NORTHWESTERN
UNIVERSITY

Computer Science Department

Technical Report
Number: NU-CS-2026-15

June, 2026

Computational Methods in Economics

Chaofeng Wu

Abstract

This dissertation develops and applies computational methods to three problems at the frontier of empirical economics, united by a common objective: to operationalize abstract economic concepts using modern machine learning tools, to uncover new empirical knowledge through those measurements, and to evaluate the conditions under which such tools can be trusted.

The first essay asks what makes a scientific paper novel, and whether novelty of different kinds matters differently for impact. Treating novelty as a multidimensional construct rather than a single attribute, the essay leverages Large Language Models (LLMs) and embedding models to measure novelty directly from the full text of scientific papers. It finds that specific dimensions of novelty, defined by a paper's distinctiveness from its intellectual neighbors versus its engagement with the current frontier, predict fundamentally different outcomes. This finding reveals a strategic tension in knowledge production that citation-based measures of impact cannot capture. The second essay asks why strategic behavior systematically departs from game-theoretic predictions in complex environments. Drawing on a massive dataset of hundreds of millions of chess games, it develops an interpretable measure of when a position is genuinely difficult for a boundedly rational agent, and uncovers a new stylized fact: complexity arises endogenously more often among higher-skilled players. A structural model reveals that this pattern reflects experts' ability to sustain complexity, rather than a preference for risk, a finding with broader implications for understanding how people create and exploit strategic complexity.

The third essay asks how much we should trust a predictive model when it is applied outside the context in which it was estimated. It formalizes this out-of-domain prediction problem, derives coverage-guaranteed forecast intervals for transfer error, and finds suggestive evidence that the in-domain superiority of black-box algorithms does not reliably generalize across domains, a cautionary note with direct relevance for the use of machine learning in situations requiring generalizability.

Together, these essays argue that the value of computational tools in economics lies not only in their raw predictive power, but more importantly in their capacity, when disciplined by economic theory, to render previously unobservable quantities measurable and to rigorously evaluate the limits of predictive models.

Keywords

Computational Social Science, Game Theory, Innovation, Generalizability

NORTHWESTERN UNIVERSITY

Computational Methods in Economics

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Computer Science

By

Chaofeng Wu

EVANSTON, ILLINOIS

June 2026

ABSTRACT

This dissertation develops and applies computational methods to three problems at the frontier of empirical economics, united by a common objective: to operationalize abstract economic concepts using modern machine learning tools, to uncover new empirical knowledge through those measurements, and to evaluate the conditions under which such tools can be trusted.

The first essay asks what makes a scientific paper novel, and whether novelty of different kinds matters differently for impact. Treating novelty as a multidimensional construct rather than a single attribute, the essay leverages Large Language Models (LLMs) and embedding models to measure novelty directly from the full text of scientific papers. It finds that specific dimensions of novelty, defined by a paper's distinctiveness from its intellectual neighbors versus its engagement with the current frontier, predict fundamentally different outcomes. This finding reveals a strategic tension in knowledge production that citation-based measures of impact cannot capture.

The second essay asks why strategic behavior systematically departs from game-theoretic predictions in complex environments. Drawing on a massive dataset of hundreds of millions of chess games, it develops an interpretable measure of when a position is genuinely difficult for a boundedly rational agent, and uncovers a new stylized fact: complexity arises endogenously more often among higher-skilled players. A structural model reveals that this pattern reflects experts' ability to sustain complexity, rather than a preference for risk, a finding with broader implications for understanding how people create and exploit strategic complexity.

The third essay asks how much we should trust a predictive model when it is applied outside the context in which it was estimated. It formalizes this out-of-domain prediction problem, derives coverage-guaranteed forecast intervals for transfer error, and finds suggestive evidence that the in-domain superiority of black-box algorithms does not reliably generalize across domains, a

cautionary note with direct relevance for the use of machine learning in situations requiring generalizability.

Together, these essays argue that the value of computational tools in economics lies not only in their raw predictive power, but more importantly in their capacity, when disciplined by economic theory, to render previously unobservable quantities measurable and to rigorously evaluate the limits of predictive models.

ACKNOWLEDGEMENTS

I am deeply indebted to my advisors, Annie Liang and Ben Golub, for their unwavering support, vision, and patience. As someone coming primarily from a computer science background, I was incredibly fortunate to have two brilliant economists as advisors, who wholeheartedly welcomed me into their field and patiently guided my transition into this interdisciplinary space. They were always willing to entertain my half-baked or wild ideas, consistently offering invaluable feedback from the perspective of seasoned economists. Their rigorous standards and deep intuition have profoundly shaped my growth as a researcher at the intersection of computer science and economics. This was especially true during the development of my job market paper, when they served as both demanding critics and patient mentors, listening tirelessly as I worked through my narrative and offering their candid assessments each time. It was a difficult period, but they never gave up on me. They are the best advisors I could have possibly asked for. Simply put, they made my time as a PhD student a truly wonderful experience.

I am also extremely grateful to my committee members. Special thanks to Jessica Hullman for her invaluable advice and detailed feedback, which significantly improved my job market paper. I also want to extend my gratitude to Jason Hartline and Aravindan Vijayaraghavan for their guidance during my master's defense and for their constructive suggestions on my earlier work.

A significant portion of the research in this dissertation is the result of collaborative efforts. I am deeply thankful to my co-authors, Drew Fudenberg, Isaiah Andrews, Lihua Lei, Jeff Ely, as well as Annie and Ben, for being fantastic collaborators, for everything I have learned from working alongside them, and for generously allowing me to include our joint work in this dissertation.

I also want to express my special thanks to the department's administrative team, who helped me navigate the bureaucratic maze of the university and always kept their doors open. On a more

technical note, I am grateful to the high-performance computing cluster at Northwestern University for all the computing resources and support.

I owe my deepest gratitude to my parents. Thank you for your unconditional love, for supporting my choices, and for always believing in me. Your support has been the quiet force propelling me forward every step of the way.

Last but certainly not least, to my partner, Anqi Zhang: thank you for being my rock. Your companionship, your attentive care, your patience and comfort in difficult moments, and your constant encouragement made this journey not just possible, but profoundly meaningful. You were the most loyal audience and most trusted sounding board for my research. I will not forget the hours we spent sitting side by side, writing code, drafting papers, and talking through results late into the night. You carved out a corner of certainty for me in a highly unpredictable world.

To the years spent far from home.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION AND BACKGROUND

The digitalization of economic activity and the exponential growth of high-dimensional data have fundamentally expanded the analytical toolkit available to economists. While traditional economic analysis has relied heavily on structured data and concise theoretical models, the emergence of Machine Learning (ML), Natural Language Processing (NLP), and Large Language Models (LLMs) offers new avenues to investigate longstanding questions in the discipline. These advances can be organized into three complementary domains that form the intellectual backdrop of this dissertation: the high-dimensional measurement of latent concepts from unstructured text, the use of algorithmic benchmarks to study strategic behavior, and the inferential challenges that arise when predictive models are deployed outside their estimation environment. Each domain motivates one of the three self-contained research papers that constitute this work.

High-Dimensional Measurement: Text and Semantics. A central contribution of the modern empirical toolkit is the ability to extract quantitative information from unstructured text. Pioneering work has transformed textual archives into rigorous measures of policy uncertainty (Baker et al., 2016) and media slant (Gentzkow and Shapiro, 2010). These approaches have been generalized through the adoption of high-dimensional statistical models: Kelly et al. (2021) develop an economically motivated selection model for learning from text, with applications ranging from partisan speech to macroeconomic forecasting. Ash and Hansen (2023) provide a systematic overview of how text algorithms are applied in economics, covering embedding methods, supervised classification, and topic models. More recently, LLMs have enabled finer-grained semantic measurement.

This stream of literature directly motivates the measurement framework developed in Chapter 2, where I employ LLMs to quantify the position of a scientific paper within an evolving intellectual landscape.

Machine Learning, Decision Making, and Strategic Behavior. Machine learning has also reshaped how economists model rationality and decision-making. Kleinberg et al. (2015) established that many policy problems are fundamentally prediction problems in which ML outperforms traditional econometrics. A natural extension of this logic is to use algorithmic tools not merely to predict outcomes, but to operationalize latent economic concepts that are otherwise difficult to quantify. Kleinberg et al. (2018) demonstrate the power of this approach in the context of judicial bail decisions, showing that comparing human choices to an algorithmic standard can reveal the structure of information frictions and implicit biases. Chapter 3 applies a related but conceptually distinct logic to the measurement of strategic complexity. The key insight is that a chess position is *complex* if its game-theoretically optimal value cannot be recovered from human-understandable features using a simple, interpretable model (a decision tree). This definition is deliberately grounded in cognitive accessibility: a position is complex precisely when straightforward analysis is insufficient to identify the correct course of action. Using this measure, I study when and why complexity arises endogenously in high-level play.

Inference, External Validity, and Model Transferability. The adoption of black-box ML algorithms has sparked renewed interest in the conditions under which these models generalize. Frameworks such as Double/Debiased Machine Learning (Chernozhukov et al., 2018) provide principled ways to estimate causal parameters in the presence of high-dimensional nuisance terms. Yet a critical and underexplored frontier concerns the *out-of-domain* performance of predictive models: what happens when a model estimated in one context is applied to another. Mullainathan and

Obermeyer (2017) offer an early illustration of this concern, showing that algorithms that appear to perform well within their training domain can embed systematic errors that remain invisible to standard validation metrics. Banerjee et al. (2017) and Chassang and Kapon (2022) discuss external validity, which seeks to measure the extent to which results obtained in one domain hold more generally. Chapter 4 confronts this problem directly, formalizing the concept of transfer error and providing a theoretical foundation for out-of-domain inference by deriving finite-sample forecast intervals that guarantee coverage when domains are exchangeable.

1.1 The Scope of this Dissertation

This dissertation explores the intersection of these computational frontiers with economic theory. While the preceding section outlined three broad methodological domains, this work specifically targets three critical challenges within that landscape: the measurement of latent constructs from text, machine learning for strategic behavior, and the external validity of predictive models.

Despite the distinct substantive focus of each chapter—ranging from the science of science to behavioral game theory and econometric methodology—this dissertation is unified by a single overarching objective: to operationalize abstract economic concepts using modern machine learning tools, to leverage those measurements to uncover new empirical knowledge, and to evaluate the conditions under which those tools can be trusted.

- **Bridging Measurement and Theory (Chapter 2).** Building on advances in high-dimensional text measurement, Chapter 2 leverages LLMs not merely to classify papers, but to quantify the geometry of scientific novelty. I decompose novelty into two conceptually distinct dimensions: *spatial novelty* measures a paper's intellectual distinctiveness from its neighbors and *temporal novelty* measures its engagement with the current research frontier. I show that

these dimensions predict systematically different impact outcomes, where temporal novelty primarily predicts citation counts, while spatial novelty predicts disruptive impact.

- **From Prediction to Mechanism (Chapter 3).** Extending the use of algorithmic benchmarks, Chapter 3 uses the superhuman predictive power of chess engines to establish the objective value of positions. I operationalize complexity as the extent to which this objective value remains opaque to simple, interpretable heuristics. Moving beyond the existing literature's focus on single-agent errors, this chapter explores how individuals *strategically manage* complexity in dynamic interactions. The analysis uncovers a novel stylized fact—that strategically complex environments are endogenously generated more frequently by higher-skilled players—and develops a structural model showing that this reflects experts' ability to sustain complexity to exploit opponents, rather than a preference for risk.

- **Testing the Limits (Chapter 4).** Chapter 4 provides a theoretical and econometric examination of when and why powerful predictive models fail to transfer across economic contexts. By formalizing the concept of transfer error and constructing coverage-guaranteed forecast intervals, I provide both a diagnostic tool and empirical evidence of a trade-off: black-box algorithms excel within their estimation domain but may generalize more poorly than theory-driven models across domains.

1.2 Chapter Summaries

This dissertation is organized into three core chapters, each representing a self-contained research paper.[1]

[1]Chapters 2, 3, and 4 are written as independent research papers. Chapter 3 is based on joint work with Jeff Ely, Ben Golub, and Annie Liang, and Chapter 4 is based on joint work with Isaiah Andrews, Drew Fudenberg, Lihua Lei, and Annie Liang. As such, each chapter contains its own introduction, literature review, and notation definitions. Some overlap in introductory concepts is intentional to ensure each chapter stands on its own.

Chapter 2: Novelty and Impact of Economics Papers. This chapter addresses the fundamental challenge of measuring scientific progress. While "novelty" is a central concept in the economics of innovation, it has traditionally been treated as a unidimensional attribute. I propose a framework that recasts novelty as a geometric position within an evolving intellectual landscape, decomposed into two conceptually distinct dimensions: *spatial novelty*, which measures a paper's intellectual distinctiveness from its neighbors, and *temporal novelty*, which captures its engagement with a dynamic research frontier.

To operationalize these concepts, I employ LLMs and embedding models to develop semantic isolation metrics that quantify a paper's location relative to the full-text literature. Applying this framework to a large corpus of economics articles, I uncover a fundamental trade-off: these two dimensions predict systematically different outcomes. Temporal novelty primarily predicts citation counts, whereas spatial novelty predicts disruptive impact—the tendency to render prior work obsolete. This distinction further allows for the construction of a typology of semantic neighborhoods, identifying four archetypes associated with distinct and predictable impact profiles. The chapter demonstrates that novelty is a multidimensional construct whose different forms have measurable and fundamentally distinct consequences for scientific progress.

Chapter 3: Managing Strategic Complexity. This chapter shifts focus from the macro-structure of science to the micro-structure of strategic interaction. Standard game-theoretic analysis often yields incomplete descriptions of behavior in complex games because it assumes perfect rationality. I investigate the limits of these assumptions using a large database of chess games.

The central methodological contribution is a novel, interpretable index of positional complexity. A position is complex if its game-theoretically optimal value cannot be recovered from human-understandable features by a simple decision tree. The underlying intuition is that complexity is a

property of cognitive accessibility—a position is complex precisely when straightforward feature-based analysis is insufficient to identify the correct course of action, mirroring the cognitive challenge faced by human players.

I show that this complexity index effectively predicts deviations from standard game-theoretic play. The analysis then reveals a new stylized fact regarding the endogenous emergence of complexity: strategically complex positions occur more frequently among higher-rated players. To understand this pattern, I develop a tractable model that captures the key strategic incentives for and against complex play, solve for equilibrium, and structurally estimate the model to assess the underlying forces. The estimation reveals that the cross-sectional pattern is driven not by experts seeking risk, but by their ability to sustain complex positions without committing errors. This chapter illustrates how a carefully designed interpretable ML tool can serve as a structural yardstick for estimating behavioral models of bounded rationality.

Chapter 4: The Transfer Performance of Economic Models. This chapter provides a critical econometric perspective on the challenge of deploying predictive models outside their estimation environment. I formalize the "out-of-domain" prediction problem, which arises whenever a model estimated on data from one domain is used to make predictions in another.

I define the *transfer error* of a model based on its performance in a new domain, and derive finite-sample forecast intervals guaranteed to cover realized transfer errors with a user-selected probability when domains are exchangeable. Applying these intervals to the prediction of certainty equivalents, I find suggestive evidence of a systematic trade-off: black-box algorithms outperform theory-grounded economic models for in-domain prediction, but generalize more poorly across domains. This chapter serves as a necessary methodological reflection on the stability of predictive inference in economics, and provides practitioners with concrete tools for quantifying

and communicating the risks of model transfer.

CHAPTER 2
NOVELTY AND IMPACT OF ECONOMICS PAPERS

2.1 Introduction

A central challenge for researchers, funding agencies, and academic institutions is to identify which new ideas are most likely to generate significant scientific impact. While novelty is widely regarded as a key driver of impact, its complex nature and the mechanisms through which it operates remain poorly understood. This challenge reflects a fundamental tension in research strategy: should efforts focus on work with high uniqueness, characterized by profound originality in questions and methods, or on work with good timing that taps into "hot" topics currently capturing the scientific community's attention? Although both uniqueness and timing are recognized as critical dimensions of novelty, existing measures typically fail to disentangle them. Citation-based approaches (Uzzi et al., 2013; Lee et al., 2015) rely on backward-looking references, overlooking the paper's actual semantic content. Meanwhile, keyword-based methods (Azoulay et al., 2011; Mishra and Torvik, 2016) are often too coarse to capture subtle conceptual shifts or distinguish between a buzzword and a substantive contribution. Consequently, we lack a framework that can simultaneously quantify a paper's distinctiveness from the past and its alignment with the future.

In this study, we introduce a new conceptual framework that recasts novelty not as an intrinsic property, but as a *positional good* within the evolving intellectual landscape. We decompose this position into two conceptually different dimensions. **Spatial novelty** measures a paper's intellectual distinctiveness by quantifying its semantic isolation from all prior work at the time of publication. It asks: *is this paper positioned in an isolated neighborhood?* In contrast, **Temporal**

novelty measures a paper's engagement with a dynamic research frontier by quantifying the rate of change in its immediate semantic neighborhood. It asks: *is this paper positioned in a fast-changing neighborhood?* By treating papers as coordinates in a high-dimensional space, we can map their positions relative to both the accumulated stock of knowledge and the shifting currents of scientific attention.

To operationalize this framework, we design a set of semantic isolation metrics leveraging recent advances in Large Language Models (LLMs). Unlike prior studies that rely on titles or abstracts, we develop a pipeline to analyze the *full text* of research papers. This allows us to capture granular details of arguments, methodologies, and results that shorter summaries often omit. Using a custom dataset containing the full text of 105,082 economics papers, we construct comprehensive semantic embeddings for each document. Before analyzing research strategies, we rigorously validate these metrics. We demonstrate that our semantic isolation measures possess strong predictive power for two distinct forms of impact: citation counts (reflecting community attention) and the disruption index (reflecting the extent to which a paper renders prior work obsolete). This validation confirms that our embedding-based approach successfully extracts latent signals of value from the text that are complementary to traditional indicators.

Building on these validated metrics, we employ Principal Component Analysis (PCA) to construct composite scores for spatial and temporal novelty. Our analysis uncovers a critical strategic trade-off: the two scores exhibit a negative correlation, suggesting that it is empirically difficult for a single work to be simultaneously highly distinctive (far from the consensus) and perfectly timed (within a surging trend). Based on this trade-off, we classify the intellectual landscape into four archetypes. Papers with low novelty on both dimensions are **Consolidating** serving to reinforce existing paradigms. Those with high spatial but low temporal novelty are **Outlying** representing unique but potentially premature or niche contributions. Papers with low spatial but high temporal

novelty are **Trendy** riding the wave of current hot topics without significantly deviating from the norm. Finally, papers that achieve high scores on both dimensions are **Trailblazing** effectively balancing distinctiveness with relevance.

Applying this typology, we analyze the relationship between novelty strategies and scientific impact. We report three main discoveries. First, the drivers of impact are distinct: temporal novelty is the primary predictor of high citation counts, suggesting that being on time leads to popularity. In contrast, spatial novelty is the main predictor of disruptive influence, suggesting that uniqueness leads to fundamental shifts in the literature. Second, this distinction validates our archetypes: Trendy papers tend to be highly cited but less disruptive, while Outlying papers are disruptive but often receive fewer immediate citations. Third, and perhaps most importantly, we find that Trailblazing papers exhibit a synergistic interaction. By successfully navigating the trade-off between uniqueness and timing, these papers have a disproportionately high probability of being *both* highly cited and highly disruptive. Our findings demonstrate that novelty is not a monolithic concept, but a multidimensional strategy with measurable and systematically different consequences for scientific progress.

The remainder of the paper is organized as follows. Section 2.2 details our conceptual framework and the design of the semantic isolation metrics. Section 2.3 describes the dataset and the technical construction of these metrics using LLMs. Section 2.4 presents the validation results, demonstrating the metrics' predictive power for citations and disruption. Section 2.5 describes the construction of the composite novelty scores and the resulting typology. Section 2.6 discusses our empirical findings on the relationship between these novelty strategies and scientific impact.

2.1.1 Related Work

Measures of Novelty. Approaches to measuring novelty fall into two primary categories: citation-based and text-based. Citation-based methods typically quantify the atypicality of reference combinations (Uzzi et al., 2013; Lee et al., 2015; Trapido, 2015). While influential, these metrics overlook the semantic content of the work. Text-based approaches address this by analyzing keywords, such as the age of MeSH terms (Azoulay et al., 2011) or new word combinations (Mishra and Torvik, 2016; Arts et al., 2025). However, keyword methods face limitations due to polysemy and often miss deeper semantic connections. More recent work has begun to use embedding techniques on titles to measure semantic distance (Shibayama et al., 2021). Our method advances this line of inquiry by leveraging Large Language Models (LLMs) to analyze the *full text*. Unlike proxies based on titles or abstracts, our approach generates a comprehensive semantic representation, allowing us to measure novelty as a paper's precise position and movement relative to the entire literature. To validate our metrics, we follow the established practice of indirect validation (Foster et al., 2021; Arts et al., 2025), assessing their predictive power for citation counts and disruption rather than relying on subjective expert labels.

Conceptualization of Novelty. Scientific novelty has been conceptualized through various theoretical lenses (Zhao and Zhang, 2025). A foundational dichotomy distinguishes between the introduction of entirely new elements (Berlyne, 1960; Yan et al., 2020; Luo et al., 2022) and the novel recombination of existing elements, a concept rooted in Schumpeterian innovation theory . Other frameworks conceptualize novelty in terms of network structures, where new connections bridge previously disconnected knowledge communities (Foster et al., 2015; Hofstra et al., 2020), or from a psychological perspective, where novelty is a function of surprise relative to a reader's cognitive schema (Foster et al., 2021). In contrast to these mechanism-based views, we concep-

tualize novelty as a *positional good* within an evolving intellectual landscape. This perspective captures two distinct aspects: the static semantic distance from prior work (distinctiveness) and the dynamic rate of change in the paper's neighborhood (timing).

Typologies of Novelty. Existing typologies generally classify novelty along two axes: *mechanism* (e.g., recombination vs. creation of new entities) and *domain* (e.g., new theory, methodology, or data) (Leahey et al., 2023). Our work introduces a new typology—spatial versus temporal novelty—that acts as a dimension orthogonal to the domain of contribution. While previous classifications ask *what* part of a paper is novel, our framework asks *how* the paper positions itself relative to the existing scientific literature. This allows us to distinguish between work that is intellectually distant (Outlying) and work that is perfectly timed with a surging trend (Trendy), a distinction often blurred in uni-dimensional novelty measures.

NLP and LLMs in text analysis. Computational analysis of scientific texts has evolved significantly. Early work relied on sparse vector representations such as term frequency–inverse document frequency (TF–IDF) and probabilistic topic models like Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to map research trends (Gupta et al., 2022; Egger and Yu, 2022). More recent methods leverage dense vector representations (embeddings) to model document semantics more effectively. Embedding-based models like SPECTER (Cohan et al., 2020) and domain-specific word embeddings (Tshitoyan et al., 2019) have improved performance on tasks such as classification, clustering, and citation prediction by capturing nuanced contextual relationships.

Building on these foundations, large language models (LLMs) now enable direct synthesis and representation of scientific text at an unprecedented scale. Current applications include screening for systematic reviews (Dennstädt et al., 2024), generating retrieval-augmented literature surveys (Han et al., 2024), classifying research articles (Raja et al., 2023), and predicting bibliometric out-

comes (Vital Jr et al., 2025). LLMs have also proven effective in summarization tasks, producing outputs for both expert and lay audiences (Pakull et al., 2024; Ding et al., 2024) and even drafting review articles (Wang et al., 2024). Our work contributes to this literature by using LLMs not just for summarization but to create the rich semantic representations that underpin our measurement of novelty.

2.2 Framework

This section develops the conceptual and measurement framework for our analysis. We first describe our two measures of scientific impact—citation count and the disruption index—which capture distinct dimensions of a paper's influence. We then reconceptualize novelty in positional terms, introducing spatial and temporal novelty to capture a paper's location in the intellectual landscape. Finally, we illustrate the semantic isolation metrics used to operationalize our framework.

2.2.1 Measuring Multidimensional Impact

Research papers can show impact in different forms. A paper might attract widespread attention, reflected in high citation counts, while another might receive modest immediate attention yet fundamentally redirect its field's trajectory. To capture these distinct pathways, we employ two complementary metrics that capture distinct dimensions of scientific impact: citation count and the disruption index.

Citation count: measuring scholarly attention. Our first measure, cumulative citation count, serves as a standard proxy for scholarly attention and influence. Citations reflect the research community's collective assessment of a paper's value and its direct utility in subsequent scholarship.

