of bandwidth $1/h$ and scale factor l. The Gaussian kernel is a standard choice, e.g. it is the default kernel for Scikit-learn's Gaussian Process Regressor (Pedregosa et al., 2011). Taking the i, j-th entry of the covariance

[5] Schölkopf et al. (2012) and Janzing and Schölkopf (2010) employ this assumption without being explicitly Bayesian).

matrix $\Sigma_{h,l}$ to be $k_{h,l}(X_i, X_j)$, we then have

$$\mathrm{P}(f) \propto \frac{\exp\left(-\frac{1}{2}(f(\{X\}_n) - \mathrm{E}[f(\{X\}_n)])^T \Sigma_{h,l}^{-1}(f(\{X\}_n) - \mathrm{E}[f(\{X\}_n)])\right)}{|\Sigma_{h,l}|^{1/2}}. \tag{4.10}$$

To be marginal effect-agnostic (Definition 2), we take $\mathrm{E}[f] = 0$ so that $E[Y|M_X] = 0 = \bar{Y}_n$. While some nonzero mean functions also satisfy this condition, we wish to make no unwarranted assumptions about the relationship between X and Y. Equation (4.10) then reduces to

$$\mathrm{P}(f) \propto \frac{\exp\left(-\frac{1}{2} f(\{X\}_n)^T \Sigma_{h,l}^{-1} f(\{X\}_n)\right)}{|\Sigma_{h,l}|^{1/2}}.$$

To complete the model, we take $\epsilon \sim \mathrm{N}(0, \sigma_Y^2)$. While other types of noise models are possible, a Gaussian noise model provides major computational benefits when coupled with a Gaussian Process, as we shall illustrate below.

Next, we specify the model prior. For $\mathcal{P}_f$, we take $h \sim |\mathrm{N}(0, \sigma_h^2 = 1)|$ and $l \sim |\mathrm{N}(0, \sigma_l^2 = 1)|$. These priors encourage smoother process realizations through larger bandwidths $1/h$ and smaller l, so that the prior exerts a regularizing influence. We employ Scott's Rule (Scott, 2015) with a constant multiplier, taking $\sigma_h = \frac{1}{4} n^{1/5}$. Note that $\mathrm{E}[h] = \sqrt{2/\pi}\sigma_h$ and $\mathrm{Var}(h) = (1 - \frac{2}{\pi})\sigma_h^2$, so the location and scale of the distribution of h are both proportional to $n^{1/5}$ as desired. For $\mathcal{P}_{P_\epsilon}$, we take $\sigma_Y \sim \mathrm{U}(0, 1)$: because $\{Y\}_n$ has unit sample variance after rescaling, the residual standard deviation vary between 0 and 1 depending on how informative X is about Y.

Finally, we derive the evidence for Bayesian ANMs. Given that $f(\{X\}_n) \in \mathbb{R}^n$, the parameter space for this model is $(n+3)$-dimensional (including h, l, and σ_Y), posing potential problems for computationally approximating the evidence (see Section 4.2.2). However, the combination of GP prior for f and Gaussian distribution for ϵ allows us to marginalize analytically over the n GP realization parameters (Rasmussen and Williams, 2006). Denote $\phi(x|\mu, \Sigma)$ the density at $x \in \mathbb{R}^n$ of a multivariate Normal distribution with mean μ and

covariance Σ. Then the marginalization over f is

$$\int_{\mathbb{F}} p(\{Y\}_n|\{X\}_n, f, P_\epsilon)p(f|h,l)\,df = p(\{Y\}_n|\{X\}_n, h, l, \sigma_Y^2) = \phi(\{Y\}_n|0, \Sigma_{h,l} + \sigma_Y^2\mathrm{I}). \tag{4.11}$$

We can then compute the numerator of (4.8) as

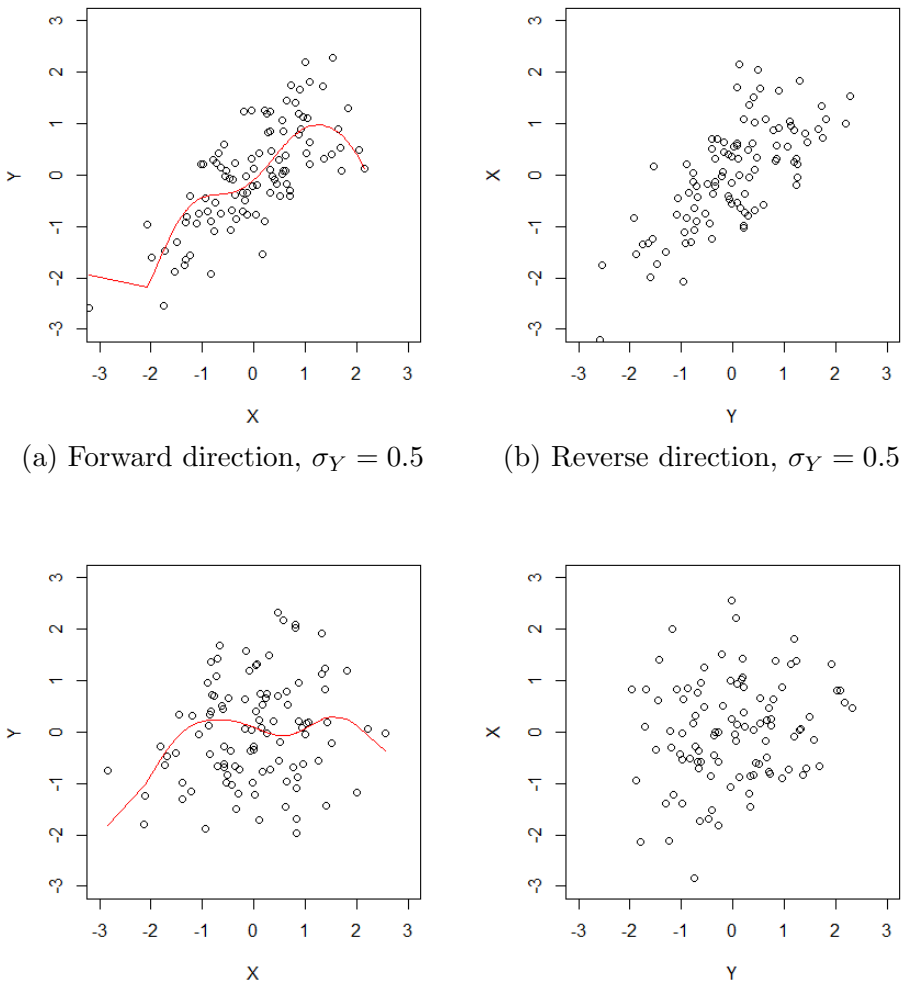
$$\begin{aligned}
&\int_{\mathbb{F},\mathbb{P}_\epsilon} p(\{Y\}_n|\{X\}_n, f, P_\epsilon)\,d\mathcal{P}_{f,P_\epsilon|P_X} \\
&= \int_{f\in\mathbb{R}^n, h,l,\sigma_Y\in\mathbb{R}^+} p(\{Y\}_n|\{X\}_n, f, P_\epsilon)p(f|h,l)p(h,l,\sigma_Y)\,df\,dh\,dl\,d\sigma_Y \\
&= \int_{h,l,\sigma_Y\in\mathbb{R}^+} \phi(\{Y\}_n|0, \Sigma_{h,l} + \sigma_Y^2\mathrm{I})p(h,l,\sigma_Y)\,dh\,dl\,d\sigma_Y,
\end{aligned}$$

leaving us to integrate over just h, l, and σ_Y—few enough parameters for a reliable evidence approximation.

Simulation Study: Noise vs. Identifiability in Bayesian ANMs

We conclude this section by investigating the relationship between noise and ease of causal discovery under the Bayesian Gaussian Process ANM. We simulate data from a Gaussian Process with additive Gaussian noise of varying standard deviation. For expositional simplicity, the simulation model parameters are a fixed sample from the mother distribution—only the realizations of the Gaussian Process and the data vary. We take $X_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_X = 0, \sigma_X = 1)$, the kernel bandwidth to be $\frac{1}{6}n^{1/5} \approx \text{median}(|\mathrm{N}(0, \left[\frac{1}{4}n^{1/5}\right]^2)|) = \text{median}(h)$, the kernel scale $l = 1$, and the error standard deviation σ_Y to be 0.5 in the "low-noise" case and 1 in the "high-noise" case. Figure 4.2 plots 100 simulated data points in the low-noise regime and high-noise regime both in the forward and reverse causal direction. The sampled GP realization f is plotted with linear interpolation in red, in the forward direction only. In neither case can one determine the causal direction by eye, as the noise is large enough relative to the scale of the data to obscure the fact that the GP realization is non-invertible in both cases.

We replicate the data generation and evidence approximation calculations for models M_X and M_Y 100 times for $n = 100$, reporting the fraction of those 100 repetitions for which

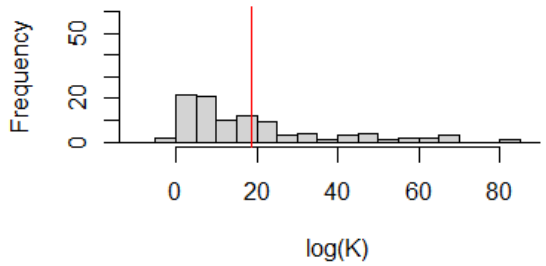


(a) Forward direction, $\sigma_Y = 0.5$ (b) Reverse direction, $\sigma_Y = 0.5$

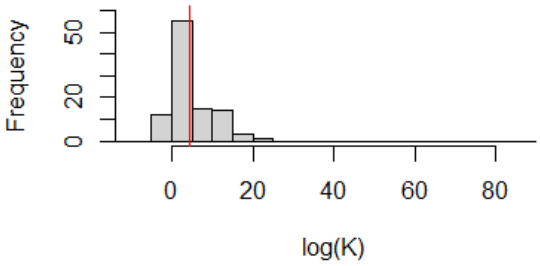
(c) Forward direction, $\sigma_Y = 1$ (d) Reverse direction, $\sigma_Y = 1$

Figure 4.2: Example data ($n = 100$), low-noise ($\sigma_Y = 0.5$) and high-noise ($\sigma_Y = 1$) regimes, forward and reverse causal directions. Sampled GP mechanism f is plotted in red in the forward direction only. $\{X\}_n$ and $\{Y\}_n$ are both centered and scaled.

the correct model M_X was chosen using the decision rule (4.3). Results are presented in Figure 4.3 as histograms of the log Bayes Factor over the 100 replications for the low- and high-noise cases. We take 10,000 Monte Carlo samples from the model prior over h, l, and σ_Y, which we argue is adequate from the discussion in Section 4.2.2 and the cumulative evidence plots in Figure 4.4.

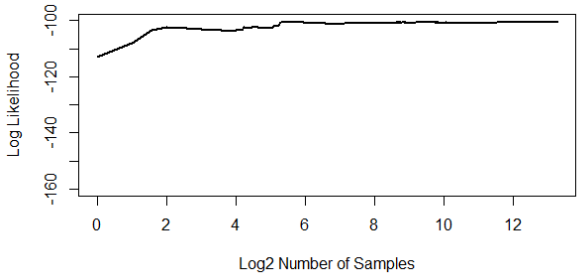


(a) Low noise ($\sigma = 0.5$); 98% discovery accuracy.

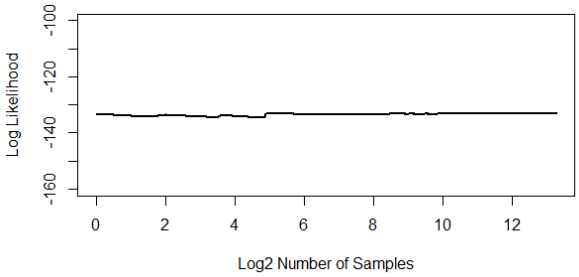


(b) High noise ($\sigma = 1$); 88% discovery accuracy.

Figure 4.3: Histograms of log Bayes Factors over 100 replications using Monte Carlo evidence approximation (Algorithm 1) with 10,000 samples from the prior for the Gaussian Process model with $n = 100$ observations. Red vertical lines mark average log Bayes Factors.



(a) Cumulative evidence for M_X.



(b) Cumulative evidence for M_Y.

Figure 4.4: Cumulative evidence approximation for a single simulation under each causal model over $B = 10,000$ Monte Carlo samples from the model prior in the low-noise regime.

In both the high- and low-noise regimes, greater evidence is found for M_X than M_Y ($K > 1$) the vast majority of the time—though less often in the high-noise regime, as expected. As expected, the magnitude of the Bayes Factor tends to be notably higher in the low-noise regime than in the high-noise regime, conveying more certainty in the true direction of causality. Section 4.3 demonstrates that high accuracy and meaningful Bayes Factors also occur in more challenging settings when the model is misspecified and the true mechanism is invertible.

Finally, for comparison with the identifiable GP data-generating models, we simulated bivariate Gaussian data from the linear model $Y = X + \mathrm{N}(0, \sigma_Y)$ in which $P_X = \mathrm{N}(0, 1)$ and $\sigma_Y \sim \mathrm{U}(0, 1)$. Proposition 4 shows that this model is causally unidentifiable under our assumptions. Reassuringly, Bayesian ANMs using the decision rule (4.3) picked the correct causal direction in 43 out of 100 simulations ($p = 0.2$ for the null hypothesis that the decision rule (4.3) picks the correct direction 50% of the time).

4.2.4 Bayesian LiNGAMs

In this section, we discuss Bayesian Linear non-Gaussian Additive Models, which have also been explored in Shimizu and Bollen (2014) and Hoyer and Hyttinen (2009). We assume $Y = \beta X + \epsilon$, where $\epsilon \sim \mathrm{Laplace}(0, \sigma_Y)$—a standard and effective error model for non-Gaussian modeling (Hyvärinen et al., 2004; Hyvärinen and Smith, 2013). To be marginal cause-agnostic, we center and scale $\{X\}_n$ and $\{Y\}_n$ to have zero sample mean and unit sample mean absolute deviation from the mean ($\sum_{i=1}^n |X_i - \bar{X}|$), and define $\mathcal{P}_{P_X}$ and $\mathcal{P}_{P_Y}$ to be $\mathrm{Laplace}(0, 1)$ so that

$$\mathrm{P}(\{X\}_n | M_X) = \mathrm{P}(\{Y\}_n | M_Y) = \frac{1}{2^n} \exp(-n).$$

Our Bayesian LiNGAM is similar to the `BayesLiNGAM` model of Hoyer and Hyttinen (2009), though we use a simpler error model that keeps the parameter space small and obeys marginal cause-agnosticism (Definition 2). Our approach is also similar to that of Shimizu and Bollen (2014), with the primary difference being that we do not model latent variables or the accompanying parameter coefficients and hyperparameters. While both of these existing approaches are fully Bayesian, they do not analyze Bayes Factors or their relationship to causal identifiability.

As for ANMs, we take $\sigma_Y \sim \mathrm{U}(0, 1)$ so that the model prior is minimally informative about the strength of the association between X and fY. Even if $\beta \perp\!\!\!\perp \sigma_Y$ before rescaling, $|\beta|$ is inversely correlated with σ_Y after rescaling the data. In Appendix C.5, we find that when centering and scaling by the true population mean and mean absolute deviation from

the mean ($\mathrm{E}|Y - \mathrm{E}[Y]|$),

$$\begin{aligned} |\beta| &= \frac{1}{2}\left(\sqrt{-3\sigma_Y^2 + 2\sigma_Y + 1} - \sigma_Y + 1\right) \\ &\equiv q_\beta(\sigma_Y). \end{aligned} \tag{4.12}$$

However, this relationship can in practice be noisy because we must standardize the data by sample quantities. We use a slack parameter s to account for this noise. Taking ξ to be an independent Rademacher sample (1 or -1, each with probability 1/2), we specify $\mathcal{P}_{\beta|\sigma_Y}$ as

$$\beta|\sigma_Y \sim \xi \cdot \mathrm{U}(\max\{0, q_\beta(\sigma_Y) - s\}, \min\{1, q_\beta(\sigma_Y) + s\}) \tag{4.13}$$

with $s = 0.3$. Simulation accuracies and log Bayes Factors are robust to the choice of s.

4.3 Strength of Evidence in Bayesian Causal Discovery

This section demonstrates, via simulation studies, that the magnitude of Bayes Factors predicts the accuracy of Bayesian Causal Discovery under the decision rule (4.3). We also show that Bayes Factor magnitudes are determined by the signal-to-noise ratio and by the strength of nonlinearity (for ANMs) or non-Gaussianity (for LiNGAMs) in the data-generating model. We find that Bayes Factors are informative even when the model prior is misspecified and the true joint distribution is close to a causally unidentifiable bivariate Gaussian distribution (see Proposition 4). We also compare the Bayesian models' discovery accuracy—the rate at which they make correct decisions under decision rule (4.3)—to that of their Frequentist counterparts, finding that the Bayesian formulations are more accurate than the Frequentist ones in these scenarios.

Bayesian Causal Discovery is partly motivated by the fact that scientists need confidence in the correctness of the DAG underlying a causal analysis and that most non-Bayesian discovery algorithms provide no interpretable estimate of predictive uncertainty for a given single data set. This section shows that Bayes Factors, which are computed for a single sample $\{X, Y\}_n$, provide such an uncertainty estimate.

4.3.1 Degree of Nonlinearity in ANMs

For Bayesian ANMs, we investigate how the degree of nonlinearity of a data-generating mechanism influences Bayesian Causal Discovery accuracy and log Bayes Factor magnitude. We employ the same model prior and evidence approximation procedure as in Section 4.2.3. We estimate Frequentist ANMs (Hoyer et al., 2009) by fitting a Gaussian Process regression using Scikit-learn's implementation of Gaussian Process regression (Pedregosa et al., 2011), which maximizes the marginal likelihood to choose parameter values (and thus is not fully Bayesian). For testing independence of the residuals and the cause variable, we use the Matlab implementation of the Hilbert Space Independence Criterion test provided by the authors of Gretton et al. (2008). p-values from both the bootstrap and Gamma-approximated null distribution—the two approaches to statistical testing provided in Gretton et al. (2008)—were nearly always numerically zero and could not be used. However, because the quantities in equations (2) and (5) of Gretton et al. (2008) are symmetric in X, Y and completely determine the null distribution of the test statistic, the test statistics have the same null distribution and could be compared directly. We choose the model (M_X or M_Y) with the smaller test statistic, which indicates more plausible independence between the cause variable and the residuals.

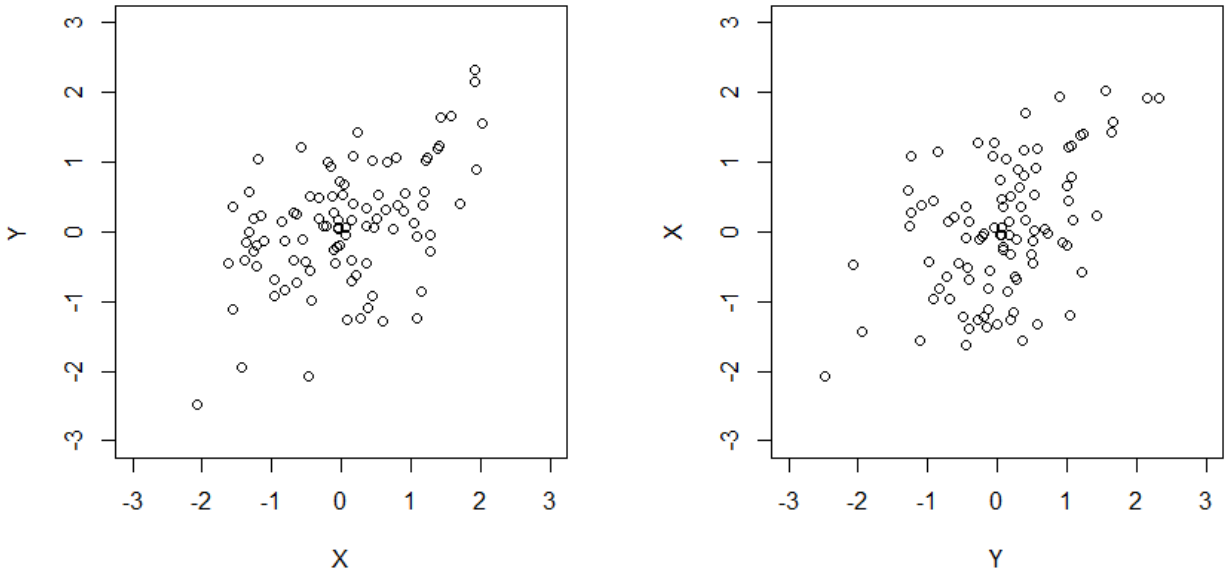
To generate varying degrees of nonlinearity and noise in the simulated data, we let $X_i \sim \mathrm{N}(0, 1)$ for $i \in \{1, \ldots, n\}$. Denoting $m_X(\gamma)$ and $s_X(\gamma)$ the empirical mean and standard deviation of $\mathrm{sign}(X_i)|X|^\gamma$, we then set

$$Y_i = \left(\mathrm{sign}(X)|X|^\gamma - m_X(\gamma)\right)/s_X(\gamma) + \epsilon,$$

with $\epsilon \sim \mathrm{N}(0, \sigma_Y^2)$. $m_X(\gamma)$ and $s_X(\gamma)$ allow us to more precisely control the signal-to-noise ratio in the data and standardize across varying values of γ. Observations are i.i.d.

We vary the simulation parameters as $\gamma \in \{1/3, 1/2, 1, 2, 3\}$ and $\sigma_Y \in \{0.5, 1, 2\}$. When $\gamma = 1$, the model is linear with Gaussian errors and is therefore unidentifiable under the ANM. The larger $|\log \gamma|$ is, the more nonlinear the model. Stronger nonlinearity is offset by larger noise variance. Figure 4.5 displays a simulated data set from this model with $\gamma = 3$ and

$\sigma_Y = 1$. With a monotonic mechanism and moderate noise, it is impossible to distinguish by eye from Figure 4.5 which direction satisfies the independent noise assumption.



(a) Simulated data in the forward causal direction.

(b) Simulated data in the reverse causal direction.

Figure 4.5: Examples of simulated data from the power law model ($\gamma = 3$, $\sigma_Y = 1$) on which we test Bayesian and Frequentist ANMs. $\{X\}_n$ and $\{Y\}_n$ are both centered and scaled.

Table 4.1 displays the discovery accuracies for Frequentist ANM, while Table 4.2 displays the proportion of correct discoveries for the Bayesian ANM with the average log Bayes Factor in parentheses. In our color palette, warmer (cooler) colors represent higher (lower) discovery accuracy, to help the reader easily identify the associations between discovery accuracy and the simulation parameters. 100 experimental replications were used for both algorithms.

Comparing Tables 4.1 and 4.2, Bayesian ANMs yield uniformly better accuracy than Frequentist ANMs under this data-generating model. As expected, for the linear ($\gamma = 1$) model, accuracies and log Bayes Factors are not significantly different from 50% and 0. When the mechanism is only modestly nonlinear ($\gamma \in \{1/2, 2\}$), low noise yields high accuracy, but high noise reduces it to scarcely better than guessing. For highly nonlinear mechanisms

Table 4.1: Frequentist ANM simulation results.