Disruption Index: measuring field redirection. Our second measure captures a paper's capacity to redirect a field's intellectual trajectory. We employ the Disruption Index (DI) developed by Funk and Owen-Smith (2017) and Wu et al. (2019), which we classify as an impact measure rather than a novelty measure (Leibel and Bornmann, 2024). Whereas citations measure the magnitude of influence, the DI measures its nature: whether a paper's contribution displaces prior work or builds cumulatively upon it. The index is derived from the citation network surrounding a focal paper (FP) and is defined as:

$$DI = \frac{N_F - N_B}{N_F + N_B + N_R}$$

where N_F represents papers citing only the focal paper, N_B represents papers citing both the focal paper and its references, and N_R serves as a normalizing term for papers citing the focal paper's references but not the focal paper itself.

A high positive score indicates a "disruptive" paper, as subsequent work cites it as a new origin point, displacing its intellectual predecessors. A negative score suggests a "consolidating" paper that integrates prior knowledge, prompting subsequent work to cite both the focal paper and its antecedents. This dual-measurement strategy allows us to test whether different dimensions of novelty systematically produce different forms of impact.

2.2.2 Novelty as Semantic Position

The study of scientific progress hinges on the concept of novelty, yet its measurement remains a central challenge. Prevailing methods mainly rely on proxies, either inferred from its inputs (e.g., references) or its internal content (e.g., new keywords). This view, while intuitive, limits our ability to understand the strategic and ecological dynamics of scientific discovery.

We propose a fundamental shift in perspective, reconceptualizing novelty not as an intrinsic property but as a relational construct. In our view, a paper's novelty is a function of its location

within a high-dimensional semantic space, characterized by both its static isolation and its dynamic evolution.

To characterize a paper's position, we decompose it into two conceptually different dimensions. These dimensions describe its relationship with the dynamic "semantic neighborhood" it inhabits:

- **Spatial novelty** measures a paper's static position in the semantic space. It quantifies intellectual distinctiveness by measuring a paper's isolation from the accumulated stock of prior knowledge.
- **Temporal novelty** measures the dynamics of its position. It captures strategic timing by measuring the change in a paper's immediate semantic over the period leading up to publication, reflecting its engagement with an evolving research frontier.

These two dimensions are central to our analysis, as we hypothesize they correspond to distinct research pathways with fundamentally different consequences for scientific impact. To operationalize this framework, we develop a family of semantic isolation metrics, which are designed to quantify a paper's static isolation (for spatial novelty) and the change in its isolation over time (for temporal novelty) using text embeddings.

2.2.3 Semantic Isolation Metrics

We construct semantic isolation metrics to quantify both spatial and temporal novelty. Each metric is based on the semantic distance between papers, computed by embedding every paper's full content into a high-dimensional vector space (see Section 2.3.2) and measuring pairwise distances. For spatial novelty, we define *point-in-time isolation metrics*, each capturing a distinct facet of a paper's semantic distinctiveness relative to the existing literature. For temporal novelty, we derive *dynamic isolation metrics* by tracking how each point-in-time measure evolves across successive

time windows around a paper's publication (e.g., three years pre-publication versus publication). The resulting slopes serve as our indicators of well-timed engagement with emerging research frontiers. We synthesize these measures into two composite scores, as detailed in Section 2.5.1.

In this section, we provide a conceptual illustration of semantic isolation metrics. A full technical description of these metrics appears in Appendix A.1.

Point-in-Time Isolation Metrics. Point-in-time metrics assess a paper's semantic isolation at a given moment. As illustrated in Figure 2.1, for each focal paper (blue point), we consider its relationship to other work (green points), as measured by four complementary metrics:

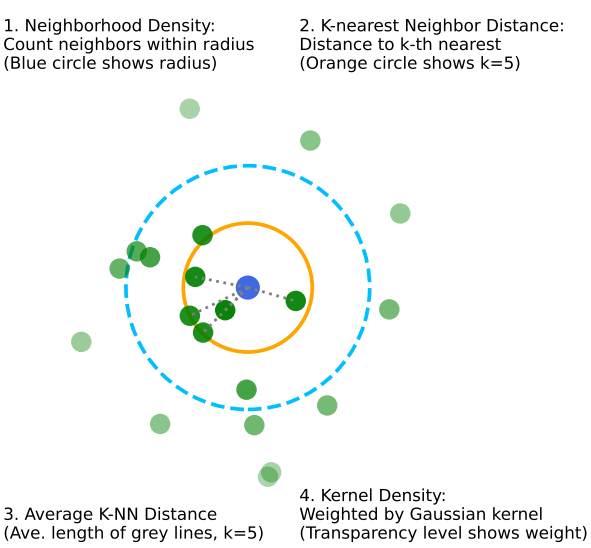


Figure 2.1: *Illustration of point-in-time isolation metrics. The blue point is the focal paper, and the green points are other papers.*

1. **Neighborhood Count:** The number of neighboring papers within a given radius.

2. $k-$**Nearest Neighbor (**k**-NN) Distance:** The distance to the k-th nearest neighbor. A larger distance implies that a paper must reach further into the semantic space to find even its closest peers.

3. **Average k-NN Distance:** The average distance to the k nearest neighbors. This provides a more robust measure of local isolation than the single k-NN distance, as it is less sensitive to the position of a single neighbor.

4. **Kernel Density Estimate:** A smooth, weighted measure of local density computed using a Gaussian kernel. This method provides a more continuous measure of isolation than discrete neighbor counts by down-weighting more distant papers.

These complementary metrics provide a complete picture of a paper's semantic neighborhood. This approach allows us to distinguish whether a paper is truly isolated from all prior work, or if it belongs to a small, nascent cluster that is itself distant from the dense core of the literature. Collectively, these metrics provide a rich characterization of a paper's spatial novelty.

Dynamic Isolation Metrics. While point-in-time metrics capture static isolation, our dynamic metrics measure the evolution of a paper's semantic neighborhood. These metrics are designed to quantify temporal novelty by assessing whether a paper is situated in a stable intellectual area or a rapidly evolving research frontier. We construct these by calculating a point-in-time isolation metric, $M(t)$, at different time horizons relative to the focal paper's publication date ($t = 0$), as shown in Figure 2.2. We measure change from two perspectives:

- **Retrospective Change**: We compute the change in an isolation metric over the period leading up to the focal paper's publication (e.g., from $t = -3$ to $t = 0$). For instance, the change in k-NN distance, $M_2(k, -3) - M_2(k, 0)$, measures how much denser the neighborhood became in the three years prior to publication. A large positive change indicates that the paper is entering a rapidly growing research frontier. This is our primary measure for constructing the temporal novelty score, as it relies only on information available at publication.

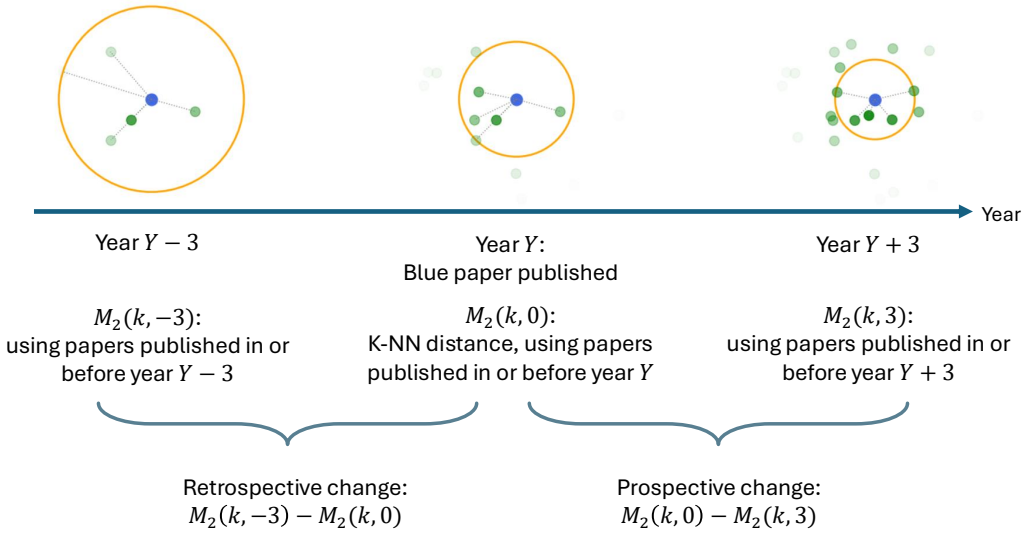


Figure 2.2: *Illustration of dynamic isolation metrics. Given the focal paper published in year* Y*,* $M_2(k,t)$ *represents the* $k-$*NN distance using papers that published in or before year* $Y+t$*. For example,* $M_2(k,-3)$ *is calculated with papers that published in or before year* $Y-3$*.*

- **Prospective Change:** We also compute the change in the years following publication (e.g., from $t=0$ to $t=3$). The difference $M_2(k,0)-M_2(k,3)$ measures how the neighborhood evolves after publication of the focal paper. Because perspective change relies on future information unavailable at the time of publication, we exclude it from our main results and instead use it to validate our temporal novelty concept (see Appendix A.3.4).

For example, a paper on behavioral economics published in 2005 exhibits high temporal novelty if many similar papers appeared between 2002 and 2004 (high retrospective change), indicating it is well-timed to an emerging frontier, even if its spatial novelty is modest.

In addition to these two main types of semantic isolation metrics, we also include baseline paper-level mean distance, which captures basic features of semantic distance. More details are in Appendix A.1.4.

2.3 Data

Our empirical analysis relies on a large corpus of full-text economics articles processed through a custom natural language understanding pipeline. In this section, we (1) describe our data sources and explain how we constructed the analysis sample, and (2) summarize the steps used to generate our semantic isolation metrics. A key contribution of this paper is the creation of a novel, integrated dataset containing papers' full text and their semantic embeddings. We also contribute a reproducible pipeline for extracting these embeddings from article full text.

2.3.1 Data Source and Dataset Construction

Our analysis draws on a comprehensive corpus of scholarly articles in economics. The initial dataset comprises 105,082 articles published between 1980 and 2020, sourced from RePEc, Google Scholar, and SciSciNet (Lin et al., 2023). From SciSciNet, we obtain key bibliometric variables, including the disruption index and the atypicality score. We supplement this with citation counts collected from Google Scholar in December 2024.

To ensure a robust measurement of novelty and impact, we partition our sample by time. Papers published between 1980 and 1995 form a "knowledge base", providing the necessary historical context to measure the novelty of subsequent work. Our analysis focuses on a "focal period" cohort of papers published between 1996 and 2015. This design allows us to observe impact outcomes over a meaningful time horizon (at least 9 years post-publication) while grounding our novelty measures in a deep historical context. After filtering for data quality and completeness, our final analysis dataset contains 52,766 papers.

The data reveal substantial heterogeneity across fields in both citation patterns and the prevalence of disruptive work. For instance, Development and Agricultural Economics have the highest

median citation counts and disruptive rates, respectively, while Microeconomic Theory exhibits lower values on both dimensions. As expected, the distribution of citations is highly right-skewed across all fields. The detailed descriptive statistics is in Table A.2 in Appendix .

2.3.2 Semantic Content Extraction and Representation

Our primary data consist of full-text research articles in PDF format. We developed a multi-stage pipeline to convert these documents into structured semantic representations suitable for calculating our isolation metrics.

First, we convert the PDFs to plain text using an LLM-based Optical Character Recognition (OCR) model. Second, we use LLMs to generate comprehensive summaries of each paper. A key innovation of our approach is the extraction of multiple, structured representations of each paper's intellectual content. Rather than relying on a single summary, we prompt the LLM to decompose each paper into five distinct semantic dimensions:

- **Paragraph:** A concise, high-level overview of the paper's content.
- **Topic:** The paper's subfield, primary subject matter, and specific issues addressed.
- **Question-Conclusion:** The core research questions and key findings.
- **Methodology:** Data sources, analytical models, and technical approaches employed.
- **Theory:** The theoretical frameworks, conceptual models, and analytical innovations.

This decomposition allows us to measure novelty along different dimensions of a paper's contribution - for example, distinguishing methodological novelty from theoretical novelty.

Third, to ensure the robustness of our semantic representations, we embed each of the five summary types using three complementary embedding models: SciBERT (Beltagy et al., 2019), a

BERT model pre-trained on a large corpus of scientific literature; SPECTER (Cohan et al., 2020), a model specifically designed to generate document-level embeddings for scientific papers; and Qwen (Li et al., 2023), a general-purpose state-of-the-art LLM. This triangulation approach mitigates the risk of model-specific biases.

Finally, we calculate pairwise distances between all paper embeddings using both Euclidean distance and cosine similarity. This yields a comprehensive feature space derived from the combination of five summary types, three embedding models, and two distance metrics. This systematic process results in 3240 candidate isolation features for each paper, providing a rich basis for our analysis. Further details are provided in Appendix A.2.

Addressing Look-Ahead Bias. The use of modern LLMs trained on vast internet corpora introduces a potential for "look-ahead bias", where the model might inadvertently incorporate information unavailable at the time a paper was written (Sarkar and Vafa, 2024). We mitigate this concern through several strategies. First, our LLM task is summarization, not prediction. Different from prediction task such as predicting the citation count of a paper will receive by 2024 (which is likely to be in a LLM's training set), in our task the model is prompted to condense the provided text, not to infer future outcomes or context. Second, our prompts are designed to be content-focused, explicitly instructing the model to avoid external contextualization. Third, our metrics are based on relative semantic distances within a historical cohort, making them less sensitive to absolute shifts in the embedding space.

2.4 Validating Semantic Isolation Metrics

Before employing our semantic isolation metrics to construct a theoretical framework of novelty, we need to first establish their empirical relevance. This section demonstrates that these metrics are

powerful predictors of a paper's future impact. We show that they not only outperform established novelty measures but also provide complementary information, confirming that they capture a distinct and meaningful dimension of scientific contribution.

2.4.1 Experimental Setup

We validate the semantic isolation metrics through two predictive tasks: forecasting long-term citation counts and identifying disruptive papers. Our experimental design aims to demonstrate that these metrics (1) possess significant predictive power and (2) capture unique information complementary to standard bibliometric controls and existing novelty proxies (e.g., atypicality). We define several feature groups for our predictive models:

- **Controls:** Basic metadata including publication year, reference count, document length, word count, and sentence count.
- **Subfield (Baseline):** Indicator variables for the LLM-classified subfields.
- **Journal:** Encoded journal identifiers and top-5 journal indicators.
- **Textual Features and topic models:** Standard NLP features including word counts, part-of-speech tag frequencies, and other text-derived metrics. Latent Dirichlet Allocation (LDA) and Term Frequency–Inverse Document Frequency (TF-IDF) representations. They contains future information that is not available at the time a paper is published, as they are calculated with the whole dataset.
- **Atypicality:** The combination-based novelty scores developed by Uzzi et al. (2013), including 10th percentile, median, and number of combinations, as provided by SciSciNet (Lin et al., 2023).

- **Isolation:** Our proposed set of semantic isolation metrics, including point-in-time isolation metrics and retrospective dynamic isolation metrics[1].

To fully capture the nonlinear relationship between features, we employ XGBoost for prediction tasks. We use consistent hyperparameters across all feature combinations to ensure performance differences reflect information content rather than optimization artifacts. These hyperparameters are identified by tuning the model on the most complex feature set (the union of all feature groups), as parameters that perform well on this high-dimensional space are the most likely to be robust and generalize effectively to simpler feature subsets. All results are averaged over 100 bootstrap iterations for robust statistical inference.

2.4.2 Predicting Log Citation Counts

We first assess the ability of our isolation metrics to predict long-term citation counts, a standard proxy for scientific impact. Our primary evaluation metric is the out-of-sample Mean Squared Error (MSE). We also report the "completeness score" from Fudenberg et al. (2022b), which measures the proportional reduction in MSE relative to a baseline feature set (Controls + Subfield). The larger the completeness (maximum is 1), the more error the tested model reduce from the base, which means the better the model's performance. For example, using all features together gives a completeness score of 34.92%, meaning that this large set of features can reduce 34.92% of our baseline prediction error. This suggests that predicting log citation counts is a hard task.

Results. The top panel of Table 2.1 presents the standalone predictive power of the atypicality and isolation feature groups. When added to the control variables, the atypicality score offers no

[1]We exclude prospective dynamic isolation metrics from our main analysis to avoid inadvertently incorporating future information into the prediction of future outcomes. The version with perspective dynamic isolation metrics is in Appendix A.3.4.

improvement over the baseline. In stark contrast, our isolation metrics achieve a completeness score of 9.93%, demonstrating substantial predictive power. This is also surprising as only using isolation features can reach nearly 30% of the completeness score of using all features (34.92%), reassuring that isolation features predict citation count well.

Feature set	MSE	Completeness
control+subfield (baseline)	1.7044	0.0
	(0.00154)	
control+atypicality	1.7110	-0.0039
	(0.00156)	
control+isolation	1.5352	**0.0993**
	(0.00133)	
control+subfield+atypicality	1.6646	0.0234
	(0.00154)	
control+subfield+atypicality+isolation	1.5090	**0.1146**
	(0.00133)	
all features (excl. isolation)	1.1465	0.3273
	(0.00111)	
all features (incl. isolation)	1.1092	**0.3492**
	(0.00105)	

Table 2.1: *Performance of predicting log citation counts using different feature group combinations. The top panel shows individual feature groups; the middle and bottom panel show the additional value of isolation features. Standard errors (in parentheses) are calculated from 100 bootstrap iterations. Completeness is the proportional reduction in MSE relative to the baseline.*

This result raises a key question: do isolation and atypicality measure the same underlying construct, with isolation being a more precise proxy? Or do they capture fundamentally different dimensions of novelty? The middle panel of Table 2.1 provides a clear answer. Adding atypicality to the baseline provides a modest 2.34% improvement in completeness. However, subsequently adding our isolation metrics increases completeness to 11.46% - an incremental gain of 9.12 percentage points. This strong additional effect provides compelling evidence that isolation and atypicality are complementary, measuring distinct aspects of novelty. Atypicality captures the novelty

of a paper's inputs (unusual combinations of references), while our metrics capture the novelty of its output (the semantic content of its contribution).

Finally, we test whether the predictive power of our isolation metrics is robust to the inclusion of a comprehensive set of controls, including sophisticated textual representations like TF-IDF and LDA. The bottom panel of Table 2.1 shows that even when added to a model containing all other feature groups, our isolation metrics provide a statistically, significant improvement, increasing completeness by an additional 2.2 percentage points. This improvement is especially meaningful given that LDA and TF-IDF (in all features) contain future information that is not available at the time of the paper is published.

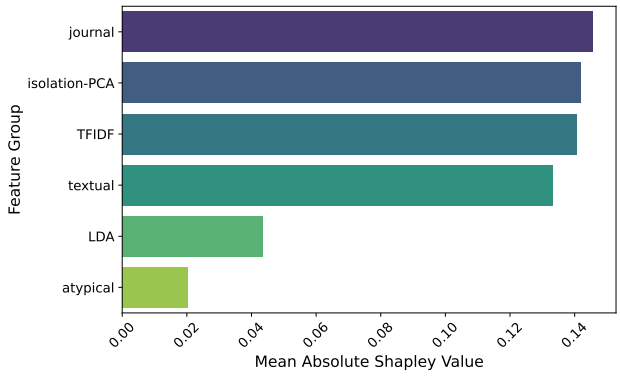


Figure 2.3: *Feature importance (mean absolute SHAP) for predicting log citations.*

To assess the relative importance of each feature group, we compute SHAP (SHapley Additive exPlanations) values[2]. SHAP values provide a unified measure of feature importance by attributing each prediction to individual features. They are grounded in Shapley values from cooperative game theory, ensuring a fair and consistent allocation of contribution across features. As shown in Figure 2.3, the isolation metrics (reduced to 200 dimensions via PCA for comparability) contribute more to the model's predictions than any other feature group, including TFIDF features. This confirms

[2]We apply 5-fold cross validation, and combine the prediction on test folds to calculate the SHAP contribution.

that semantic isolation is not a peripheral signal but a central predictor of scientific impact.

2.4.3 Predicting Disruptive Papers

We next turn to a more nuanced prediction task: identifying disruptive papers. This allows us to test whether our isolation metrics are merely a proxy for future attention (citations) or if they capture a more fundamental aspect of a paper's contribution. We frame this as a binary classification task, classifying a paper as disruptive if its Disruption Index (DI) is greater than zero (Wu et al., 2019).

Given the class imbalance (16.3% of papers are disruptive), we focus on precision as our primary evaluation metric. Our goal is showing that semantic isolation is a high-quality signal that reliably predicting disruptive papers. Precision answers the question: "When our model predicts a paper is disruptive, how often is it correct?" A high precision score provides strong evidence that semantic isolation is a reliable signal of disruptive potential. On the other hand, recall, which measures ability to identify every disruptive paper, is of secondary importance. We also report the F1-score and the completeness score (where error is defined as $(1 - precision)$. We use balanced class weights during training and report mean F1 and precision across 100 times bootstrapping.

Results. The results for disruption prediction, presented in the top panel of Table 2.2, mirror the patterns observed for citations. Our isolation metrics substantially outperform the atypicality score, achieving a 13.47% completeness score on precision compared to just 1.33% for atypicality. This consistency across different prediction tasks strengthens the evidence that isolation captures fundamental aspects of research novelty.

Moreover, the metrics are again complementary, as shown in the middle panel Table 2.2. Adding isolation metrics to a model that already includes the baseline and atypicality features increases precision completeness from 4.34% to 15.13%. This additional effect confirms that our

Feature set	Precision	Comp-prec	F1	Comp-F1
control+subfield	0.3140	0.0	0.4151	0.0
	(0.00053)		(0.0057)	
control+atypicality	0.3231	0.0133	0.4220	0.0118
	(0.00057)		(0.00058)	
control+isolation	0.4064	**0.1347**	0.4699	0.0937
	(0.00064)		(0.00059)	
control+subfield	0.3140	0.0	0.4151	0.0
	(0.00053)		(0.00057)	
control+subfield+atypicality	0.3437	0.0434	0.4424	0.0466
	(0.00054)		(0.00051)	
control+subfield+atypicality+isolation	0.4178	**0.1513**	0.4788	0.1089
	(0.00072)		(0.00064)	
all features (excl. isolation)	0.4270	0.1647	0.4915	0.1305
	(0.00070)		(0.00060)	
all features (incl. isolation)	0.4517	**0.2007**	0.5063	0.1559
	(0.00081)		(0.00067)	

Table 2.2: *Performance of predicting disruptive papers. The top panel shows results using individual feature groups; the middle and bottom panel show the additional value of isolation features. Standard errors (in parentheses) are calculated from 100 bootstrap iterations. Comp-prec shows the completeness of error in precision, and Comp-F1 shows the completeness of error in F1.*

measure of semantic distance captures a dimension of novelty relevant for disruption that is distinct from the recombinative novelty captured by atypicality.

The bottom panel of Table 2.2 shows that isolation metrics continue to provide significant predictive value even in the presence of all other features, increasing precision completeness from 16.47% to 20.07%. This 3.6 percentage point improvement demonstrates the consistent and non-redundant value of isolation across different impact dimensions.

The SHAP analysis[3] in Figure 2.4 further confirms that our isolation features are a key driver of the model's predictions. This confirms that isolation captures meaningful signals for identifying potentially disruptive research, further validating our approach across different dimensions of

[3]We apply 5-fold cross validation, and combine the prediction on test folds to calculate the shap contribution.

research impact.

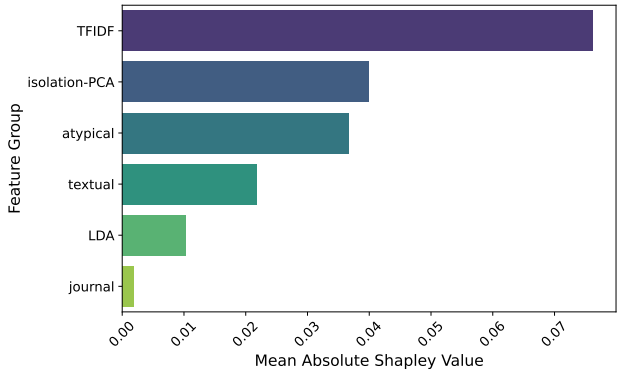


Figure 2.4: *Feature importance (mean absolute SHAP) for predicting disruption.*

2.4.4 Summary of Validation

Our predictive analyses provide robust evidence for the empirical relevance between our semantic isolation metrics and future impacts. Across two distinct impact measures, these metrics demonstrate substantial predictive power, significantly outperforming the established atypicality benchmark. Crucially, their predictive content is both complementary to atypicality and non-redundant with a comprehensive set of textual and bibliometric controls, confirming that they capture a novel dimension of scientific contribution rooted in a paper's semantic positioning and temporal dynamics. Notably, these features are calculated with information available at the publication time of a paper, making it a useful tool to assess the future impact at a paper's early stage.

As detailed in Appendix A.3.3, these findings are robust across economic subfields. Decomposition (see Appendix A.3.2) reveals that features related to papers' core contributions (paragraph summaries and question-conclusion framing) provide the strongest predictive signals. Appendix A.3.4 shows that including prospective metrics can significantly improve the predicting performance, serving as a validation of our temporal novelty concept.