$\gamma \backslash \sigma_Y$	0.5	1	2
1/3	0.94	0.67	0.59
1/2	0.89	0.60	0.55
1	0.48	0.61	0.60
2	0.95	0.70	0.62
3	0.88	0.74	0.59

Discovery accuracies for Frequentist ANM over 100 replications and $n = 100$. Warmer (cooler) shades indicate higher (lower) accuracy. The degree of mechanism nonlinearity is represented by $|\log(\gamma)|$, while σ_Y is the exogenous noise level in the outcome.

Table 4.2: Bayesian ANM simulations results.

$\gamma \backslash \sigma_Y$	0.5	1	2
1/3	1.0 (12.8)	0.80 (2.6)	0.67 (0.8)
1/2	1.0 (7.7)	0.72 (1.2)	0.70 (0.7)
1	0.52 (0.1)	0.58 (0.2)	0.51 (0.0)
2	1.0 (15.5)	0.95 (3.7)	0.48 (0.3)
3	1.0 (28.9)	1.0 (8.6)	0.65 (0.7)

Discovery accuracies (average log Bayes Factor) for Bayesian ANM over 100 replications and $n = 100$. Warmer (cooler) shades indicate higher (lower) accuracy. The degree of mechanism nonlinearity is represented by $|\log(\gamma)|$, while σ_Y is the exogenous noise level in the outcome.

($\gamma \in \{1/3, 3\}$), higher noise also yields lower accuracy, though the signal is not completely lost at $\sigma_Y = 2$.

While the error distribution of our Bayesian ANMs is correctly specified here, the Gaussian process mechanism is much more complex than the true power-law mechanism, which

one might expect to yield poor model fit at the small sample size of $n = 100$. Nevertheless, Bayes Factors increase with the degree of nonlinearity and decrease with σ_Y, and are monotonically related to discovery accuracy. Note that there is no deterministic function linking Bayes Factors and discovery accuracy, even when the model prior is correct; very different distributions of K can correspond to the same $\mathrm{P}(K > 0)$. García-Donato and Chen (2005) propose a "calibrating value" for decisions made using Bayes factors that equalizes error rates, but it is not a Bayes decision rule.

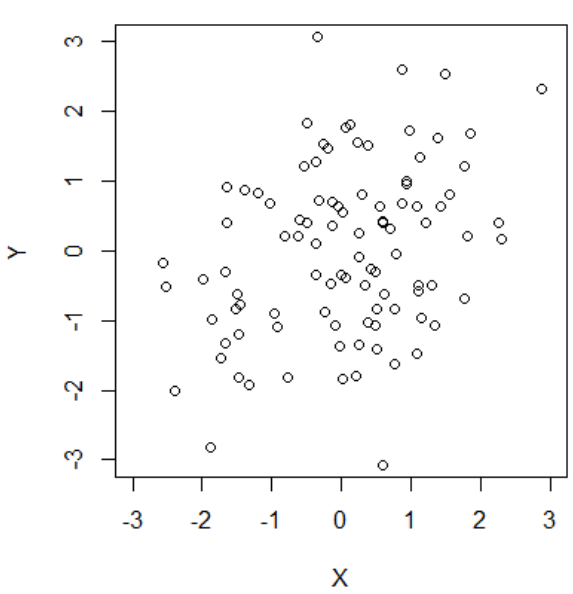
4.3.2 Degree of Non-Gaussianity in LiNGAMs

Next, we analyze Bayesian Causal Discovery accuracy and Bayes Factors for simulated data with a linear mechanism and error distributions that range from "far from" to "close to" non-identifiability (multivariate Gaussianity), with a comparison to the classical Frequentist LiNGAMs of Shimizu et al. (2006).

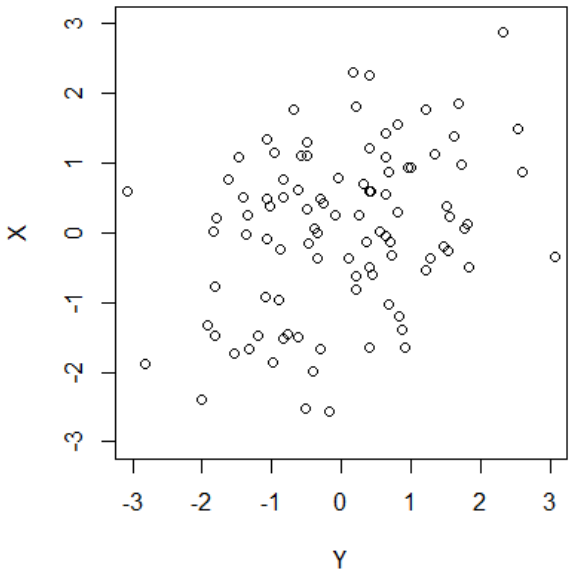
We simulate from the linear model $Y = \beta X + \epsilon$, so that non-Gaussianity of X and/or ϵ identifies causality. We generate both X and ϵ from the t distribution with df degrees of freedom and excess kurtosis $6/(df - 4)$ for $df > 4$. For the error term, we scale the t distribution, dividing by its expected mean absolute deviation from the mean for the specified df and multiplying by σ_Y so that the expected mean absolute deviation from the mean is σ_Y.

In our simulations, we take $df \in \{5, 6, 10, 20, 50\}$ so that the excess kurtosis of the residuals ranges over $\{6, 3, 1, 0.38, 0.13\}$; as the t-distribution's df increases, the joint distribution of (X, Y) approaches Normality and the model approaches causal non-identifiability under LiNGAM. Due to the inverse relationship between $|\beta|$ and σ_Y that scaling the data induces (see Appendix C.5), we fix $\beta = 0.5$ and let σ_Y vary in $\{1, 2, 4\}$. Figure 4.6 shows simulated data from this model with $df = 6$ and $\sigma_Y = 1$. By eye, it is unclear which direction satisfies the LiNGAM assumptions.

We then approximate the Bayes Factor using Monte Carlo approximation with 10,000 samples from the model prior employed in Section 4.2.4, which uses a Laplace error distribu-



(a) Simulated data in the true causal direction.



(b) Simulated data in the reverse causal direction.

Figure 4.6: Examples of simulated data from the linear non-Gaussian model ($df = 6$, $\sigma_Y = 1$) on which we test Bayesian and Frequentist LiNGAMs. $\{X\}_n$ and $\{Y\}_n$ are both centered and scaled.

tion and inversely correlates the linear regression coefficient β and the error variance σ_Y in the model prior. For Frequentist LiNGAM, we use the default implementation of the `lingam` function from the `pcalg` package (Kalisch et al., 2019). In the bivariate case, this implementation outputs an estimate of the true causal linear model coefficients, corresponding to M_X, M_Y, or $X \perp\!\!\!\perp Y$. Because we assume that X and Y have been pre-screened for independence, we report Frequentist LiNGAM accuracy as the percentage of correct selections of M_X plus half the percentage of $X \perp\!\!\!\perp Y$ selections.

Table 4.3 displays Frequentist LiNGAM accuracies over 100 simulated data sets of size $n = 100$ for all combinations of the parameters σ_Y and df, while Table 4.4 does the same for Bayesian LiNGAM but additionally includes average log Bayes Factors. $n = 100$ was chosen because it facilitates clear illustration of how non-Gaussianity and noise in the data

affect causal discovery. Larger (smaller) n lead to more (less) decisive Bayes Factors, and the impact of σ_Y is less clear for sample sizes substantially different from $n = 100$.

Table 4.3: Frequentist LiNGAM simulation results.

$df \backslash \sigma$	0.5	1	2
5	0.81	0.76	0.54
6	0.76	0.71	0.55
10	0.67	0.63	0.53
20	0.49	0.49	0.53
50	0.49	0.56	0.52

Discovery accuracies for Frequentist LiNGAM over 100 replications and $n = 100$. Warmer (cooler) shades indicate higher (lower) accuracy. Larger df yield lower excess kurtosis and more closely approximate Gaussian noise, while σ_Y is the exogenous noise level in the outcome.

Table 4.4: Bayesian LiNGAM simulation results.

$df \backslash \sigma_Y$	0.5	1	2
5	0.92 (6.2)	0.85 (3.3)	0.71 (1.1)
6	0.91 (5.5)	0.83 (3.2)	0.73 (1.0)
10	0.83 (3.8)	0.76 (2.9)	0.67 (0.9)
20	0.80 (3.5)	0.72 (2.3)	0.61 (1.0)
50	0.77 (3.2)	0.69 (1.6)	0.61 (1.1)

Discovery accuracies (average log Bayes Factors) for Bayesian super-Gaussian LiNGAM over 100 replications and $n = 100$. Warmer (cooler) shades indicate higher (lower) accuracy. Larger df yield lower excess kurtosis and more closely approximate Gaussian noise, while σ_Y is the exogenous noise level in the outcome.

Table 4.4 illustrates that the accuracy of Bayesian LiNGAM (Bayesian and Frequentist) decreases as the error distribution approaches Normality or the noise level increases. For Bayesian LiNGAM, the strength of the evidence decreases commensurately as well—and Bayesian LiNGAM is generally more accurate than Frequentist LiNGAM for this simulation setup.

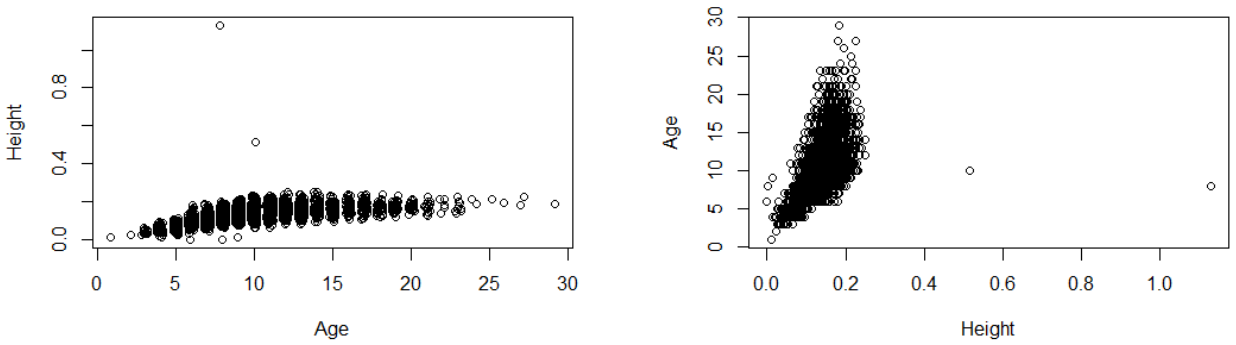
In contrast to the ANM simulation study of Section 4.3.1, the LiNGAM mechanism model is correctly specified and the error distributions are not. However, the results are similar: accuracy—which can only be assessed over repeated experiments—is highly dependent on the signal-to-noise ratio and the extent to which the data-generating mechanism deviates from joint Normality. Bayes Factors—which can be used to quantify uncertainty for individual scientific studies—also predict causal identifiability and noise levels, and closely track accuracy. In both simulation settings, Bayesian Causal Discovery outperformed Frequentist algorithms in accuracy.

Finally, decisions based on Bayes Factors are robust to modest discrepancies between the model and true data-generating mechanism. This quality is particularly important in causal discovery since we view the modeling and identification assumptions we make as plausible but rough approximations to the truth. Our ANM Gaussian Process model is highly flexible precisely because we do not know the true nature of the data-generating mechanism, which is likely to be simpler than a typical GP realization. Conversely, for LiNGAMs, we assume linearity of f, which we hope is a sufficient approximation to the almost certainly nonlinear true mechanism.

4.4 Application to the Tuebingen Cause-Effect Pairs

We now turn from simulated to real data. In Figure 4.7, we plot the age and height of a collection of $4,177$ abalone sea snails. We know that age causes height, but because the mechanism linking age and height appears monotonic—with the exception of two outliers—this causal relationship is not clear from the scatter plot alone. It is clear, however, that the mechanism is nonlinear, and in the correct age-causes-height direction (Figure 4.7a) the

errors appear roughly independent of age, whereas in the incorrect direction (Figure 4.7b) the error variance appears larger for taller abalone. We therefore expect a Bayesian ANM to yield a Bayes Factor that favors the age-causes-height model. Indeed, an ANM yields $K = 1.6 \times 10^{34}$ in favor of age causing height, a much more convincing Bayes Factor than the $K = 3.1 \times 10^{1}$ produced by a LiNGAM.



(a) Forward causal direction.

(b) Reverse causal direction.

Figure 4.7: Tuebingen Pair 8, featuring abalone snail age X which causes height Y.

The abalone data come from the Tuebingen Cause-Effect Pairs, a collection of data sets with known causal structure used to test causal discovery algorithms (Mooij et al., 2016). Since collecting many data sets with known causal structure is vastly harder than the usual task of assembling a single test data set, the Pairs are the primary standard by which many causal discovery algorithms' real-data performance is judged. Versions of the Pairs were used in, e.g., Lopez-Paz et al. (2015) and Janzing et al. (2012). See Appendix D in Mooij et al. (2016) for a detailed description of every Pair. We perform Bayesian Causal Discovery on the Pairs using ANMs and LiNGAMs, assessing the procedure's discovery accuracy and uncertainty quantification via Bayes Factors. For comparison, we also apply Frequentist versions of ANMs and LiNGAMs to the Pairs.

The latest version of the Pairs contains 108 data sets, 102 of which are bivariate. Three of those 102 Pairs have one binary variable, making our methods inapplicable. Additionally, we

exclude two Pairs that feature "hour of the day" as a covariate to avoid modeling periodicity. Sample sizes for the remaining 97 Pairs range from 94 to 16,382. Table 4.5 displays a summary of the distribution of sample sizes in these Pairs. A visual inspection of the Pairs' scatter plots showed that they vary widely in structure. 39 Pairs include at least one discrete variable with sufficiently many regularly-spaced levels to warrant continuous approximation. 12 Pairs contain zero-inflated variables, substantial outliers, or other conspicuous fluctuations in the bivariate density. The Pairs also vary in the degree of linearity of the mechanism: while a linear relationship appears to accurately describe some, others clearly display a nonlinear association. This variation motivates the use of both ANMs and LiNGAMs in our analysis.

Table 4.5: Sample sizes of the 97 Tuebingen Pairs used.

Pairs with $n \leq 400$	51
Pairs with $n \in (400, 1000]$	13
Pairs with $n \in (1000, 5000]$	21
Pairs with $n > 5000$	12

Our general approach of excluding only Pairs with data types that are completely incompatible with our models is motivated by the dearth of Pairs, and follows Mooij et al. (2016); Janzing et al. (2012); Peters et al. (2014).

4.4.1 Choosing an Identification Strategy

Researchers typically choose causal identification assumptions—in this case, ANMs or LiNGAMs—on the basis of their scientific knowledge. However, selecting causal identification assumptions on the basis of scientific knowledge for 99 data sets is impractical. As such, we compared exploratory and algorithmic approaches to choosing between ANMs and LiNGAMs. During visual inspection of the Pairs, we categorized each Pair's mechanism as appearing either linear or monotonic in the first derivative. Only two Pairs' mechanisms ap-

peared to not fit into this dichotomy—numbers 85 and 88—and since both were borderline cases, we characterized them as monotonic in the first derivative.

Since we assume $\epsilon \perp\!\!\!\perp X$ and we wish to decide whether to assume linearity or nonlinearity, Sen and Sen (2014)'s test for combined independence of errors and goodness-of-fit of the linear model is in theory a perfect algorithmic approach to deciding between ANMs and LiNGAMs. The assumption of independent errors (specifically, homoscedasticity) appears likely to be invalid for many of the Tuebingen Pairs, though. Therefore, a rejection of the null hypothesis of linearity and independent errors from this test could occur when linearity is a poor assumption, but also when linearity holds with heteroscedastic errors.

We instead opt for a faster, prediction-based approach to choosing a causal identification assumption for each Pair: Algorithm 2, `LinearityCheck`. Because the mechanisms appear to fit the linear vs. first-derivative-monotonic dichotomization well, `LinearityCheck` assumes that if the mechanism relating X and Y is nonlinear, a linear regression model with a quadratic term will have higher predictive power than a simple linear regression in at least one causal direction. The tuning parameter r_0 decides how much stronger the predictions from the quadratic model must be, in terms of out-of-sample R^2, for ANMs to be chosen over LiNGAMs. We chose $r_0 = 0.95$ as a compromise between competing concerns: r_0 too close to 1 would require us to use ANMs even when linearity is a very reasonable approximation, while smaller r_0 would yield very few uses of ANMs.

`LinearityCheck` chooses ANMs for 31 of the 97 Pairs and LiNGAMs for the remaining 66. Table 4.6 illustrates that these choices align closely with our visual inspection of the Pairs, though `LinearityCheck` tends to pick LiNGAMs more frequently.

4.4.2 Model Specifications and Evidence Computation

Once either a LiNGAM or ANM has been chosen, one must specify the model prior. The paucity of test data sets prohibits parameter tuning or model selection approaches that require splitting the data, such as cross-validation, so the model specifications were completely determined a priori. We applied the LiNGAM model prior from Section 4.2.4 as it does not

Algorithm 2 Discovery Model Choice for the Tuebingen Pairs

procedure LINEARITYCHECK($\{X, Y\}_n$, $k = 5$, $r_0 = 0.95$)
  $r_X^1 \leftarrow k$-fold cross-validated RMSE estimate from model $Y = \beta_0 + \beta_1 X + \epsilon$
  $r_X^2 \leftarrow k$-fold cross-validated RMSE estimate from model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$
  if $r_X^2 / r_X^1 < r_0$ **then**
    return ANM
  else
    $r_Y^1 \leftarrow k$-fold cross-validated RMSE estimate from model $X = \beta_0 + \beta_1 Y + \epsilon$
    $r_Y^2 \leftarrow k$-fold cross-validated RMSE estimate from model $X = \beta_0 + \beta_1 Y + \beta_2 Y^2 + \epsilon$
    if $r_Y^2 / r_Y^1 < r_0$ **then**
      return ANM
    else
      return LiNGAM
    end if
  end if
end procedure

Table 4.6: Choice of identification assumptions.

	`LinearityCheck` selects LiNGAM	`LinearityCheck` selects ANM
Mechanism appears linear	49	1
Mechanism derivative appears monotonic	17	30

Comparison of `LinearityCheck`'s identification assumption selections to our visual inspection of Pairs' mechanisms.

require tuning. For ANMs, we chose σ_h—the most influential tuning parameter—to allow for more flexible GPs than the value used in Section 4.3, since the Tuebingen Pairs almost certainly contain a much wider range of mechanisms than our simulations. However, we needed parameters such that GPs from the model were typically approximately monotonic in the first derivative, in order to align with our visual inspection of the Pairs. Balancing these considerations yielded $\sigma_h = \frac{1}{2} n^{1/5}$, i.e. a bandwidth distribution with half the mean

and standard deviation as that used in Section 4.3.

Finally, computing the evidence when n is large is difficult for two reasons. First, values of the integrand in (4.4) are often within machine precision of zero for $n > 400$ (for both ANMs and LiNGAMs). This presents a problem for both the numerical integration and the Monte Carlo approaches to approximating the evidence.[6] Second, for Gaussian Process ANMs, standard multivariate Normal density computations use Cholesky decompositions of $\Sigma_{h,l} + \sigma_Y \mathrm{I}$ that are $\mathcal{O}(n^3)$ and must be performed for each Monte Carlo sample (or numerical integration function evaluation) for all 99 data sets for both M_X and M_Y.[7] These exact density computations become cumbersome when $n > 1000$. Following Janzing et al. (2012); Stegle et al. (2010), we randomly subsample each Tuebingen Pair to have a maximum sample size of 400.

4.4.3 Decision Theory for Bivariate Bayesian Causal Discovery

In this section, we allow for three decisions: M_X, M_Y, and M_0: indecision, which is used when the evidence for M_X or M_Y is not sufficiently strong. The optimal decision d is determined by the posterior model probabilities and by misclassification costs w_X (the cost of deciding M_X when M_Y is the true model), w_Y (the cost of deciding M_Y when M_X is the true model), and w_0 (the cost of abstaining from a decision). The optimal decision rule minimizes the risk

$$R(d) = w_X \mathrm{P}(M_Y, d = M_X | \{X, Y\}_n) + w_Y \mathrm{P}(M_X, d = M_Y | \{X, Y\}_n) + w_0 \mathrm{P}(d = M_0 | \{X, Y\}_n). \tag{4.14}$$

In our application to the Tuebingen Pairs, we assume that $p_{M_X} = 1/2$ and $w_X = w_Y$

[6] Even numerical stabilization tricks, such as computing $\frac{1}{B}\sum_i L_i$ as $\exp\left(\log\left(L^{(B)}\right) + \log\left(1 + \sum_{i=1}^{B-1} \exp\left(L^{(i)} - L^{(B)}\right)\right)\right)$, do not enable Algorithm 1 to avoid this numerical precision problem since the largest Monte Carlo sample log likelihood $\log(L^{(B)})$ tends to be much closer to 0 than to most of the other $L^{(i)}$.

[7] We use R's mvtnorm package (Genz et al., 2020) for multivariate Normal density computations (it is well-maintained and performed similarly to competitors in testing), and the cubature package (Narasimhan et al., 2020) for numerical integration.

so that we treat the models equally in terms of probability and cost. In Appendix C.2, we prove that the optimal decision rule under these assumptions is completely characterized by a decision threshold c:

$$\begin{aligned} d = M_X \text{ if } \log(K) > c \\ d = M_Y \text{ if } \log(K) < -c \\ d = M_0 \text{ if } |\log(K)| \leq c. \end{aligned} \tag{4.15}$$

For the Tuebingen Pairs, we make a discovery decision via (4.15), testing various values of c. For these real data, our model priors will always be somewhat misspecified; additionally, errors may be heteroscedastic and variables are only approximately continuous in many cases, violating our baseline assumptions. Thus, we do not expect log Bayes Factor magnitude to be as closely related to discovery accuracy as in our Section 4.3 simulations. We do expect discovery accuracy to generally increase with c, however, particularly for small c that may convey genuine uncertainty about the true causal direction.

4.4.4 Causal Discovery on the Tuebingen Pairs

In Figure 4.9, we display the discovery accuracies of Bayesian Causal Discovery via Monte Carlo and numerical evidence approximation on the Tuebingen Pairs. These accuracies are computed over $c \in \{0, 10, \ldots, 90\}$, corresponding to 10 Bayes decision rules with increasing p_{M_0}. Accuracy is the most appropriate metric for bivariate causal discovery, as false positives/negatives are meaningless since the X and Y labels of variables are arbitrary. We attempt to avoid the concern (Mooij et al. (2016), p32n19) that "it is easy to visually over-interpret the significance of [accuracy curves] in the low decision-rate region" by simultaneously plotting the accuracy needed for statistical significance at the $p < 0.1$ level, which is a (roughly) decreasing function in the sample size as shown in Figure 4.8.