2.5 A Two-Dimensional Framework of Novelty

Having established that our semantic isolation metrics robustly predict scientific impact, we now employ them as the measurement foundation for our two-dimensional framework of novelty. This section moves beyond treating novelty as a monolithic concept, demonstrating how distinct pathways to intellectual contribution can be empirically measured and systematically categorized.

2.5.1 Constructing Spatial and Temporal Novelty Scores

We operationalize our framework by constructing two composite scores that correspond to the conceptual dimensions of novelty outlined previously. These scores are calculated using only information available at the time of publication, ensuring clear interpretation of temporal relationships.

While we have a large set of semantic isolation features, we need to derive scalar values for spatial novelty and temporal novelty. Simple aggregation methods such as taking the mean are straightforward but would lose the nuances captured by individual features. For example, averaging the 30-NN distance and 3-NN distance would either be meaningless or dominated by the larger 30-NN values. Therefore, we employ Principal Component Analysis (PCA) to construct a composite score while preserving most of the variation in the features.

- **Spatial Novelty Score:** We apply PCA to the set of contemporaneous point-in-time isolation metrics (i.e., metrics calculated using all prior work, $t = 0$) and the baseline paper-level mean distance[4]. The first principal component serves as our spatial novelty score. This score captures a paper's static semantic isolation from the existing body of knowledge at its moment of publication.

[4]Basic features of semantic distance, introduced in Appendix A.1.4.

- **Temporal Novelty Score:** We apply PCA to the set of retrospective dynamic isolation metrics, which measure the change in a paper's semantic neighborhood in the years leading up to its publication (e.g., the change from $t = -3$ to $t = 0$). The first principal component serves as our temporal novelty score, capturing the degree to which a paper is situated within a rapidly evolving or "hot" research frontier.

The use of contemporaneous point-in-time metrics and retrospective dynamic metrics makes sure that our framework are not using any future information. By using the first principal component for each, we create summary indices that capture the largest shared dimension of variation across our granular metrics, reducing measurement noise while retaining the core signal of each novelty type. Further details on the PCA construction are provided in Appendix A.4.1.

2.5.2 A Typology of Research Neighborhoods

We classify each paper based on the state of its semantic neighborhood, defined by its position in the two-dimensional novelty space. A neighborhood is classified as "High" on a given dimension if the paper's score on that dimension is above the median. Crossing these two classifications yields a four-quadrant typology, with each quadrant representing a distinct archetype of research environment:

- **Consolidating Neighborhoods (Low Spatial / Low Temporal):** These are stable, well-established research areas where contributions are semantically close to existing work. Papers in these neighborhoods typically refine or incrementally extend established paradigms.
- **Outlying Neighborhoods (High Spatial / Low Temporal):** These are intellectually distant and slow-moving or static research areas. Papers here represent unconventional ideas disconnected from contemporary research trends.

- **Trendy Neighborhoods (Low Spatial / High Temporal)**: These are rapidly evolving research frontiers where contributions, while not semantically distant from their immediate peers, are part of a concentrated wave of scientific interest.

- **Trailblazing Neighborhoods (High Spatial / High Temporal)**: These are nascent, intellectually distinctive areas of inquiry that are also rapidly gaining momentum. Papers in these neighborhoods are positioned at the start of new, fast-growing research directions.

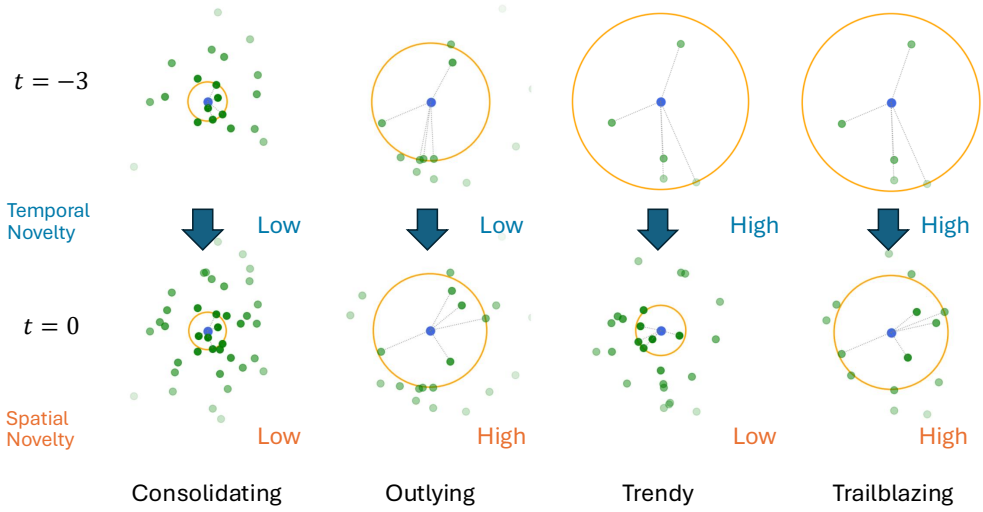


Figure 2.5: *A stylized illustration of the four novelty archetypes. The blue point is the focal paper; green points are prior work. The orange circle represents the paper's isolation level (e.g., its 5-NN distance) at publication ($t = 0$) and three years prior ($t = -3$). High spatial novelty is indicated by a large radius at $t = 0$. High temporal novelty is indicated by a large change in the radius between $t = -3$ and $t = 0$.*

It is crucial to emphasize that this typology characterizes the context in which a paper is published, not the intrinsic attributes of the paper itself. A paper is thus described by the type of neighborhood it inhabits (e.g., a paper in a Trailblazing neighborhood). This distinction is fundamental to our analysis, as it shifts the focus from a paper's identity to its strategic position within the scientific landscape.

Figure 2.5 provides a stylized illustration of these four archetypes in semantic space. High spatial novelty corresponds to a large isolation radius at the time of publication ($t = 0$). High temporal novelty is visualized by a significant decrease in the isolation radius between a prior period ($t = -3$) and the publication date, indicating a rapid densification of the paper's intellectual neighborhood.

2.5.3 The Structure of the Novelty Space

An analysis of the joint distribution of the two novelty scores reveals important structural features of the scientific landscape.

A Structural Trade-Off. The spatial and temporal novelty scores exhibit a statistically significant negative correlation $r = -0.26$. This finding provides quantitative evidence of a structural trade-off between the two pathways to novelty. In other words, it is empirically challenging for a single work to be both radically different from prior art (high spatial novelty) and situated at the heart of a burgeoning research trend (high temporal novelty). This trade-off is reflected in the distribution of papers across our typology (Figure 2.6). The off-diagonal quadrants (Trendy and Outlying neighborhoods) are the most populated (each containing about 30% of papers), representing the more common research strategies. The relative scarcity of papers in the Trailblazing neighborhoods (about 20%) underscores the difficulty of simultaneously achieving high spatial and temporal novelty. This negative correlation points to a stylized fact of scientific production: a structural trade-off exists between intellectual distinctiveness and timely engagement.

Mapping Economic Subfields. Figure 2.7 plots the median novelty scores for each economic subfield. The patterns align with what one might expect given these fields' methodological approaches. Economic History, with its emphasis on context-specific analysis, falls predominantly

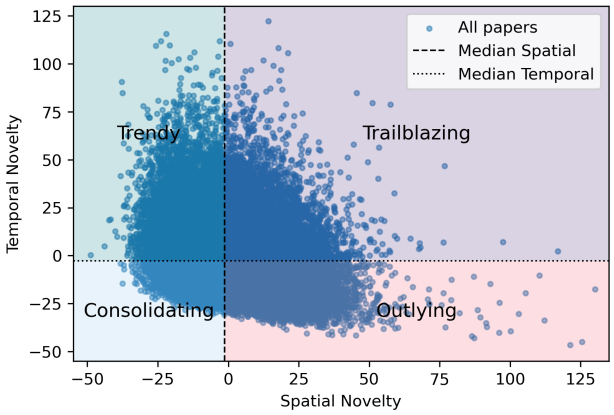


Figure 2.6: *The joint distribution of spatial and temporal novelty. The negative correlation suggests a structural trade-off.*

into the Outlying quadrant. Microeconomic Theory, building on a stable set of foundational models, centers in the Consolidating quadrant. Fields associated with recent rapid growth, such as Behavioral and Health Economics, appear in the Trailblazing quadrant.

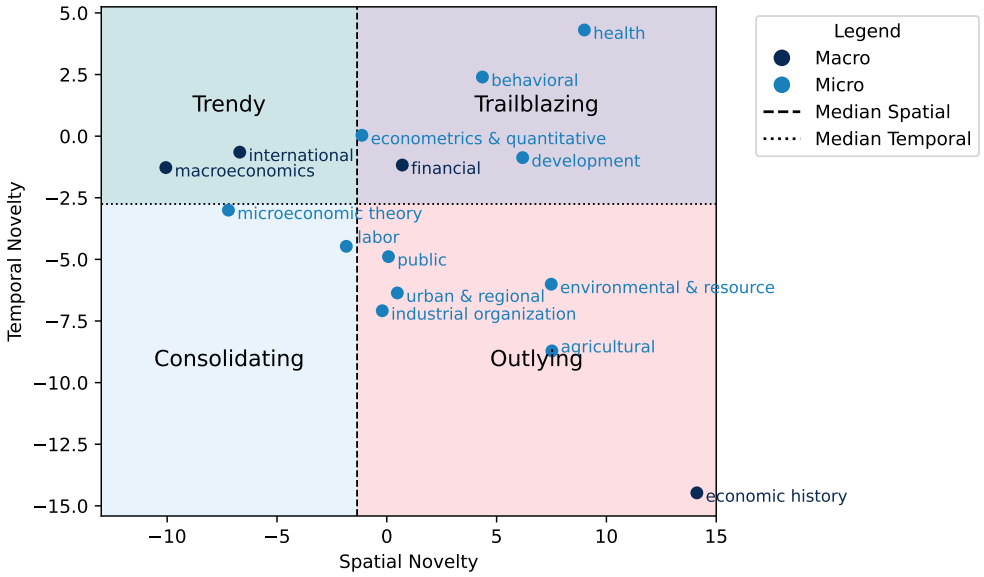


Figure 2.7: *Median novelty profiles by economic subfield. Each point represents the median spatial and temporal novelty score for all papers in a given subfield.*

While different fields tend to cluster in different quadrants, these mappings reflect only me-

dian tendencies and mask considerable within-field heterogeneity. Individual papers within every field span multiple archetypes. For instance, though Behavioral Economics exhibits higher novelty on average, it includes substantial Consolidating work that refines foundational insights alongside Trailblazing explorations of new domains. This diversity underscores that our framework characterizes research strategies, not research quality: each quadrant represents a legitimate scholarly approach with distinct contributions to knowledge accumulation. Importantly, our framework also reveals substantial overlap across fields with distinct median profiles. Appendix A.4.2 presents both the inter-quartile range version and the full distributions for Macroeconomics and Economic History, confirming that our novelty scores identify research archetypes that transcend field boundaries, and provides detailed distributions for all subfields that substantiate the structural trade-off and field-level patterns.

2.6 Novelty and Impacts

Having established our two-dimensional novelty framework, we now test its central prediction: that different pathways to novelty produce systematically different forms of scientific impact. We examine whether our four neighborhood archetypes predict differential patterns in two impact outcomes — high citation rates and disruptive influence. To ensure robustness, all results are averaged over 100 bootstrap iterations[5].

2.6.1 Temporal Novelty Predicts High Citations

Our first analysis examines the likelihood of a paper becoming highly cited, defined as reaching the top 25% of citations within its publication-year-subfield cohort. The results, presented in Figure 2.8, reveal a clear and powerful pattern: temporal novelty is the primary predictor of citation

[5]We provide standard errors in Appendix A.5.

success.

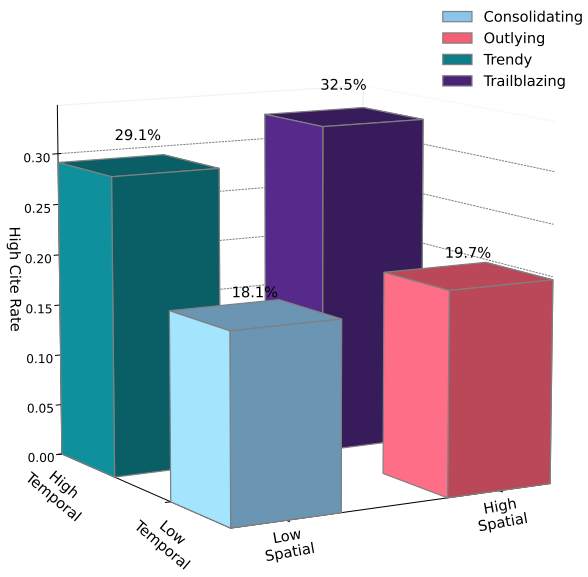


Figure 2.8: *Probability of Being a Highly Cited Paper (Top 25%) by Novelty Type. The baseline probability of being a top-25% cited paper is 24.77%.*

Increasing temporal novelty from low to high dramatically increases a paper's probability of being highly cited. Among papers with low spatial novelty, papers in Trendy neighborhoods are 1.6 times more likely to be highly cited than papers in Consolidating neighborhoods (29.1% vs. 18.1%). Similarly, among papers with high spatial novelty, papers in Trailblazing neighborhoods are 1.65 times more likely than papers in Outlying neighborhoods (32.5% vs. 19.7%). This high-temporal-novelty advantage of over 10 percentage points is substantial and consistent across both levels of spatial novelty.

In contrast, the effect of spatial novelty is modest. Moving from low to high spatial novelty increases the high-citation probability by only 1.6 percentage points for low-temporal-novelty papers and 3.4 percentage points for high-temporal-novelty papers. While statistically significant[6],

[6]Detail numbers appear in Table A.19, Appendix A.5.1.

this effect is an order of magnitude smaller than that of temporal novelty.

This pattern suggests that positioning a contribution within a dynamic research frontier confers a powerful attentional advantage. Papers that engage with rapidly evolving conversations benefit from the field's momentum, the influx of new researchers, and the dense citation networks that form around emerging topics.

2.6.2 Spatial Novelty Predicts Disruption

The relationship between novelty and impact reverses when we examine disruption. As shown in Figure 2.9, spatial novelty emerges as the dominant predictor of disruptive influence.

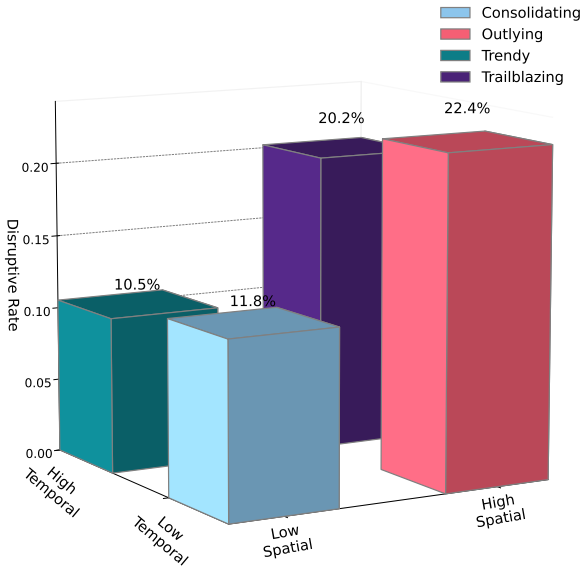


Figure 2.9: *Probability of being a disruptive paper by novelty type. The baseline probability of being disruptive is 16.30%.*

Increasing spatial novelty from low to high nearly doubles a paper's probability of being disruptive. Papers in Outlying neighborhoods are 1.9 times more likely to be disruptive than papers in Consolidating neighborhoods (22.4% vs. 11.8%), and papers in Trailblazing neighborhoods are

1.9 times more likely than papers in Trendy neighborhoods (20.2% vs. 10.5%). This high-spatial-novelty advantage is large and consistent across both levels of temporal novelty.

Conversely, temporal novelty is weakly negatively associated with disruptive potential. High-temporal-novelty papers are approximately 1.3 percentage points less likely to be disruptive than their low-temporal-novelty counterparts. This finding is conceptually intuitive: disruption, by definition, involves rendering prior work obsolete. Such a stark break is more likely to occur when a paper introduces a truly distinct perspective (high spatial novelty) rather than when it participates in a crowded, cumulative conversation at an evolving frontier (high temporal novelty).

2.6.3 Navigating the Trade-off: The Trailblazing Archetype

Our framework identifies Trailblazing as a unique archetype that successfully navigates the trade-off between spatial and temporal novelty. While these are the rarest neighborhoods, their impact profile is exceptional. Papers in Trailblazing neighborhoods are not only strong performers on both individual impact dimensions (32.5% high-citation rate, 20.2% disruption rate), but they also exhibit a disproportionate likelihood of achieving both forms of impact simultaneously.

As shown in Figure 2.10, Papers in Trailblazing neighborhoods have a 5.5% probability of being both highly cited and disruptive. This is nearly three times the rate of papers in Consolidating neighborhoods (1.9%) and significantly higher than that of papers in Outlying are (4.0%) or Trendy neighborhoods (3.1%). This high dual-impact probability suggests a powerful synergistic effect: being both intellectually distinct (spatial) and well-timed (temporal) does not merely add to a paper's impact, but amplifies its transformative potential, creating a rare pathway to both immediate relevance and lasting influence.

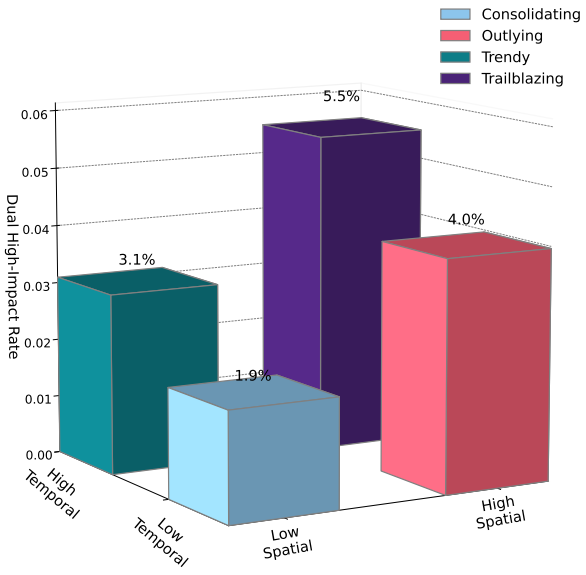


Figure 2.10: *Fraction of papers that are both top 25% high-cited and disruptive, categorized by novelty type. The base rate of being both high-cited and disruptive is 3.59%.*

2.6.4 Robustness and Alternative Specifications

The core patterns identified in our analysis prove robust across a wide range of alternative specifications. As detailed in the appendix, the dominance of temporal novelty for achieving high-citation status and of spatial novelty for generating disruptive influence holds when using stricter citation thresholds (top 10%, top 1%) and when examining disruption rates within different subsamples of highly-cited papers (Appendices A.5.1 and A.5.3). An analysis of median citation counts adds further nuance: while confirming the primary role of temporal novelty, it also shows that spatial novelty provides a significant positive boost to citation levels, particularly for papers already high in temporal novelty (Appendix A.5.2). The synergistic advantage in achieving dual impact is similarly robust for papers in Trailblazing neighborhoods (Appendix A.5.4).

Furthermore, the framework's integrity is confirmed by its insensitivity to key methodological choices. The fundamental relationships are preserved when using a more restrictive 75th percentile

threshold to define our novelty archetypes (Appendix A.5.5). Crucially, field-by-field analyses demonstrate that the observed patterns hold within subfields, confirming that our findings represent general pathways to scientific influence rather than mere artifacts of disciplinary differences (Appendix A.5.6).

This body of evidence strongly supports our multidimensional framework. It challenges unidimensional approaches to novelty assessment by demonstrating that the pathways to scientific impact are not singular but plural, each requiring distinct intellectual strategies and offering different rewards.

2.7 Conclusion

This paper challenges a monolithic view of novelty, proposing instead a multidimensional framework that distinguishes a paper's intellectual distinctiveness (spatial novelty) from its engagement with an emerging frontier (temporal novelty). We develop a scalable toolkit to measure these dimensions and provide robust evidence that they are tied to fundamentally different forms of scientific impact: citations and disruption. Our central finding is that the pathway to scientific influence is not singular, but plural. Recognizing these diverse pathways has direct implications for science policy and evaluation. Funding and promotion systems that rely heavily on citation-based metrics risk systematically favoring contributions in well-timed, Trendy neighborhoods at the expense of intellectually distant, Outlying work with disruptive potential. Our framework offers a more nuanced vocabulary and a practical toolkit for cultivating a more diversified and ultimately more innovative research portfolio. The conceptual shift from novelty as an intrinsic attribute to a positional good opens up numerous avenues for future research. It invites a new program of inquiry focused on the strategic dynamics of scientific discovery: how do these trade-offs manifest across researchers, institutions, and funding contexts? By providing a quantitative map of the in-

tellectual landscape, our work lays the groundwork for a richer, more ecological understanding of how science progresses.

CHAPTER 3
MANAGING STRATEGIC COMPLEXITY

3.1 Introduction

> "*The question [is] whether a simple attack, or one more carefully prepared ... will produce greater effects. [...] Together with the expediency of a complicated attack we must consider all the dangers which we run during its preparation, and should only adopt it if there is no reason to fear that the enemy will disconcert our scheme. Whenever this is the case we must ourselves choose the simpler, i.e., quicker way.* (Clausewitz, 1832)

In conflicts ranging from war to business to politics, the complexity of strategic interactions is determined in part by the players within them.[1] Managing complexity can be a way to gain a strategic advantage, but game theory has had little to say about how and why. What are the tradeoffs as adversaries decide whether to complicate or simplify decisions (their opponent's as well as their own)? How should an analyst model the endogenous evolution of the complexity of a strategic situation? We study these questions both theoretically and empirically in the context of a strategically rich game for which there is an enormous amount of data: chess.

From a theoretical perspective, chess is an attractive setting to study because it is a vivid instance of how standard game theory fails to account for the role of complexity. Zermelo's theorem says that since chess is a finite, two-player, extensive-form game without chance moves, every

[1] This chapter is based on joint work with Jeff Ely, Ben Golub, and Annie Liang.

position has a minimax value: a unique prediction of the final outcome under rational play.[2] The application of this result to chess is provocative because, in many positions, players neither know the minimax value nor expect either themselves or their opponents to play the strategies presupposed by Zermelo's theorem. This fact is important in practical play. For instance, chess coach Bruce Pandolfini advises, "If winning, clarify; if losing, complicate [the position]" (Pandolfini, 1992). The standard game-theoretic model of chess lacks a way to justify or analyze this intuitively appealing advice, and one of our aims is to develop a framework capable of such an analysis.

From an empirical perspective, chess is convenient because it is a rare example of an extensive-form game that is played in the wild by players who are experienced, motivated, deliberate, and sophisticated, where the game played by the players is precisely the game that is described by the extensive-form representation. Moreover, play of this game on Internet platforms generates a huge amount of empirical data that can be used to develop and estimate theories. We take advantage of this, drawing our data from the open database of over 3 billion games played on the popular platform Lichess.

We begin with a simple test of the standard game-theoretic model of chess: How well does the standard model predict average game outcomes in actual play, starting from a given position? We restrict attention to endgame positions, which occur in the final stage of a chess game when few pieces remain on the board. These positions have been completely solved through brute-force analysis, and so the minimax value of the game starting at any endgame position is known. We compare the prediction of the standard game-theoretic model of chess—i.e., the minimax value of the position—to the actual empirical average outcome.

To interpret the raw predictive performance of the model, we normalize a performance of 0 to be the performance of a naive benchmark, and a performance of 1 to be perfect prediction. Follow-

[2] In fact, Ewerhart (2000) shows that chess is dominance solvable in two steps.

ing Fudenberg et al. (2022a), we report the *completeness* of a model based on where in this range it falls: the fraction of possible improvement over the naive benchmark that the model achieves. We find that the game-theoretic model has a completeness of 72%, demonstrating that the standard backward induction solution of chess endgames is an important predictor of behavior. Nevertheless, there remains a sizable gap between the received game-theoretic (minimax) prediction and perfect prediction.

To assess how much of the remaining variation in outcomes is predictable, we next train a machine learning algorithm to predict the average game outcome given not only the position's minimax value, but additionally a large number of other features of the position. We find that this algorithm bridges much of the remaining gap, achieving a completeness of 93%.[3] Thus additional positional information, beyond what is captured by standard game-theoretic model, is important for predicting behavior.

In what follows, we focus on the role of one particular factor that matters for empirical play, but is not accounted for by the standard theory: the complexity of a position. Intuitively, among positions with the same minimax values, some are harder to understand and play correctly than others. Complexity is a multifaceted concept, and rather than attempting to analyze it fully, we define a simple heuristic measure of the complexity of a position. Our definition is based on whether the minimax value of the position can be predicted using apparent features of the position, or whether this value must be calculated through a more complicated brute-force analysis of the game tree. Specifically, we fix a heuristic procedure for assessing the value of a position, and define a position to be *strategically simple* if our heuristic prediction of the minimax value of the position is correct, and *strategically complex* otherwise. We find that this measure of complexity

[3]Several recent papers use machine learning to understand human behavior in games. For example, McIlroy-Young et al. (2020) trains an algorithm to mimic human play in chess, and Fudenberg and Liang (2019) and Hartford et al. (2016b) train machine learning algorithms to predict behavior in normal-form games.

does indeed predict deviations of actual play from the minimax prediction, where real behavior is closer to the standard game-theoretic prediction in simple positions than in complex positions.

We next use our measure of complexity to document an important cross-sectional fact: complex positions occur more often between higher-rated players. More precisely, complex positions occur 73% more often in games between the top tercile of players than in games between the bottom tercile.[4] Explaining why such variation occurs and interpreting what it implies about strategic management of complexity requires a model that takes into account the various incentives for complex play.

In the second part of the paper, we develop such a model. Our goal is to tractably capture some key forces relevant to endogenous choice of complexity in a zero-sum game. The model has two players who take turns moving, and an evolving *position state* variable—a coarse description of the position on the board. The position state encodes (among other things) whether the position is *simple* or *complex*. Moves available to a player induce distributions over the next-period position states. Players can choose to make simple moves that aim to produce simple states or complex moves that aim to produce complex states for the opponent. However, in moving, players may unwittingly make *blunders*, defined as moves that reduce the value of the game for them—for example, changing an objectively drawn position to one where the opponent has a forced win. This corresponds to misperceiving the position state that a move will lead to. Players are more likely to make blunders when they move in complex positions, and also when they make more complex moves. The key tradeoff is the one discussed by Clausewitz (1832) in the epigraph: a more complex move creates a harder situation for the opponent when it succeeds, at the cost of increasing the probability of one's own blunder in implementing the strategy. At the time of a move, a player has a stochastic signal, called an *insight*, which informs him of how likely he is to

[4]Complex positions account for 14.8% of positions between high-rated players and 8.5% of positions among low-rated players.

avoid blundering if he attempts a complex move. The random insight draw reflects factors such as experience with the position, time available to think, etc. We study Markov-perfect equilibria, focusing on symmetric equilibria played by equally-skilled players.

We first fully analyze a special case of our model with two position states (simple and complex), and use this characterization to highlight some of the key forces shaping endogenous complexity choice. We show that any Markov-perfect equilibrium is associated with a cutoff: how strong players' insights have to be in order for them to choose complex moves. We characterize the equilibrium cutoff and study local comparative statics of behavior as we change players' skill, defined as the player's blunder probability in both simple and complex positions. The main message of this exercise is that increasing the skill of both players has a theoretically ambiguous effect on the prevalence of complex positions. Intuitively, the reason is that making players more skilled has two effects. On the one hand, a more skilled opponent blunders less in all positions, and therefore the value to putting a skilled opponent in a complex situation is smaller. On the other hand, a more skilled player is less worried about the persistence of complexity, since he himself is better at playing such positions; this reduces the cost of taking the game in a complex direction. Thus, stronger players may be more or less likely to complicate a position (relative to weaker players), and which possibility obtains is an empirical question.

To investigate this question, we structurally estimate a richer version of our model. To do this, we use our classifications of positions based on complexity to define position states, and measure transitions among these states in the data. We recover a structural estimate of the cutoff for playing complex moves in different rating groups. The estimates show that strong players are in fact more cautious than weak players given the same information. Interpreting this finding, we conclude that high-rated players spend a lot of time in complex positions not because they are more willing to take risks, but because they are able to maintain complex positions for many moves without

blundering.

3.1.1 Related literature

We build on a literature that studies the role of complexity for decision-making. This literature has primarily focused on single-agent decision problems, for example finding that people make more errors when the environment is complex (Caplin et al., 2011; Abeler and Jäger, 2015; Oprea, 2020) and that the complexity of the decision environment predicts which behavioral biases manifest (Ba et al., 2023). Recent papers (like us) take advantage of the enormous Lichess database to study decision-making through chess. For example, Russek et al. (2022) and Howard (2023) provide evidence that chess players are rationally attentive over how to allocate their thinking time, and Salant and Spenkuch (2023) show that players make more errors in more complex positions (i.e., ones with a longer depth-to-mate).[5]

Relative to the above papers, our interest is in the role of complexity for strategic interactions. This is a conceptually different problem because the complexity of a strategic environment is often determined by the players. Thus a first-order question is when (and why) a player should choose to complicate the environment. There has been relatively little study of how individuals strategically manage complexity to either take advantage of an opponent's bounded rationality or protect against their own bounded rationality.[6] This is especially so in extensive-form games, where the study of bounded rationality is often intractable. Different from classic papers in behavioral game theory (see (Crawford et al., 2013) for a survey), we do not explicitly model the players' bounded rationality, but instead propose a stylized model that focuses on strategic management of complex-

[5] Although we use a different definition of complexity, our finding that the standard game-theoretic model predicts worse in complex positions echoes this result.

[6] Some notable exceptions include Gabaix and Laibson (2006)'s models of firms' choices to shroud attributes when interacting with myopic consumers, and Foarta and Morelli (2023)'s model of endogenous complexity choice for policy reforms.

ity.

Finally, a small literature tests game theoretic predictions "in the field" using real-world data from professional sports games (Walker and Wooders, 2001; Palacios-Huerta, 2003; Chiappori et al., 2002). These papers consider simultaneous normal-form games, e.g., penalty kicks where each player has two possible actions. Our paper builds on this literature by contributing one of the first field tests of game-theoretic predictions in a rich extensive-form game, finding the standard game-theoretic model to be quite predictive but incomplete. The feasibility of our analysis is due to an unusual combination of properties of chess: (1) the game involves rich dynamics, (2) the game-theoretic prediction can be computed, and (3) there is a large quantity of data.

3.2 Data analysis

3.2.1 Description of data

Our data is built using the open database of the popular online chess platform Lichess. We include all standard rated games with Bullet, Blitz, or Rapid time controls played between January, 2013 and October, 2022.

We test the standard game-theoretic model of chess by predicting how play starting at a given position will end. Our dataset is $\mathcal{D} = \{(p_i, y_i)\}_{i=1}^{n}$, where p_i is the FEN code for a position (which fully specifies the state of the board and the player to move), and y_i is the average outcome in games in which position p_i occurred. That is,

$$y_i = \frac{1}{|G(p_i)|} \sum_{g \in G(p_i)} o_g$$

where $G(p_i)$ is the set of games g in which position p_i appears, and $o_g \in \{0, 1/2, 1\}$ is the outcome of game g (where 0 corresponds to a win for Black, $1/2$ corresponds to a draw, and 1 corresponds

to a win for White). We predict y_i given the position p_i, using mean-squared error as the loss function, i.e., $\ell(\hat{y}_i, y_i) = (\hat{y}_i - y_i)^2$ where $\hat{y}_i$ is the predicted average outcome.

We apply a few restrictions to which positions are considered. First, we include only positions with 4-6 pieces, i.e., endgame positions. These positions have the advantage of being completely "solved," i.e., it is known whether White can force a win, Black can force a win, or if either player can force a draw at this position. We thus know the standard game-theoretic prediction (which we want to test) and do not need to rely on proxies such as chess engine evaluations. Besides this restriction, we include in $G(p_i)$ only games in which the player to move at position p_i has at least 30 seconds on the clock, and require that $G(p_i)$ include at least 100 games.

These restrictions leave us with a total of 604,081 positions from 172,543,394 games, where each position appears in an average of 285.6 games. Figure 3.1 reports a histogram of average game outcomes across all positions in our data.

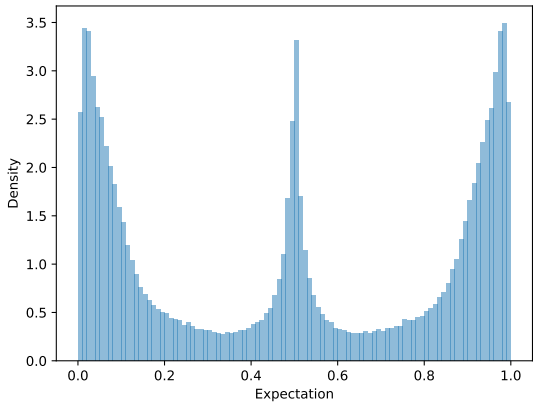


Figure 3.1: *Distribution of the average game outcome across all positions in our data set.*

3.2.2 Testing the standard game-theoretic prediction

The minimax value of the game starting from a specific position is known as the *tablebase* value of the position (named after a database comprising these values). We denote the tablebase value of

position p_i by $v_{TB}(p_i) \in \{1, 1/2, 0\}$. According to the textbook model of chess, all games in which position p_i arise will result in the outcome $v_{TB}(p_i)$. Thus, the prediction rule we call Tablebase is the following.

Tablebase: In each position p_i, predict $v_{TB}(p_i)$.

To understand the size of the absolute error of the Tablebase prediction rule, we compare it to the performance of a naïve benchmark, which we interpret as a "worst-case" prediction rule for this problem. The benchmark we use is informed by a common heuristic in chess: Each piece is assigned a point value,[7] and the strength of each side is computed as its total point count. A player with a higher point count is said to hold a "material advantage." For each position p_i, we define $v_{MA}(p_i)$ to take value 1 if White has a material advantage, 0 if Black has a material advantage, and $1/2$ if the two sides are equal in material value.

Material Advantage. For each position p_i predict $v_{MA}(p_i)$.

Table 3.1 reports the mean-squared error of each of these prediction rules. We find that Tablebase predicts the average game outcome much better than our material advantage benchmark, demonstrating that game-theoretic properties are highly predictive of the final game outcomes.[8]

Nevertheless, there remains a substantial gap between the performance of this rule and perfect prediction.[9] We now quantify this. The completeness of a prediction rule is defined to be its improvement over Material Advantage, divided by the maximum improvement possible (i.e., what would be achieved by perfect prediction). Table 3.1 reports the completeness of Tablebase to be

[7] Pawns are each worth one point, knights and bishops are each worth three, rooks are worth five points, and the queen is worth nine points (Shannon, 1950).

[8] Since our Tablebase and Material Advantage rules are each non-nested within one another, in principle Tablebase could under-perform our benchmark for predicting actual play.

[9] In Appendix B.1.1, we report the performance of a more flexible variant on our Tablebase model, in which positions are partioned based on their tablebase value, but the prediction for each partition is empirically estimated from data.

72% (i.e., Tablebase reduces 72% of the error of the material advantage benchmark relative to perfect prediction). Thus while the tablebase value is an informative statistic of the position for predicting the average game outcome, it is not the only statistic that matters: There is additional predictive signal in the position.

We can attempt to uncover some of that signal by defining features to describe other strategic properties of the position. These features cover a broad range of categories. Some are summary statistics of the position, such as each player's total piece count and the difference between these counts (thus nesting the Material Advantage benchmark model). Some are more detailed descriptions of the position, such as whether the position involves only kings and pawns, and the area of the smallest rectangle that encloses all of the pieces. Some are descriptions of the complexity of the position—for example, the fraction of feasible moves that are non-blunders (i.e., that maintain the current tablebase value), and the minimum number of moves required to force a win. We also include the tablebase value of the position as a feature (thus nesting Tablebase as a model). The complete list of 66 features can be found in Appendix B.3.

We train a random forest algorithm on these features, and evaluate its five-fold cross-validated mean-squared error.[10] We find that this machine learning algorithm removes much of the remaining error, achieving a completeness of 93%. The strong performance of this algorithm makes it clear that the imperfect predictiveness of Tablebase is not due to average game outcomes being fundamentally unpredictable from positional information. Rather, there is systematic variation in game outcomes that cannot be predicted from the tablebase value of a position, but can be predicted from other features of the position. Table 3.1 is thus a strong refutation of the strawman model of chess as a game of perfect information played by unboundedly rational players.

[10]That is, we split the data into five equally sized sets of positions. In each round of cross-validation, we select one of the subsets to use as a test set, train the model on the remaining data, and evaluate the average mean-squared error on the test observations (the *test error*). The five-fold cross-validated mean-squared error is the average test error across these rounds.

	Error	Comp
Perfect Prediction	0	1
ML Algorithm	0.0035 (1.1×10^{-5})	93%
Tablebase	0.0138 (3.3×10^{-5})	72%
Material Advantage	0.0494 (12.2×10^{-5})	0

Table 3.1: *Performance of the different prediction methods.*

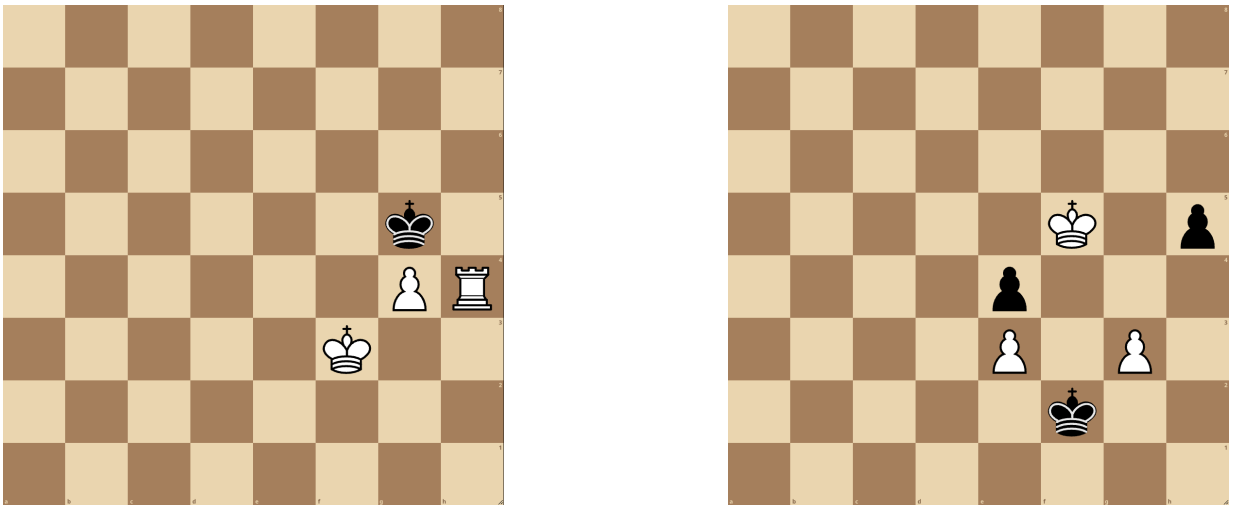
3.2.3 Strategic complexity

One feature that Tablebase does not account for is the complexity of the position. For example, both positions in Figure 3.2 are White wins according to Tablebase, but the win is easy to see in the first position (Panel (a)), and not so easy to see in the second (Panel (b)). In actual game play, the first position leads to a White win 90% of the time, while the second position leads to a White win only 17% of the time.

Rather than attempt to precisely capture what complexity means in chess, we instead define a heuristic measure. Roughly speaking, we think of the position as simple if the tablebase value of the position is easy to discern based on a small number of interpretable features of the position. To implement this, we train a short decision tree (depth-5) to predict the tablebase value of a position given other interpretable features of the position.[11] Let $\hat{v}_{TB}(p_i) \in \{0, 1/2, 1\}$ denote the prediction of the decision tree.

Definition 1. Say that a position is *strategically simple* if the decision tree prediction is correct,

[11] A decision tree is a popular machine learning algorithm that partitions the feature space and learns a constant prediction for each partition element. We implement a version of this algorithm in which the tree recursively partitions the feature space using a True/False question about the value of some feature, which is chosen to greedily minimize mean squared error. We constrain the decision tree to only include 5 levels of features (resulting in a depth-5 decision tree), and train it using our entire dataset of positions. Even with 5 levels of features, the trained model is not interpretable, so we do not include it here.



(a) *Example "simple" position*

(b) *Example "complex" position*

Figure 3.2: *In both positions, the tablebase value is a White win. But empirical play diverges substantially from this prediction in the left panel and not in the right. We may intuitively consider the left panel an example of a "simple" position and the right panel as an example of a "complex" position, and the classification rule we develop agrees with this labeling.*

i.e., $\hat{v}_{TB}(p_i) = v_{TB}(p_i)$, and otherwise say that the position is *strategically complex*.

To interpret this definition, observe that the model learned by the decision tree algorithm is optimized to predict well *in general* across positions. When this model fails to predict well, suggesting that the position at hand is atypical, we call the position complex. The position in Panel (a) of Figure 3.2 is labeled simple by our measure, while the position in Panel (b) is labeled complex.

Since our complexity measure is a function only of the position (and not based on empirical data of play), we can meaningfully ask how it correlates with actual play. We do this in Table 3.2, where we compare the performance of Tablebase across simple and complex positions. We find that, as expected, Tablebase is a less accurate predictor of final game outcomes in complex positions.

We next study the aggregate distribution of positions according to our measure, which reflects the endogenous choices of players regarding whether to pursue more complicated or simpler lines

	Simple Positions	Complex Positions
Tablebase Error	0.0125	0.0206
	(3.4×10^{-5})	(10.5×10^{-5})

Table 3.2: *Tablebase is a less accurate predictor in complex positions.*

of play. We consider the distribution of positions separately for players in three rating categories: The *low-rated* group consists of players whose ratings are below the 33rd percentile, the *medium-rated* group consists of players whose ratings are between the 33rd and 66th percentile, and the *high-rated* group consists of players whose ratings are above the 66th percentile.[12,13]

Figure 3.3 reports the aggregate incidence of complex positions in games between players of each of our three rating groups. In games between low-rated players, approximately 10.87% of the positions that occur are complex. This number rises to 13.83% for games between medium-rated players, and 18.45% for games between high-rated players. The stylized fact that emerges is thus that complex positions occur more often in games between higher-rated players.

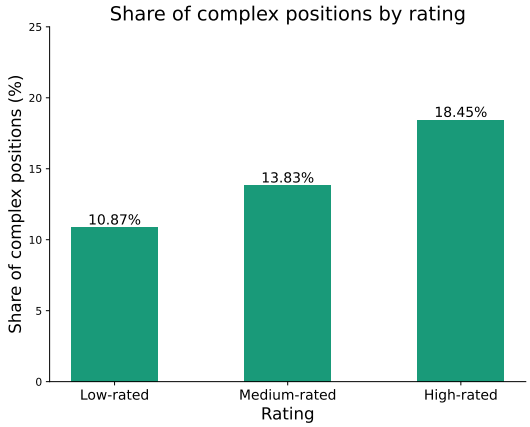


Figure 3.3: *Complex positions emerge more often in games between higher-rated players.*

[12]The Elo rating of a player is a constantly-updated statistical estimate designed to satisfy the property that two equally rated players are expected to be equally matched. We use the reported Lichess rating distribution, according to which the 33rd percentile rating is approximately 1280, and the 67th percentile is approximately 1670.

[13]See Appendix B.1.3 for a reproduction of Table 3.1 for players in each of these three rating groups.

This cross-sectional comparison shows that more skilled players bring about a different distribution of complexity through their moves. But our aggregate statistics reported in Figure 3.3 confound several possible reasons for this variation.

One possibility is that players have a stronger incentive to pursue complex moves against higher-rated opponents, because this is the only way to induce those opponents to blunder. Thus, players are more likely to risk a complex line of play in these higher-rated games, even when they are not confident that this line will be to their advantage. This story suggests a higher *benefit* to complex moves for higher-rated players. Alternatively, it could be that higher skilled players are simply better at thinking through complex lines, and hence less likely to blunder when executing these lines. In this case complex moves are implemented at a lower *cost* for higher-rated players.

A third possible force, which could operate alongside the channels just discussed, is that better players actually choose complex moves relatively rarely, but that those moves lead to *no-exit* complex positions, i.e., hard lines with few opportunities to transition out into simple play as long as players do not blunder. In this possibility, the high incidence of complex positions is a consequence of higher-rated players being able to play correctly in persistent no-exit lines, rather than playing speculatively often.

We do not have good a priori reasons to forecast which of these different explanations is more important. To tease out these different strategic forces and better structure our interpretation of the data, we develop a model that will allow us to more carefully analyze the various channels.

3.3 Theoretical Analysis

We now present a simple dynamic game, with the purpose of studying how strategic choices shape the complexity of a game. We describe a general framework in Section 3.3.1 to introduce key terminology and ideas. In Section 3.3.2, we specialize to a binary state case to illustrate some

important strategic forces in the model. We show that the effect of skill on players' strategic choices is theoretically ambiguous: In equilibrium, more skilled players may be more or less inclined to try complex moves given the same subjective probability that they are blunders. In the subsequent Section 3.4, we extend this basic model to include other, more realistic, features of actual play, and structurally estimate this model.

3.3.1 General framework

There are two players, $i = 1, 2$. Time is discrete. Player 1 moves at odd periods and 2 moves at even periods.

There is a set of *position states* $\mathcal{X}$. (All sets will be taken to be finite.) Each position state $x \in \mathcal{X}$ is associated with a set of possible moves, M_x. Each move $m \in M_x$ is associated with a transition distribution $\tau_{x,m} \in \Delta(\mathcal{X})$ over the next position state. The moving player knows the current state x and receives a signal about the vector $(\tau_{x,m})_{m \in M_x}$, which gives the player a subjective belief over the consequences of various moves.[14] We denote the signal, or *insight*, received by the player moving at time t by I_x^t; the distribution of the signal may depend on the current position state x. The signal takes values in some space $\mathcal{I}_x$.

A subset $\mathcal{E} \subseteq \mathcal{X}$ of position states are *terminal* states. Each such state $x \in \mathcal{E}$ is assigned a value $v_x \in \{0, \frac{1}{2}, 1\}$, corresponding, respectively, to whether that state yields player 1 a loss, a draw, or a win.[15] The game is zero-sum; it gives player 1 the value v_x at the first terminal state x reached, and player 2 a value of $1 - v_x$.

We will focus on Markov strategies. A Markov strategy prescribes, for every state $x \in \mathcal{X} \setminus \mathcal{E}$, an x-dependent map $\sigma_x : \mathcal{I}_x \to M_x$. A Markov strategy profile is a Markov-perfect equilibrium

[14] We omit notation for the prior over this variable.

[15] We will work with specifications under which a terminal state is reached with probability 1, making it immaterial how we define what happens otherwise.

if, given the other's strategy, neither player can find any other dynamic strategy (not necessarily Markov) that yields a higher expected payoff.

3.3.2 Binary position states

The general model allows for position states to be a rich description of the position, for example determining the kind of moves that are available at that position. We now specialize to binary position states that simply encode whether the position is simple or complex. We demonstrate that even in this minimal model, the incentives for complexity are nuanced.

Sections 3.3.2.1 and 3.3.2.2 describe the game; Section 3.3.2.3 solves for equilibrium; and Section 3.3.2.4 considers a comparative static of players' strategies with respect to player ability.

3.3.2 Setup

There are two possible non-terminal position states $\mathcal{X} = \{s, c\}$ called "simple" and "complex." At every position, the set of possible moves is $M_x = \{S, C\}$; that is, there is always a simple action and a complex action available.

At each time t, if the current state $x_t \in \{s, c\}$ is non-terminal, the player to move, i, chooses a move $m_t \in \{S, C\}$. Then Nature determines a random transition to the next position state x_{t+1}. One possibility is that the game transitions to the terminal state $\mathfrak{b}_i$ where player i loses; this is called a blunder. The probability of a blunder depends on the position state x_t and the player's move m_t, as described in Figure 3.4. All else equal, the probability of blundering is higher in complex positions and after complex moves. Specifically, a simple move in position x avoids a blunder with probability α_x, where $\alpha_c > \alpha_s$.

A complex move in a position x instead avoids a blunder with probability $\alpha_x I_x^t$. This random variable, the *insight draw* I_x^t, takes values in $[0, 1]$, and represents the moving player's information

about the transitions. We interpret it as a reduced form for various factors, such as how much time the moving player has to think through the consequences of the currently available complicating tactic, or how well the player has studied this particular line of play. Higher realizations of insight mean that the relative cost of a complex move (compared to a simple move) is lower. When $I_x^t = 1$ then simple and complex moves lead to equal probabilities of a blunder.

The realizations of insight I_x^t are i.i.d. across periods t. We assume that each I_x^t has an atomless density, and use F_x to denote the CDF of I_x^t. Finally, we assume that the distribution of I_s^t first-order stochastically dominates the distribution of I_c^t; that is, insights are stochastically higher and thus complex moves are less risky in simple positions than in complex positions. We focus on a symmetric version of this game, where all parameters and distributions are identical across the two players.