Figure 4.9 shows that the accuracy of decision rule (4.15), which is no better than guessing at $c = 0$, increases with the decision threshold for small values of c before stabilizing between 63% and 71% and $p < 0.1$ for $c \geq 30$. The relationship between accuracy and $|\log(K)|$ is very

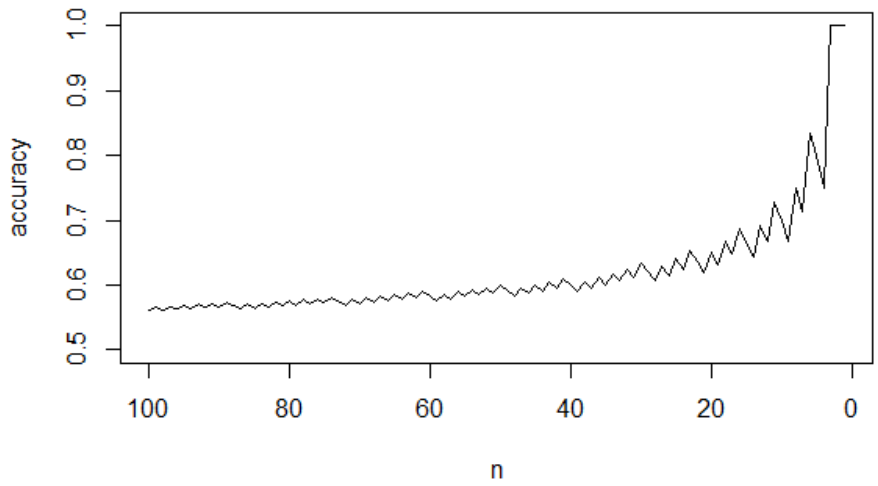
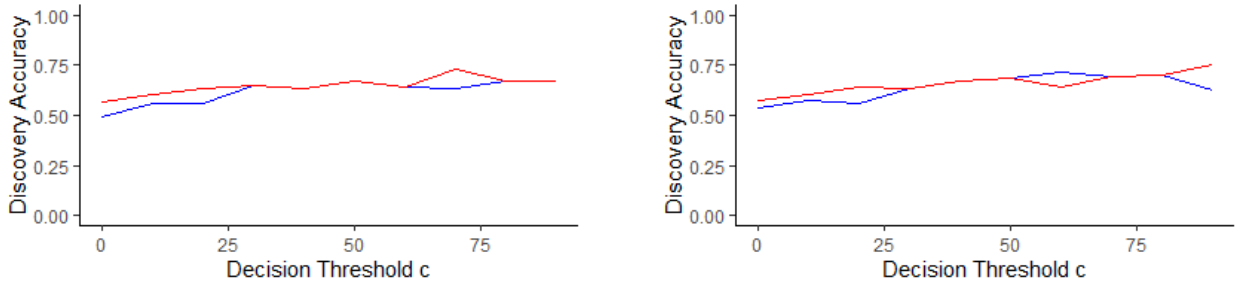


Figure 4.8: Accuracy required to achieve $p < 0.1$ significance for sample sizes $\{1, \ldots, 99\}$. For a given n and $X \sim \text{Binom}(n, 0.5)$, the smallest q such that $1 - \text{P}(X \leq qn) < 0.1$.



(a) Monte Carlo evidence approximation.

(b) Numeric evidence approximation.

Figure 4.9: Discovery accuracy (in blue) vs. decision threshold c for Bayesian Causal Discovery. Also shown: discovery accuracy needed to achieve $p < 0.1$ significance, in red.

different here than in our Section 4.3 simulations: accuracy is modest even as the posterior odds ratio (4.1) nears exp(100). And, because of the small number of Pairs, it is difficult to say how well the relationship between c and accuracy will generalize to other causal discovery applications.

The paucity of Pairs also prevents deep subgroup analysis due to low power and multiple testing concerns. However, we do stratify our analysis along two important dimensions: sample size and causal identification strategy. Figure 4.10 displays discovery accuracy for Bayesian Causal Discovery with numeric evidence approximation for the Pairs with $n > 400$, for which subsampling to $n = 400$ was used. Bayes Factors take on a wide range for these pairs due to the larger sample sizes, and accuracy grows swiftly as c increases from 0, exceeding 80% for $c \in [30, 70]$.

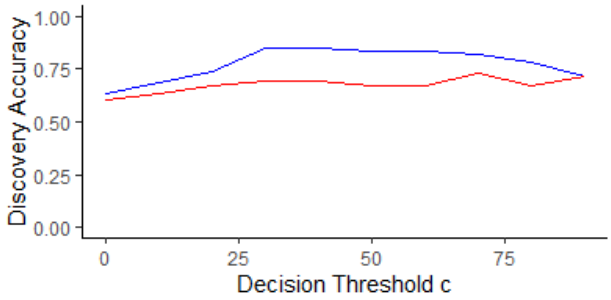
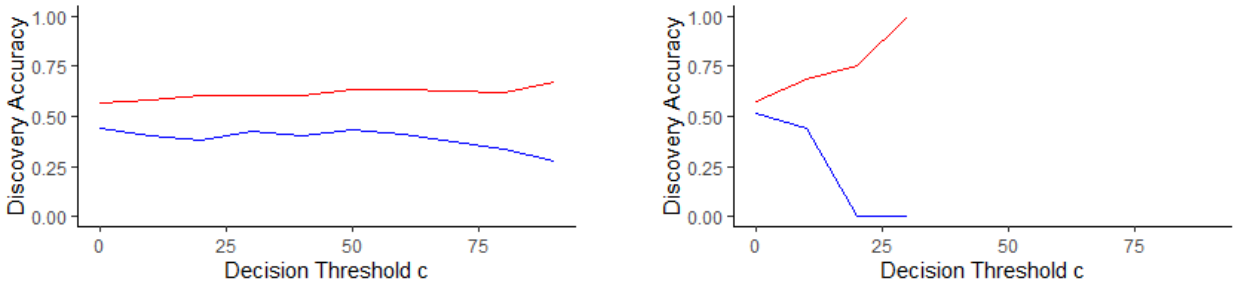


Figure 4.10: Discovery accuracy (in blue) vs. decision threshold c for Bayesian Causal Discovery, subsampled Pairs with $n > 400$. Also shown: accuracy needed to achieve $p < 0.1$ significance, in red.

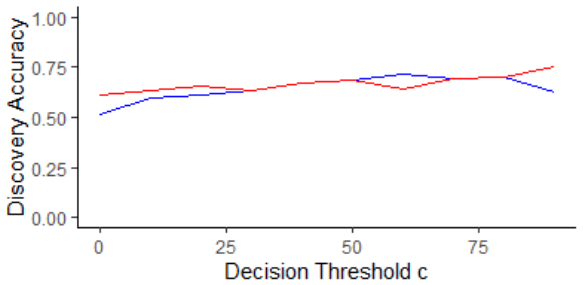
Figure 4.11 shows the discovery accuracies for ANMs and LiNGAMs separately, both for all of the Pairs and the Pairs for which each model was selected by `LinearityCheck`. LiNGAM log Bayes Factors were small in magnitude relative to those of the ANM, exceeding 20 for just four pairs (which were classified as having nonlinear mechanisms by `LinearityCheck`).

While LiNGAMs performed poorly in both cases, the use of `LinearityCheck` to choose an identification strategy for each pair clearly improved the accuracy of ANMs and of Bayesian Causal Discovery overall. LiNGAM's lackluster discovery accuracy is unsurprising: the parametric Laplace model for the errors (whose non-Gaussianity is assumed to identify the causal structure) is far less accommodating than the highly flexible Gaussian Process that models the ANM mechanism. And, while ANMs were selected by `LinearityCheck` when

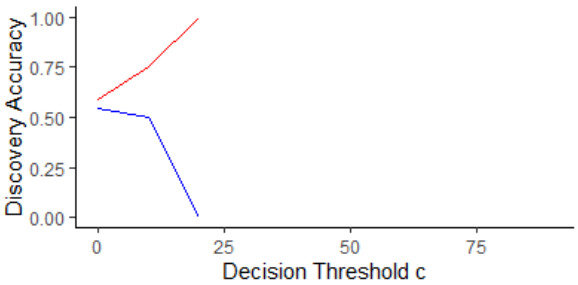


(a) ANMs used for **all Pairs**.

(b) LiNGAMs used for **all Pairs**.



(c) ANMs chosen by `LinearityCheck`.



(d) LiNGAMs chosen by `LinearityCheck`.

Figure 4.11: Discovery accuracy (in blue) vs. decision threshold c for Bayesian ANMs and LiNGAMs. Also shown: accuracy needed to achieve $p < 0.1$ significance, in red.

the mechanism was determined to be nonlinear, we did not screen for non-Gaussianity of errors, putting LiNGAMs at a disadvantage.

These two subgroup analyses together suggest that Bayesian ANMs, applied to relatively large data sets for which nonlinearity of the mechanism can be assumed, can yield reliable causal discovery when the Bayes Factor is decisive. Indeed, while ANMs saw just 69% accuracy for Pairs with $n > 400$ with a threshold $c = 0$ ($p = 0.1$), accuracy rises to 85% with a threshold of $c = 40$ ($p = 0.01$).

4.4.5 Comparing Causal Discovery Algorithms

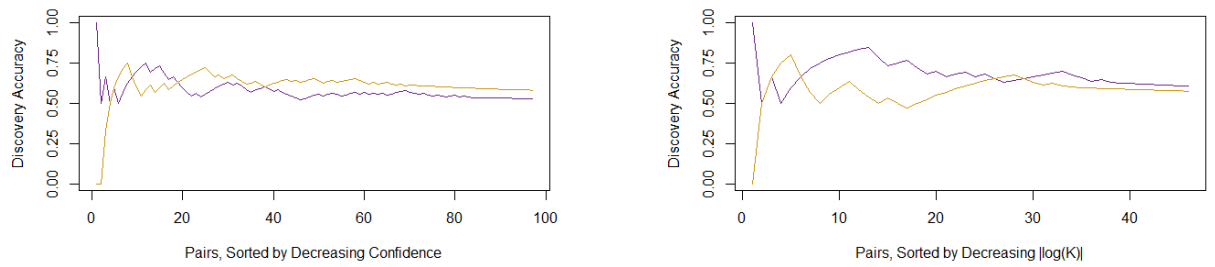
The Bayesian Causal Discovery accuracies presented above are broadly comparable to those of Frequentist ANMs (Hoyer et al., 2009) and IGCI (Janzing et al., 2012) found in the survey of Mooij et al. (2016), which used 100 Tuebingen Pairs (overlapping almost completely with our 97 Pairs). Those authors tested 16 versions of ANMs and 21 versions of IGCI based on, e.g., different ways of choosing the Gaussian Process kernel bandwidth for ANMs and different entropy estimation techniques for IGCI. These experiments revealed a wide range of discovery accuracies, with estimates ranging from approximately 20% to 70–75% for each algorithm. As Spiegelhalter (2003) notes, we cannot reliably distinguish between any estimated accuracies when the confidence intervals for these estimates all overlap the grand mean of the estimates. In Mooij et al. (2016), for both ANMs and IGCI, the 95% confidence intervals of all implementations achieving greater than 50% accuracy overlap the average accuracy of these implementations, meaning we cannot make a distinction among versions of an algorithm with high confidence based on results from that paper.

The high degree of confidence interval overlap is driven by two factors: the small number of Tuebingen Pairs ensures that confidence intervals are relatively wide, and the difficulty of the causal discovery problem means that accuracies of the various implementations are all relatively close to 50%. Additionally, multiple testing concerns would make it difficult to confidently identify the best implementations even if the confidence intervals did not overlap the overall mean accuracy for some implementations.

To directly compare Bayesian Causal Discovery with established methods, we test Frequentist ANMs and LiNGAMs on the 97 Pairs summarized in table 4.5. We used the model (ANM or LiNGAM) chosen by `LinearityCheck` for the Frequentist approach as well as Bayesian Causal Discovery, facilitating a fair comparison between the two approaches. For both algorithms, we used the same out-of-the-box implementations as in Section 4.3: the `pcalg` implementation of LiNGAMs (Kalisch et al., 2019), and Scikit-learn's Gaussian Process Regression (Pedregosa et al., 2011) with Gretton et al. (2008)'s HSIC implementation

for ANMs. We specify the models fully before looking at the data to ensure that accuracy estimates are not inflated (see Section 4.4.1) and only compare their performance on all Pairs and those with $n > 400$ to avoid multiple testing concerns.

Figure 4.12 compares the Frequentist and Bayesian algorithms for all Pairs and for those with $n > 400$. The y-axis displays cumulative discovery accuracy for the Pairs ordered by decreasing absolute log Bayes Factor.



(a) All Pairs.

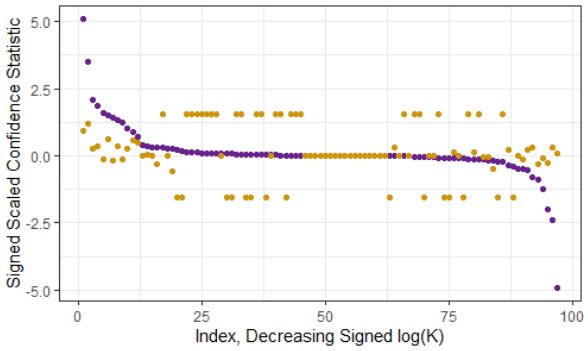
(b) Pairs with $n > 400$.

Figure 4.12: Cumulative accuracy of Bayesian (purple) vs. Frequentist (gold) causal discovery algorithms. For a given n on the x-axis, y-value reflects accuracy over the n Pairs with the largest values of $|\log(K)|$.

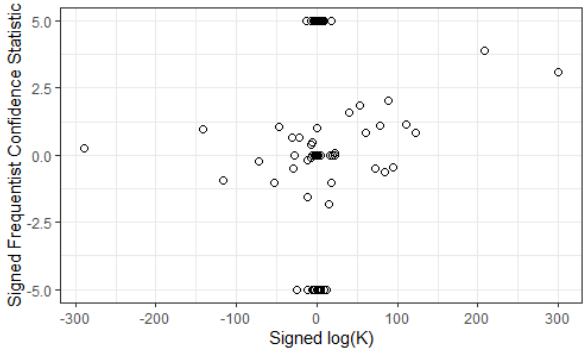
While the Frequentist algorithms perform better across all 97 Pairs, the Bayesian versions appear to have an edge for larger sample size of 400—particularly when confidence is high.

A different way of visualizing this comparison uses "signed" log Bayes Factors and Frequentist test statistics (jointly referred to as "confidence statistics"), which are positive (negative) when they yield a correct (incorrect) discovery under decision rule (4.3). For Frequentist ANMs, the test statistic is the ratio of HSIC independence test statistics. For Frequentist LiNGAMs, the log test statistic is 0 if independence is inferred, and positive (negative) infinity if M_X (M_Y) is inferred. In Figure 4.13, we plot signed log confidence statistics explicitly. When a log confidence statistic is infinite (as is the case for Bayesian Causal Discovery when the evidence for exactly one of M_X or M_Y is zero, or when Frequentist LiNGAM does not decide independence), we set it to be just outside the range of finite

observed values for plotting purposes only.



(a) Signed, scaled confidence statistic for Pairs sorted by decreasing signed log Bayes Factor, Bayesian (purple) and Frequentist (gold) approaches.



(b) Signed log Bayes Factor vs. signed Frequentist confidence statistic.

Figure 4.13: Comparisons of Frequentist and Bayesian causal discovery using signed confidence statistics.

In Figure 4.13a, we sort the Pairs in order of decreasing signed log Bayes Factor, comparing the signed confidence statistics for Frequentist[8] and Bayesian approaches (after scaling for comparability). Note that for the log Bayes Factors closest to zero, for which LiNGAM was chosen by `LinearityCheck`, Frequentist LiNGAM infers independence, while for log Bayes Factors with slightly larger magnitudes it does not.

The accuracy of a method for a given scaled decision threshold c_s can be found by comparing the number of correctly discovered Pairs above the line $y = c_s$ to the number of incorrectly discovered Pairs below the line $y = -c_s$. A similar approach can be used for both axes of Figure 4.13b, which compares signed log Bayes Factors on the x axis to signed Frequentist confidence statistics on the y axis. In both figures, the relatively weak correlation between Bayesian and Frequentist confidence is clear.

[8] All infinite Frequentist LiNGAM log confidence statistics were set to be just outside the range of the other Frequentist confidence statistics; these artifacts appear as gold rows of points in Figure 4.13a and as rows at $y = 5$ and $y = -5$ in Figure 4.13b.

Any differences between the Frequentist and Bayesian algorithms' performance are unlikely to generalize beyond these 97 Pairs. We test the difference in discovery accuracy between the Bayesian and Frequentist versions as follows: a correct classification is scored as 1, an indecision (produced when the evidence for M_X and M_Y are both numerically zero, or when Frequentist LiNGAM identifies independence) as 1/2, and an incorrect classification as zero. We then perform a paired t-test on these scores. For all Pairs, the Frequentist approach scores 0.05 points better on average, with $p = 0.38$. For just the subsampled Pairs with $n > 400$, the Bayesian approach is 0.03 points better ($p = 0.685$).

Additionally, we cannot determine whether the relationship between confidence and accuracy differs between Frequentist and Bayesian implementations: for both types of algorithm, we fit logistic regressions predicting whether a Pair was classified correctly from the centered and scaled estimated confidence statistic. Testing whether the coefficient of interest differed between the Bayesian and Frequentist regressions yielded $p = 0.21$ for all Pairs and $p = 0.57$ for the large-sample Pairs, so there is little evidence that one type of algorithm shows a stronger relationship between confidence and accuracy than the other.

In summary, the low availability of test data prohibits us from drawing strong, general conclusions regarding Frequentist vs. Bayesian approaches or the strength of the relationship between the decision threshold c and discovery accuracy. We do find Bayesian Causal Discovery to be moderately effective when n is not small and Bayes Factors are large. Our inspection of the Pairs revealed that the assumptions of ANMs and LiNGAMs are likely invalid for many of the Pairs. Additionally, the Tuebingen Pairs are may be confounded. For example, Pair 13: horsepower $\rightarrow$ fuel consumption may be confounded by car design choices influencing both horsepower and fuel consumption. Both of these factors help explain the modest accuracies seen by Frequentist and Bayesian versions of both algorithms.

4.5 Discussion

In Section 4.2, we demonstrated the adaptability of Bayesian Causal Discovery, developing Bayesian versions of two complementary causal discovery algorithms: ANMs, which use

nonlinear mechanisms and Gaussian errors to identify causality, and LiNGAMs, which use linearity and non-Gaussianity. We showed that Bayes Factors quantify confidence in causal discovery: in Section 4.3, we found that the Bayes Factor conveys the degree of identifiability and amount of noise in the data, while in Section 4.4 we used the Bayes Factor to construct an optimal Bayes decision rule, equating degree of confidence with an explicit misclassification cost.

Crucially, non-Bayesian approaches do not yield inferences that can be directly used to construct optimal decision rules. Many non-Bayesian causal discovery approaches output statistics with no probability interpretation that only provide directional information about the causal direction. Even the p-values yielded by bootstrap approaches (Friedman et al., 1999; Komatsu et al., 2010) require nontrivial assumptions to be adapted to the Bayesian decision theory context (Berger and Sellke, 1987). While our analysis in Section 4.4.5 found no significant difference between the ability of Bayes Factors and Frequentist confidence measures to predict classification accuracy for the Tuebingen Pairs, optimal decision-making from a cost-based perspective remains a theoretical advantage of the Bayesian approach.

Additionally, we found that the discovery accuracy of Bayesian ANMs and LiNGAMs is superior to Frequentist versions' accuracy in our simulation settings, and competitive on the real-world Tuebingen Pairs. Combined, our empirical results suggest that Bayesian Causal Discovery may be successful when model priors are reasonable approximations to the truth, when sample sizes are large, and when Bayes Factors are decisive.

Our approach is not without limitations: the Tuebingen Pairs present numerous practical difficulties that Bayesian Causal Discovery—like other causal discovery algorithms—does not always successfully address, including irregularities of joint densities and error dependence. Additionally, while our Bayesian approach quantifies uncertainty, it also faces greater computational challenges than Frequentist causal discovery algorithms.

One thrust of future methodological work should be to maintain the fully Bayesian context but relax the assumptions of this chapter. Modeling error dependence—or, equivalently, latent variables—would increase the practical utility of Bayesian Causal Discovery, but the

increased complexity will create additional computational challenges. An extension to the multivariate case is theoretically viable since Bayes Factors generalize straightforwardly to more than two models. However, the number of DAGs on p nodes is super-exponential, and multivariate models necessitate more parameters over which we must integrate to compute the evidence. Both of these factors increase computational demands substantially.

Second, future research should investigate uncertainty quantification and robustness in Frequentist causal discovery algorithms. Frequentist uncertainty quantification approaches are currently few and limited to a subset of existing models (Friedman et al., 1999; Komatsu et al., 2010). The robustness of causal discovery algorithms to departures from modeling assumptions has also not been widely studied, analytically or via simulation.

For causal discovery to grow as a field and be more widely used in scientific studies, algorithms must be robust enough to succeed in the realm of messy, real-world data, and researchers must be able to quantify the uncertainty in learned causal structures. By introducing Bayesian Causal Discovery, we hope to open a new direction of inquiry that brings the field of causal discovery closer to meeting scientists' needs.

Chapter 5

DISCUSSION

Public research funding institutions like NIH are constantly engaged in an iterative process of self-examination and improvement; they owe it to their constituencies to equitably fund research that serves the public good. In Chapter 2 of this dissertation, we performed one such examination for the NIH, finding that criterion scores (introduced in 2009) fully explain racial disparities in preliminary Overall Impact scores. This study highlighted a major impediment to causal analyses of peer review data: the uncertain directions of causality among the peer review scores.

While many causal discovery algorithms for learning causal structures from observational data exist, most do not quantify the uncertainty in the discovery process, and are therefore of limited practical scientific use. Chapter 4 attacked this problem, developing a novel Bayesian framework for causal discovery that naturally quantifies uncertainty and can be adapted to discovery algorithms that use different assumptions to identify causality from an observational distribution.

Finally, in Chapter 3, we develop a novel approach that quantifies the information in peer review scores. Such an approach is merited because funding institutions like NIH use peer review to identify and fund "good science"—but no gold standard for measuring scientific quality exists. Rather than analyzing peer review outcomes via a proxy like bibliometrics, we assess the informativeness of scores independently of research outcomes. Inspired by recent NIH studies on scoring scales (NIH Staff, 2019c), we develop the concept of refinement, a measure of rating informativeness contextualized by rounding behavior induced by the underlying scoring scale. We analyze refinement at AIBS, finding that merit (overall) scores are more refined on average than criterion scores.