3.3.2 Discussion

The basic strategic incentives in this model are as follows. Choosing a complex move has three implications: First, it leads to a higher probability of blunder, which reduces the incentive to choose a complex move. Second, conditional on not blundering, it puts the opponent in a complex position next period (thus facing a higher probability of blunder), which increases the incentive to choose a complex move. Finally, it affects the subsequent evolution of complexity of future positions, with implications that depend on what the equilibrium looks like.

We collect below some further comments on the model:

The interpretation of a period. It is possible to think of each "move" in our model as a stand-in for a sequence of play. In this case a complex move means engagement in a complex sequence, for example a tricky exchange. The probability of blunder is then actually the total probability of a blunder at some point in this sequence. When we introduce a richer model suited to structural

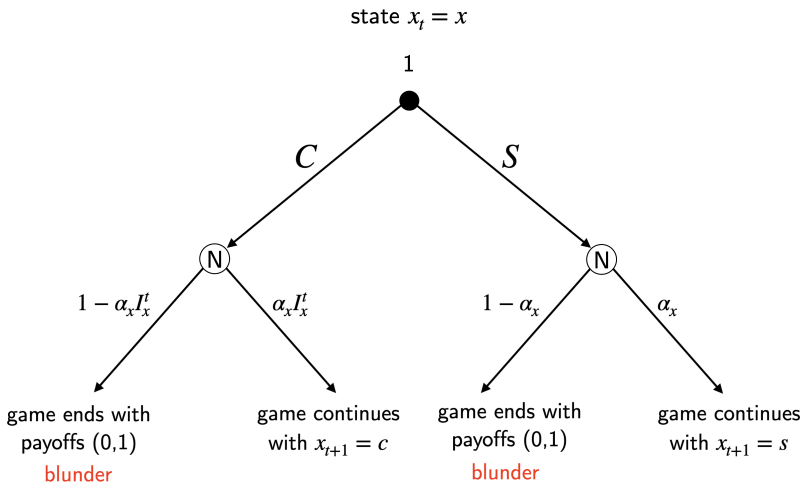


Figure 3.4: *Description of the game*

estimation in the next section, we will explicitly model this aspect.

Reduced form model. We don't explicitly model the specific chess moves that player might make as part of an extensive-form game. We view this simplification as a key strength of the model, since it makes the study of boundedly rational players possible even in this (very rich) extensive-form game. We make progress by qualitatively categorizing moves into "blunders" and "non-blunders," and focus on decisions for whether to go down complex versus simple paths in the game tree.

Decisiveness of blunders. In this model, blundering transitions to an objectively lost position with a payoff of 0. A richer model would add more states to capture uncertainty over whether the opponent will actually be able to achieve a win. We abstract from this possibility for a cleaner analysis, although the general framework from Section 3.3.1 can accommodate such an elaboration.

3.3.2 Equilibrium existence and characterization

Our solution concept is Markov-perfect equilibrium (subsequently simply *equilibrium*), so that each player's strategy is a function of that period's position state x_t and insight draw $I^t_{x_t}$. Our first result says that in any Markov-perfect equilibrium, player strategies are characterized by threshold rules.

Proposition 1. *Consider any Markov equilibrium. In this equilibrium, players' strategies are characterized by a threshold $\rho \in [0,1]$, such that at any period t with position x and insight draw I^t_x a player plays Complex if $I^t_x > \rho$, and Simple if $I^t_x < \rho$ (with ties broken arbitrarily).*

To characterize equilibrium thresholds, it will be useful to keep track of some statistics of play. For any cutoff ρ, let $p_x(\rho) = 1 - F_x(\rho)$ be the equilibrium probability of playing a complex move in position x. Moreover, define the *successful gambit probability* of state x as

$$g_x(\rho) = \alpha_x \int_\rho^1 r \, \mathrm{d}F_x(r).$$

This is the equilibrium probability that the player plays C and does not blunder. Given these quantities, define

$$\mathcal{R}(\rho) := \frac{1 - g_s(\rho) + g_c(\rho)}{1 + \alpha_s(1 - p_s(\rho)) - \alpha_c(1 - p_c(\rho))} \tag{3.1}$$

Using this function, we can write down a simple, one-dimensional characterization of Markov-perfect equilibria.

Proposition 2. *A Markov-perfect equilibrium in threshold strategies exists. Every equilibrium with threshold ρ satisfies $\rho = \mathcal{R}(\rho)$ and is interior.*

From now on, we focus on equilibria that are stable to best-response dynamics.[16]

[16]This focus is not crucial to the main message of this section—that comparative statics in ability are theoretically

Definition 2. A Markov-perfect equilibrium with threshold ρ is *locally stable* if $\mathcal{R}'(\rho) < 1$.

This definition can be motivated with a standard learning dynamics. The condition implies that small common perturbations of the threshold ρ are eliminated by adjustment dynamics[17] and the system converges back to ρ (see Figure 3.5).

Proposition 3. *A Markov-perfect, locally stable equilibrium in threshold strategies exists and is interior.*

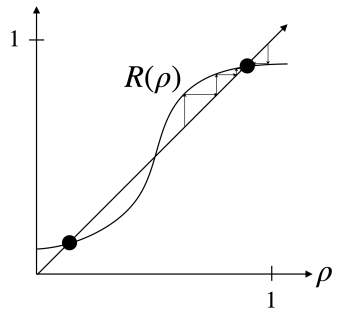


Figure 3.5: *For this example function $\mathcal{R}(\rho)$, there are two stable equilibria and one unstable equilibrium.*

3.3.2 Comparative statics in ability

We now examine comparative statics for how play at stable equilibria changes in player ability, which we model as improvement in the accuracy levels (α_s, α_c). We analyze a perturbation that increases both α_s and α_c.

Proposition 4. *Fix initial parameters (α_c^0, α_s^0) and an interior locally stable equilibrium with threshold ρ_0. Letting $(\dot{\alpha}_c, \dot{\alpha}_s)$ denote changes in these parameters, there is a unique continuous function $\rho(\dot{\alpha}_c, \dot{\alpha}_s)$ defined on a neighborhood of* **0** *describing the comparative statics of the*

ambiguous. However, focusing on stable equilibria makes that message more robust, since unstable equilibria often have strange comparative statics.

[17]For completeness, we specify the dynamics we have in mind. Each player's threshold is public. In continuous time, players have the opportunity to revise their thresholds at some rate. When they do so, they revise them myopically, best-responding to the threshold of the last game they played (or an average of recent games).

equilibrium threshold, with $\rho(\mathbf{0}) = \rho_0$*. Then the threshold* ρ *decreases and play of* C *increases,* $\rho(\dot{\boldsymbol{\alpha}}) < \rho_0$*, if*

$$\frac{\dot{\alpha}_s}{\dot{\alpha}_c} > \gamma, \tag{3.2}$$

where $\gamma > 0$ *depends on the equilibrium. The threshold* ρ *increases if the inequality is reversed.*

As a special case of this result, suppose players improve only in simple positions ($\dot{\alpha}_s > 0, \dot{\alpha}_c = 0$) or only complex positions ($\dot{\alpha}_s = 0, \dot{\alpha}_c > 0$). Proposition 4 says that (holding all else equal), the equilibrium cutoff ρ decreases in the first case and increases in the latter. That is, players are more willing to pursue complex lines as their play (uniformly) improves in simple positions, and players are less willing to pursue complex lines as their play (uniformly) improves in complex positions.

In general, when players simultaneously improve in both simple and complex positions, then the impact on incentives for complexity can go either way. The condition in (3.2) says that play of C increases if players get better at simple positions sufficiently fast relative to their improvement in complex positions. Intuitively, this condition compares the current risk of the complex move relative to the gain in the probability of the opponent blundering next period. When the ratio in (3.2) is large, then the perturbation makes it more worthwhile to put the opponent in a complex position.

The quantity γ on the right-hand side of (3.2) is explicitly computed in Section B.2.3. An interesting limit case occurs when the players' cutoff ρ for complex moves is close to 1. This is the case if complex moves are chosen only when players are highly confident; we will see below that this is a relevant case in the data. Then γ is approximately 1, so what matters for the direction of change in ρ is whether players improve more in simple positions than in complex ones.

3.4 Structural Estimation

In the data, we found that higher-rated players end up in complex positions more often than lower-rated players. As the simple model makes clear, both skill and strategic choice have a role to play in the observed prevalence of complexity. Moreover, we saw that the strategic channels are subtle—whether higher-rated players are more inclined to choose complex positions is not determined by theoretical considerations, and depends on the details of the environment. To better understand the strategic forces that shape the prevalence of complexity in our data, we now structurally estimate a richer version of our model.

The model in the previous section was highly simplified for illustrative purposes, and did not include many features relevant to real play. For example, in reality, insight distributions are not held fixed as we vary player skill levels, as we assumed in the comparative statics studied in Section 3.3.2.4. Another important feature in the data is that, at many moves, players do not have a choice between simple and complex positions. In particular, complex positions often beget complex positions, with no possibility for a (non-blunder) exit to a simple position. This is clearly a potentially important feature in explaining the prevalence of complex positions in reality, and should be assessed alongside the strategic channels.

To accommodate these features, in this section, we develop a richer version of the model, and subsequently structurally estimate from the data the strategic choices of complexity of players at various rating levels. The estimates will imply that higher-rated players actually have more stringent cutoffs for choosing complex moves.

3.4.1 A richer model

We begin by describing the expanded set of position states $\mathcal{X}$. Each position is characterized first by whether the position is simple or complex, and second by whether the set of available moves is $\{S\}$ (only simple moves are available), $\{C\}$ (only complex moves are available), or $\{S, C\}$ (both simple and complex moves are available). We describe each position by the pair $x = (x_p, x_f) \in \{s, c\} \times \{S, C, SC\}$, where the first component, x_p, refers to the present state of the position, and the second component, x_f, refers to the reachable futures. Thus there are 6 non-terminal states.

We now describe the transitions between position states, i.e. the distributions $\tau_{x,m}$. When a player is called to move at a state $x = (x_p, x_f)$, he avoids a blunder and implements the desired type of position m_t with a certain probability: this probability is α_{x_p} if he plays S and $\alpha_{x_p} I^t_{x_p}$ if he plays C. The numbers α_{x_p} are known constants, while the insight draw $I^t_{x_p}$ has a known distribution F_{x_p} that admits an atomless density. We will assume that $\alpha_{x_p} < 1$ and $\alpha_{x_p} I^t_{x_p} < 1$ for both values of x_p, but do not otherwise restrict these numbers or distributions. If the moving player i does not avoid a blunder, then the next position state is $x_{t+1} = \mathfrak{b}_i$, i.e., a loss for player i.

Conditional on a blunder not happening, the move $m \in M_x$ induces a distribution $\mu_{x,m} \in \Delta(\mathcal{X})$ over states for the other player.[18] This distribution must be consistent with m in the sense that if $m = S$ (the move is simple), then $\mu_{x,m}$ is supported on position states x' with $x'_p = S$ (the next position is simple), while if $m = C$ (the move is complex) then $\mu_{x,m}$ is supported on position states with $x'_p = C$ (the next position is complex). What μ determines is the distribution over x'_f for the next position state, i.e., what types of futures are reachable from the next position.

The payoffs are just as described in Section 3.3.2. As before, we focus on symmetric equilibria.

[18] Note that this is the distribution τ in the general model conditioned on not blundering.

3.4.1 Discussion

The main generalization relative to the binary-state model from Section 3.3.2 is that making a move successfully (i.e., without a blunder) can induce different distributions over next-period position states, depending on the value of the current position state x. For example, the model allows for the complex state to be more persistent, in the sense that it is more likely that the only move subsequently available to the opponent maintains the complexity of the position (corresponding to the state $(c, \{C\})$).

The non-terminal positions should be interpreted as objectively drawn positions, because a player can only transition to a terminal tate by making a blunder that gives the other player a win.[19] A non-blunder move should thus be interpreted as maintaining the draw and giving the other player the move. When we estimate the model, we will work with positions which satisfy these conditions. In reality, even in this set of positions games can also end through a draw by mechanisms such as insufficient material or stalemate. We omit these to simplify the analysis, though we do not expect that adding transitions to such states would significantly change the subsequent results.

3.4.2 Equilibrium cutoffs and observed play

The following result is a straightforward extension of our previous solution to the binary-state model (Proposition 1).

Proposition 5. *There exists a Markov-perfect equilibrium of the game. It is characterized by a cutoff ρ_x at those states $x = (x_p, x_f)$ such that $x_f = \{S, C\}$, (i.e., where the moving player has a choice). The moving player chooses $m = C$ if $I^t_{x_p} > \rho_x$ and $m = S$ if $I^t_{x_p} < \rho_x$, again with ties broken arbitrarily.*

[19] In winning positions, there are blunders of two types—ones that result in a tabelebase draw, as well as ones that result in a loss, as well as moves that immediately end the game with a win.

We now derive the cutoffs that are optimal in a given equilibrium as functions of observables, in preparation for structurally estimating the model.

Play of an equilibrium yields, for each $x \in \mathcal{X}$, a transition distribution $\boldsymbol{t}_x \in \Delta(\mathcal{X})$ to non-terminal states. We write $t_{x,x'}$ for the probability of transitioning from game state x to x', and $\boldsymbol{T}$ for the matrix whose row indexed by x is $\boldsymbol{t}_x$. We also write b_x for the probability of a blunder in state x. Note that $b_x = 1 - \sum_{x'} T_{x,x'}$.

In a given equilibrium, every game state $x \in \mathcal{X}$ is associated with an equilibrium value v_x. Letting $\mathcal{X}_{NT} = \mathcal{X} \setminus \mathcal{E}$ be the set of non-terminal states, the vector $(v_x)_{x \in \mathcal{X}_{NT}}$ satisfies a recursive equation:

$$\boldsymbol{v} = \underbrace{\boldsymbol{b} \cdot 0}_{\text{blunder}} + \underbrace{\boldsymbol{T}(\mathbf{1} - \boldsymbol{v})}_{\text{successful move}} . \tag{3.3}$$

The moving player receives 0 if he blunders, and otherwise he receives 1 minus the value of the state to his opponent. Recursively, this system determines the value of any state from the known values of the terminal (blunder) states: solving the system, we have[20]

$$\boldsymbol{v} = (\boldsymbol{I} + \boldsymbol{T})^{-1}\boldsymbol{T}\mathbf{1}. \tag{3.4}$$

Let $\mathcal{X}_C \subseteq \mathcal{X}$ be "choice nodes," i.e., the set of position states x such that x_f is nonsingleton—i.e., so that there are non-blunder moves creating both simple and complex positions for the opponent. Recall that the move $m \in M_x$ induces a distribution $\mu_{x,m}$ over states for the other player conditional on not blundering. Given $\mu_{x,m}$, we can compute the conditional expected value of

[20] By the assumption stated before Proposition 5, $\boldsymbol{T}$ has all rows summing to strictly less than 1, so the inverse exists by the Gershgorin circle theorem.

making each move at state $x \in \mathcal{X}_C$ to be

$$u_{x,m}^{\neg b} = \sum_{x' \in \mathcal{X}} \mu_{x,m}(x')(1 - v_{x'}). \tag{3.5}$$

The cutoff for making a complex move can then be calculated to be

$$\rho_x = \frac{u_{x,s}^{\neg b}}{u_{x,c}^{\neg b}}. \tag{3.6}$$

This formula, which is the main point of this subsection, relates the optimal cutoff to the payoff quantities on the right-hand side. In equilibrium, a complex move is attractive insofar as it makes the opponent worse off when it is successfully executed, putting the opponent in a state with a lower value v_x. The degree of this appeal is quantified by the formula in (3.6).

3.4.3 Estimating the model

We assume that observed play follows a Markov-perfect equilibrium of the above game. Fixing a given population of games (for example, games between low-rated players), our structural model allows us to estimate cutoffs ρ_x in different position states x.

In more detail, we first fix a rating tercile and look at all positions that appear in at least 100 games between players within this rating tercile. In accordance with the model, we also restrict attention to starting positions that are tablebase draws, and from which legal moves result in either a tablebase loss for the moving player, or a position at which the other player has the move. We call the resulting set of positions $\hat{\mathcal{P}}$. A blunder is a move that changes the tablebase value of the position to a loss for the moving player.[21] We define $\hat{\mathcal{P}}'$ to consist of all positions reachable by

[21] Note that, given the set of positions we are working with, any move is either this type of blunder, or a move that maintains the tablebase value of a draw.

making a non-blunder legal move from some position in $\hat{\mathcal{P}}$. Using our complexity measure from Section 3.2.3, we classify all positions in $\hat{\mathcal{P}} \cup \hat{\mathcal{P}}'$ as simple or complex. We then assign each position $p \in \hat{\mathcal{P}}$ a label (x_p, x_f), with x_p reflecting whether the position itself is simple or complex, and x_f reflecting whether the non-blunder moves result in only simple positions for the opponent, only complex ones, or at least one of both. This yields a label for every position in $\hat{\mathcal{P}}$ and a label for every possible move as simple or complex. We can then compute empirical frequencies of transitions $\hat{P}(x_t, m_t, x_{t+1})$ from position state x_t to position state x_{t+1} following move m_t. Each observed transition is weighted by the number of times it occurs in the data.

A consistent estimator $\hat{\boldsymbol{T}}$ for $\boldsymbol{T}$ is readily constructed by using empirical frequencies of transition:

$$\hat{T}_{x,x'} = \sum_{m \in M_x} \hat{P}(x, m, x').$$

Plugging this into (3.4), we obtain an estimate $\hat{\boldsymbol{v}}$ of the vector of position values. We can also directly use the empirical transition matrix $\hat{P}$ to estimate $\mu_{x,m}$, the conditional distribution over $\mathcal{X}$ after move m in position state x.[22]

Finally, our estimates of $\hat{\mu}$ and $\hat{\boldsymbol{v}}$ allow us to estimate $u^{\neg b}_{x,m}$ using (3.5). We thus obtain an estimator of the insight cutoff ρ_x for choosing to play C rather than S in state x, given by

$$\hat{\rho}_x = \frac{\hat{u}^{\neg b}_{x,c}}{\hat{u}^{\neg b}_{x,s}}.$$

3.4.4 Structural estimates

We define the *acceptable complexity risk* at any position state x to be $\delta_x \equiv 1 - \rho_x$. This is the discount in execution probability for complex moves that makes a player indifferent between

[22] Letting $n_{x,(x'_p,x'_f)}$ be the number of observations in the data with a given value of (x'_p, x'_f) following x, we define $\hat{\mu}_{x,m}(x'_p, x'_f) = n_{x,(m,x'_f)} / \sum_{x''_f} n_{x,(m,x''_f)}$ if $x'_p = m$ and $\hat{\mu}_{x,m}(x'_p, x'_f) = 0$ otherwise.

choosing a simple and a complex move.[23] It is a measure of (equilibrium) willingness to make complex moves, with higher values meaning a higher inclination to make such moves given the same information.

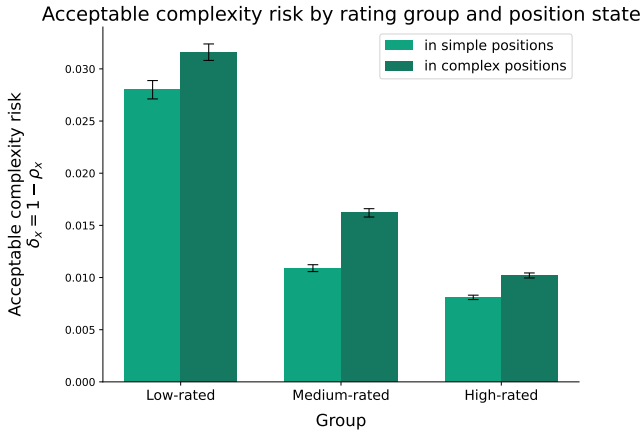


Figure 3.6: *The estimated acceptable complexity risk $\hat{\delta}_x$ for each rating group in simple and complex positions.*

We estimate the model separately on each of the three rating tercile groups described in Section 3.2.3.[24] Figure 3.6 reports our estimates of acceptable complexity risk for each rating group and type of starting position. (Appendix B.4 documents the details behind the estimation.)

The estimates show that in both simple and complex positions, the acceptable complexity risk decreases monotonically in player ratings. That is, higher-rated players need to be more sure that they will not blunder following a complex move in order to attempt it. These differences are statistically significant.[25]

This finding is perhaps surprising in light of our previous result that play of complex positions *increases* in frequency for higher-rated players. These two findings are reconciled by the different

[23] In principle this number could be negative, but it is always positive in the data.

[24] We have 24,343,705 games and get 7,235 positions for low-rating group; for medium-rating group, we have 75,463,973 games and get 40,326 positions; for high-rating group, we have 82,817,091 games and get 77,027 positions.

[25] Bootstraped standard errors are computed by redrawing games with replacement from all games included in the relevant player group, and then constructing the data set of positions from the sampled set of games.

persistence of states for players of different rating levels. In particular, complex moves often lead to complex positions in which the only non-blunder moves are ones that maintain the complexity of the position. Low-rated players are more likely to blunder in such positions, resulting in an early termination of the complex sequence. In contrast, high-rated players blunder less, which yields longer sequences of complex positions. This persistence of complex positions compensates for the lower rate with which higher-rated players transition to such positions, and results in a higher prevalence of complex positions in their play.[26]

We conclude with a discussion of why the acceptable complexity risk decreases as player ratings increase. To understand this, recall the fundamental tradeoffs in our model. First, there is the immediate consideration that complex moves are harder to implement and increase the probability of a blunder. This force dissuades complex play. Second, there is a strategic consideration: If the value v_x to starting a turn in position x is lower at complex positions (as we find it is), then there is a strategic benefit to putting the opponent in a complex position next turn. This force encourages complex play. Finally, there is a dynamic consideration: Putting the opponent in a complex position may boomerang back, amplifying the impact of the first force (i.e., a higher probability of blundering in complex positions). This again dissuades complex play.

Naively considering the first force—the relative probability of blunder in the two states—would suggest that low-rated players are less likely to engage in complex play, since we find that low-rated players blunder much more frequently in complex positions compared to simple positions. But it turns out that this direct force is overturned by two things. First, we find that v_x is substantially more sensitive to the position state for low-rated players than high-rated players (see Section B.4.3); that is, low-rated players are more disadvantaged by being placed in a complex position compared to high-rated players. This means that strategically, it is more worthwhile for

[26] Section B.4 details the transition rates in all rating groups.

low-rated players to put their low-rated opponents in such positions.

The greater variability of v_x for lower-rated players also means they are more disadvantaged if the position boomerangs back to them in a complex state. But the second crucial observation is that this dynamic consideration is only relevant if the opponent does not in fact blunder. Since the low-rated opponent is more likely to blunder in a complex position,[27] the relative weight of the third force is smaller. Together, these differences result in low-rated players being more willing to implement complex moves in our data.

3.5 Concluding discussion

There are many natural next steps on the questions this paper examines.

Other complexity measures. Our modeling approach is flexible with respect to the definition of the complexity of a situation facing a player. We have presented one operational definition to implement the empirical analysis, but one could use others. For example, the depth of strategic reasoning required in various positions could be a natural factor for such analyses to take into account.

The strategic implications of time. Another factor that we have not focused on is time on the clock, which affects how long players have to think about their moves. Time is relevant for how complexity affects players, and the strategic management of the clock is a realistic feature of chess. In reporting our empirical results, we group games with different time controls, but one could disaggregate these. Our structural model could be extended by allowing the insight distributions to depend on an information acquisition decision, i.e., how long to think. A fully strategic model

[27] Indeed, for high-rated players in such positions the blunder rate is 0.04, half of the value of 0.08 observed among the high players (Section B.4.1).

of time-management is a formidable task, but one that might make for rich extensions of the basic model.

Richer position states and asymmetric abilities. Our structural exercise used a particular set of positions to estimate players' cutoffs for playing complex moves. It would be straightforward and interesting to study a much larger class of position states. For example, we could include position states where either player is ahead—i.e., has a tablebase win.[28] This would directly test the Pandolfini advice quoted in our introduction, to complicate losing positions and simplify winning positions.

Relatedly, we have focused on games between closely-matched players, partly because these are most common. However, it would be feasible to cut the data more finely and estimate the strategic implications of ability asymmetries.

[28]This would entail more transitions to keep track of, and the structural estimation would have more parameters. For example, from a position that is a tablebase win for the moving player, a blunder could transition the state to either a tablebase draw or a tablebase loss for that player.

CHAPTER 4
THE TRANSFER PERFORMANCE OF ECONOMIC MODELS

4.1 Introduction

Economists routinely make predictions in environments where data are unavailable, relying on evidence from related but distinct contexts.[1] For example, an economist at a development agency may need to predict diffusion of microfinance takeup in one Indian village given data on diffusion in others. Or an economist at an insurance company may need to predict willingness-to-pay for certain insurance plans given data on willingness-to-pay for others.