5.1 Future Work: Peer Review

Further research should extend the novel refinement and Bayesian Causal Discovery approaches developed here so that they can be used to contextualize and improve the original NIH peer review study that underlies Chapter 2. I am part of a UW research team that is performing causal discovery on NIH peer review data with the aim of learning the causal relationships among the criterion and overall scores. If this study is conclusive, one could update the analysis in Chapter 2 using the public data set (Erosheva et al., 2020a) by controlling only for criterion scores found to be determined prior to the overall score. Such an analysis would reveal whether racial disparities in criterion scores actually generate disparities in overall scores or are simply artifacts of overall score disparities. Because criterion scoring is thought to focus reviewers' attention on factors related to merit (Thorngate et al., 2009), disparities arising via criterion scores may be more fair—or easier to mitigate—than disparities that first appear in the overall score and are simply reflected in the criterion scores.

If Bayesian Causal Discovery can be extended to multivariate settings in a computationally feasible manner (see Section 5.2.1), it is another attractive approach to learning the causal structure of NIH peer review scoring. In particular, it would allow us to quantify our confidence in the learned structure and model latent confounding. Accounting for unobserved confounding is particularly important for causal discovery in peer review because many aspects of the peer review process that inform scores are difficult to model (e.g., the contents of a grant proposal).

The main limitation of the refinement metric developed in Chapter 3 is that it is specific to scales with a hierarchical structure in which some levels are widely considered to be "rounder" than others. For the AIBS 1.0–5.0 decimal scale, Figure 3.1 illustrates that integer scores are the "roundest" scores, with other multiples of 0.5 still being more frequently used than the other score levels. What makes a refined collection of ratings, however, on NIH's 1–9 integer scale, for which no such hierarchy exists? Is entropy, which captures the unpredictability of

the scores, meaningful by itself?

Another extension might consider the refinement of aggregate scores, such as NIH panel Overall Impact scores, that are directly used to make funding decisions. At NIH, these aggregate scores are converted into a percentile score which conveys only the application's relative rank among other applications recently reviewed by the SRG. While percentile scores adjust for differences in how different SRGs tend to use the NIH's scoring scale, they also erase information contained in Overall Impact scores. We wish to quantify the information lost in percentile scoring and make rigorous comparisons to other score standardization approaches.

While the refinement and Bayesian Causal Discovery methods were motivated by open problems in the statistical analysis of peer review data, their applicability reaches far beyond peer review. Both methods are limited in the types of data to which they can be applied: Entropic Refinement requires a scale on which scores can be rounded, and our analysis of Bayesian Causal Discovery is currently limited to the bivariate realm. Thus, future work based on this dissertation need not be constrained to peer review or even to data applications.

5.2 Future Work: Beyond Peer Review

The refinement concept developed in Chapter 3 is applicable anytime ratings on a numeric scale are used in the absence of a gold standard. It is especially useful when the scores are not used solely to rank but also to distinguish items from one another or convey a notion of quality. Thus, refinement could shed light on restaurant ratings (e.g. Yelp, Google), dysphagia evaluations (O'Neil et al., 1999), and even some athletic competitions in which scoring is subjective (e.g. the NBA Dunk Contest). As for peer review, refinement must be generalized to non-decimal scales for such analyses to proceed.

Bayesian Causal Discovery, the causal discovery framework developed in Chapter 4, can be extended in a number of ways. I now discuss the extensions with the most potential impact, and the difficulties in making them.

5.2.1 Multivariate Bayesian Causal Discovery

A generalization to the multivariate case faces three broad categories of challenges. First, the number of possible DAGs under consideration can explode combinatorially: the number of DAGs on k variables is $O\left(\exp\left(\binom{k}{2}\right)\right)$, complicating the comparison of all possible causal models. Second, additional variables bring additional parameters into the model. To the extent that the evidence integral (4.4) cannot be factored into separate integrals, the difficulty of approximating the evidence accurately increases greatly with the number of parameters (see Appendix C.3). Finally, accounting for latent confounders presents a separate, unique challenge.

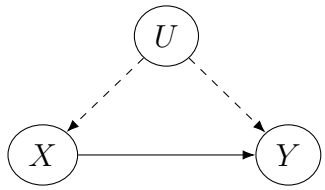
In multivariate settings, we do not have to compute the evidence for each DAG separately, because the likelihood of an observation $D \in \mathbb{R}^k$ under model M factors as

$$\mathrm{P}(D|M) = \prod_{i=1}^{k} \mathrm{P}(D_i|pa_M(D_i)) \tag{5.1}$$

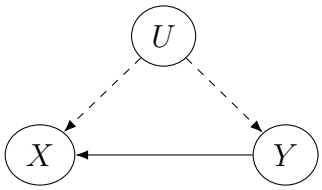
where $pa_M(D_i)$ refers to the parents of variable D_i under causal model M (Pearl et al., 2016). The components of the product in (5.1) can be computed separately and then recycled across different models M. Hoyer and Hyttinen (2009) propose a greedy algorithm that leverages this decomposition by iteratively adding the edge yielding the highest marginal likelihood to a DAG and recycling computations for the $k-1$ variables whose parents do not change (this is only feasible under independent priors for the edge strengths and error terms). Alternatively, one could reduce the set of DAGs under consideration via preprocessing, for example by using expert/scientific knowledge, or an inexact causal discovery algorithm like the `pc` algorithm (Spirtes et al., 1993; Kalisch et al., 2019) to obtain a smaller class of DAGs.

Shimizu and Bollen (2014) extend the Bayesian LiNGAMs of Hoyer and Hyttinen (2009) to include latent variables, finding that their algorithm learns causal relationships among General Social Survey variables better than LiNGAM approaches that do not model latent variables or the non-Bayesian latent LiNGAM of Hoyer et al. (2006). Future work should extend Bayesian ANMs similarly. Rothenhäusler et al. (2016) show that a bivariate causal

direction can be identified in the presence of linear confounding and Gaussian errors (Figure 5.1). If U is observed, then the dashed edges in Figure 5.1 can be parameterized and the approach in Chapter 4 can be used. When U is unobserved, we can model it as inducing dependence between X and ϵ (under model M_X) and allowing the noise distribution $P_{\epsilon|X}$ to be additionally parameterized by X.



(a) M_X with a latent linear confounder.



(b) M_Y with a latent linear confounder.

Figure 5.1: Bivariate causal discovery with a latent confounder U. Dashed (solid) edges correspond to linear (nonlinear) mechanisms. Per Rothenhäusler et al. (2016), M_X and M_Y cannot generate the same joint distribution on X, Y, U when errors are Gaussian.

5.2.2 New Applications of Bayesian Causal Discovery

In Chapter 4, we developed Bayesian versions of Additive Noise Models (Hoyer and Hyttinen, 2009) and Linear Non-Gaussian Additive Models (Shimizu et al., 2006) which quantify uncertainty in causal discovery and accurately choose causal models in some settings. Appendix C.1 additionally shows that Bayesian Causal Discovery can be applied to Information-Geometric Causal Inference (Janzing et al., 2012), though less successfully. These three algorithms are just a subset of the myriad approaches to causal discovery, however; future work should consider Bayesian versions of other algorithms. For example, the algorithms of Kocaoglu et al. (2016) and Sun et al. (2006, 2007) assume minimal complexity of the distribution of the effect given the cause; Bayesian versions could penalize complexity/entropy explicitly in $\mathcal{P}_F$ and $\mathcal{P}_{P_\epsilon}$. Yu et al. (2020b) prove identifiability for certain zero-inflated graphical models, which could be estimated Bayesianly. Finally, Chen et al. (2019) identify

DAGs when all variables' error variances can be assumed equal; this assumption could be naturally incorporated into a multivariate Bayesian Causal Discovery framework.

More applications to real-world data—particularly beyond the Tuebingen Pairs test data sets (Mooij et al., 2016)—are needed for causal discovery to develop credibility and be better-understood by the scientific community. In particular, performance tests on larger corpora of data sets will allow us to determine—with statistical guarantees—what discovery techniques are best in given situations. Comparing causal discovery and interventional techniques in settings where both observational and experimental data is available (e.g., economic data with and without policy interventions) would provide another type of validation.

Bayesian Causal Discovery, which requires specifying a joint distribution for the data, should be applied in contexts where the functional forms of the mechanism and/or error distribution are better-understood than in the Tuebingen Pairs. Peer review scoring may be one such situation: because scores are all on the same scale and overall scores can be conceptualized as weighted averages of criterion scores, linearity of mechanisms is a plausible assumption. fMRI data, which captures neuron activations, is another realm where the similarity of the covariates and scientific context could allow us to choose causal identification assumptions and specify a model prior in an informed manner.

5.3 Our Contributions

Chapter 2 of this dissertation builds on a significant body of work about racial disparities in NIH peer review which began with Ginther et al. (2011). In Erosheva et al. (2020b), on which that chapter is based, we are the first to analyze individual reviewer-level preliminary review scores and investigate the role of criterion scores in review score racial disparities; in this chapter, I provide the first (to my knowledge) explicitly causal analysis of these disparities.

While it helps to round out the collection of research on peer review outcome disparities, Chapter 2 also provides a springboard into two novel approaches to the analysis of peer review data. Chapter 3 navigates the lack of a gold standard for the quality of scientific research proposals by quantifying the informativeness of grant proposal peer review scores

via a refinement metric, a new tool with which reviewing agencies may analyze and improve their funding processes. Next, motivated by the crucial assumption we make regarding the causal ordering of review scores in Chapter 2, Chapter 4 develops a Bayesian approach to causal discovery that facilitates the use of Bayes decision rules to make discoveries with certainty. Both of these research frontiers are largely undeveloped, especially in terms of applications to real peer review data. Thus, this dissertation marks a starting point for new, impactful avenues of research in statistical methods and applications in peer review.

BIBLIOGRAPHY

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541. Publisher: Institute of Mathematical Statistics.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. Google-Books-ID: ztXL21Xd8v8C.

Athey, S. (2018). The Impact of Machine Learning on Economics. In *NBER Chapters*, pages 507–547. National Bureau of Economic Research, Inc.

Bailar, J. C. I. and Patterson, K. (1985). Journal Peer Review. *New England Journal of Medicine*.

Bar, H. Y. and Lillard, D. R. (2012). Accounting for heaping in retrospectively reported event data – a mixture-model approach. *Statistics in Medicine*, page 19.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., Boeck, P. D., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Ho, T. H., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Zandt, T. V., Vazire, S., Watts,

D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1):6.

Berger, J. O. and Sellke, T. (1987). Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence. *Journal of the American Statistical Association*, 82(397):112–122. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Bertrand, M. and Mullainathan, S. (2003). Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. Working Paper 9873, National Bureau of Economic Research.

Blau, F. D., Currie, J. M., Croson, R. T. A., and Ginther, D. K. (2010). Can Mentoring Help Female Assistant Professors? Interim Results from a Randomized Trial. *American Economic Review*, 100(2):348–352.

Bopp, M. and Faeh, D. (2008). End-digits preference for self-reported height depends on language. *BMC Public Health*, 8(1):342.

Borgman, C. L. and Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, 36(1):2–72. _eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/aris.1440360102.

Bornmann, L., Mutz, R., and Daniel, H.-D. (2010). A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants. *PLoS ONE*, 5(12).

Carlin, B. P. and Chib, S. (1995). Bayesian Model Choice Via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3):473–484. ad,m.

Casella, G. and Berger, R. L. (1987). Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem. *Journal of the American Statistical Association*, 82(397):106–111. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Chen, W., Drton, M., and Wang, Y. S. (2019). On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980. Publisher: Oxford Academic.

Chen, Z. and Chan, L. (2013). Causality in Linear Nongaussian Acyclic Models in the Presence of Latent Gaussian Confounders. *Neural Computation*, 25(6):1605–1641.

Cinelli, C. and Hazlett, C. (2019). Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, n/a(n/a).

Coates, T.-N. (2013). What We Mean When We Say 'Race Is a Social Construct'. Section: U.S.

Coleman, N. (2020). Why We're Capitalizing Black. *The New York Times*.

Cooper, G. F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347.

Coveney, J., Herbert, D. L., Hill, K., Mow, K. E., Graves, N., and Barnett, A. (2017). 'Are you siding with a personality or the grant proposal?': observations on how peer review panels function. *Research Integrity and Peer Review*, 2(1):19.

Cover, T. M. and Thomas, J. A. (2012). *Elements of Information Theory*. John Wiley & Sons. Google-Books-ID: VWq5GG6ycxMC.

Cunningham, W. H., Cunningham, I. C. M., and Green, R. T. (1977). The Ipsative Process to Reduce Response Set Bias. *The Public Opinion Quarterly*, 41(3):379–384. Publisher: [Oxford University Press, American Association for Public Opinion Research].

Curtice, J. (2009). Neither Representative nor Accountable: First-Past-the-Post in Britain. In *Duverger's Law of Plurality Voting: The Logic of Party Competition in Canada, India, the United Kingdom and the United States*, Studies in Public Choice, pages 27–45. Springer, New York, NY.

Cyert, R. M. and DeGroot, M. H. (1987). Bayesian Decision Theory. In Cyert, R. M. and DeGroot, M. H., editors, *Bayesian Analysis and Uncertainty in Economic Theory*, pages 7–26. Springer Netherlands, Dordrecht.

Diamond, G. A. and Forrester, J. S. (1983). Clinical Trials and Statistical Verdicts: Probable Grounds for Appeal. *Annals of Internal Medicine*, 98(3):385–394. Publisher: American College of Physicians.

Dovidio, J. F. and Gaertner, S. L. (2000). Aversive Racism and Selection Decisions: 1989 and 1999. *Psychological Science*, 11(4):315–319.

Drury, C. G. and Sinclair, M. A. (1983). Human and Machine Performance in an Inspection Task. *Human Factors*, 25(4):391–399.

Ebert-Uphoff, I. and Deng, Y. (2012). Causal Discovery for Climate Research Using Graphical Models. *Journal of Climate*, 25(17):5648–5665. Publisher: American Meteorological Society Section: Journal of Climate.

Eblen, M. K., Wagner, R. M., RoyChowdhury, D., Patel, K. C., and Pearson, K. (2016). How Criterion Scores Predict the Overall Impact Score and Funding Outcomes for National Institutes of Health Peer-Reviewed Applications. *PLOS ONE*, 11(6):e0155060.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Institute of Mathematical Statistics Monographs. Cambridge University Press, Cambridge.

Erosheva, E., Grant, S. L., Chen, M.-C., Lindner, M. D., Nakamura, R., and Lee, C. J. (2020a). Data and Reproducibility Code Supplement to "NIH Peer Review: Criterion Scores Completely Account for Racial Disparities in Overall Impact Scores". Publisher: OSF.

Erosheva, E., Walton, E. C., and Takeuchi, D. T. (2007). Self-Rated Health among Foreign- and US-Born Asian Americans: A Test of Comparability. *Medical care*, 45(1):80–87.

Erosheva, E. A., Grant, S., Chen, M.-C., Lindner, M. D., Nakamura, R. K., and Lee, C. J. (2020b). NIH peer review: Criterion scores completely account for racial disparities in overall impact scores. *Science Advances*, 6(23):eaaz4868. Publisher: American Association for the Advancement of Science Section: Research Article.

Espeland, W. N. and Stevens, M. L. (1998). Commensuration as a Social Process. *Annual Review of Sociology*, 24(1):313–343.

Evans, M. and Swartz, T. (1995). Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems. *Statistical Science*, 10(3):254–272.

Fang, F. C., Bowen, A., and Casadevall, A. (2016). NIH peer review percentile scores are poorly predictive of grant productivity. *eLife*, 5:e13323.

Feurer, I. D., Becker, G. J., Picus, D., Ramirez, E., Darcy, M. D., and Hicks, M. E. (1994). Evaluating Peer Reviews: Pilot Testing of a Grading Instrument. *JAMA*, 272(2):98–100. Publisher: American Medical Association.

Fleurence, R. L., Forsythe, L. P., Lauer, M., Rotter, J., Ioannidis, J. P., Beal, A., Frank, L., and Selby, J. V. (2014). Engaging Patients and Stakeholders in Research Proposal Review: The Patient-Centered Outcomes Research Institute. *Annals of Internal Medicine*, 161(2):122.

Friedman, N., Goldszmidt, M., and Wyner, A. (1999). On the Application of The Bootstrap for Computing Confidence Measures on Features of Induced Bayesian Networks. *AISTATS*, page 6.

Gallo, S. (2021). Grant Peer Review Scoring Data with Criteria Scores. Publisher: figshare type: dataset.

Gallo, S. A., Sullivan, J. H., and Glisson, S. R. (2016). The Influence of Peer Reviewer Expertise on the Evaluation of Research Funding Applications. *PLOS ONE*, 11(10):e0165147.

García-Donato, G. and Chen, M.-H. (2005). CALIBRATING BAYES FACTOR UNDER PRIOR PREDICTIVE DISTRIBUTIONS. *Statistica Sinica*, 15(2):359–380. Publisher: Institute of Statistical Science, Academia Sinica.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A., Lee, D., and Guo, J. (2015). Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543. Publisher: American Educational Research Association.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Bornkamp, B., Maechler, M., and Hothorn, T. (2020). mvtnorm: Multivariate Normal and t Distributions.

Ginther, D. K., Basner, J., Jensen, U., Schnell, J., Kington, R., and Schaffer, W. T. (2018). Publications as predictors of racial and ethnic differences in NIH research awards. *PLOS ONE*, 13(11):e0205929.

Ginther, D. K., Haak, L. L., Schaffer, W. T., and Kington, R. (2012). Are race, ethnicity, and medical school affiliation associated with NIH R01 type 1 award probability for physician investigators? *Academic Medicine: Journal of the Association of American Medical Colleges*, 87(11):1516–1524.

Ginther, D. K., Kahn, S., and Schaffer, W. T. (2016). Gender, Race/Ethnicity, and National Institutes of Health R01 Research Awards: Is There Evidence of a Double Bind for Women of Color? *Academic medicine : journal of the Association of American Medical Colleges*, 91(8):1098–1107.

Ginther, D. K., Schaffer, W. T., Schnell, J., Masimore, B., Liu, F., Haak, L. L., and Kington, R. (2011). Race, Ethnicity, and NIH Research Awards. *Science*, 333(6045):1015–1019.

Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press. Google-Books-ID: Gr0wCwAAQBAJ.

Goldstein, H. (2011). *Multilevel Statistical Models*. John Wiley & Sons.

Gráda, C. Ó. (2006). Dublin Jewish Demography a Century Ago. *THE ECONOMIC AND SOCIAL REVIEW*, page 26.

Grant, D. (2016). The essential economics of threshold-based incentives: Theory, estimation, and evidence from the Western States 100. *Journal of Economic Behavior & Organization*, 130:180–197.

Grant, S., Meila, M., Erosheva, E., and Lee, C. (2020). Refinement: Measuring Informativeness of Ratings in the Absence of a Gold Standard. page 54.

Green, J. G., Calhoun, F., Nierzwicki, L., Brackett, J., and Meier, P. (1989). Rating intervals: an experiment in peer review. *The FASEB Journal*, 3(8):1987–1992.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464.

Greiner, D. J. and Rubin, D. B. (2011). Causal Effects of Perceived Immutable Characteristics. *Review of Economics and Statistics*, 93(3):775–785.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008). A Kernel Statistical Test of Independence. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 585–592. Curran Associates, Inc.

Gronau, Q. F., Singmann, H., and Wagenmakers, E.-J. (2020). bridgesampling: An R Package for Estimating Normalizing Constants. *Journal of Statistical Software*, 92(1):1–29. Number: 1.

Hargens, L. and Herting, J. (1990). Neglected considerations in the analysis of agreement among journal referees. *Scientometrics*, 19(1-2):91–106.

Harsanyi, J. C. (1978). Bayesian Decision Theory and Utilitarian Ethics. *The American Economic Review*, 68(2):223–228. Publisher: American Economic Association.

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, 46(6):1251–1271.

Higginson, A. D. and Munafò, M. R. (2016). Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions. *PLOS Biology*, 14(11):e2000995.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, 15(3):199–236.

Hodson, G., Dovidio, J. F., and Gaertner, S. L. (2002). Processes in Racial Discrimination: Differential Weighting of Conflicting Information. *Personality and Social Psychology Bulletin*, 28(4):460–471.

Hoppe, T. A., Litovitz, A., Willis, K. A., Meseroll, R. A., Perkins, M. J., Hutchins, B. I., Davis, A. F., Lauer, M. S., Valantine, H. A., Anderson, J. M., and Santangelo, G. M. (2019). Topic choice contributes to the lower rate of NIH awards to African-American/black scientists. *Science Advances*, 5(10):eaaw7238.

Hoyer, P. O. and Hyttinen, A. (2009). Bayesian Discovery of Linear Acyclic Causal Models. page 9.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 689–696. Curran Associates, Inc.

Hoyer, P. O., Shimizu, S., and Kerminen, A. J. (2006). Estimation of linear, non-gaussian causal models in the presence of confounding latent variables.

Hu, L. and Kohler-Hausmann, I. (2020). What's Sex Got To Do With Fair Machine Learning? *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 513–513. arXiv: 2006.01770.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. (2012). Learning Linear Cyclic Causal Models with Latent Variables. *Journal of Machine Learning Research*, page 53.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634.

Hyvärinen, A., Karhunen, J., and Oja, E. (2004). *Independent Component Analysis*. John Wiley & Sons. Google-Books-ID: 96D0ypDwAkkC.

Hyvärinen, A. and Smith, S. M. (2013). Pairwise Likelihood Ratios for Estimation of Non-Gaussian Structural Equation Models. *Journal of Machine Learning Research*, 14(Jan):111–152.

Iacus, S. M., King, G., and Porro, G. (2012). Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis*, 20(1):1–24.

Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Steudel, B., and Schölkopf, B. (2012). Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31.

Janzing, D. and Schölkopf, B. (2010). Causal Inference Using the Algorithmic Markov Condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194. Conference Name: IEEE Transactions on Information Theory.

Jayasinghe, U. W., Marsh, H. W., and Bond, N. (2001). Peer Review in the Funding of

Research in Higher Education: The Australian Experience. *Educational Evaluation and Policy Analysis*, 23(4):343–364.

Jayasinghe, U. W., Marsh, H. W., and Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 166(3):279–300.

Johnson, V. E. (2008). Statistical analysis of the National Institutes of Health peer review system. *Proceedings of the National Academy of Sciences of the United States of America*, 105(32):11076–11080.

Kahneman, D. (2013). *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, 1st pbk. ed edition. OCLC: ocn834531418.

Kahneman, D. and Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, 64(6):515–526.

Kalisch, M., Hauser, A., Maechler, M., Colombo, D., Entner, D., Hoyer, P., Hyttinen, A., Peters, J., Andri, N., Perkovic, E., Nandy, P., Ruetimann, P., Stekhoven, D., Schuerch, M., and Eigenmann, M. (2019). pcalg: Methods for Graphical Models and Causal Inference.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

Klesges, R. C., Debon, M., and Ray, J. W. (1995). Are self-reports of smoking rate biased? Evidence from the Second National Health and Nutrition Examination Survey. *Journal of Clinical Epidemiology*, 48(10):1225–1233.

Kocaoglu, M., Dimakis, A. G., Vishwanath, S., and Hassibi, B. (2016). Entropic Causal Inference. *arXiv:1611.04035 [cs, math, stat]*. arXiv: 1611.04035.

Komatsu, Y., Shimizu, S., and Shimodaira, H. (2010). Assessing Statistical Reliability of LiNGAM via Multiscale Bootstrap. In Diamantaras, K., Duch, W., and Iliadis, L. S., editors, *Artificial Neural Networks – ICANN 2010*, Lecture Notes in Computer Science, pages 309–314, Berlin, Heidelberg. Springer.

Lamont, M. (2009). *How Professors Think*. Harvard University Press. Google-Books-ID: slK0xmSu33MC.

Langfeldt, L. (2001). The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome. *Social Studies of Science*, 31(6):820–841.

Lauer, M. (2020). Anonymizing Peer Review for the NIH Director's Transformative Research Award Applications – NIH Extramural Nexus.

Lauer, M. S., Danthi, N. S., Kaltman, J., and Wu, C. (2015). Predicting Productivity Returns on Investment. *Circulation Research*, 117(3):239–243.

Lauer, M. S. and Nakamura, R. (2015). Reviewing Peer Review at the NIH. *New England Journal of Medicine*, 373(20):1893–1895.

Lee, C. J. (2012). A Kuhnian Critique of Psychometric Research on Peer Review. *Philosophy of Science*, 79(5):859–870.

Lee, C. J. (2015). Commensuration Bias in Peer Review. *Philosophy of Science*, 82(5):1272–1283.

Lee, C. J. and Moher, D. (2017). Promote scientific integrity via journal peer review data. *Science*, 357(6348):256–257.

Lee, C. J., Sugimoto, C. R., Zhang, G., and Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.22784.

Lenk, P. (2009). Simulation Pseudo-Bias Correction to the Harmonic Mean Estimator of Integrated Likelihoods. *Journal of Computational and Graphical Statistics*, 18(4):941–960.

Li, D. and Agha, L. (2015). Research funding. Big names or big ideas: do peer-review panels select the best science proposals? *Science (New York, N.Y.)*, 348(6233):434–438.

Lindner, M. D. and Nakamura, R. K. (2015). Examining the Predictive Validity of NIH Peer Review Scores. *PLoS ONE*, 10(6).

Lindner, M. D., Torralba, K. D., and Khan, N. A. (2018). Scientific productivity: An exploratory study of metrics and incentives. *PLOS ONE*, 13(4):e0195321.

Little, R. J. A. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons. Google-Books-ID: BemMDwAAQBAJ.

Llorente, F., Martino, L., Delgado, D., and Lopez-Santiago, J. (2021). Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *arXiv:2005.08334 [cs, stat]*. arXiv: 2005.08334.

Lopez-Paz, D., Muandet, K., and Recht, B. (2015). The Randomized Causation Coefficient. *J. Mach. Learn. Res.*, 16(1):2901–2907.

Lynn, M., Flynn, S. M., and Helion, C. (2013). Do consumers prefer round prices? Evidence from pay-what-you-want decisions and self-pumped gasoline purchases. *Journal of Economic Psychology*, 36:96–102.

Maine State Legislature Staff (2019). Ranked Choice Voting in Maine | Maine State Legislature.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. University of Minnesota Press, Minneapolis, MN, US. Pages: x, 149.

Minton, P. D., Raiffa, H., and Schlaifer, R. (1961). Applied Statistical Decision Theory.

Mitrovic, J., Sejdinovic, D., and Teh, Y. W. (2018). Causal Inference via Kernel Deviance Measures. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 6986–6994. Curran Associates, Inc.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Scholkopf, B. (2016). Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *Journal of Machine Learning Research*, page 102.

Morera, O. F. and Dawes, R. M. (2006). Clinical and statistical prediction after 50 years: a dedication to Paul Meehl. *Journal of Behavioral Decision Making*, 19(5):409–412. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.538.

Nakamura, R. (2019). Testing of 2 Application Ranking Approaches at the National Institutes of Health Center for Scientific Review | Peer Review Congress.

Narasimhan, B., Koller, M., Johnson, S. G., Hahn, T., Bouvier, A., Kiêu, K., and Gaure, S. (2020). cubature: Adaptive Multivariate Integration over Hypercubes.

NIH Staff (2009). Get a Handle on Changes from the Enhancing Peer Review Process.

NIH Staff (2012). Scoring System and Procedure.

NIH Staff (2013). Additional Scoring Guidance for Research Applications.

NIH Staff (2016a). Definitions of Criteria and Considerations for Research Project Grant (RPG/R01/R03/R15/R21/R34) Critiques.

NIH Staff (2016b). Scoring Guidance.

NIH Staff (2019a). Enhancing Peer Review at NIH - Scoring and Review Changes.

NIH Staff (2019b). List of NIH Institutes, Centers, and Offices.

NIH Staff (2019c). A Pilot Study of Half-Point Increments in Scoring | NIH Center for Scientific Review.

NIH Staff (2020). Anonymizing Peer Review for the NIH Director's Transformative Research Award Applications – NIH Extramural Nexus.

NIH Staff (2021a). All Other CSR Special Emphasis Panels | NIH Center for Scientific Review.

NIH Staff (2021b). CSR Data & Evaluations | NIH Center for Scientific Review.

NIH Staff (2021c). Integrated Review Groups | NIH Center for Scientific Review.

NIH Staff (2021d). NIH Research Project Grant Program (R01).

NIH Staff (2021e). Understand Paylines and Percentiles | NIH: National Institute of Allergy and Infectious Diseases.

Norton, M. I., Sommers, S. R., Vandello, J. A., and Darley, J. M. (2006). Mixed motives and racial bias: The impact of legitimate and illegitimate criteria on decision making. *Psychology, Public Policy, and Law*, 12(1):36–55.

Norton, M. I., Vandello, J. A., and Darley, J. M. (2004). Casuistry and Social Category Bias. *Journal of Personality and Social Psychology*, 87(6):817–831.

O'Neil, K. H., Purdy, M., Falk, J., and Gallo, L. (1999). The Dysphagia Outcome and Severity Scale. *Dysphagia*, 14(3):139–145.

Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*, 82(4):669–688.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

Pearl, J. and Verma, T. (1987). The Logic of Representing Dependencies by Directed Graphs. *AAAI*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., and Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*, page 6.

Peters, J. and Bühlmann, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228. Publisher: Oxford Academic.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053.

Plug, C. (1977). Number Preferences in Ratio Estimation and Constant-Sum Scaling. *The American Journal of Psychology*, 90(4):699–704.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass. OCLC: ocm61285753.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. SAGE.

Rothenhäusler, D., Ernest, J., and Bühlmann, P. (2016). Causal inference in partially linear structural equation models. *arXiv:1607.05980 [math, stat]*. arXiv: 1607.05980.

Schmidt, F. L. and Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, 1(2):199–223. Place: US Publisher: American Psychological Association.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., and Clark, L. (1991). RATING SCALES NUMERIC VALUES MAY CHANGE THE MEANING OF SCALE LABELS. *Public Opinion Quarterly*, 55(4):570–582.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012). On Causal and Anticausal Learning. *International Conference on Machine Learning*, page 8.

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons. Google-Books-ID: pIAZBwAAQBAJ.

Sen, A. and Sen, B. (2014). Testing independence and goodness-of-fit in linear models. *Biometrika*, 101(4):927–942. Publisher: Oxford Academic.

Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K., and Wainwright, M. (2014). When is it Better to Compare than to Score? *arXiv:1406.6618 [cs, stat]*. arXiv: 1406.6618.

Shimizu, S. and Bollen, K. (2014). Bayesian Estimation of Causal Direction in Acyclic Structural Equation Models with Individual-specific Confounder Variables and Non-Gaussian Distributions. *Journal of Machine Learning Research*, 15:2629–2652.

Shimizu, S., Hoyer, P. O., Hyvarinen, A., and Kerminen, A. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, page 28.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvarinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. (2011). DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. *Journal of Machine Learning Research*, page 24.

Shortliffe, E. H. and Sepúlveda, M. J. (2018). Clinical Decision Support in the Era of Artificial Intelligence. *JAMA*, 320(21):2199–2200. Publisher: American Medical Association.

Simonsohn, U. (2013). Just Post It: The Lesson From Two Cases of Fabricated Data Detected by Statistics Alone. *Psychological Science*, 24(10):1875–1888.

Smaldino, P. E. and McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(9):160384.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., and Woolrich, M. W. (2011). Network modelling methods for FMRI. *NeuroImage*, 54(2):875–891.

Spiegelhalter, D. (2003). Ranking institutions. *The Journal of Thoracic and Cardiovascular Surgery*, 125(5):1171–1173. Publisher: Elsevier.

Spirtes, P., Glymour, C., and Scheines, R. (1993). Discovery Algorithms for Causally Sufficient Structures. In Spirtes, P., Glymour, C., and Scheines, R., editors, *Causation, Prediction, and Search*, Lecture Notes in Statistics, pages 103–162. Springer, New York, NY.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, Prediction, and Search*. MIT Press.

Spirtes, P. and Zhang, K. (2016). Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1):3.

Stegle, O., Janzing, D., Zhang, K., Mooij, J. M., and Schölkopf, B. (2010). Probabilistic latent variable models for distinguishing between cause and effect. *Neural Information Processing Systems*, page 9.

Stevenson, M. T. and Doleac, J. L. (2019). Algorithmic Risk Assessment in the Hands of Humans. page 72.

Sun, X., Janzing, D., and Schölkopf, B. (2006). Causal Inference by Choosing Graphs with Most Plausible Markov Kernels. In *ISAIM*.

Sun, X., Janzing, D., and Schölkopf, B. (2007). Distinguishing between cause and effect via kernel-based complexity measures for conditional distributions. In *ESANN 2007 Proceedings*.

Teh, Y. W., Newman, D., and Welling, M. (2007). A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. In Schölkopf, B., Platt, J. C., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 1353–1360. MIT Press.

Thorngate, W., Dawes, R. M., and Foddy, M. (2009). *Judging merit*. Psychology Press, New York. OCLC: ocn263497897.

Uhlmann, E. L. and Cohen, G. L. (2005). Constructed Criteria: Redefining Merit to Justify Discrimination. *Psychological Science*, 16(6):474–480.

Uhlmann, E. L. and Cohen, G. L. (2007). "I think it, therefore it's true": Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, 104(2):207–223.

van Rooyen, S., Black, N., and Godlee, F. (1999). Development of the Review Quality Instrument (RQI) for Assessing Peer Reviews of Manuscripts. *Journal of Clinical Epidemiology*, 52(7):625–629.

Wang, J., Veugelers, R., and Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8):1416–1436.

Wang, Y. S. and Drton, M. (2018). High-Dimensional Causal Discovery Under non-Gaussianity. *arXiv:1803.11273 [stat]*. arXiv: 1803.11273.

Warner, E. T., Carapinha, R., Weber, G. M., Hill, E. V., and Reede, J. Y. (2016). Faculty Promotion and Attrition: The Importance of Coauthor Network Reach at an Academic Medical Center. *Journal of General Internal Medicine*, 31(1):60–67.

Wilcoxon, F. (1946). Individual Comparisons of Grouped Data by Ranking Methods. *Journal of Economic Entomology*, 39(2):269–270. Publisher: Oxford Academic.

Wolpert, R. L. and Schmidler, S. C. (2012). Alpha-Stable Limit Laws for Harmonic Mean Estimators of Marginal Likelihoods. *Statistica Sinica*, 22(3):1233–1251. Publisher: Institute of Statistical Science, Academia Sinica.

Yu, R., Silber, J. H., and Rosenbaum, P. R. (2020a). Matching Methods for Observational Studies Derived from Large Administrative Databases. *Statistical Science*, 35(3):338–355. Publisher: Institute of Mathematical Statistics.

Yu, S., Drton, M., and Shojaie, A. (2020b). Directed Graphical Models and Causal Discovery for Zero-Inflated Data. *arXiv:2004.04150 [stat]*. arXiv: 2004.04150.

Zhang, K. and Hyvärinen, A. (2009). On the Identifiability of the Post-nonlinear Causal Model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 647–655, Arlington, Virginia, United States. AUAI Press. event-place: Montreal, Quebec, Canada.

Appendix A

The Chapter 2 project originated in 2014 when Drs. Carole Lee and Elena Erosheva won the Peer Review Challenge hosted by NIH and CSR, leading to a contract granting access to a limited, confidential, de-identified data set constructed from NIH records. These data were released specifically for the purpose of studying the relationship between preliminary criterion scores and preliminary Overall Impact scores at the level of individual reviewers. In Chapter 2, we evaluate whether racial disparities in preliminary Overall Impact scores of assigned reviewers can be explained by criterion scores, other application and applicant characteristics, and differences in commensuration practices that are based on race. This Chapter of the Appendix relays fine-grained details of the study data and its selection, the statistical methodology used, as well as reproducibility results based on the data that were made available to the public (Erosheva et al., 2020a).

A.1 Variable Definitions

Table A.1: Study variables.

Type	Name	Description
Dependent Variable	Preliminary Overall Impact	Integer Score from 1 to 9; smaller is better
Variables of Interest		
Race	PI Black	1 for Black, 0 for white; self-reported
Preliminary Criteria		

	Significance	Integer Score from 1 to 9; smaller is better
	Investigator	Integer Score from 1 to 9; smaller is better
	Innovation	Integer Score from 1 to 9; smaller is better
	Approach	Integer Score from 1 to 9; smaller is better
	Environment	Integer Score from 1 to 9; smaller is better
Structural Covariates		
CSR Peer Review		
	IRG	Integrated Review Group
	SRG	Scientific Review Group
	Institute/Center	NIH Institute/Center making funding decisions
Other Indicators		
	Application ID	Encrypted application indicator
	Applicant ID	Encrypted applicant/PI indicator
	Reviewer ID	Encrypted reviewer indicator
Other Covariates		
Applicant-Specific		
	Gender	F/M, self-reported
	Ethnicity	Hispanic/Latino or not, self-reported

	Career Stage	Early Stage (ES), Experienced, or Non-ES New Investigator
	Educational Degree	PhD, MD, MD/PhD, Other
	Terminal Degree Year	Year of most recent degree
	NIH Funding History	First NIH application, previously applied, or previously funded
	Geographic Location	Location of institution: Central, East, South, or West
	NIH Funding Bin	FY 2014 total institution NIH funding; 5 bins
	Institution Sector	Public, Private, or Other
	Graduate Education	1 if institution provides graduate education, 0 if not
	IPEDS Lookup	1 if institution in IPEDS database, 0 if not
	MSI Type	Minority Serving Institution type: HBCU, HSI, or otherwise
Application-Specific		
	Application Type	New or Renewal
	Solicitation Type	Request for Application, Program Announcement, Others
	Amended Status	Amended or not
	Multiple PIs	Yes or no
	Requested Costs	Funding dollars requested

Support Years	Support years requested, from 1 to 5
Council Year	2014-2016; year of review councils
Review Group Type	Standing Study Section, Recurring SEP or Nonrecurring SEP
Human Subjects	Acceptable, unacceptable, or inapplicable
Animal Subjects	Acceptable, unacceptable, or inapplicable
Child Code	Acceptable, unacceptable, or inapplicable
Gender Code	Acceptable, unacceptable, or inapplicable
Minority Code	Acceptable, unacceptable, or inapplicable

Table A.2: NIH's Descriptions for Overall Impact and Five Review Criteria (NIH Staff, 2016a).

Score	Description
Overall Impact	Reviewers will provide an Overall Impact/priority score to reflect their assessment of the likelihood for the project to exert a sustained, powerful influence on the research field(s) involved, in consideration of the following five core review criteria, and additional review criteria (as applicable for the project proposed).

Scored Criteria

Significance	Does the project address an important problem or a critical barrier to progress in the field? If the aims of the project are achieved, how will scientific knowledge, technical capability, and/or clinical practice be improved? How will successful completion of the aims change the concepts, methods, technologies, treatments, services, or preventative interventions that drive this field?
Investigators	Are the principal investigators, collaborators, and other researchers well suited to the project? If early stage investigators or new investigators are in the early stages of independent careers, do they have appropriate experience and training? If established, have they demonstrated an ongoing record of accomplishments that have advanced their field(s)?
Innovation	Does the application challenge and seek to shift current research or clinical practice paradigms by utilizing novel theoretical concepts, approaches or methodologies, instrumentation, or interventions? Are the concepts, approaches or methodologies, instrumentation, or interventions novel to one field of research or novel in a broad sense? Is a refinement, improvement, or new application of theoretical concepts, approaches or methodologies, instrumentation, or interventions proposed?
Approach	Are the overall strategy, methodology, and analyses well reasoned and appropriate to accomplish the specific aims of the project? Are potential problems, alternative strategies, and benchmarks for success presented? If the project is in the early stages of development, will the strategy establish feasibility and will particularly risky aspects be managed?

Environment	Will the scientific environment in which the work will be done contribute to the probability of success? Are the institutional support, equipment, and other physical resources available to the investigators adequate for the project proposed? Will the project benefit from unique features of the scientific environment, subject populations, or collaborative arrangements?

A.2 Study Data

This section describes the study data in full, including key information from Chapter 2 for completeness. The data come from the NIH IMPAC II (Information for Management, Planning, Analysis, and Coordination) grant data system from the council years 2014–2016. This study focused on Black-white disparities; we did not include 1,771 applications submitted by PIs whose race was American Indian or Alaskan, Asian, Native Hawaiian or Pacific Islander, or who indicated more than one race, as well as 8,648 applications for which PI race was withheld or unknown. Table A.1 summarizes all variables used in our analyses and their definitions. At the time of application, PI demographics are voluntarily reported by applicants.

In the full set of 54,740 applications, approximately 15% of the applications from Black and white PIs were missing information on PI gender, ethnicity (Hispanic/Latino or not), and educational degree, and were excluded from the study. Specifically, 232 were missing gender information, 7,409 were missing ethnicity information, and 1,639 were missing degree information. The remaining 46,226 applications—1,015 (or 2.2%) from Black PIs and 45,211 (or 97.8%) from white PIs—were evaluated by 19,197 unique reviewers who wrote 139,216 reviews. 73.7% of the reviewers reviewed in just one of the three council years, 2014-2016, for which we have data, while 22.9% reviewed in two and 3.4% reviewed in all three years. Because PIs can amend each application that is not funded initially and submit multiple applications for different projects, there are fewer unique PIs than applications—500 (2.5%) PIs are Black and 19,653 (97.5%) are white. Among these applications with no missing

data, there were 1,015 applications from Black PIs, which received 3,064 reviews from 2,322 unique reviewers (Table A.3). There were 45,211 applications from white PIs, which received 136,152 reviews from 19,100 unique reviewers.

Table A.3: Summary statistics for the FY 2014-2016 applicant pool.

Applicant Race	Applications	PIs	Reviewers	Review Records
Black	1,015	500	2,322	3,064
White	45,211	19,653	19,100	136,152
Total	46,226	20,153	19,197	139,216

Study codes—Human Subjects, Animal Subjects, Child, Gender, and Minority—are categorical variables that take on a number of values. For our analyses, all study codes were re-coded/coarsened to "Acceptable", "Unacceptable", or "Inapplicable", in order to avoid numerical estimation problems with rare categories and for ease of interpretability. Below, we describe only codes that occurred in our study data (for example, code 20—no exemption designated, so award cannot be processed—never occurs in our data and is not discussed). Links to the current NIH study codes are provided as hyperlinked URLs in the text for ease of reference.

For Human Subject codes (https://www.niaid.nih.gov/grants-contracts/human-subjects-involvement-codes), code 10 (no human subjects involved) was re-coded to "Inapplicable," code 44 (human subjects involved, SRG concerns) to "Unacceptable," and other codes (30—certified with no SRG concerns; 54—previous concerns resolved; and exemptions) to "Acceptable." For Animal Subjects (https://www.niaid.nih.gov/grants-contracts/research-animals-involvement-codes), code 10 (no animal subjects involved) was re-coded to "Inapplicable," code 44 (animal subjects involved, SRG concerns) was re-coded to "Unacceptable," and others (30—animals involved with no SRG concerns; 32—animals involved with SRG comments; 48—conditional award with terms and conditions; 54—previous concerns resolved)

to "Acceptable."

For Gender codes (https://www.niaid.nih.gov/grants-contracts/human-subjects-inclusion-codes), the categories of interest indicated whether or not women were knowingly included in the proposed study. Codes "1A" and "2A" were re-coded as "Acceptable," because they represent studies in which the researchers knowingly included women in the study design that were deemed acceptable. Codes "1U" and "2U" were re-coded as "Unacceptable," as they represent studies in which the researchers knowingly included women in the study design that were deemed unacceptable. The remaining applications—those whose proposed studies did not include human subjects, or did not knowingly include women—were re-coded as "Inapplicable."

The Minority codes (https://www.niaid.nih.gov/grants-contracts/human-subjects-inclusion-codes) and Child codes (the NIH Child Subjects codes for the data used in this study have since been updated to "Age codes" and can be found at https://grants.nih.gov/grants/funding/lifespan/review_codes.doc) are structured similarly to the Gender code and were re-coded analogously, with those applications in which minority subjects or child subjects were knowingly included being separated from the others in the re-coding. Note that, because the Human Subjects code indicates whether or not human subjects were included in the proposed study, the human subjects studies at the "Inapplicable" level of the Gender, Minority, and Child Subject codes are still recognized by our models as distinct from those without human subjects.

Finally, the NIH Funding Bin variable is determined by the amount of NIH R01 funding given to **all investigators** at an institution in 2014, and then split into 5 bins with roughly equal numbers of **Black PIs** in each bin. These bins are delineated in Table A.4.

A.2.1 Matching and Study Subsets Selection

While the observational units in our study are reviews, matching occurred at the application level because all covariates other than scores vary at the application or applicant level. We matched exactly on eight variables thought to be related to preliminary scores and

Table A.4: Dollar award ranges for NIH funding bins.

Institution Bin	NIH Awarded Dollars		Number of Black
	Minimum	Maximum	Awardees
1	$360,448,763	$593,400,359	200
2	$149,626,530	$360,448,762	203
3	$63,082,330	$149,626,529	206
4	$23,982,606	$63,082,329	194
5	$0	$23,982,605	212

award rates. Exact matching is a version of Coarsened Exact Matching (CEM, Iacus et al. (2012)); see proof in Appendix A.2.1. The matching variables, summarized in Table 2.2, are: contact PI's gender, ethnicity, career stage, educational degree, institution's NIH funding bin, application type, application's amended status (first submission or resubmission), and area of science as represented by the Integrated Review Group (IRG).

CEM has several desirable properties including congruence (i.e., matching is performed on the data space rather than in a space of some metric such as the propensity score), relatively easy and flexible implementation, and Monotonic Imbalance Bounding (specifying the coarsening level for each variable automatically bounds the imbalance allowed for each covariate) (Iacus et al., 2012). It is recommended that coarsening levels be chosen based on subject matter knowledge about the measurement and the likely importance of different covariates (Iacus et al. (2012), p. 16). Due to the high number of categorical covariates, our choice was to carry out exact matching on eight key variables and implement complete coarsening for the rest. This choice achieved a desirable trade-off between improved balance and sample size: tests revealed that matching on an additional covariate—even a coarsened one—led to substantial constraints on sample size. Our matching procedure improved balance on all the matching variables and on most other applicant- and application-specific

covariates (Table A.5).