Traditionally, economists address this challenge by estimating economic models on existing data and using the estimated parameters to make predictions in new domains. But the predictive success of machine learning algorithms in several economics applications (e.g. Hartford et al., 2016a; Hofman et al., 2021; Banerjee et al., 2024) raises the question of whether the economic model is essential in this process, or whether predictions would be improved by training and porting a flexible machine learning algorithm instead. This is an empirical question: On the one hand, machine learning methods are capable of uncovering novel patterns that existing models miss (Fudenberg and Liang, 2019; Peterson et al., 2021; Ludwig and Mullainathan, 2023); on the other, many researchers believe that structured economic models capture fundamental regularities that generalize more reliably across domains (Coveney et al., 2016; Athey, 2017; Beery et al., 2018; Manski, 2021). Understanding whether economic models do in fact transfer better across domains is important to understanding their future role within economic analysis and policymaking.

[1] This chapter is based on joint work with Isaiah Andrews, Drew Fudenberg, Lihua Lei, and Annie Liang.

Our paper provides a conceptual framework that formalizes the "out-of-domain" transfer problem, and develops econometric tools for systematic comparison between economic modeling and machine learning algorithms. The work on external validity described in Section 4.1.1 considers either an ex post perspective (directly applying the model on a new dataset) or an interim perspective (conditioning on partial data from the target domain). In contrast, we take an ex-ante perspective where the economist is aware of a class of relevant domains, such as Indian villages, but does not know which specific transfer problem will realize. For example, instead of assessing whether a structural model of information diffusion will transfer well from a specific Indian village to another, we ask whether the structural model transfers well "in general" across Indian villages. Our focus is not on the average or expected performance of a model across domains, but on the *realized transfer error* that arises when a researcher, using a fixed procedure and limited training domains, applies the model to a new domain drawn from a population of domains. This realized transfer error is itself a random variable, reflecting randomness in both the training domains and the target domain.

Our main theoretical contribution is the construction of finite-sample forecast intervals that characterize how well a model or algorithm performs in new domains.[2] We provide intervals that are valid for a benchmark setting where all domains are equally likely to realize for training and testing (corresponding to an assumption that domains are exchangeable), as well as when the testing domain is qualitatively different from the training domains. (Appendix C.4.1 generalizes our approach and results even further.) The definition of transferability is left to the user, and our results apply for a large class of measures, including predictive accuracy or how well a qualitative or quantitative finding based on parameter estimates generalizes across domains. In general, our method can be applied to assess and compare the transferability through parameter estimates or

[2] We use the term "forecast interval," rather than "confidence interval," to emphasize that the target is a random realized transfer error rather than a fixed population quantity.

counterfactual predictions across economic models that share a common structural parameter or generate the same class of counterfactuals.

Finally, we use our framework to compare the generalizability of economic models and black box machine learning methods when predicting certainty equivalents for lotteries. We focus on predictive accuracy in this application because machine learning methods do not provide structural parameter estimates that are comparable to those in the economic models. While economic models of risk preference have been studied extensively from the perspectives of how well they fit economic data (Harless and Camerer, 1994; Hey and Orme, 1994; Bruhin et al., 2010; Bernheim and Sprenger, 2020) and how their predictive performance compares to that of machine learning algorithms (Peysakhovich and Naecker, 2017; Plonsky et al., 2019; Fudenberg et al., 2022a), the existing tests have primarily been within the context of a single domain.[3] We offer a new perspective on how well these models transfer across domains, finding that although machine learning algorithms outperform the economic models when trained and tested on (disjoint) data generated under the same conditions, the economic models generalize better. Our analysis suggests the primary reason for this is not that the machine learning algorithms overfit, but rather that economic models of risk preference are more effective at extrapolating from observed cases to new ones.

We now describe the main parts of the paper in more detail. Section 4.2 describes our conceptual framework, which extends the familiar notion of out-of-sample evaluation to out-of-domain evaluation. In the standard out-of-sample test, a model's free parameters are estimated on a training sample, and the predictions of the estimated model are evaluated on a test sample, where the observations in the training and test samples are disjoint but drawn from the same distribution. We depart from this framework by supposing that the distribution of the data varies across a set of "domains," e.g. different subject pools or choice frames.

[3] An important exception is Einav et al. (2012), which examines how general risk preferences are across different domains of choice under uncertainty.

We adopt the perspective of an external *analyst* who wants to assess the efficacy of a procedure that a *researcher* uses to make predictions in a new domain. For example, the researcher's procedure might be to estimate an economic model on data from one domain and use the estimated model to make predictions in another. Alternatively, the researcher's procedure might involve training and porting a black box algorithm, or pooling data across many domains and estimating a model on this aggregated dataset. The procedure's exact performance depends on the domains that are used for estimation—the *training domains*—and the domain that is realized for prediction—the *target domain*. The analyst's goal is to develop a forecast interval for the random performance of this procedure across the many possible realizations of those domains. Unlike ex-post assessment on a single new dataset, this forecast interval summarizes the range of plausible transfer performance and hence provides a more comprehensive description of generalizations.

Section 4.3 constructs forecast intervals for the baseline setting where the distributions governing different domains are exchangeable. This means that while there may be ex-post differences between the domains on which the model is estimated and on which it is applied, these differences are not ex-ante known to the analyst. We propose the following meta-analysis protocol: The analyst first collects data across as many domains as possible, henceforth the analyst's *meta-data set*. The analyst then repeatedly splits these into training and test domains, and runs the researcher's procedure for each of these domain realizations. For example, if the researcher's procedure involves transferring an economic model, then the analyst would estimate the researcher's model on the training domains and use it to make predictions in the test domain. Intuitively, the transfer performance between domains in the analyst's meta-data set can serve as a proxy for the transfer performance to the new (unobserved) target domain. Pooling these transfer performances across different choices of training and test domains yields an empirical distribution of transfer errors. We show that the quantiles of this distribution can be used to compute valid finite-sample forecast

intervals for the transfer error on the target domain.

Section 4.4 extends our methods for the scenario where the training domains are known to be systematically different from the target domain. Specifically, we suppose that the training domains are exchangeable, but allow the target domain to be governed by a qualitatively different distribution. We fully generalize our procedure and results when the analyst either knows or can bound a likelihood ratio relating the target and training distributions. We further develop two novel partial orders for comparing how well models generalize in environments where the analyst lacks any knowledge at all about the likelihood ratio. Though this ordering is inherently incomplete, we demonstrate that it can be used to compare economic models and black box algorithms in our subsequent application. Each of these procedures (and all other methods described in this paper) are implemented in an `R` package (`transferUQ`), available on Github.[4]

Our statistical framework and results extend the recent literature on conformal inference (e.g. Vovk et al., 2005; Tibshirani et al., 2019; Lei and Candès, 2021) and randomization inference (e.g. Ritzwoller et al., 2024) by replacing the assumption of exchangeable observations with the more general assumption of exchangeable domains.[5] Importantly, our framework does not rely on parametric restrictions, smoothness across domains, common support conditions, or assumptions that domains are close in any metric. The identifying content comes entirely from symmetry at the domain level, or from explicit bounds on deviations from this symmetry. Conceptually, we show that out-of-domain performance can be evaluated using finite-sample methods that require neither structural assumptions nor asymptotics if one treats domains, rather than individual observations, as the exchangeable objects.

Importantly, while conformal inference aims to predict a single data point from other data

[4]https://github.com/lihualei71/transferUQ

[5]Thus conformal inference corresponds to the special case of our model where there is a point mass on a single domain.

points, our goal is to generate forecast intervals on a function of multiple data points, involving both observed and unobserved data. Our problem thus involves a novel multi-dimensional structure that requires new arguments both in the baseline exchangeable case and the extension to distribution shift.

Finally, Section 4.5 applies our procedures to compare the transferability of economic models and black box algorithms for predicting certainty equivalents for binary lotteries. In this application, we evaluate how well a model (or algorithm) predicts certainty equivalents reported by one subject pool when estimated on data from another. We evaluate two models of risk preferences (expected utility and cumulative prospect theory) and two popular black box machine learning algorithms (random forest and kernel regression). We find that although the black box algorithms outperform the economic models out-of-sample when trained and tested on data from the same domain, the economic models generalize more reliably across domains. Specifically, while the forecast intervals for the black box algorithms and economic models overlap, the forecast intervals for the black box methods are wider, and their upper bounds are substantially higher.

One possible explanation for this finding, based on intuition from conventional out-of-sample testing, is that black boxes are very flexible and hence learn idiosyncratic details that do not generalize across subject pools. But when we restrict the analysis to a subset of samples involving the same set of lotteries, the economic models and black box algorithms have nearly indistinguishable forecast intervals. Thus black box algorithms are not universally worse at transfer; instead, their relative performance depends on specific characteristics of the transfer problem. In particular, we find that black boxes seem to transfer worse when the primary source of variation across samples is a shift in the marginal distribution over features (i.e., which lotteries appear in the sample), rather than a shift in the distribution of outcomes conditional on features (i.e., the distribution of certainty equivalents given fixed lotteries).

4.1.1 Related Literature.

4.1.1 Evaluating economic models

Hofman et al. (2021) calls for more work in the social sciences addressing the question "how well does a predictive model fit to one data distribution generalize to another?" Our results speak directly to this challenge by providing a framework for estimating transfer errors across domains.

Predictive accuracy is far from the only criterion that matters when evaluating models, but has historically played an important role in the development of economic theories.[6] Recent papers consider not just the model's fit on a given dataset, but also how well predictions transfer across domains. For example, Külpmann and Kuzmics (2022) estimates various game-theoretic models on 2×2 normal-form games and evaluates their predictive performance on 3×3 normal-form games; Natenzon (2019) estimates discrete choice models on data for four choice menus and evaluates their predictive performance on a fifth menu; and Fudenberg and Karreskog Rehbinder (2024) evaluates how well learning models for repeated games transfer across classes of stage game payoffs. This paper provides a general framework that nests these transfer problems, and develops formal statistical results for assessing transfer performance.

4.1.1 Comparing economic models and black box algorithms

Several papers have compared the predictive performance of machine learning algorithms with that of more structured economic models in out-of-sample tests (Peysakhovich and Naecker, 2017; Plonsky et al., 2019; Camerer et al., 2019; Fudenberg and Liang, 2019; Hirasawa et al., 2022; Hsieh et al., 2023; Ellis et al., 2024a,b), typically finding that machine learning algorithms achieve higher

[6] As discussed by Harless and Camerer (1994), the poor predictive performance of expected utility theory was a primary motivation for the development of alternative models in behavioral economics. Both Harless and Camerer (1994) and Hey and Orme (1994) provide early assessments of alternative theories based on predictive performance.

predictive accuracy. In contrast, our empirical results suggest that machine learning algorithms can be less effective than economic models when trained and tested on data from different domains. These findings echo Gechter et al. (2019), which shows that structural models deliver better policy recommendations for conditional cash transfer policies in new contexts than black box methods do.

Collectively these papers lend empirical support to the folk intuition that one of the key values of economic theory is its generalizability across economic domains, and they show the need for a way to compare economic models and machine learning algorithms on out-of-domain tests. Our statistical framework and theoretical results in Sections 4.2-4.4 provide the tools to do this.

4.1.1 Generalizability

Several literatures in economics, computer science, and statistics consider generalizability from different perspectives. Our paper is most closely related to the literature on external validity (Banerjee et al., 2017; Chassang and Kapon, 2022), which seeks to measure the extent to which results obtained in one domain hold more generally. Unlike much of this literature, we consider generalizability across random domains without assuming that the distributions governing behavior in different domains are close in some distance metric (Adjaho and Christensen, 2022), share a common support over $\mathcal{X}$ or $\mathcal{Y}$ (Sahoo et al., 2022; Lei et al., 2023), or can be estimated using background covariates (Tipton and Olsen, 2018). Relative to these assumptions, our approach has the advantage of allowing for arbitrary and unknown relationships between the realized distributions governing domains, but rules out ex-ante predictable patterns in how the joint distribution varies across samples (such as time trends).

The literature on meta-analysis focuses on the related but different goal of synthesizing results across different domains (Card and Krueger, 1995; DellaVigna and Pope, 2019; Embrey et al.,

2018; Imai et al., 2020; Vivalt, 2020). The most closely related work in this area is Meager (2019) and Meager (2022), which provide posterior predictive intervals for new domains under parametric assumptions about the distribution of effects across domains.

A third distinct goal is to build models that generalize well to new domains. This is the objective of the literature on domain generalization (Zhou et al., 2021), as well as papers such as Hotz et al. 2005 and Dehejia et al. 2021, which leverage knowledge about the distribution of covariates to extrapolate out-of-domain. Our focus is not on developing new models or algorithms with good out-of-domain guarantees, but rather on developing forecast intervals for the out-of-domain performance of existing models and algorithms.

4.1.1 Conformal inference and exchangeability

Our statistical framework builds on the literature on conformal inference by extending it to cross-domain predictions. Classic conformal methods assume that observations are exchangeable, and construct prediction intervals for a scalar outcome Y_i. We instead seek a prediction interval for the bivariate function $v(S_i, S_j)$, namely the transfer error from sample S_i to sample S_j. Standard conformal inference arguments do not apply in our setting because $\{v(S_i, S_j)\}_{i,j}$ is not exchangeable. When there are more than two domains, even when the samples $\{S_i\}$ are i.i.d., $v(S_1, S_2)$ is generally dependent on $\{v(S_1, S_j) : j \geq 3\}$.

Section 4.4 and Appendix C.4.1 relax our exchangeability assumption to allow bounded deviations from exchangeability. Our results there connect to the literature on sensitivity analysis (e.g. Rosenbaum, 2005; Aronow and Lee, 2013; Andrews and Oster, 2019; Sahoo et al., 2022), though our theoretical analysis is more involved because of result of the multivariate structure.

4.2 Framework

In economics, it is common to transfer quantitative conclusions from a single domain to another, e.g., for parameter calibration in structural models (Greenwood et al., 1997; McKay et al., 2016; Oswald, 2019) and extrapolation of treatment effect estimates beyond the experimental population (Mogstad and Torgovitsky, 2018; Tipton and Olsen, 2018; Cattaneo et al., 2021; Maeba, 2022). In this case, the relevant question is whether extrapolating from one sample leads to good predictions in the new domain. If instead data is gathered from multiple domains and the observations are aggregated and used to estimate a model (as in the meta-analyses of Meager 2019, 2022), the relevant question may be how well the estimated model on the aggregated data generalizes to a new domain.

4.2.1 The Analyst's Problem

We adopt the perspective of an external *analyst* who wants to assess the efficacy of a *researcher's* procedure for making predictions in new domains. The analyst has access to a collection of n domains, while the researcher uses only $r < n$ of these domains for training. The distinction between analyst and researcher is central to our approach: it allows the analyst to construct valid inference about the researcher's transfer performance without requiring assumptions beyond exchangeability. In other words, the analyst can evaluate whether the researcher's procedure generalizes well across domains *in general*, rather than assessing performance on one specific realization of the data.

For example, an economist at a development agency may have data from many Indian villages (the analyst's metadata), but wants to assess whether a structural diffusion model estimated on data from a single village (the researcher's training domain) will transfer well when applied to a

new village selected from the same region. Or an economist at an insurance company may have data on willingness-to-pay for various insurance products from multiple subject pools, but wants to evaluate whether a model trained on one subject pool will predict well for another.

Write $\mathcal{T}$ for the ordered set of r training domains, or equivalently a vector of length r with distinct elements from $\{1, \ldots, n\}$ [7] and $S_{\mathcal{T}}$ for the data from these domains. The researcher fits a model and aims to apply it to a different domain S', reporting an output such as a parameter estimate or a prediction error. The analyst's goal is to construct a forecast interval for the *realized transfer error* when the researcher's training domains and target domain are drawn at random from the population of domains.

4.2.2 Transfer Error: Definition and Scope

Let $\mathcal{X}$ be a set of covariate vectors and $\mathcal{Y}$ a set of outcomes. An *observation* is a pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and a *domain* is a set of observations $S = \{(x_i, y_i)\}_{i=1}^{m}$. We now describe the class of transfer errors we consider.

Example 1 (Different Subject Pools). Each sample S_d corresponds to data from a given subject pool, where subject pools may differ in demographic characteristics. For example, S_1 may contain data from Caltech undergraduates, while S_2 contains data from a representative Prolific subject pool.

Example 2 (Different Choice Frames). Each sample S_d corresponds to data collected under a particular framing of choice questions. For example, S_1 might contain reported certainty equivalents for compound lotteries, and S_2 the reported certainty equivalents for equivalent simple lotteries.

Example 3 (Different Choice Menus). There is a finite set of goods $a_1, a_2, \ldots, a_m$, and each sample S_d includes observed choices from a different subset of available goods. For example, S_1 might

[7] We do not assume the training procedure is invariant to the ordering of the training domains.

contain all choices from binary menus and S_2 all choices from menus that include a_1.

4.2.2 Prediction-based Transfer Errors

Our leading examples measure how well a model's predictions transfer across domains. Suppose the training data $S_{\mathcal{T}}$ is used to select a *prediction rule* $f_{S_{\mathcal{T}}} : \mathcal{X} \to \mathcal{Y}$. The accuracy of this rule is evaluated using a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, where

$$e(f, S) = \frac{1}{\#S} \sum_{(x,y)\in S} \ell(f(x), y)$$

denotes the average loss when predicting outcomes in sample S.

Raw Transfer Error: The simplest transfer error is

$$e_{\mathcal{T},d^*} = e(f_{S_{\mathcal{T}}}, S_{d^*}) \tag{4.1}$$

i.e., the average loss when the model estimated on training domains is used to predict outcomes in the target domain.

Example 4 (Transferring Models of Risk Preferences). Covariates $\mathcal{X}$ describe lotteries (prizes and probabilities), and outcomes $\mathcal{Y}$ are reported certainty equivalents. A researcher estimates a risk preference model (e.g., Expected Utility) on data from one subject pool and uses the estimated parameters to predict certainty equivalents for another subject pool. The transfer error measures prediction accuracy across subject pools.

Normalized Transfer Errors: It is often useful to normalize the raw transfer error. For in-

stance, we might normalize by the in-sample error on the target domain,

$$e^{\text{norm}}_{\mathcal{T},d^*} = \frac{e(f_{S_{\mathcal{T}}}, S_{d^*})}{e(f_{S_{d^*}}, S_{d^*})}, \tag{4.2}$$

which measures how much less accurate the transferred model is compared to a model retrained on the target domain. This is useful for distinguishing whether transfer problems reflect overfitting or lack of generalization.

Example 5 (The Value of Re-Estimating Structural Models). An economist estimates a diffusion model on data from one village and predicts takeup in another village. The normalized transfer error indicates how much worse the prediction is than if the economist could re-estimate the model on the new village. A value near 1 suggests good transfer; a value much greater than 1 suggests the model misses important village-specific factors.

More generally, transfer errors can measure parameter stability across domains (e.g., $e(\mathcal{T}, d^*) = \|\hat{\theta}_{\mathcal{T}} - \hat{\theta}_{d^*}\|$) or whether qualitative findings replicate across domains. Appendix C.2 discusses these extensions in detail.

Remark 1. When $r = 1$ and the model has only a single parameter θ of interest, a common practice is to estimate $\hat{\theta}_j$ on the domain S_j and report the range of $\{\hat{\theta}_1, \ldots, \hat{\theta}_n\}$ to quantify transferability. One can interpret the range as an interval estimate for $\hat{\theta}_{n+1}$. However, this approach does not generalize to multivariate parameter estimates or more complex transfer errors.

4.2.3 Ex-Ante Assessment of Transfer Performance

To evaluate the researcher's procedure, the analyst takes an *ex-ante* perspective: before observing which specific training and target domains materialize, what is the distribution of transfer errors?

Formally, the analyst has access to metadata $\mathbf{M} = \{S_1, \ldots, S_n\}$ consisting of n domains. We

treat the researcher's set of r training domains as a random subset $\mathbf{T}$ drawn uniformly from all possible r-subsets of $\{1, \ldots, n\}$. The unobserved target domain is S_{n+1}. The quantity of interest is the random transfer error $e_{\mathbf{T},n+1}$, which depends on both the randomness in which training domains are selected and the randomness in the data generating process.

The analyst's goal is to construct a *forecast interval* $[L(\mathbf{M}), U(\mathbf{M})]$ that covers $e_{\mathbf{T},n+1}$ with high probability:

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \in [L(\mathbf{M}), U(\mathbf{M})]\right) \geq 1 - \alpha, \tag{4.3}$$

where α is a user-selected miscoverage level (e.g., $\alpha = 0.1$ for 90% coverage). This differs from a confidence interval, which covers a fixed population quantity; here, we are forecasting a random realized value.

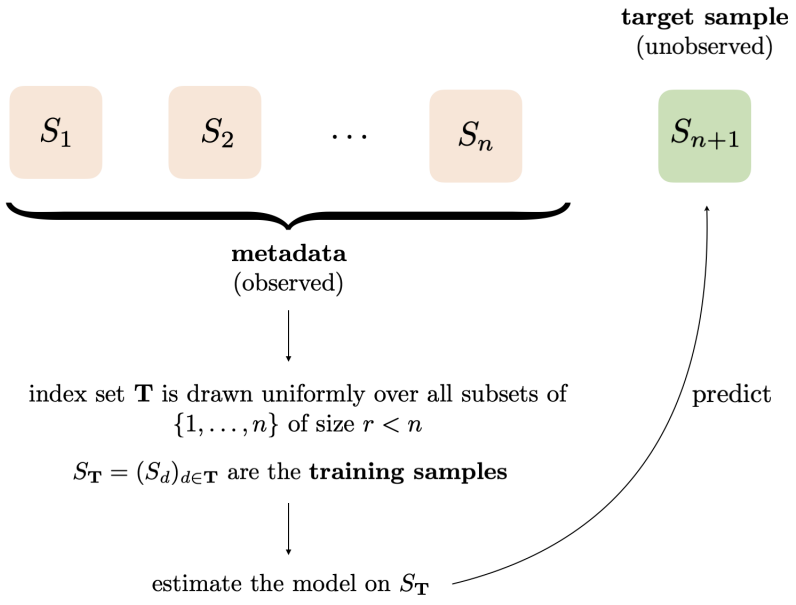


Figure 4.1: *The transfer error $e_{\mathbf{T},n+1}$ is the prediction error of a model estimated on training domains $S_{\mathbf{T}}$ (randomly selected from metadata $\mathbf{M}$) and evaluated on an unobserved target domain S_{n+1}. Both the selection of training domains and the target domain are random.*

4.2.4 Why Ex-Ante Assessment?

One might ask: why evaluate the researcher's procedure across all possible domain selections, rather than conditioning on a specific training domain? The most common way to assess model transferability is to evaluate an *ex-post* error after the data S' from the target domain is fully observed. While such an ex post error can be computed exactly, it does not generalize to other target domains that could have been realized, due to either the researcher's idiosyncratic choice target domain or randomness in the data collection process. In particular, near-zero ex-post errors may arise by chance and do not imply perfect transfer to other slightly different domains.

Our approach is instead to conduct an *ex-ante* assessment of model transferability before the target domain S' is realized. Ex-ante assessment supports replicability, as the analysis applies to any researcher who follows the same procedure (e.g., "estimate a CRRA model on one randomly selected domain and apply it to a new domain"). It also avoids cherry-picking from assessing transfer performance on a researcher-chosen domain pair, which may be unrepresentative.[8]

Ex-ante evaluation requires assumptions which relate the training domains and the unseen target domain, because otherwise the observed data can give no guidance as to the possible error on the target. A common assumption in both the theory and econometrics literature is that $(S_{\mathcal{T}}, S')$ are ex ante exchangeable, which formalizes the idea that the target domain is not "special" relative to the training domains, and holds by construction if the target domain is selected at random (Menzel, 2023). This is our baseline approach (Section 4.3), and we relax it in Section 4.4.

[8] In some applications, the researcher observes covariates from the target domain even though outcomes are unobserved. For example, in predicting treatment effects in a new location, researchers may know which households participate (covariates) but not their treatment responses (outcomes). The same theory applies if we define transfer error based on predictions which account for this covariate information. This is sometimes called an *interim assessment*.

4.3 Baseline Setting: Exchangeability

We begin with a baseline setting in which the different domains are governed by exchangeable distributions. Specifically, we suppose that in the analyst's statistical model, the samples $S_1, S_2, \ldots$ are generated in the following way:

Assumption 1. *There is a fixed (but unknown) meta-distribution $\mu \in \Delta(\mathcal{P} \times \mathbb{N})$ over joint distributions $\mathcal{P} \equiv \Delta(\mathcal{X} \times \mathcal{Y})$ and sample sizes $\mathbb{N}$, where each sample S_d is generated by first drawing a distribution and sample size $(P_d, m_d) \sim \mu$, and then independently drawing m_d observations (x, y) from P_d.*[9]

In Examples 1-3, this assumption implies that (from the analyst's perspective) the subject pools, choice frames, or choice menus differentiating the samples are themselves drawn i.i.d. from a fixed distribution.[10] Section 4.3.1 presents our forecast intervals for this setting, and Section 4.3.2 proves the validity and tightness of these intervals.