Matched Subset Selection

This section relates the details of the matched subset selection algorithm, which were constructed to:

1. Maximize the number of applications in the matched data set;

2. Maximize the number of reviews of applications in the matched data set;

3. Enforce exact matching on the 8 matching variables (the remainder are "fully coarsened" and thus trivially matched in a CEM); and,

4. Respect the constraint that no more than four applications in the entire matched data set may come from the same reviewer. This constraint was implemented due to the sensitive nature of the data, in order to ensure the privacy and confidentiality of reviewers. We refer to this constraint as the "confidentiality constraint."

Prior to selection of the matched subset, a near 1:2 matching was performed: each application from a Black PI was matched with up to two applications from white PIs on the eight matching variables (see Table 2.2). Algorithm 3 details the greedy Coarsened Exact Matching algorithm used. For each Black application, this algorithm seeks a perfect match from at least one white application, though it will accept an imperfect match on 6 or more variables as a second match. The large number of applications in the full data set (Table A.3) prohibited searching through numerous matchings for an optimal matching that might have admitted a better trade-off between sample size and matching strictness. Future work using peer review data with many covariates and large sample sizes should take advantage of recent advances in matching algorithms, such as Yu et al. (2020a).

Next, review records were selected for each set of matched applications using Algorithm 4. This second-stage algorithm was demanded by the confidentiality constraint—including all

Algorithm 3 Matching

Generate a dictionary D with Black application IDs as keys and lists (length 0–2) of white applications IDs as values. Let X be the table with rows corresponding to all applications in the full data set, B (W) the set of Black (White) application row IDs in the full data set, and M the set of column indices of the 8 matching covariates.

```
1: procedure ExactMatching(X, B, W, M)
2:     Create a dictionary D with B as keys and values initialized as empty lists
3:     for i in B do
4:         l ← length(D[i])
5:         while length(D[i]) < 2 do
6:             for j in W do
7:                 m ← ∑(X[i, M] = X[j, M])
8:                 if m = 8 OR (l = 1 AND m ≥ 6) then
9:                     D[i] ← [D[i], j]
10:                    W ← W \ j
11:                    l ← l + 1
12:                end if
13:            end for
14:            if l = length(D[i]) then
15:                NEXT (iteration of loop)
16:            end if
17:        end while
18:    end for
19:    return D
20: end procedure
```

reviews of all applications selected by Algorithm 3 would yield more than four repetitions of many reviewers in the selected data.

This algorithm attempts to maximize the number of records in the data set by minimizing the number of tuples of matched applications that must be discarded while respecting the confidentiality constraint. It does this by initially selecting only one review for as many applications in a matched tuple as possible (to ensure each application can be represented in the matched subset) before greedily selecting other reviews. It prioritizes the selection of reviews of Black applications, since they are scarce. The algorithm replaces matched white applications whose reviews are all ineligible under the confidentiality constraint if possible. As a result, no Black applications and just nine white applications were discarded (and replaced) by this algorithm when the study data were selected.

Random Subset Selection

The random subset selection algorithm, Algorithm 5, was designed to generate a representative set of reviews of white applications while respecting the confidentiality constraint. This constraint prevents us from simply selecting all reviews of a set of white applications chosen uniformly at random.

A naive approach to random subset selection would replace each application with no reviewers satisfying the confidentiality constraint with another randomly selected application—or, equivalently, sample applications under the confidentiality constraint until $2n$ applications are represented. However, this would systematically bias the sample by including applications that were reviewed by less prolific reviewers at higher rates. By replacing applications with no reviewers satisfying the confidentiality constraint with applications that were also reviewed by at least one experienced reviewer with five or more reviews in the data set, we mitigate this bias. Additionally, our random white sample may not include reviewers that reviewed numerous Black applications. However, because there are so few Black applications, we judged it more important to maximize the number of reviews of Black applications in the sample.

Algorithm 4 Matched Subset Selection

Select the matched subset from the matching dictionary D produced by Algorithm 3. Let R be the set of reviewer IDs and W the set of white applications in the full data.

1: **procedure** MATCHEDSUBSETSELECTION(D, R, W)
2: Create a dictionary E with R as keys and values initialized to zero
3: Create a list of review IDs L
4: **for** i in keys(D) **do**
5: $bflag \leftarrow FALSE$
6: $X \leftarrow [\,]$
7: **while** X is empty **do**
8: Consider next review of application i; call this review k and its reviewer l
9: **if** $E[l] < 4$ **then**
10: $E[l] \leftarrow E[l] + 1$
11: $X \leftarrow [X, k]$
12: $bflag \leftarrow TRUE$
13: **end if**
14: **end while**
15: **if** !$bflag$ **then** ▷ No reviews of Black application i were eligible
16: NEXT (iteration of loop on line 4) ▷ Skip the matched white applications
17: **end if**
18: $wflag1 \leftarrow FALSE$
19: **for** j in $D[i]$ **do**
20: $wflag2 \leftarrow FALSE$
21: **while** looping through reviews k with reviewers l of application j **do**
22: **if** $E[l] < 4$ **then**
23: $E[l] \leftarrow E[l] + 1$
24: $L \leftarrow [L, k]$
25: $wflag1 \leftarrow TRUE$
26: $wflag2 \leftarrow TRUE$
27: **end if**
28: **end while**
29: **if** !$wflag2$ **then**
30: Search $W \setminus$ values(D) for a new white application j that matches Black application i on at least 6 matching variables and return to line 20 if successful
31: **end if**
32: **end for**
33: **if** $wflag1$ **then**
34: **for all** reviews k with reviewers l of applications $[i, D[i]] \setminus [X, L]$ **do**
35: **if** $E[l] < 4$ **then**
36: $E[l] \leftarrow E[l] + 1$
37: $X \leftarrow [X, k]$
38: **end if**
39: **end for**
40: $L \leftarrow [L, X]$

Algorithm 4 Matched Subset Selection (Continued)

```
        else                          ▷ No reviews of matched white applicants were eligible
            Decrement the value of review X's reviewer in E by 1
        end if
    end for
    return L, E
end procedure
```

Coarsened Exact Matching with Exact Matching on a Subset of Covariates

In this section, we prove that exact matching on selected variables is a version of Coarsened Exact Matching (CEM). In this section only, we use the language of potential outcomes and treatment effects for the ease of exposition and to reflect the language used in Iacus et al. (2012). We remind the reader that our analysis uses the associative language of adjusted racial disparities, rather than the causal language of racial biases.

Exact matching on a strict subset of the covariates is CEM with "full coarsening" on the unmatched covariates. One may verify this by checking that the proofs in Iacus et al. (2012) do not assume that each coarsened variable has at least 2 coarsened levels. For completeness, we provide an in-depth example proof for the boundedness of SATT (Sample Average Treatment Effect on the Treated) estimation error under exact matching.

Equation (7) of Iacus et al. (2012) states that as long as the true treatment effect is a Lipschitz function of the observed covariates and the maximum width of a coarsening interval (or set, for categorical variables) is $\epsilon_j < \infty$ for the j-th covariate, then the SATT absolute estimation error is bounded. For this to hold with exact matching on a covariate subset, we simply require that the range of each non-matching variable be bounded by some ϵ_j for an appropriate metric in addition to the Lipschitz requirement. Bounded range is equivalent to having a bounded coarsening interval width, which is nearly always the case in practice; furthermore, if no such ϵ_j exists then no coarsening will yield finite ϵ_j and thus the requirement is not restrictive. Explicitly, for continuous non-matching variables we require the range to be finite, for ordinal non-matching variables we require a finite number of levels,

Algorithm 5 Random Subset Selection

Select the random subset from portion of the full data set remaining after matched subset selection (Algorithm 4). Let L be the list of matched subset review IDs and E be the reviewer count dictionary produced by Algorithm 4, with n the number of Black applications represented in L and F the full data set with applications represented in L removed.

1: **procedure** RandomSubsetSelection(L, E, F, n)
2: Select $2n$ white applications from L uniformly at random, call this set T
3: Initialize an empty list of reviews R
4: **for all** applications i in T **do**
5: **if** a reviewer l of a review k of application i satisfies $E[l] < 4$ **then**
6: $R \leftarrow [R, k]$
7: $E[l] \leftarrow E[l] + 1$
8: NEXT (iteration of loop on line 4)
9: **else**
10: $flag \leftarrow FALSE$
11: **while** $!flag$ **do**
12: Select a white application a uniformly at random from $F \setminus T$
13: **if** there exists a review k_1 with reviewer l_1 of application a such that $E[l_1] \geq 4$ and a review k_2 with reviewer l_2 of application a such that $E[l_2] < 4$ **then**
14: $R \leftarrow [R, k]$
15: $E[l] \leftarrow E[l] + 1$
16: $T \leftarrow [T, a] \setminus i$
17: **end if**
18: **end while**
19: **end if**
20: **end for**
21: **for all** applications i in T **do**
22: **for all** reviews k not in R with reviewers l of application i **do**
23: **if** $E[l] < 4$ **then**
24: $R \leftarrow [R, k]$
25: $E[l] \leftarrow E[l] + 1$
26: **end if**
27: **end for**
28: **end for**
29: **return** R
30: **end procedure**

and for categorical non-matching variables we impose no restriction as the distance between any two levels of a categorical variable can be said to be one. All of these conditions hold for a finite population, which is the case here as we are estimating sample treatment effects. For the exact matching variables, we have $\epsilon_j = 0$ by definition.

Let $X_1, \ldots, X_k$ be the observed variables with $X_{i1}, \ldots, X_{ik}$ the values for the i-th observation. Then let the potential outcome at treatment level 0 be $Y_i(0)$ (treatment level zero corresponds to being white) for the i-th individual. Under the conditional ignorability assumption, we can write

$$Y_i(0) = g_0(X_{i1}, \ldots, X_{ik})$$

where we have omitted possible mean-zero noise for ease of exposition (this is justifiable because conditional ignorability guarantees that this noise is conditionally independent of the treatment given the covariates, so it does not contribute any bias to our estimator; it only adds to its variability). Since we always observe $Y_i(1)$ for treated individuals, our estimate of the treatment effect for the treated is

$$\widehat{TE}_i = Y_i(1) - \widehat{Y_1(0)}$$

and the true treatment effect for the treated is

$$TE_i = Y_i(1) - Y_i(0).$$

For the difference-in-means estimator, $\widehat{Y_i(0)}$ is also the observed outcome for a matched untreated unit. We then have

$$\begin{aligned} TE_i &= \widehat{TE}_i + \widehat{Y_i(0)} - Y_i(0) \\ &= \widehat{TE}_i + g_0(\widetilde{X}_1) - g_0(X_i) \end{aligned}$$

where $\tilde{X}_i$ represents the covariate vector for the observation matched to the i-th treated unit, with observed outcome $\widehat{Y_i(0)}$. Taking an average over the treated units, we get

$$SATT = \widehat{SATT} + \frac{1}{n_T} \sum_{i=1}^{n_T} g_0(\tilde{X}_i) - g_0(X_i).$$

Now, assume that g_0 is Lipschitz in the sense that replacing X_j with any $\tilde{X}_j$ within the coarsened matching caliper of width ϵ_j and holding all other variables fixed changes the value of g_0 by at most L_j for any j. Then for any i,

$$|g_0(\widetilde{X}_t) - g_0(X_i)| \leq \sum_j L_j \epsilon_j$$

and, as desired, it immediately follows that

$$|SATT - \widehat{SATT}| \leq \sum_j L_j \epsilon_j.$$

For another example, consider Section 4.1 of Iacus et al. (2012), regarding the maximum imbalance bound. For the non-matching variables, the imbalance bound is simply 1 (although in practice the imbalance is typically much less than 1, as can be seen from Table A.5), and it remains true that specifying a coarsening for one variable does not affect the imbalance bound for other variables because the maximum possible imbalance under the L_1 distance is 1. As noted by Iacus et al. (2012), this property stands in contrast to certain Mahalanobis distance matching methods where the user demands a certain sample size from the matching, in which additionally imposing an upper bound on balance or coarsening for one variable can increase the imbalance bound for other variables.

Balance Analysis

Matching seeks to make the joint distributions of the covariates in the treated and untreated groups similar; this similarity is known *balance*. Because the curse of dimensionality prohibits visualizing or easily checking the degree of overlap between joint distributions of covariates, balance is usually assessed for marginal distributions (Ho et al., 2007). Our measure of balance is L1 overlap; letting P and Q be probability distributions on a domain $\mathcal{X}$ with respect to measure μ, the L1 overlap is on minus the total variation distance:

$$1 - \int_{\mathcal{X}} |P - Q| \, d\mu$$

Rather than simply assessing differences in means and standard deviations by covariate, the L1 overlap measures how different the entire empirical distributions of the variables are between the Black and white subsets, a distinction emphasized by Ho et al. (2007). The L1 approach to measuring overlap—recommended by Iacus et al. (2012)—is superior because the entire distribution of a covariate is of concern when a model is misspecified, and model misspecification is one of the main concerns matching is designed to address.

Table A.5 displays the L1 overlap percentage for the random and matched subsets, as well as the percentage increase in overlap for the matched subset. Exact matching variables are in bold. Note that overlap for matching variables (Table 2.2) may not be exactly 1 because the number of reviews per application varies, so the distributions in the white and Black matched subsets may differ slightly. Overall, the overlap noticeably improved for the 8 exact matching covariates, and improved moderately for most other covariates. Note that the overlap for Institution Sector and Institution Lookup variables declined slightly after matching. While this is not a major concern because the overlap is still quite high, we note that CEM does not guarantee that imbalance will improve on every covariate, merely that there is an upper bound on imbalance for each covariate.

Table A.5: L1 overlap for control covariates; exact matching variables are in bold.

Variable	Random	Matched	Increase
IRG	0.77	0.99	28%
Amended Status	0.95	1.00	5%
Application Type	0.90	1.00	11%
Gender	0.91	1.00	10%
Ethnicity	0.98	1.00	2%
Educational Degree	0.88	0.98	11%
Career Stage	0.76	0.99	31%
NIH Funding Bin	0.93	0.99	6%

Institute/Center	0.83	0.90	8%
SRG	0.89	0.91	2%
Institution Sector	0.95	0.94	-1%
Graduate Education	1.00	1.00	0%
IPEDS Lookup	0.98	0.96	-2%
MSI Type (council year 2015)	0.97	0.97	0%
Solicitation Type	0.97	0.99	1%
Multiple PIs	0.98	0.99	1%
Support Years	0.96	0.99	2%
Council Year	0.96	0.96	0%
Review Group Type	0.94	0.94	0%
Human Subjects	0.76	0.95	24%
Animal Subjects	0.85	0.99	16%
Gender Code	0.90	0.98	9%
Minority Code	0.90	0.99	9%
Child Code	0.90	0.98	9%
Terminal Degree Year	0.78	0.85	9%
NIH Funding History	0.89	0.97	9%
Geographic Location	0.87	0.89	2%
(Log) Requested Costs	0.83	0.91	9%

A.3 Multilevel Modeling

This section includes supplementary details that support the exposition of the models in Section 2.3.

A.3.1 Hausman Test for SRG Effects

In this section we detail the Hausman test that helps justify our use of random, rather than fixed, effects for SRGs. The null hypothesis for the Hausman test is that the fixed-effects coefficients are consistent in both the SRG random effects and SRG fixed effects models, and consequently that the SRG random effect model estimates are efficient. It is shown in Hausman (1978) that the covariance between an asymptotically efficient estimator and its difference from a different consistent but inefficient estimator is asymptotically zero. A simple chi-squared test can be constructed based on this result with the following statistic and null distribution

$$\left(\widehat{\beta_{RE}} - \widehat{\beta_{FE}}\right)^T \left(\text{Cov}\left(\widehat{\beta_{RB}}\right) - \text{Cov}\left(\widehat{\beta_{FB}}\right)\right)^{-1} \left(\widehat{\beta_{RB}} - \widehat{\beta_{FB}}\right) \sim \chi^2_p$$

where RE stands for random effects and FE for fixed effects, p is the number of fixed-effect coefficients estimated in the model (and the number of degrees of freedom of the chi-square distribution), and the inverse is a pseudo-inverse. Because under local alternatives to the null (i.e. slight model misspecification) the test statistic has a non-central chi-square distribution (i.e. larger expected value), we reject the null if the test statistic is too large. For the matched subset analysis of the full data set (the analysis presented in the main text), this statistic was 28.76 on 115 degrees of freedom, with p-value approximately 1. The test fails to reject the null hypothesis of no endogeneity, and since random effects align well with our substantive knowledge of the SRG, we elect to fit SRG random effects. In practice, both random effects and fixed effects models lead to the same substantive conclusions and very similar coefficient estimates for our data.

A.3.2 Model Diagnostics

We assess the fit of our commensuration model (2.8) to the data, checking for apparent violations of the linear mixed-effects modeling assumptions detailed in Section 2.3.2. For mixed-effects models, residual analysis constitutes the bulk of the diagnostics. There are

three main types of residuals in mixed-effects models: conditional residuals, marginal residuals, and BLUPs (best linear unbiased predictors). If y is the outcome, x the observed covariates, $\hat{\beta}$ the estimated fixed-effect coefficients, and $\hat{\gamma}$ the best linear unbiased predictor of the random effects, then:

- the conditional residuals are $e_c = y - \hat{\beta}x - \hat{\gamma}$,
- the marginal residuals are $e_m = y - \hat{\beta}x$, and
- the BLUPs are $\hat{\gamma} = e_m - e_c$.

Linear dependence means we need examine only two of these residuals. We examined normal quantile-quantile plots for the conditional residuals and BLUPs for the three random intercepts included in the commensuration model (not shown). The conditional residuals and BLUPs displayed wider tails than a normal distribution, indicating some excess kurtosis that was not enough to raise any concerns given the large sample size and robustness of linear regression to deviations from normality of residuals. Furthermore, no conditional residuals nor BLUPs displayed evidence of heteroscedasticity or dependence on the main covariates of interest (i.e., race, the preliminary criterion scores, requested costs—the only purely continuous variable—and terminal degree year, which was modeled as a categorical variable). For both residual types, residual analysis plots indicated that assumptions of homoscedasticity, independence between residuals and covariates, and approximate Gaussianity are not unreasonable.

A.4 Final (Post-Discussion) Scores

This analysis is for applications that have reached the SRG discussion stage. Not all reviewers change their criterion and Overall Impact scores after discussion (Table A.6). Among post-discussion reviews, 20% saw a change in both the Overall Impact score and at least one criterion score post-discussion, 27% saw a change in the Overall Impact score but not in the criterion scores, 4% saw a change in the criterion scores but not in the Overall Impact

score, and 49% saw no change in any scores. From the available data, it is impossible to tell what motivated post-discussion changes in scores, or lack thereof. We also cannot model the potentially complex dependencies between scores for the same application that arise through SRG discussion.

Table A.6: Percentage of Reviews with Scores Changed

	Criterion Scores Changed	Criterion Scores Unchanged
Impact Score Changed	0.20	0.27
Impact Score Not Changed	0.04	0.49

Score change behavior after discussion; discussed reviews only.

We replicate the racial disparity analysis for final Overall Impact scores (Table A.7), with the caveats above meaning that these results cannot be considered conclusive. Our conclusions are largely the same: controlling for final criterion scores accounts for essentially all racial disparities in final Overall Impact scores. Two important differences are that the racial disparities observed without controlling for final criterion scores are not as large as those for preliminary scores, and that reviewer random intercept and residual variability are substantially smaller than for preliminary scores. This latter phenomenon could result from SRG discussions leading reviewers to consensus (Fleurence et al., 2014).

A.5 Reproducibility

Because of the sensitive nature of individual-level data, a reduced data set that contains the same reviews but fewer covariates is available for public use (Erosheva et al., 2020a). This public-use data set includes all of the covariates of interest (applicant race, preliminary criterion and Overall Impact scores), the structural covariates (PI ID, application ID, reviewer ID, administering institute ID, IRG ID, and SRG ID), the matching variables (contact PI's gender, ethnicity, career stage, educational degree, institution's NIH funding bin, application

Table A.7: Selected parameter estimates from models 2.1–2.4, final Overall Impact scores.

	Model 2.4	Model 2.3	Model 2.2	Model 2.1
Non-Structural Covariates	$\emptyset$	X	Z	X, Z
Matching?	No	Yes	No	Yes
Race Fixed Effect				
Coefficient	0.458*	0.131	0.163	0.044
(Std. Err.)	(0.079)	(0.080)	(0.043)	(0.045)
p-value	< 0.005	0.101	< 0.005	0.329
Random Effects				
Reviewer Std. Dev.	0.247	0.184	0.178	0.190
PI Std. Dev.	0.937	0.890	0.471	0.471
SRG Std. Dev.	0.370	0.366	0.185	0.169
Residual Std. Dev.	0.907	0.917	0.586	0.595

Race coefficient estimates, their effect sizes, and variance components estimates from four hierarchical linear models for **final** Overall Impact scores. Model 1 controls for structural covariates and is fit to the random subset; Model 2 controls for structural and matching covariates and is fit to the matched subset; Model 3 controls for structural covariates and criterion scores and is fit to the random subset; Model 4 controls for structural, matching covariates, and criterion scores and is fit to the matched subset. Coefficient estimates for control variables are not shown. Significance * is reported for $p < 0.005$.

type, application's amended status, and the area of science represented by the Integrated Review Group), as well as the final Overall Impact score.

Here, we reproduce results of the multilevel analysis of racial disparities in preliminary Overall Impact scores from Table 2.3 and of commensuration practices in Table 2.4 using the public use data set. We also reproduce Figure 2.5 for the matched subset of the public-use reduced-covariates data set.

A.5.1 Racial Disparities

Table A.8 presents multilevel modeling results from the public-use data that are analogous to those reported in Table 2.3. We find that the race coefficient estimates from Models 1 and 2 (which do not control for preliminary criterion scores) obtained from public-use data are positive, statistically significant, and very similar in magnitude to those reported in Table 2.3. Once preliminary criterion scores are included (Models 3 and 4), the race coefficient estimates in Table A.8 become practically and statistically insignificant, as in the confidential data set. While the main results in the confidential and the public use data sets are strikingly similar, we note that the random intercept variability for PIs and SRGs is somewhat larger for Model 2 of the public-use data set (Table A.8) than for Model 2 of the confidential study data set (Table 2.3). This is because fewer PI-specific covariates are included in the model as there are fewer covariates available in the public data set to explain PI-specific variability.

A.5.2 Commensuration Practices

Table A.9 contains relevant parameter estimates from the linear commensuration models that were fit using the public data. For the matched subset, which is less susceptible to model misspecification (Ho et al., 2007; Erosheva et al., 2007), the signs and magnitude of coefficient estimates for the interaction coefficients are strikingly similar to our main results on commensuration practices from Table 2.4. Commensuration differences across all preliminary criterion scores remain small for the public data: expected differences for Black applications

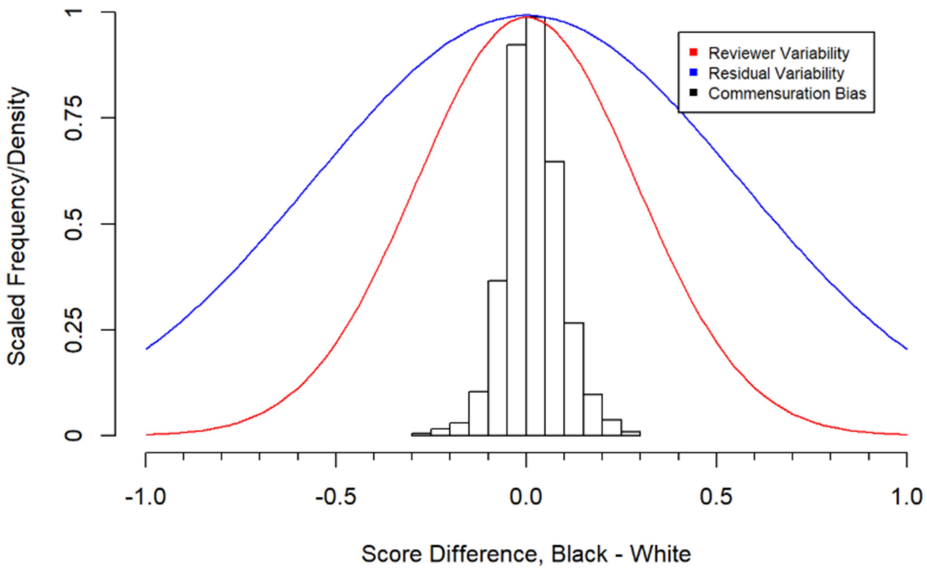


Figure A.1: Public-use data set: Distribution of estimated expected preliminary Overall Impact score differences due to commensuration (histogram) and distributions of reviewer intercepts (red line) and model residuals (blue line), under the commensuration model (2.8) for the matched subset (Table A.9).

in the preliminary Overall Impact score of 0.1 or more as result of commensuration practices are rare (see Figure A.1).

Table A.8: Selected parameter estimates from models 2.1–2.4, public-use data.

	Model 2.4	Model 2.3	Model 2.2	Model 2.1
Non-Structural Covariates	$\emptyset$	X	Z	X, Z
Matching?	No	Yes	No	Yes
Race Fixed Effect				
Coefficient	0.700*	0.431*	0.031	0.014
(Std. Err.)	(0.064)	(0.057)	(0.017)	(0.017)
p-value	< 0.005	< 0.005	0.071	0.412
Effect Size	0.533	0.333	0.054	0.025
Random Effects				
Reviewer Std. Dev.	0.490	0.501	0.274	0.288
PI Std. Dev.	0.836	0.769	0.093	0.097
SRG Std. Dev.	0.306	0.304	0.084	0.087
Residual Std. Dev.	1.312	1.296	0.567	0.563

Public-use data set: Race coefficient estimates, their effect sizes, and variance components estimates from four hierarchical linear models for preliminary Overall Impact scores. Model 1 controls for structural covariates and is fit to the random subset; Model 2 controls for structural and matching covariates and is fit to the matched subset; Model 3 controls for structural covariates and criterion scores and is fit to the random subset; Model 4 controls for structural, matching covariates, and criterion scores and is fit to the matched subset. Coefficient estimates for control variables are not shown. Significance * is reported for $p < 0.005$.

Table A.9: Selected parameter estimates, commensuration model (2.8), public-use data.

Variable	Estimate (Std. Err.)	P-Val.
Fixed Effects		
Significance	0.263 (0.008)*	< 0.005
Investigator	0.060 (0.011)*	< 0.005
Innovation	0.132 (0.008)*	< 0.005
Approach	0.604 (0.007)*	< 0.005
Environment	0.019 (0.011)	0.090
PI Race = Black	-0.031 (0.047)	0.508
Significance * PI Black	-0.035 (0.012)	0.008
Investigator * PI Black	0.017 (0.017)	0.337
Innovation * PI Black	-0.021 (0.014)	0.125
Approach * PI Black	0.045 (0.012)*	< 0.005
Environment * PI Black	-0.009 (0.018)	0.630
Random Effects		
Reviewer Intercepts Std. Dev.	0.288	
PI Intercepts Std. Dev.	0.092	
SRG Intercepts Std. Dev.	0.088	
Residual Variability Std. Dev.	0.562	

Public-use data set: Preliminary criterion, race, commensuration (race-criterion interaction) coefficients, and variance components estimates for preliminary Overall Impact scores on $n = 7471$ reviews of 2566 applications (matched subset) and $n = 8595$ reviews of 3045 applications (random subset). Control variables (coefficient estimates are not shown) are the matching variables. Significance * is reported for $p < 0.005$.

A.6 Differences from Published Paper

This section briefly summarizes the major differences between Chapter 2 and the published paper, Erosheva et al. (2020b).

- In Section 2.3.3, we explicitly establish the assumptions that would lend our model estimates causal interpretations, and discuss how the data fit and deviate from those assumptions. We frame the causal effect of interest in terms of the causal diagram in Figure 2.2 (Pearl, 1995).

- In Erosheva et al. (2020b), the random subset is treated as a sensitivity scenario in which the $\tilde{X}$ covariates are only controlled for via regression coefficients and random effects. In that paper, analyses in Tables 2.3 and 2.4 are performed on both the matched and random subsets of the data separately. In Chapter 2, we frame matching as a nonparametric regression adjustment technique that is employed in conjunction with standard linear regression controls whenever we wish to adjust for applicant- and application-specific covariates.

- We compute and interpret the Robustness Value (Cinelli and Hazlett, 2019) for model coefficients that indicate commensuration differences, showing that the modest evidence in favor of racial commensuration differences could be nullified by a weakly influential unobserved factor—or greatly strengthened by a modestly influential latent factor.

- We treat each of models 1–4 from Erosheva et al. (2020b) separately, explaining the causal interpretation of each in the context of Figure 2.2.

- We replicate the Bayesian p-value analysis of Casella and Berger (1987) for the $p < 0.005$ cutoff suggested by Benjamin et al. (2018) to help the reader understand the importance of this lower significance threshold. We also clarify that the standard for statistical significance in this paper is still less strict than that of practical significance.

- In Erosheva et al. (2020b), the analysis of commensuration practices was largely deferred to the supplement. We present the commensuration analysis in full in Section 2.4.

- Section 2.1, the introduction to Chapter 2, frames the chapter as part of our larger study of peer review. The description of NIH's peer review process from the introduction to Erosheva et al. (2020b) has been moved to Chapter 1.

- Algorithms 3, 4, and 5 have been formalized in Section A.2 of Appendix 2. They were not presented formally in the supplement to Erosheva et al. (2020b).

- We use a Bonferroni adjustment when analyzing p-values for interaction coefficients in Table 2.4. Interpretations do not differ from Erosheva et al. (2020b).

- We include an update on the status of NIH anonymization studies.

- Tables A.2 and A.1, which are in the main text of Erosheva et al. (2020b), are located in Appendix A to keep the focus of Chapter 2 on methodology.

- Effect sizes are no longer reported in tables of regression coefficients.

Appendix B

This Appendix provides proofs of propositions from Chapter 3.

B.1 Proposition 1: Refinement Decomposition

We require the following Lemma, which is a generalization of (Cover and Thomas, 2012) Chapter 2, Exercise 19:

Lemma 1. *For a finite mixture P of m distributions with mutually disjoint support, $P = \sum_{i=1}^{m} \lambda_i P_i$ with $\sum_i \lambda_i = 1$ and $\lambda_i \geq 0$ for all i,*

$$H(P) = -\sum_i \lambda_i \log \lambda_i + \sum_i \lambda_i H(P_i) \tag{B.1}$$

Proof: We have

$$\begin{aligned} H(P) &= -\sum_{i=1}^{m} \sum_{k \in supp(P_i)} \lambda_i P_i(k) \log(\lambda_i P_i(k)) \\ &= -\sum_i \lambda_i \log(\lambda_i) + \sum_i \lambda_i H(P_i). \end{aligned}$$

The second equality follows from the log of a product equaling the sum of the logs and rearrangement of terms. We can now derive the refinement decomposition.

Proof of Proposition 1: Taking $0 \log(0) = 0$ and δ_s the degenerate distribution with $\mathrm{P}(s) = 1$,

$$\begin{aligned} H(R_t(\mathbf{Y})) &= H(\sum_{s \in \mathbb{S}_t} p(B_t(s)) \delta_s) \\ &= -\sum_{s \in \mathbb{S}_t} p(B_t(s)) \log(p(B_t(s))) \end{aligned} \tag{B.2}$$

by Lemma 1 since B_t generates a partition and hence empirical distributions with disjoint support, and the entropy of degenerate distributions δ_s is zero. Applying the same lemma to the unrounded $\mathbf{Y}$ yields, in a similar fashion,

$$H(Y) = -\sum_{s\in\mathbb{S}_t} p(B_t(s)) \log(p(B_t(s))) + \sum_{s\in\mathbb{S}_t} p(B_t(s)) H(\mathbf{Y}_s).$$

Combining with (B.2) finishes the proof.

B.2 Proposition 2: Extrema and Range

Proposition B.2 follows from Proposition 1. It is clear from Proposition 1 that Entropic Refinement achieves its minimum of 0 when there is only one unique observed score value in any basin of attraction, so that $H(\mathbf{Y}_s) = 0$ for all s for which $p(B(s)) > 0$. With a bit more effort we can construct the maximum. Because within-basin entropy only depends on the scores in a single basin, the $H(\mathbf{Y}_s)$ terms in (3.4) can be independently maximized. The uniform distribution over k elements yields maximum entropy $\log(k)$ for discrete distributions over k outcomes, so we must have within-basin uniformity. Finally, (3.4) is a weighted average, so the maximum occurs when only maximum-sized basins have nonzero weight. Thus the maximizing empirical distributions put probability mass only on the maximum-sized basins. Maximum Entropic Refinement therefore occurs when the empirical score distribution is uniform on maximum-size basins and there are no observed scores in other basins.

B.3 Proposition 3: Joint vs. Average Entropic Refinement

This section proves Proposition 3 for (WLOG) $t = 1$. We first prove that $r_E^{avg} \leq r_E^{joint}$. Recall that C is the number of criterion scores, so that

$$H(\mathbf{X}) \geq \max_{i\in[C]} H\left(\mathbf{X}^i\right) \geq \frac{1}{C}\sum_{i=1}^{C} H\left(\mathbf{X}^i\right) \tag{B.3}$$

where the first inequality is equality only if the criterion scores are deterministically related.

Now let $\mathbf{X}_s \equiv \left\{x \in \mathbf{X} \cap B_1^C(s)\right\}$ where $B_1^C(s)$ is the C-dimensional basin of attraction to $s \in \mathbb{S}_1^C$. Let i index criteria, so that $\mathbf{X}_s^i$ is the set of scores on criterion i in the rounding basin of s. We then write the joint entropy in terms of the decomposition (3.4) and substitute (B.3):

$$r_E^{joint}(\mathbf{X}) = \sum_{s \in \mathbb{S}_1^C} p(\mathbf{X}_s) H(\mathbf{X}_s) \tag{B.4}$$

$$\geq \sum_{s \in \mathbb{S}_1^C} p(\mathbf{X}_s) \frac{1}{C} \sum_{i \in [C]} H\left(\mathbf{X}_s^i\right) \tag{B.5}$$

$$= \frac{1}{C} \sum_{i \in [C]} \sum_{s \in \mathbb{S}_1^C} p(\mathbf{X}_s^i) H\left(\mathbf{X}_s^i\right) \tag{B.6}$$

$$= r_E^{avg}(\mathbf{X}) \tag{B.7}$$

(the third line follows simply from distributing $p(\mathbf{X}_s)$ inside the sum over i and rearranging sums).

For the second inequality, $Cr_E^{avg} \geq r_E^{joint}$, denote p^i the empirical probability distribution associated to $\mathbf{X}^i$, for $i \in [C]$. Assume initially that $C = 2$.

Consider first the case $\mathbf{X}^1 \perp\!\!\!\perp \mathbf{X}^2$, that is, $p = p^1 p^2$. In this case, we have that $H(\mathbf{X}) = H\left(\mathbf{X}^1\right) + H\left(\mathbf{X}^2\right)$. Moreover, because in this case $R\left(\mathbf{X}^1\right) \perp\!\!\!\perp R\left(\mathbf{X}^2\right)$, we also have $H(R(\mathbf{X})) = H\left(R\left(\mathbf{X}^1\right)\right) + H\left(R\left(\mathbf{X}^2\right)\right)$. It follows then, immediately, that $r_E^{joint} = \frac{1}{2}\left(r_E\left(\mathbf{X}^1\right) + r_E\left(\mathbf{X}^2\right)\right) = r_E^{avg}$.

Let us now consider the general case, for which

$$H(\mathbf{X}) = H\left(\mathbf{X}^1\right) + H\left(\mathbf{X}^2\right) - I\left(\mathbf{X}^1, \mathbf{X}^2\right),$$

where the last term denotes the mutual information (Cover and Thomas, 2012) between the two criterion scores. Similarly,

$$H(R(\mathbf{X})) = H\left(R\left(\mathbf{X}^1\right)\right) + H\left(R\left(\mathbf{X}^2\right)\right) - I\left(R\left(\mathbf{X}^1\right), R\left(\mathbf{X}^2\right)\right),$$

and therefore

$$Cr_E^{joint} = r_E\left(\mathbf{X}^1\right) + r_E\left(\mathbf{X}^2\right) - \left(I\left(\mathbf{X}^1, \mathbf{X}^2\right) - I\left(R\left(\mathbf{X}^1\right), R\left(\mathbf{X}^2\right)\right)\right).$$

We show now that the last term, $I\left(\mathbf{X}^1, \mathbf{X}^2\right) - I\left(R\left(\mathbf{X}^1\right), R\left(\mathbf{X}^2\right)\right)$, is non-negative. For this we use the data processing inequality (Cover and Thomas, 2012), which states that if random variables X, Y, Z form a Markov chain, in other words if $X \perp\!\!\!\perp Z|Y$, then $I(X,Y) \geq I(X,Z)$. We have that $\mathbf{X}^1 \perp\!\!\!\perp R(\mathbf{X}^2)|\mathbf{X}^2$, from which we derive that $I\left(\mathbf{X}^1, \mathbf{X}^2\right) \geq I\left(\mathbf{X}^1, R\left(\mathbf{X}^2\right)\right)$. Moreover, $R\left(\mathbf{X}^2\right) \perp\!\!\!\perp R\left(\mathbf{X}^1\right)|\mathbf{X}^1$, from which we have that $I\left(\mathbf{X}^1, R\left(\mathbf{X}^2\right)\right) \geq I\left(R\left(\mathbf{X}^1\right), R\left(\mathbf{X}^2\right)\right)$, which concludes the proof. The proof for $C > 2$ follows by induction.

Appendix C

This is the Appendix to Chapter 4. It provides proofs of claims in the text, as well as supplemental analyses.

C.1 Bayesian IGCI

IGCI (Janzing et al., 2012) assumes that if $X \in [0,1]$ causes $Y \in [0,1]$ and $Y = f(X)$ (a deterministic relationship), then f is orthogonal to P_X in the sense that

$$\begin{aligned} &\int_0^1 \log|f'(x)| P_X(x)dx - \int_0^1 \log|f'(x)|dx \int_0^1 P_X(x)dx \\ =&\int_0^1 \log|f'(x)| P_X(x)dx - \int_0^1 \log|f'(x)|dx \\ =&0. \end{aligned} \tag{C.1}$$

(C.1) does not hold when regions of high/low density for X also have high/low absolute slope for f. As with Pearson's correlation, orthogonality may occur when f' and P_X appear truly unrelated, but it may also hold when, for example, a notable positive association in one part of $[0,1]$ and a notable negative association in another part cancel out.

The authors describe this relationship as "independence," which we take issue with. Independence is a property of random variables; considering f and P_X to be random, as we do in our Bayesian framework, it is very difficult to conceive of a joint distribution for f, P_X in which the two are independent and yet (C.1) is even approximately satisfied all or much of the time. Instead, as we shall illustrate, mother distributions which satisfy (C.1)—even approximately—impose strong dependence between f and P_X. The remainder of this section explores two approaches to Bayesian IGCI which encode orthogonality into the model prior in different ways.

C.1.1 Strict Orthogonality

We begin with an example of Bayesian IGCI in which the model prior requires exact orthogonality between P_X and f (note that the notation and terminology from Section 4.2 of Chapter 4 will be used going forward). Since Janzing et al. (2012) consider the deterministic case, our exposition will take $\epsilon = 0$ and the model prior to be $\mathcal{P}_{P_X,f}$.

We construct our model prior via the factorization $\mathcal{P}_{P_X,f} = \mathcal{P}_{f|P_X}\mathcal{P}_{P_X}$: nature generates a distribution P_X from $\mathcal{P}_{P_X}$, and then selects f from $\mathcal{P}_f$ such that equation (C.1) holds. That is,

$$\mathcal{P}_{f|P_X}(f|P_X) \propto \mathcal{P}_f(f)\mathbf{1}\left(\int_0^1 \log|f'(x)|P_X(x)\,dx - \int_0^1 \log|f'(x)|\,dx = 0\right). \tag{C.2}$$

Thus, requiring orthogonality of P_X and f means that they are dependent.

Continuing equation (4.4), we have

$$\begin{aligned}
\mathrm{P}(\{X,Y\}_n|M_X) &= \int \mathrm{P}(\{X,Y\}_n|P_X,f)d\mathcal{P}_{P_X,f} \\
&= \int \mathrm{P}(\{Y\}_n|\{X\}_n,f)\mathrm{P}(\{X\}_n|P_X)d\mathcal{P}_{P_X,f} \\
&= \int_{f\perp P_X} \mathrm{P}(\{X\}_n|P_X)\mathrm{P}(\{Y\}_n|\{X\}_n,f)d\mathcal{P}_{P_X}d\mathcal{P}_f
\end{aligned} \tag{C.3}$$

where the third line follows from (C.2). Equation (C.3) demonstrates that the evidence for M_X is determined by the prior probability of mechanisms f that fit the data exactly, the prior probability of P_X that are orthogonal to f, and how well P_X fits the data $\{X\}_n$.

C.1.2 Approximate Orthogonality

One might object that (C.2) is too restrictive, that in nature it is highly unlikely that f and P_X are completely orthogonal. We can loosen the requirement of strict orthogonality by replacing the indicator in equation (C.2) with a function $g(|x|) : \mathbb{R}^+ \to \mathbb{R}^+$ that is non-increasing:

$$\mathcal{P}_{f|P_X}(f|P_X) \propto \mathcal{P}_f(f)g\left(\left|\int_0^1 \log|f'(x)|P_X(x)dx - \int_0^1 \log|f'(x)|dx\right|\right). \tag{C.4}$$

Supposing g induces a proper prior distribution, this model prior places higher probability on pairs (P_X, f) that are closer to orthogonality.

C.1.3 Examples

The examples we now provide are illustrative—they are not intended for use with real data. We believe that the orthogonality conditions demanded by Bayesian IGCI are too cumbersome for practical use. The integrals in (C.2) and (C.4), which could be difficult to evaluate precisely, must be computed at every step of an evidence approximation procedure. In the examples we present, the model priors are unrealistically simplistic and there is no noise ϵ (also unrealistic), which makes the computation tractable.

Strict Orthogonality

To make calculations in the strict orthogonality setting easily verifiable, we use a contrived setup in which the P_X and f that have support in the model prior have been chosen via Gram-Schmidt to be exactly orthogonal.

Suppose that M_X holds with $P_X = \mathrm{U}(0,1)$ and $f(x) = \sqrt[6]{x}$. Thus, while X is uniform and its density is orthogonal to f, Y is heavily skewed left and its distribution is clearly dependent on f.

We define $\mathcal{P}_{P_X}$ as follows:

- $P_X(x) = 1 \equiv p_0(x)$ with probability $1/2$ and
- $P_X(x) = \frac{48}{41}(1 - (x - \frac{3}{4})^2 - \frac{11}{72}(\log(x) + 1)) \equiv p_1(x)$ with probability $1/2$.

$\mathcal{P}_f$ shall be uniform over $f(x) = x^6$, $f(x) = x$, and $f(x) = \sqrt[6]{x}$. In this case, each of these three functions is orthogonal to both p_0 and p_1[1]. The joint prior thus places probability $1/6$ on each combination (P_X, f). The model prior under M_Y is identical.

[1]For monomial f, this will always be the case, as the log absolute slope is always an affine function of $\log(x)$.

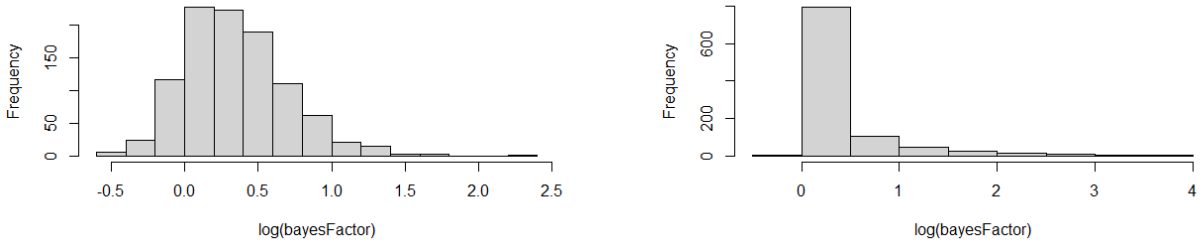
After observing data, only the true $f(x) = \sqrt[6]{x}$ fits the data deterministically under M_X. Thus

$$\begin{aligned} \mathrm{P}(\{X,Y\}_n|M_X) &= \int_{f \perp P_X} \mathrm{P}(\{Y\}_n|\{X\}_n, f)\mathrm{P}(\{X\}_n|P_X)d\mathcal{P}_{P_X}d\mathcal{P}_f \\ &= \frac{1}{6}p_0(\{X\}_n) + \frac{1}{6}p_1(\{X\}_n) \end{aligned} \tag{C.5}$$

Under M_Y, only the function $g(y) = y^6$ fits the data deterministically, so

$$\begin{aligned} \mathrm{P}(\{X,Y\}_n|M_Y) &= \int_{g \perp P_Y} \mathrm{P}(\{X\}_n|\{Y\}_n, g)\mathrm{P}(\{Y\}_n|P_Y)d\mathcal{P}_{P_y}d\mathcal{P}_g \\ &= \frac{1}{6}p_0(\{Y\}_n) + \frac{1}{6}p_1(\{Y\}_n) \end{aligned} \tag{C.6}$$

We then simulate the distribution of the Bayes Factor K_n for different sample sizes n, displayed in figure C.1.