4.3.1 Baseline procedure

Thanks to Assumption 1, the observed samples in the metadata, $\{S_1, \ldots, S_n\}$, can act as surrogates for the unseen target sample, S_{n+1}. As before, let $e^{\mathbf{M}}_{\mathcal{T},d}$ denote the (observed) transfer error from any selection of training samples $\mathcal{T}$, as an ordered set from $\{1, \ldots, n\}$, to any surrogate target sample $d \in \{1, \ldots, n\} \backslash \mathcal{T}$ from the metadata (where we now make the dependence of this quantity on $\mathbf{M}$ explicit). We use $\mathbb{T}_{r+1,n}$ to denote the set of $\frac{n!}{(n-r-1)!}$ unique pairs $(\mathcal{T}, d)$ that can be constructed in

[9] All of our results extend unchanged if samples from the different domains are ex-ante exchangeable rather than i.i.d.

[10] Assumption 1 can be understood as a hierarchical model (Meager, 2019, 2022) or a version of cluster sampling (Liang and Zeger, 1986; Bugni et al., 2023). If framed in this way, the analyst's goal is to do predictive inference for new clusters. When μ assigns probability 1 to a single distribution in $p \in \mathcal{P}$ or when μ assigns probability 1 to $m = 1$, this reduces to i.i.d. sampling of observations from a fixed joint distribution, but our focus is on settings where neither of these is the case.

this way when $\mathcal{T}$ is of length r. Then

$$F_{\mathbf{M}} = \frac{(n-r-1)!}{n!} \sum_{(\mathcal{T},d)\in\mathbb{T}_{r+1,n}} \delta_{e^{\mathbf{M}}_{\mathcal{T},d}} \tag{4.4}$$

is the empirical distribution of transfer errors in the pooled sample $\left\{e^{\mathbf{M}}_{\mathcal{T},d} : (\mathcal{T},d) \in \mathbb{T}_{r+1,n}\right\}$ as we vary which samples in the metadata are used for training and testing. (Throughout δ denotes the Dirac measure). In the case where $r = 1$, so that a single sample is used for training, the observed transfer errors can be represented as a matrix as depicted in Figure 4.2, and $F_{\mathbf{M}}$ is their empirical distribution.

train \ test	1	2	…	$n-1$	n
1	–	$e_{1,2}$	…	$e_{1,n-1}$	$e_{1,n}$
2	$e_{2,1}$	–	⋱		⋮
⋮	⋮	⋱	–	⋱	⋮
$n-1$	⋮		⋱	–	$e_{n-1,n}$
n	$e_{n,1}$	…	…	$e_{n,n-1}$	–

Figure 4.2: $e_{d,d'}$ *is the transfer error from sample* S_d *to* $S_{d'}$.

Definition 3 (Upper and Lower Quantiles). For any distribution P let $\overline{Q}_\tau(P) = \inf\{b : P((-\infty, b]) \geq \tau\}$ and $\underline{Q}_\tau(P) = \sup\{b : P([b,\infty)) \geq 1-\tau\}$ denote the upper and lower τth quantiles, respectively.

These quantiles coincide for continuously distributed variables with connected support.

Definition 4 (Quantiles of $F_{\mathbf{M}}$). For any $\tau \in (0,1)$, let $\overline{e}^{\mathbf{M}}_\tau \equiv \overline{Q}_\tau(F_{\mathbf{M}})$ and $\underline{e}^{\mathbf{M}}_\tau \equiv \underline{Q}_{1-\tau}(F_{\mathbf{M}})$ be the τth upper quantile and $(1-\tau)$th lower quantile of the empirical distribution of transfer errors

in the pooled sample.

Our forecast interval for the transfer error on the target sample is $[\underline{e}_\tau^{\mathbf{M}}, \bar{e}_\tau^{\mathbf{M}}]$.

4.3.2 Results

We first prove that $[\underline{e}_\tau^{\mathbf{M}}, \bar{e}_\tau^{\mathbf{M}}]$ is indeed a valid forecast interval.

Proposition 6. *For any* $\tau \in (0, 1)$*,*

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{e}_\tau^{\mathbf{M}}\right) \geq \tau \left(\frac{n-r}{n+1}\right), \tag{4.5}$$

and

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \in \left[\underline{e}_\tau^{\mathbf{M}}, \bar{e}_\tau^{\mathbf{M}}\right]\right) \geq (2\tau - 1) \left(\frac{n-r}{n+1}\right). \tag{4.6}$$

Thus $\left(-\infty, \bar{e}_\tau^{\mathbf{M}}\right]$ is a level-$\left(\frac{\tau(n-r)}{n+1}\right)$ one-sided forecast interval for $e_{\mathbf{T},n+1}$, and $\left[\underline{e}_\tau^{\mathbf{M}}, \bar{e}_\tau^{\mathbf{M}}\right]$ is a level-$\left((2\tau-1)\left(\frac{n-r}{n+1}\right)\right)$ forecast interval for $e_{\mathbf{T},n+1}$. Note that the coverage guarantee for the two-sided interval (4.6) is tighter than what would result from applying that for the one-sided interval (4.5) to both sides, which yields a worse bound $1 - 2\tau\left(\frac{r+1}{n+1}\right)$. While the guarantee for the standard conformal inference (e.g. Vovk et al., 2005) can be viewed as a special case of Proposition 6 with $r = 0$, our arguments are necessarily more complex because of the multi-dimensional structure illustrated in Figure 4.2. Parameter τ influences the width of the forecast interval, where larger choices of τ lead to wider forecast intervals with higher confidence guarantees. Parameter r determines how many samples in the meta-data are used for training versus testing, and is determined by the research procedure under evaluation.[11] The conservativeness of these forecast intervals is

[11] We expect that in general, larger choices of r will lead to lower but wider forecast intervals, since the model is estimated on a larger quantity of data, but there are fewer samples with which to evaluate the performance of the estimated model.

intrinsic to the problem: even with infinitely many observations per domain, realized transfer error can vary across domains, so ex-ante uncertainty cannot in general be eliminated.

The number of samples n and the sizes of these samples $(m_d)_{d=1}^n$ enter into our result in different ways: Increasing the number of observed domains n, holding fixed the distribution over sample sizes within each domain, does not change the distribution of $e_{\mathbf{T},n+1}$ but instead allows this distribution to be estimated more precisely. In contrast, increasing the number of observations per domain changes the distribution of $e_{\mathbf{T},n+1}$ and corresponds to the measurement of a different quantity. For example, in the limit of infinitely many observations per sample, the error $e_{\mathbf{T},n+1}$ measures how well the best predictor from the model class in the training domains transfers across domains, while if the number of observations is small, $e_{\mathbf{T},n+1}$ measures how well an imperfectly estimated model transfers.

The next result shows that the guarantees in Proposition 6 are tight to $O(1/n)$. We use $\mathbb{T}_{s,t}$ to denote the set of all vectors of length s that consist of distinct elements from $\{1,\dots,t\}$.

Proposition 7. *Assume that* $\left(e^{\mathbf{M}}_{\mathcal{T},d} : (\mathcal{T},d) \in \mathbb{T}_{r+1,n+1}\right)$ *almost surely has no ties. Then*

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{e}^{\mathbf{M}}_{\tau}\right) \leq \tau\left(\frac{n-r}{n+1}\right) + \frac{r+1}{n+1} + \frac{(n-r)!}{(n+1)!}.$$

and

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \in \left[\underline{e}^{\mathbf{M}}_{\tau}, \bar{e}^{\mathbf{M}}_{\tau}\right]\right) \leq (2\tau-1)\left(\frac{n-r}{n+1}\right) + \frac{r+1}{n+1} + \frac{(n-r)!}{(n+1)!}.$$

To gain intuition for the intervals in Proposition 6, we fix a realization of the unordered set $\{S_1,\dots,S_n,S_{n+1}\}$. Because all samples are exchangeable by assumption, the realization of $e_{\mathbf{T},n+1}$ (conditional on $\{S_d\}_{d=1}^{n+1}$) is a uniform draw from

$$\left\{e^{\mathbf{M}}_{\mathcal{T},d} : (\mathcal{T},d) \in \mathbb{T}_{r+1,n+1}\right\}. \tag{4.7}$$

If we let e^*_τ denote the upper τ-th quantile of this empirical distribution, then by definition

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq e^*_\tau \mid \{S_d\}_{d=1}^{n+1}\right) \geq \tau. \tag{4.8}$$

In the case $r = 1$ where precisely one sample is used for training, the set of pooled errors (4.7) is the shaded cells in Figure 4.3 (either yellow or blue), and the inequality in (4.8) says that the probability that the value of a randomly drawn cell falls below the τth upper quantile of cells is at least τ.

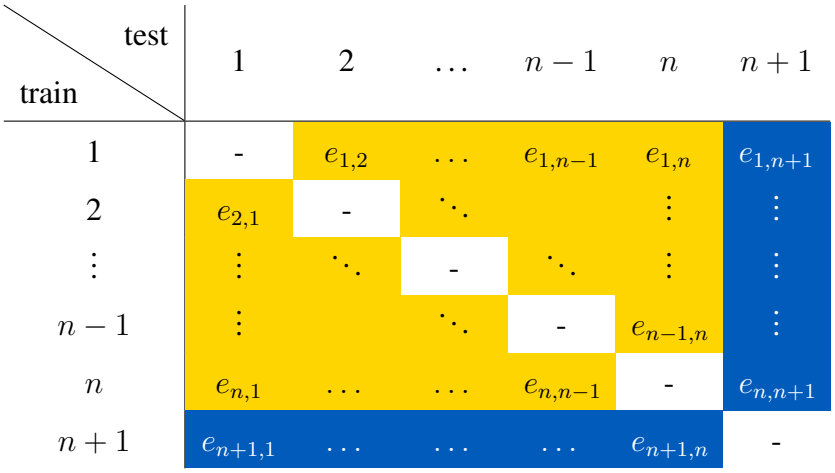


Figure 4.3: *Transfer errors when training on one domain (row) and testing on another (column).*

The analyst does not observe the target sample S_{n+1}, and so does not know e^*_τ. We instead use $\bar{e}^{\mathbf{M}}_\tau$, the τth upper quantile of the pooled sample of errors when transferring across samples in $\mathbf{M}$, to construct the forecast intervals. In Figure 4.3, the probability that $e_{\mathbf{T},n+1} \leq \bar{e}^{\mathbf{M}}_\tau$ is the probability that the value of a randomly drawn shaded cell (yellow or blue) falls below the τth quantile of the yellow cells. By a straightforward counting argument,

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{e}^{\mathbf{M}}_\tau \mid \{S_i\}_{i=1}^{n+1}\right) \geq \tau \binom{n}{r+1} \Big/ \binom{n+1}{r+1} = \tau\left(\frac{n-r}{n+1}\right).$$

Applying the law of iterated expectations (with respect to the sample $\{S_i\}_{i=1}^{n+1}$) yields the one-sided forecast interval in (4.5). The proof for the two-sided forecast interval follows a similar logic but is more involved, see Appendix C.1.3 for details.

4.4 Beyond Exchangeability

Our results so far assume that the distributions governing the different samples S_d are themselves independent and identically distributed. This assumption is not always appropriate. For example, suppose variation in domains corresponds to variation over locations, and the samples in the meta-data (but not the target sample) are from experiments run at locations chosen by experimenters. If there is selection bias over where experiments are run—for example, if the observed sites are chosen based on characteristics correlated with effect sizes (as Allcott (2015) found in the Opower energy conservation experiments)—then it may be that the target sample has fundamentally different properties from anything that is observed in the metadata. We thus now relax Assumption 1 to allow the distribution governing the training samples and the distribution governing the target sample to be drawn from different meta-distributions. Specifically, suppose that the analyst's metadata consists of samples $S_1, \ldots, S_n \sim_{iid} \mu$ as in our main model, but S_{n+1} is independently drawn from some other density ν. Let

$$\omega(S) = \frac{\nu(S)}{\mu(S)}$$

denote their likelihood ratio. We initially assume this likelihood ratio is known by the analyst (although ν and μ need not be), and subsequently consider weakenings of this assumption. As before, $e_{\mathbf{T},n+1}$ is the transfer error when training on r samples drawn uniformly at random from $\{S_1, \ldots, S_n\}$, and testing on S_{n+1}.

The following subsections consider successively weaker assumptions about what the analyst knows about the likelihood ratio ω. Sections 4.4.1 and 4.4.2 respectively generalize our previous results when the analyst either knows ω or can bound it. Section 4.4.3 provides two ways of comparing the generalizability of models in environments where the analyst knows nothing about ω. (See Appendix C.4.1 for further generalizations.)

4.4.1 The analyst knows the likelihood ratio ω

We again construct a forecast interval for $e_{\mathbf{T},n+1}$ using the pooled sample of transfer errors across samples in the metadata, that is, $\left\{e^{\mathbf{M}}_{\mathcal{T},d} : (\mathcal{T}, d) \in \mathbb{T}_{r+1,n}\right\}$. Different from the previous section, we no longer assign uniform weights to each $e^{\mathbf{M}}_{T,d}$. Intuitively, under our previous i.i.d. assumption, each sample in the metadata was equally representative of the training and target distributions, but in this relaxed model, whether a sample S_d is more representative of the training or testing distribution depends on its relative likelihood under ν and μ.

A crucial quantity is the following:

Definition 5. For every domain $d \in \{1, \ldots, n\}$, define

$$W_d = \frac{(n-r-1)!}{(n-1)!} \frac{\omega(S_d)}{\sum_{j=1}^{n} \omega(S_j)}. \tag{4.9}$$

To interpret this quantity, consider an alternative data-generating process for the metadata where for some permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$, the samples $S_{\pi(1)}, \ldots, S_{\pi(n-1)} \sim_{iid} \mu$ while $S_{\pi(n)} \sim \nu$. Fix a realization of the metadata $(S_1, \ldots, S_n)$, and suppose the analyst does not observe the permutation π. Let Π denote the set of all permutations on $\{1, \ldots, n\}$, and for any vector of sample indices $(t_1, \ldots, t_r, d)$ let

$$\Pi_{(t_1,\ldots,t_r,d)} = \{\pi \in \Pi : (\pi(1), \ldots, \pi(r)) = (t_1, \ldots, t_r) \text{ and } \pi(n) = d\}$$

denote the permutations that specify $(t_1, \ldots, t_r)$ for training and d as the target. Then conditional on a realization of the metadata $(S_1, \ldots, S_n)$, the probability that $(S_{t_i})_{i=1}^r$ are the training samples and S_d is the test sample is[12]

$$\frac{\sum_{\pi \in \Pi_{(t_1,\ldots,t_r,d)}} \left(\nu(S_{\pi(n)}) \cdot \prod_{j=1}^{n-1} \mu(S_{\pi(j)})\right)}{\sum_{\pi \in \Pi} \left(\nu(S_{\pi(n)}) \cdot \prod_{j=1}^{n-1} \mu(S_{\pi(j)})\right)} = \frac{\sum_{\pi \in \Pi_{(t_1,\ldots,t_r,d)}} \omega(S_{\pi(n)})}{\sum_{\pi \in \Pi} \omega(S_{\pi(n)})}$$
$$= \frac{(n-r-1)! \cdot \omega(S_d)}{(n-1)! \cdot \sum_{j=1}^n \omega(S_j)} = W_d.$$

This quantity depends only on the identity of the target sample d, and not on the identity of the training samples $t_1, \ldots, t_r$. Finally, let

$$F_{\mathbf{M}}^{\omega} = \sum_{(\mathcal{T},d) \in \mathbb{T}_{r+1,n}} W_d \cdot \delta_{e_{\mathcal{T},d}^{\mathbf{M}}}$$

be the weighted empirical distribution of transfer errors, where each sample d is weighted according to W_d. When the two meta-distributions μ and ν are identical as in our main model, then $W_d \equiv (n-r-1)!/n!$ for every domain d, so the distribution $F_{\mathbf{M}}^{\omega}$ is simply $F_{\mathbf{M}}$ as defined in (4.4).

Definition 6 (Quantiles of $F_{\mathbf{M}}^{\omega}$). For any likelihood ratio $\omega(\cdot)$ and quantile $\tau \in (0,1)$, define $\bar{e}_{\tau}^{\mathbf{M},\omega} = \overline{Q}_{\tau}\left(F_{\mathbf{M}}^{\omega}\right)$ and $\underline{e}_{\tau}^{\mathbf{M},\omega} = \underline{Q}_{1-\tau}\left(F_{\mathbf{M}}^{\omega}\right)$ to be, respectively, the τth upper quantile and $(1-\tau)$th lower quantile of the weighted distribution of transfer errors in the pooled sample.

Theorem 1. *For any* $\tau \in (0,1)$*,*

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{e}_{\tau}^{\mathbf{M},\omega}\right) \geq \tau \cdot \frac{n-r}{n} \mathbb{E}\left[\frac{\sum_{j=1}^n \omega(S_j)}{\sum_{j=1}^{n+1} \omega(S_j)}\right],$$

[12]This is a special case of weighted exchangeability; see Tibshirani et al. (2019). The results in this subsection continues to hold if the domains are not independent but satisfy the weighted exchangeability condition, which is more general but harder to interpret.

and

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \in \left[\underline{e}_\tau^{\mathbf{M},\omega}, \bar{e}_\tau^{\mathbf{M},\omega}\right]\right) \geq (2\tau - 1) \cdot \frac{n-r}{n} \mathbb{E}\left[\frac{\sum_{j=1}^n \omega(S_j)}{\sum_{j=1}^{n+1} \omega(S_j)}\right].$$

Furthermore, if $\left(e_{\mathcal{T},d}^{\mathbf{M}} : (\mathcal{T}, d) \in \mathbb{T}_{r+1,n+1}\right)$ *almost surely has no ties, then*

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{e}_\tau^{\mathbf{M},\omega}\right) \leq 1 - \mathbb{E}\left[\left((1-\tau)\frac{n-r}{n} - \frac{(n-r)!}{n!}\frac{\max_{k\leq n}\omega(S_k)}{\sum_{j=1}^n \omega(S_j)}\right)\frac{\sum_{j=1}^n \omega(S_j)}{\sum_{j=1}^{n+1} \omega(S_j)}\right],$$

and

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \in \left[\underline{e}_\tau^{\mathbf{M},\omega}, \bar{e}_\tau^{\mathbf{M},\omega}\right]\right) \leq 1 - \mathbb{E}\left[\left(2(1-\tau)\frac{n-r}{n} - \frac{(n-r)!}{n!}\frac{\max_{k\leq n}\omega(S_k)}{\sum_{j=1}^n \omega(S_j)}\right)\frac{\sum_{j=1}^n \omega(S_j)}{\sum_{j=1}^{n+1} \omega(S_j)}\right].$$

This result strictly generalizes Proposition 6 and Proposition 7, since when $w(\cdot)$ is the identity then $\bar{e}_\tau^{\mathbf{M},\omega} = \bar{e}_\tau^{\mathbf{M}}$ and $\underline{e}_\tau^{\mathbf{M},\omega} = \underline{e}_\tau^{\mathbf{M}}$, and the bounds in this theorem reduce to those given in Proposition 6.

4.4.2 The analyst does not know ω but can bound it

We next extend our results when the analyst does not know the likelihood ratio function ω precisely, but, as in the literature on sensitivity analysis (Rosenbaum, 2005), knows that it admits specific upper and lower bounds.

Definition 7 (Bounded Likelihood-Ratios). For any $\Gamma \geq 1$, let $\mathcal{W}(\Gamma)$ be the class of density ratios that satisfy $\omega(S) \in [\Gamma^{-1}, \Gamma]$ for all samples S.

Define the following worst case bounds for $\underline{e}_\tau^{\mathbf{M},\omega}$ and $\bar{e}_\tau^{\mathbf{M},\omega}$:

$$\bar{e}_\tau^{\mathbf{M}}(\Gamma) = \sup_{\omega \in \mathcal{W}(\Gamma)} \bar{e}_\tau^{\mathbf{M},\omega}, \quad \underline{e}_\tau^{\mathbf{M}}(\Gamma) = \inf_{\omega \in \mathcal{W}(\Gamma)} \underline{e}_\tau^{\mathbf{M},\omega} \tag{4.10}$$

As shown in Appendix C.4, these quantities can be computed from data in $O(n^{r+1})$ time.

Corollary 1. *Suppose* $\omega \in \mathcal{W}(\Gamma)$. *Then*

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \overline{e}_{\tau}^{\mathbf{M}}(\Gamma)\right) \geq \tau\left(\frac{n-r}{n+\Gamma^2}\right),$$

and

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \in [\underline{e}_{\tau}^{\mathbf{M}}(\Gamma), \overline{e}_{\tau}^{\mathbf{M}}(\Gamma)]\right) \geq (2\tau-1)\left(\frac{n-r}{n+\Gamma^2}\right).$$

4.4.3 The analyst knows nothing about ω

Finally, we provide two ways for comparing the transferability of two models $i = 1, 2$ when the analyst cannot bound ω. We do not provide formal results about these orders, but show that they have bite in our subsequent application (see Section 4.5).

Let $\overline{e}_{i,\tau}^{\mathbf{M}}(\Gamma)$ and $\underline{e}_{i,\tau}^{\mathbf{M}}(\Gamma)$ again denote the worst case bounds for model i, as defined in (4.10).

Definition 8 (Worst-Case Dominance). Say that model 1 worst-case-dominates model 2 at the τ-th quantile if

$$\overline{e}_{1,\tau}^{\mathbf{M}}(\Gamma) \leq \overline{e}_{2,\tau}^{\mathbf{M}}(\Gamma) \quad \forall \Gamma \in [1, \infty).$$

That is, model 1 worst-case-dominates model 2 at the τ-th quantile if for every Γ, the worst-case upper bound for model 1 is less than the worst-case for upper bound for model 2.

We can strengthen this comparison by requiring the upper bound of the forecast interval for model 1 to be smaller than the upper bound of the forecast interval for model 2 pointwise for each $\omega \in \mathcal{W}(\Gamma)$, rather than simply comparing worst-case upper bounds.

Definition 9 (Everywhere Dominance). Say that model 1 everywhere dominates model 2 at the τ-th quantile if

$$\overline{e}_{1,\tau}^{\mathbf{M},\omega} \leq \overline{e}_{2,\tau}^{\mathbf{M},\omega} \quad \forall \Gamma > 1 \forall \omega \in \mathcal{W}(\Gamma).$$

Many decision rules will not be comparable under either of these definitions, but we show they are empirically relevant in our application. The even stronger requirement that $\overline{e}_{1,\tau}^{\mathbf{M},\omega} \leq \underline{e}_{2,\tau}^{\mathbf{M},\omega}$ uniformly in ω, i.e., that the upper bound of model 1's forecast interval is always smaller than the lower bound of model 2's forecast interval, is likely too stringent to be useful in practice.[13]

4.5 Application

To illustrate our methods, we evaluate the transferability of predictions of certainty equivalents for binary lotteries across domains that have different subject pools and also differ in other ways. We focus on this application for several reasons: First, there are many public data sources that we can use to construct our metadata. Second, the associated economic models have been extensively examined from the perspective of predictive performance (Harless and Camerer, 1994; Hey and Orme, 1994; Bruhin et al., 2010; Bernheim and Sprenger, 2020), and recent work evaluates how well these models predict relative to black box algorithms (Peysakhovich and Naecker, 2017; Plonsky et al., 2019; Fudenberg et al., 2022a). Finally, as Einav et al. (2012) points out, given how models of risk preferences are often used in practice, it is also important to evaluate how well they transfer across domains. We thus view this application as a natural setting to illustrate our methods.

Section 4.5.1 describes our metadata, and Section 4.5.2 describes the decision rules we consider. Section 4.5.3 conducts "within-domain" out-of-sample tests, where the training and test data are drawn from the same domain. Section 4.5.4 compares transfer performance across domains by constructing forecast intervals for three different definitions of transfer error.

[13]This stronger order has bite only when the transfer error for model 1 across "the most dissimilar" training and testing domains is lower than the transfer error for model 2 for "the most similar" training and testing domains.

4.5.1 Data

Our metadata consists of samples of certainty equivalents from 44 subject pools, which we treat as the domains. These data are drawn from 14 papers in experimental economics, with twelve papers contributing one sample each, one paper contributing two, and a final paper (a study of risk preferences across countries) contributing 30 samples. Our samples range in size from 72 observations to 8906 observations, with an average of 2752.7 observations per sample.[14] Besides the difference in subject pools, these samples may differ in other details, such as whether the lotteries were restricted to the gain domain. We convert all prizes to dollars using purchasing power parity exchange rates (from OECD 2023) in the year of the paper's publication.