(a) $n = 100$. (b) $n = 1000$.

Figure C.1: Distribution of log Bayes Factors for the strict orthogonality Bayesian IGCI model, over 1000 simulations.

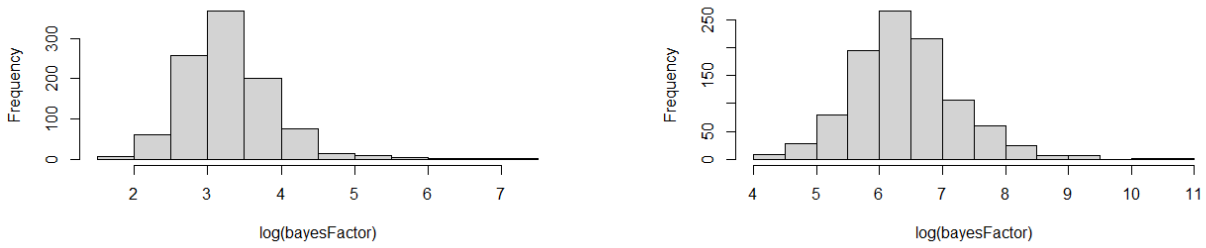
In this case, the true model prevails the majority of the time, though by a narrow margin (the log Bayes Factor is almost always less than 2 even for $n = 1000$). Other model priors yield log Bayes Factors that are usually less than zero even when M_X is the truth. We believe that requiring strict orthogonality will inevitably lead to contrived model prior specifications such as these, which one should not expect to yield accurate causal discovery.

Approximate Orthogonality

Imposing approximate orthogonality (C.4) facilitates a more realistic choice of model prior. We define $\mathcal{P}_{P_X}$ as follows: let $\mu \sim \mathrm{U}(0,1)$ and $\sigma \sim \mathrm{U}(0,1)$ independently; then define $P_X(x) \propto e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$ for $x \in [0,1]$ and zero otherwise. That is, X is a Gaussian truncated to the unit interval. We take $f(x) = x^{\exp(\beta)}$ with $\beta \in [-2,2]$, and—conditional on P_X—β is chosen with probability proportional to $C \equiv g(|\int_0^1 \log|f'(x)|P_X(x)dx - \int_0^1 \log|f'(x)|dx|)$ with $g(x) = (1+x)^{-2}$. In the deterministic realm, we can learn f exactly from a single sample with probability one (as long as $X_1 \neq 0$, (X_1, Y_1) uniquely identifies β).

With such a simple model, C can be approximated efficiently. Since $\log(f'(x)) = \beta + (e^\beta - 1)\log(x)$, we must approximate $\mathrm{E}[\log(X)]$ when X has a truncated Normal distribution in the unit interval. Via Teh et al. (2007), we can do this via a Taylor series expansion to arrive at $\mathrm{E}[\log(X)] \approx \log(\mathrm{E}[X]) - \frac{\mathrm{Var}(X)}{2\mathrm{E}[X]^2}$.

Figure C.2 shows the histogram of the log Bayes Factors for samples of size 10 and 100.



(a) $n = 10$.

(b) $n = 100$.

Figure C.2: Distribution of log Bayes Factors for the approximate orthogonality Bayesian IGCI model, over 1000 simulations and 1000 Monte Carlo samples from the prior for each simulation.

In both cases, the correct direction of causality is always inferred, with appropriately larger Bayes Factors in the larger-n case. In the correct causal direction, the distributions P_X that fit $\{X\}_n$ well are also those that yield high values of C. Under M_Y, the wrong model,

the P_Y that fit $\{Y\}_n$ well yield small values of C, because they are far from orthogonal to g. Thus the evidence for M_X tends to be stronger than that for M_Y.

C.2 Decision Theory and Thresholds

Proposition 5. *Given $w_X > 0$, the cost of selecting M_X when M_Y generated the data, and $w_Y > 0$ defined similarly, the Bayes Decision Rule chooses $d = M_X$ when $\frac{P(M_X|\{X,Y\}_n)}{P(M_Y|\{X,Y\}_n)} > \frac{w_X}{w_Y}$ and $d = M_Y$ otherwise.*

Proof: we seek to minimize

$$\begin{aligned}
&w_X \mathrm{P}(M_Y, d = M_X|\{X,Y\}_n) + w_Y \mathrm{P}(M_X, d = M_Y|\{X,Y\}_n) \\
=&w_X \mathrm{P}(M_Y|\{X,Y\}_n)\mathrm{P}(d = M_X|\{X,Y\}_n) + w_Y \mathrm{P}(M_X|\{X,Y\}_n)\mathrm{P}(d = M_Y|\{X,Y\}_n).
\end{aligned}$$

It immediately follows that we set $\mathrm{P}(d = M_X|\{X,Y\}_n) = 1$ when

$$\begin{aligned}
&w_X \mathrm{P}(M_Y|\{X,Y\}_n) \leq w_Y \mathrm{P}(M_X|\{X,Y\}_n) \\
\iff &\frac{\mathrm{P}(M_X|\{X,Y\}_n)}{\mathrm{P}(M_Y|\{X,Y\}_n)} \geq \frac{w_X}{w_Y} \qquad \text{(C.7)}
\end{aligned}$$

and $\mathrm{P}(d = M_Y|\{X,Y\}_n) = 1$ otherwise. We can randomize d in the unlikely event that equality holds in (C.7).

Now suppose we admit the option to abstain, M_0, in addition to M_X and M_Y. Define $w_0 > 0$ the cost of abstaining from a causal discovery decision. We now seek to minimize

$$R(d) = w_X \mathrm{P}(M_Y, d = M_X|\{X,Y\}_n) + w_Y \mathrm{P}(M_X, d = M_Y|\{X,Y\}_n) + w_0 \mathrm{P}(d = M_0|\{X,Y\}_n)$$

Similar calculations yield the following proposition:

Proposition 6. *In addition to w_X and w_Y as defined in Proposition 5, let $w_0 > 0$ be the*

cost of abstaining. Then the Bayes Decision Rule is

$$\begin{aligned} d = M_0 &\iff w_0 < \min\{w_X P(M_Y|\{X,Y\}_n), w_Y P(M_X|\{X,Y\}_n)\} \\ d = M_X &\iff w_X P(M_Y|\{X,Y\}_n) < \min\{w_0, w_Y P(M_X|\{X,Y\}_n)\} \\ d = M_Y &\iff w_Y P(M_X|\{X,Y\}_n) < \min\{w_0, w_X P(M_Y|\{X,Y\}_n)\}. \end{aligned}$$

We require

$$w_0 < \frac{1}{2}\min\{w_X, w_Y\} \tag{C.8}$$

so that each of the three decisions is possible. When $w_X = w_Y \equiv w_1$, we abstain only when

$$\begin{aligned} & \frac{w_0}{w_1} < \min\{\mathrm{P}(M_Y|\{X,Y\}_n), \mathrm{P}(M_X|\{X,Y\}_n)\} \\ \implies & \frac{w_0/w_1}{1 - w_0/w_1} < \min\{\frac{\mathrm{P}(M_Y|\{X,Y\}_n)}{\mathrm{P}(M_X|\{X,Y\}_n)}, \frac{\mathrm{P}(M_X|\{X,Y\}_n)}{\mathrm{P}(M_Y|\{X,Y\}_n)}\} \\ \implies & \log\left(\frac{w_0/w_1}{1 - w_0/w_1}\right) < \log\left(\min\{\frac{\mathrm{P}(M_Y|\{X,Y\}_n)}{\mathrm{P}(M_X|\{X,Y\}_n)}, \frac{\mathrm{P}(M_X|\{X,Y\}_n)}{\mathrm{P}(M_Y|\{X,Y\}_n)}\}\right) \\ \implies & c \equiv -\log\left(\frac{w_0/w_1}{1 - w_0/w_1}\right) > \left|\log\left(\frac{\mathrm{P}(M_X|\{X,Y\}_n)}{\mathrm{P}(M_Y|\{X,Y\}_n)}\right)\right|. \end{aligned} \tag{C.9}$$

In the second line, we apply the monotonically increasing odds transformation to both sides and use the fact that the two posterior probabilities sum to 1. In the final line, we use the following facts:

- $w_0 < \frac{1}{2}w_1 \implies \frac{w_0/w_1}{1-w_0/w_1} < 1$, and
- when p, q are probabilities that sum to 1, $\log(p/q) = -\log(q/p)$.

When $p_{M_X} = p_{M_Y}$, equation (C.9) reduces to

$$c > |\log(K)|\,, \tag{C.10}$$

i.e. we only abstain when the log Bayes Factor magnitude is smaller than some threshold c.

C.3 Gaussian Pseudo-Bias

In this section of the appendix, we investigate the relationships among sample size, model complexity, and pseudo-bias of the evidence. In order to make the problem analytically tractable, the setup will be slightly unrealistic (though still of interest—Wolpert and Schmidler (2012) analyzes the same model in just one dimension and shows that the harmonic mean marginal likelihood estimator (Kass and Raftery, 1995) will typically converge very slowly in this model). We model the data as a set of n i.i.d. samples from a multivariate Gaussian distribution with known identity-multiple covariance and multivariate Gaussian prior with identity-multiple covariance for the mean (a Gaussian random intercepts model, but with many of the parameters of interest assumed known). That is, $X_i \overset{i.i.d.}{\sim} \mathrm{N}(\mu, \sigma^2 \mathrm{I}_d)$ with $\mu \sim \mathrm{N}(0, \gamma^2 \mathrm{I}_d)$ and σ^2, γ^2 known. Since the variances are known, we consider the data to be simply the sample mean $\bar{X}_n$, a sufficient statistic. Finally, we center the data per Section 4.2.1 so that $\bar{X}_n = 0$.

We can compute the evidence analytically for this model, allowing us to compare Monte Carlo approximation (Algorithm 1) to the truth. Analysis of the full data $\{X\}_n$ does not yield such clarity because if the samples X_i are independent conditional on μ, then they are dependent marginally, and vice-versa. This complicates characterization of the probability that a Monte Carlo approximation is downwardly pseudo-biased, as do here.

Clearly $\bar{X}_n|\mu \sim \mathrm{N}(\mu, \sigma^2/n)$. Noting that $\mathrm{E}[X_i] = 0$, $\mathrm{Var}(X_i) = \sigma^2 + \gamma^2$, and

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= \mathrm{E}[\mathrm{Cov}(X_i, X_j|\mu)] + \mathrm{Cov}(\mathrm{E}[X_i|\mu], \mathrm{E}[X_j|\mu]) \\ &= \gamma^2, \end{aligned}$$

we have $\bar{X}_n \sim \mathrm{N}(0, \gamma^2 + \sigma^2/n)$. Because the conditional likelihood will be highly peaked around its maximum (Kass and Raftery, 1995), we are concerned that Monte Carlo will fail to obtain samples from the prior distribution on μ ($P_\mu = \mathrm{N}(0, \gamma^2)$) near where this peak occurs and thus underestimate the true marginal likelihood. We therefore derive a lower bound on the probability that the Monte Carlo approximation of Algorithm 1 is smaller

than the truth and investigate when this lower bound is close to 1—i.e., when downward pseudo-bias is essentially assured.

The Monte Carlo approximation is guaranteed to be too small if every Monte Carlo sample $\mu^{(j)}$ satisfies $\mathrm{P}(\bar{X}_n|\mu^{(j)}) < \mathrm{P}(\bar{X}_n)$. Given $\bar{X}_n = 0$,

$$\mathrm{P}(\bar{X}_n|\mu) < \mathrm{P}(\bar{X}_n) \iff \mu^T\mu > d\left(\frac{\sigma^2}{n}\right)^d \log\left(\frac{\gamma^2 + \sigma^2/n}{\sigma^2/n}\right). \tag{C.11}$$

This result squares with our intuition: as n grows, $\mathrm{P}(\bar{X}_n|\mu)$ becomes more peaked around $\mu = 0$, so $\mu \neq 0$ yield relatively smaller conditional likelihoods as n grows. The expected value of $\mu^T\mu$ grows linearly in d, but since each component of $\bar{X}_n$ is independent and the likelihood's gradient is extremely steep in n, even a single component of μ that is far from zero is likely to cause (C.11) to be satisfied, hence the exponential decay of the right-hand side in d.

Since $\mu^T\mu \sim \gamma^2\chi^2_d$, we can exactly evaluate the probability that all B Monte Carlo samples from P_μ satisfy (C.11) under Algorithm 1. This is a sufficient condition for downward pseudo-bias, so said probability is a lower bound on the probability of downward pseudo-bias.

The problem is parameterized by B, the number of samples from the prior on μ; n, the number of observed data points; d, the dimension of the model prior; σ, and γ. Figure C.3 shows the lower bound on the probability of an evidence under-estimate as a function of B and n for the one-dimensional case when $\sigma = \gamma = 1$. Because for $d = 1$ the right-hand side of (C.11) is $O(\log(n)/n)$, the contour of the lower bound in B grows very slowly as a function of n; even just $B = 500$ samples from the prior appears sufficient for samples of $n = 1000$ or even larger.

However, letting d vary in addition to n provides an different picture, as Figure C.4 attests. When noise is relatively low ($\sigma = 1$), downward pseudo-bias becomes assured around $d \approx 3$, unless $n < 100$. When the noise level is substantially higher, the trade-off between n and d becomes more clear, though for samples in the high hundreds $d \approx 3$ is still the point at which pseudo-bias becomes a problem.

The pseudo-bias problem raises many more questions than we can answer here: how

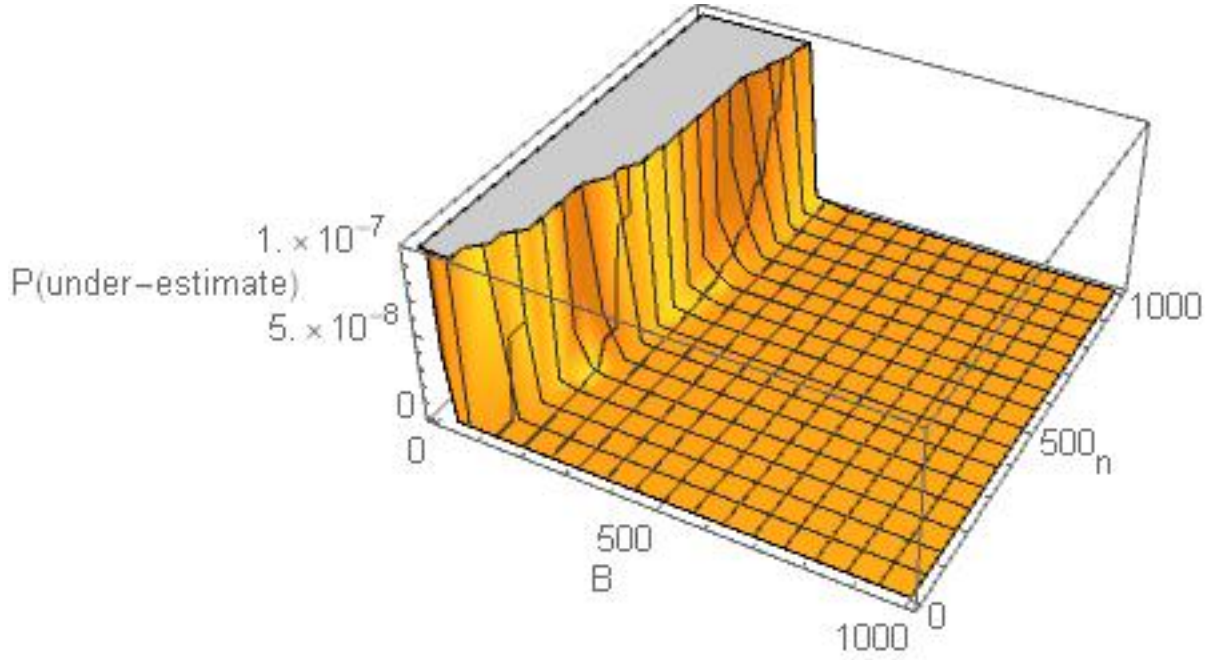
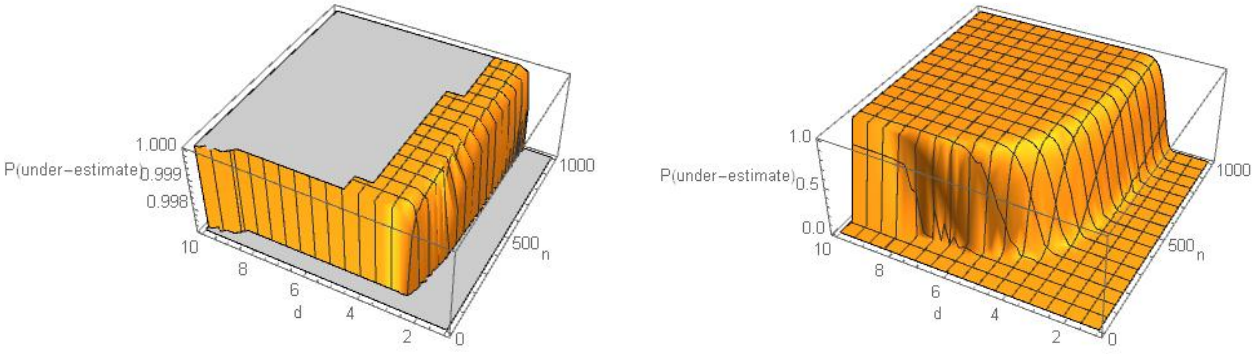


Figure C.3: Pseudo-bias in one dimension.



(a) $\sigma = 1$, $\gamma = 1$.

(b) $\sigma = 10$, $\gamma = 1$.

Figure C.4: Pseudo-bias in multiple dimensions with $B = 1000$.

tight is the lower bound on the probability of a downward-pseudo-biased approximation? Do non-Gaussian probability models yield substantially different results? Can other evidence approximation techniques avoid the pseudo-bias problem? For the last question, see Kass

and Raftery (1995); Evans and Swartz (1995); Lenk (2009); Wolpert and Schmidler (2012); in short, approximating marginal likelihoods remains a challenging problem. In Chapter 4, we marginalize over 3 parameters for ANMs and 2 for LiNGAMs, work mostly with sample sizes between 100 and 1000, and choose priors carefully.

C.4 Bridgesampling

The `bridgesampling` R package (Gronau et al., 2020) uses samples from the posterior distribution of the model parameters to approximate marginal likelihoods. We tuned a STAN (Gelman et al., 2015) Hamiltonian Monte Carlo sampler based on the Gaussian Process model prior outlined in Section 4.2.3 and used it along with `bridgesampling` to estimate the Bayes Factor for data generated under a bivariate Gaussian model for which the log Bayes Factor is analytically known to be zero (see Section 4.2.1). We ran MCMC chains long enough to yield roughly 500 effective samples and compared the resulting Bayes Factors to those of Algorithm 1 with 1000 independent Monte Carlo samples. Despite the runtime of the STAN-`bridgesampling` approach being over 10 times longer, Algorithm 1's performance was superior. The STAN-`bridgesampling` approach yielded highly variable log Bayes Factors: as large in magnitude as 36, and with a mean of -2.7 and sample standard deviation of 9.1 over 10 replications. Algorithm 1 yielded a mean of 0.3 and a sample standard deviation of 1.5 over 100 replications. Therefore, while `bridgesampling` may be superior in more complex/high-dimensional models or for data sets with more observations or variables, we do not use it. Our low-dimensional models and less complex approximation methods keep the focus on causal discovery.

C.5 LiNGAM Model Prior Specification

Under the LiNGAM $Y = \beta X + \epsilon$ with $X \sim \text{Laplace}(0, 1)$ and $\epsilon \sim \text{Laplace}(0, \sigma_Y)$, we develop the prior $\mathcal{P}_{\sigma_Y|\beta}$ as follows. Suppose briefly that the error models were Gaussian—i.e. $X \sim \text{N}(0, 1)$ and $\epsilon \sim \text{N}(0, \sigma_Y^2)$—and that instead of centering and scaling $\{Y\}_n$ by the sample mean and standard deviation, we centered it by the true mean (0) and scaled by the true

standard deviation $\sqrt{\beta^2 + \sigma^2}$. Then, because the variance of Y is 1 after rescaling,

$$\beta^2 + \sigma_Y^2 = 1 \implies \sigma_Y = \sqrt{1 - \beta^2}. \tag{C.12}$$

At the extremes, $\sigma_Y = 1 \iff \beta = 0$ while $\sigma_Y = 0$ corresponds $\beta = 1$. A similar relationship holds for the LiNGAM, though it is more complicated to analyze since the centered and scaled data have unit mean absolute deviation from the mean $\mathrm{E}|Y - \mathrm{E}[Y]|$, which is not additive (in contrast to variance).

It can be shown that the density of $Z = \text{Laplace}(0, \alpha) + \text{Laplace}(0, 1)$ is

$$\frac{\alpha}{2(\alpha^2 - 1)} \left[\exp(-|z|) + \exp(-|z|/\alpha)\right] \tag{C.13}$$

with $\mathrm{E}[Z] = 0$ by symmetry and

$$\mathrm{E}|Z - \mathrm{E}[Z]| = \frac{\alpha^2 + \alpha + 1}{\alpha + 1}. \tag{C.14}$$

(C.14) and the fact that $\mathrm{E}[Y] = 0$ (by symmetry) imply that

$$\mathrm{E}|Y - \mathrm{E}[Y]| = \frac{\beta^2 + \beta\sigma_Y + \sigma_Y^2}{\beta + \sigma_Y} \tag{C.15}$$

so that after scaling by $\mathrm{E}|Y - \mathrm{E}[Y]|$ the relationship between β and σ_Y is defined by

$$\frac{\beta^2 + \beta\sigma_Y + \sigma_Y^2}{\beta + \sigma_Y} = 1. \tag{C.16}$$

However, we scale by sample statistics in practice, so the relationship between β and σ_Y is a noisy version of (C.16). For each pair (β, σ) taking values on a grid of 100 solutions to (C.16), we simulated $1,000$ data sets of size $n = 100$ and computed the sample mean absolute deviation from the mean for each. These varied from 0.6 to 1.4, i.e. by as much as 0.4 from expectation.

We then further simulated $n = 100$ data sets using parameters from the grid $(\beta, \sigma_Y) \in \{0, 0.01, \ldots, 1\}^2$ and found the range of parameter values that yielded an outcome mean absolute deviation from the mean in $[0.6, 1.4]$. For a fixed σ_Y, β varied from the solution to (C.16) by as much as ≈ 0.3 on either side. Thus, the choice of $s = 0.3$ in (4.12) captures the extreme departures from (C.14) that may occur due to sampling variability.

VITA

Sheridan Grant has been to 44 states and 29 national parks, competed in the inaugural Beer Mile World Classic, and seen every episode of *Seinfeld*. He would like to have the quirkiness of Kramer, the wiliness of George, the wit of Elaine, and the apartment of Jerry. Unfortunately, he has the apartment of Kramer, the wit of George, the (lack of) wiliness of Elaine, and the (lack of) quirkiness of Jerry. His CV and much more can be found on his website.