Within each sample, observations take the form $(z_1, z_2, p; y)$, where z_1 and z_2 denote the possible prizes of the lottery (and we adopt the convention that $|z_1| > |z_2|$), p is the probability of z_1, and y is the reported certainty equivalent by a given subject. Thus our feature space is $\mathcal{X} = \mathbb{R} \times \mathbb{R} \times [0, 1]$, the outcome space is $\mathcal{Y} = \mathbb{R}$, and a prediction rule is any mapping from binary lotteries into predictions of the reported certainty equivalent. We use squared-error loss $\ell(y, y') = (y - y')^2$ to evaluate the error of the prediction, but for ease of interpretation we report results in terms of root-mean-squared error, which puts the errors in the same units as the prizes.[15] Since different subjects report different certainty equivalents for the same lottery, the best achievable error is generally bounded away from zero.

[14] Online Appendix C.5.1 describes our data sources in more detail.

[15] This transformation is possible because none of the results in this paper change if we redefine $e(\sigma, S) = g\left(\frac{1}{\#S}\sum_{(x,y)\in S}\ell(\sigma(x), y)\right)$ for any function g. Root-mean-squared error corresponds to setting $g(x) = \sqrt{x}$.

4.5.2 Models and black boxes

We consider two parametric economic models of certainty equivalents and two off-the-shelf black box algorithms.

Economic models. First we consider an expected utility agent with a CRRA utility function parameterized by $\eta \geq 0$ (henceforth EU). For $\eta \neq 1$, define

$$v_\eta(z) = \begin{cases} \frac{z^{1-\eta}-1}{1-\eta} & \text{if } z \geq 0 \\ -\frac{(-z)^{1-\eta}-1}{1-\eta} & \text{if } z < 0 \end{cases}$$

and for $\eta = 1$, set $v_\eta(z) = \ln(z)$ for positive prizes and $v_\eta(z) = -\ln(-z)$ for negative prizes. For each $\eta \geq 0$, define the prediction rule σ_η to be

$$\sigma_\eta(z_1, z_2, p) = v_\eta^{-1}\big(p \cdot v_\eta(z_1) + (1-p) \cdot v_\eta(z_2)\big).$$

That is, the prediction rule σ_η maps each lottery into the predicted certainty equivalent for an EU agent with utility function v_η.

Next we consider the set of prediction rules Σ_{CPT} derived from the parametric form of Cumulative Prospect Theory (CPT) first proposed by Goldstein and Einhorn (1987) and Lattimore et al. (1992). Fixing values for the model's parameters $(\alpha, \beta, \delta, \gamma)$, each lottery (z_1, z_2, p) is assigned a utility

$$w(p)v(z_1) + (1 - w(p))v(z_2)$$

where

$$v(z) = \begin{cases} z^\alpha & \text{if } z \geq 0 \\ -(-z)^\beta & \text{if } z < 0 \end{cases} \tag{4.11}$$

is a value function for money, and

$$w(p) = \frac{\delta p^\gamma}{\delta p^\gamma + (1-p)^\gamma} \tag{4.12}$$

is a probability weighting function.

For each $\alpha, \beta, \gamma, \delta$, the prediction rule $\sigma_{(\alpha,\beta,\gamma,\delta)}$ is defined as

$$\sigma_{(\alpha,\beta,\gamma,\delta)}(z_1, z_2, p) = v^{-1}\big(w(p)v(z_1) + (1 - w(p))v(z_2)\big).$$

That is, the prediction rule maps each lottery into the predicted certainty equivalent under CPT with parameters $(\alpha, \beta, \gamma, \delta)$. Following the literature, we impose the restriction that the parameters belong to the set $\Theta = \{(\alpha, \beta, \gamma, \delta) : \alpha, \beta, \gamma \in [0, 1], \delta \geq 0\}$.

We also evaluate restricted specifications of CPT that have appeared elsewhere in the literature: CPT with free parameters α and β (setting $\delta = \gamma = 1$) describes an expected utility decision-maker whose utility function is as given in (4.11); CPT with free parameters α, β and γ (setting $\delta = 1$) is the specification used in Karmarkar (1978); and CPT with free parameters δ and γ (setting $\alpha = \beta = 1$) describes a risk-neutral CPT agent whose utility function over money is $u(z) = z$ but who exhibits nonlinear probability weighting. Additionally, we include CPT with the single free parameter γ (setting $\alpha = \beta = \delta = 1$), which Fudenberg et al. (2023) found to be an especially effective one-parameter specification.

Black Box Algorithms. We consider two popular machine learning algorithms. First, we train a *random forest* (RF), which is an ensemble learning method consisting of a collection of decision trees.[16] Second, we train a *kernelized ridge regression* model (KR), which modifies OLS to weight

[16] A decision tree recursively partitions the input space, and learns a constant prediction for each partition element. The random forest algorithm collects the output of the individual decision trees, and returns their average as the prediction. Each decision tree is trained with a sample (of equal size to our training data) drawn with replacement

observations at nearby covariate vectors more heavily, and additionally places a penalty term on the size of the coefficients. Specifically, we use the radial basis function kernel $\kappa(x, \tilde{x}) = e^{-\gamma\|x-\tilde{x}\|_2^2}$ to assess the similarity between covariate vectors x and $\tilde{x}$. Given training data $\{(x_i, y_i)\}_{i=1}^N$, the estimated weight vector is $\vec{w} = (\mathbb{K} + \lambda I_N)^{-1}\vec{y}$, where $\mathbb{K}$ is the $N \times N$ matrix whose (i, j)-th entry is $\kappa(x_i, x_j)$, I_N is the $N \times N$ identity matrix, and $\vec{y} = (y_1, \ldots, y_N)'$ is the vector of observed outcomes in the training data. The estimated prediction rule is $\sigma(x) = \sum_{i=1}^N w_i \kappa(x, x_i)$.

There are at least two approaches for cross-validating hyper-parameters such as the size of the trees in the random forest algorithm. First, when there are multiple training domains one can cross-validate across them; we use this in Appendix C.5.6. Second, one can cross-validate across observations within the training domains. Since we are interested in cross-domain performance, rather than within-domain performance, it is not guaranteed that this will improve performance, and indeed we find that choosing the hyper-parameters via within-domain cross-validation leads to worse transfer performance than using default values. Thus in our main analysis with a single training domain, we set all hyper-parameters to default values.[17]

Discussion. There is no established definition of what constitutes an economic model versus a black box algorithm, but one way of distinguishing between the two approaches is whether the prediction method is tailored to a general application or a general-purpose method of prediction. EU and CPT model the risk preferences of economic agents; we would not expect these models to predict well if we changed our problem to image classification. In contrast, random forest algorithms and kernel regression have been successfully applied across a wide array of prediction problems. In this sense, EU and CPT are economic models, while RF and KR are not. Our

from the actual training data. At each decision node, the tree splits the training samples into two groups using a True/False question about the value of some feature, where the split is chosen to greedily minimize mean squared error.

[17]Specifically, we set $\lambda = 1$ and $\gamma = 1/(\#\text{covariates}) = 1/3$ in the kernel regression algorithm. See Pedregosa et al. (2011) and Chapter 14 of Murphy (2012) for further reference. For the random forest model, we set the maximum depth to none, so the tree is extended until outcomes are homogeneous within each leaf.

approach and results can, however, equally be applied to evaluate prediction methods that are a hybrid of the two approaches. For example, Plonsky et al. (2019) and Hsieh et al. (2023) consider black box algorithms whose inputs are based on prior economic theory. We leave investigation of the transfer performance of such methods to future work.

We note finally that although black box algorithms are traditionally perceived as more flexible than economic models, whether this is in fact the case is something that has to be determined case-by-case. In particular, Fudenberg et al. (2023) shows that although CPT uses only four parameters, it imposes very few restrictions on mappings from binary lotteries to certainty equivalents.

4.5.3 Within-domain performance

We first evaluate how these models perform when trained and evaluated on data from the same subject pool. We compute the tenfold cross-validated out-of-sample error for each decision rule in each of the 44 domains.[18] The two black box methods (random forest and kernel regression) each achieve lower cross-validated error than EU and CPT in 38 of the 44 domains, although the improvement is not large. To obtain a simple summary statistic for the comparison between the economic models and black boxes, we normalize each economic model's error (in each domain) by the random forest error. Table 4.1 averages this ratio across domains and shows that on average, the cross-validated errors of the economic models are slightly larger than the random forest error. That is, the CPT error is on average 1.06 times the random forest error, and the EU error is on average 1.21 times the random forest error.[19]

[18] We split the sample into ten subsets at random, choose nine of the ten subsets for training, and evaluate the estimated model's error on the final subset. The tenfold cross-validated error is the average of the out-of-sample errors on the ten possible choices of test set.

[19] The numbers in Table 4.1 are very similar if we normalized by the kernel regression error instead.

Model	Normalized Error
EU	1.21
CPT variants	
γ	1.12
α, β	1.22
δ, γ	1.08
α, β, γ	1.07
$\alpha, \beta, \delta, \gamma$	1.06

Table 4.1: *Average ratio of out-of-sample errors relative to random forest.*

These results suggest that the different prediction methods we consider are comparable for within-domain prediction, with the black boxes performing slightly better. But the results do not distinguish whether the economic models and black boxes achieve similar out-of-sample errors by selecting approximately the same prediction rules, or if the rules they select lead to substantially different predictions out-of-domain. We also cannot determine whether the slightly better within-domain performance of the black box algorithms is achieved by learning generalizable structure that the economic models miss, or if the gains of the black boxes are confined to the domains on which they were trained. We next separate these explanations by evaluating the transfer performance of the models.

4.5.4 Transfer error

We use the results in Section 4.3.2 to construct forecast intervals for the two specifications of transfer error defined in (4.1) and(4.2), which we will subsequently call *raw transfer error* and *transfer shortfall* respectively. We also consider another normalization of the raw transfer error

with respect to a proxy for the best achievable error on the target sample. Let $m \in \mathcal{M}$ index a set of models that each prescribe rules f^m for mapping data to prediction rules. Then *transfer shortfall*

$$\frac{e(f_{S_{\mathbf{T}}}, S_{n+1})}{\min_{m \in \mathcal{M}} e\left(f^m_{S_{n+1}}, S_{n+1}\right)} \tag{4.13}$$

reveals how much lower the accuracy of the transferred model $f_{S_{\mathbf{T}}}$ is compared to the best in-sample accuracy using a model from $\mathcal{M}$.[20] One advantage of this specification relative to (4.1) is that the raw error is very sensitive to the predictability of y given x in the target sample, which may differ across domains but is not directly related to the model's transferability.

In our meta-data there are $n = 44$ domains, and we choose $r = 1$ of these to use as the training domain which corresponds to the question, "If the researcher draws one domain at random, and then tries to generalize to another domain, how well will they do?" Figure 4.4 displays two-sided forecast intervals for transfer performance, transfer deterioration, and transfer shortfall (where R includes all decision rules shown in the figure). These forecast intervals use $\tau = 0.95$, so the upper bound of the forecast interval is the 95th percentile of the pooled transfer errors (across choices of the training and test domains), and the lower bound of the forecast interval is the 5th percentile of the pooled transfer errors. (See Table C.2 in Appendix C.5.3 for the exact numbers.) Applying Proposition 6, these are 86% forecast intervals. Choosing larger τ results in wider forecast intervals that have higher coverage levels, and we report some of these alternative forecast intervals in Online Appendix C.5.4, including a 96% forecast interval.

Our main takeaway from Figure 4.4 is that although the prediction methods we consider are very similar from the perspective of within-domain prediction, they have very different out-of-

[20] This quantity (subtracted from 1) is similar to the "completeness" measure introduced in Fudenberg et al. (2022a), without the use of a baseline model to set a maximal reasonable error, and adapted for the transfer setting by training and testing on samples drawn from different domains.

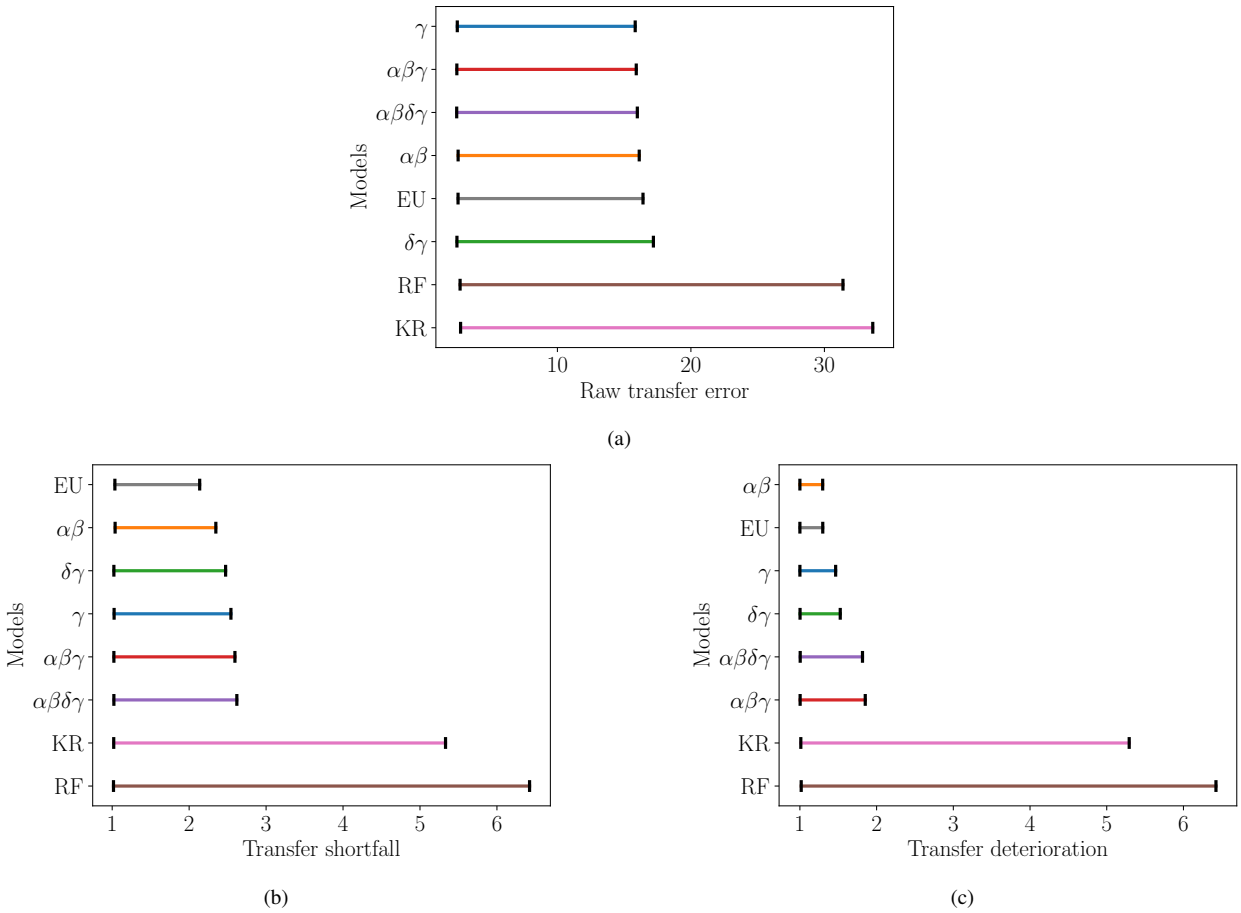


Figure 4.4: *86% (n=44, $\tau = 0.95$) forecast intervals for (a) raw transfer error, (b) transfer shortfall (with $\mathcal{R}$ consisting of the decision rules shown in the figure), and (c) transfer deterioration.*

domain implications. Panel (a) of Figure 4.4 shows that the black box forecast intervals for raw transfer error have upper bounds that are roughly twice those of the economic models. Panel (b) shows that the contrast between the economic models and the black boxes is even larger for transfer shortfall, which removes the common variation across models that emerges from variation in the predictability of the different target samples. Panel (c) of Figure 4.4, which reports transfer deterioration, shows that it is less important to re-estimate the economic models on new target domains than to retrain the black-box algorithms.

All of the forecast intervals overlap for each of the three measures. This is not surprising, as variation in the transfer errors due to the random selection of training and target domains cannot be eliminated even with data from many domains. [21] Section C.3 provides confidence intervals for different population quantities, including quantiles of the transfer error distribution and the expected transfer error, whose width we do expect to vanish as the number of domains grow large. There, we find similar conclusions with regards to the relative performance of the black box algorithms and economic models.

The appendix provides several robustness checks and complementary analyses. Online Appendix C.5.4 plots the τ-th percentile of pooled transfer errors as τ varies, demonstrating that forecast intervals constructed using other choices of τ (besides $\tau = 0.95$) would look similar to those shown in the main text. Online Appendix C.5.5 provides 86% forecast intervals for the ratio of the raw random forest transfer error to the raw CPT transfer error, and finds that the random forest error is sometimes much higher than the CPT error, but is rarely much lower. Online Appendix C.5.6 considers an alternative choice for the number of training domains, setting $r = 3$ instead of $r = 1$. While the results are similar, the contrast between the economic models and black boxes is not as large, suggesting that the relative performance of the black boxes improves given a larger number of training domains. Online Appendix C.5.2 provides forecast intervals when each of the 14 papers is treated as a different domain; once again the black box methods transfer worse than the economic models do.

We next use our theoretical results from Section 4.4 to study the consequences of relaxing the i.i.d. assumption in our comparison of CPT$(\alpha, \beta, \delta, \gamma)$ and RF. Since the main differences observed above concerned the upper bounds of our forecast intervals, we limit attention to $\tau \geq 0.5$, and

[21] We expect the black box intervals and the economic model intervals to overlap so long as the economic model errors on "upper tail" training and target domain pairs exceed the black box errors on "lower tail" training and target domain pairs.

compare the methods in terms of worst-case and everywhere dominance with respect to all three measures of transfer performance. These results are summarized in Table 4.2.

Type	raw transfer error	transfer shortfall	transfer deterioration
Worst-case dominance	$\tau \geq 0.5$	$\tau \geq 0.5$	$\tau \geq 0.5$
Everywhere dominance	$\tau \geq 0.954$	$\tau \geq 0.866$	$\tau \geq 0.647$

Table 4.2: *Comparison between CPT and RF in terms of worst-case and everywhere dominance. Each cell gives the range of τ at which CPT dominates RF.*

Table 4.2 shows that CPT worst-case-dominates RF at all quantiles $\tau \geq 0.5$ and for all three transfer error measures. Hence, our finding that the upper tail of transfer errors is larger for RF than for CPT is robust to relaxing the assumption that the training and test domains are drawn from the same distribution, provided that we are comfortable comparing the upper bound for one method to the upper bound for the other. In Appendix C.5.8, we provide a more detailed view of worst-case dominance by plotting $\bar{e}^{\mathbf{M}}_{\tau}(\Gamma)$ as functions of τ and Γ, respectively.

We can also consider the more demanding everywhere dominance criterion, which asks what happens if we relax our i.i.d. sampling assumption in a way which is as favorable to RF (and as unfavorable to CPT) as possible. We find a substantial degree of robustness even under this highly demanding criterion: CPT everywhere dominates RF in raw transfer error for all $\tau \geq 0.954$, everywhere-dominates in transfer shortfall for $\tau \geq 0.866$, and everywhere dominates in transfer deterioration for $\tau \geq 0.647$.

4.5.5 Do black boxes transfer poorly because they are too flexible?

One tempting explanation of why the black box algorithms transfer less well is that they may overfit to idiosyncratic details of the training samples that do not generalize across subject pools.

For example, suppose some subject pools tend to value lotteries depending on the specific digits they contain.[22] This regularity could not be captured by the economic models, because they do not include parameters for individual digits, but could be learned by a random forest algorithm. If so, the random forest would have better within-domain prediction for those subject pools, but worse transfer performance (assuming this regularity does not generalize across subject pools).

While the flexibility of black box algorithms is likely an important determinant of their transfer performance, a second analysis shows that this cannot be a complete explanation of our result. One of the papers we use is based on samples of certainty equivalents from 30 countries (l'Haridon and Vieider, 2019). Of the 30 samples from this paper, 29 samples report certainty equivalents for the same 28 lotteries, and the remaining sample reports certainty equivalents for 24 of those lotteries. We repeat our analysis using these 30 samples as the metadata, and find that the forecast intervals for raw transfer error are indistinguishable across the prediction methods (Panel (a) of Figure 4.5). There is some separation between the forecast intervals for the remaining two measures, but in both cases the CPT and random forest forecast intervals are substantially more similar than in the original data. If overfitting were the main explanation of our previous results, we would expect the black box algorithms to overfit here as well.

In contrast, we find that the economic models again outperform the black box algorithms when we consider a different definition of domains for the l'Haridon and Vieider (2019) data. Specifically, we aggregate all observations for the 24 lotteries that are shared in all 30 samples, and split these observations into 24 samples, where each sample includes all reported certainty equivalents (across subject pools) for a given lottery. For this new definition of domains, our transfer measures evaluate how well a model estimated on data from certain lotteries predict certainty equivalents

[22]For example, Fortin et al. (2014) find that in neighborhoods with a higher than average percentage of Chinese residents, homes with address numbers ending in "4" are sold at a 2.2% discount and those ending in "8" are sold at a 2.5% premium.

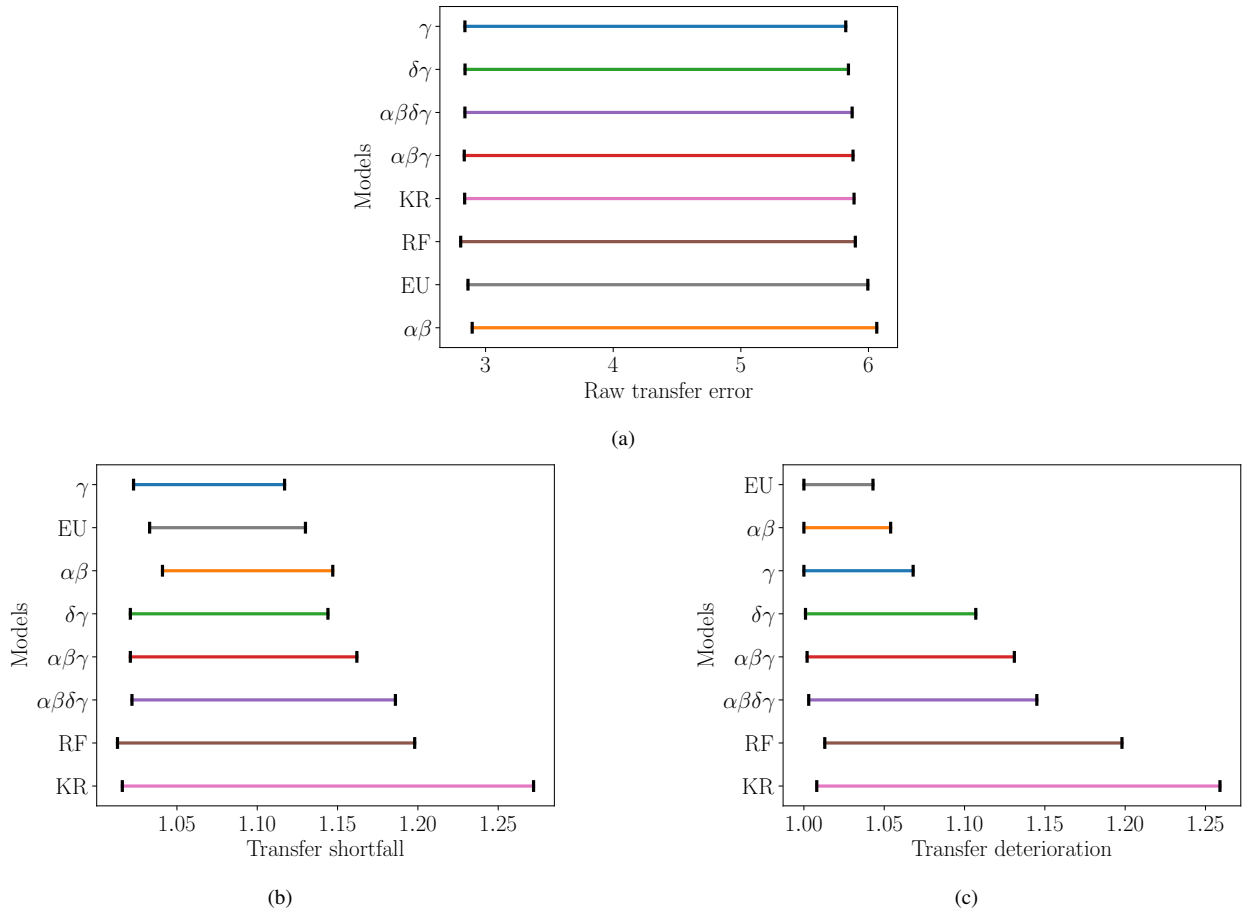


Figure 4.5: *This figure reports 84% (n=30, τ=0.95) forecast intervals when we transfer across subject pools in the l'Haridon and Vieider (2019) data.*

for other lotteries. Figure 4.6 reports 83-level forecast intervals, and we find that the economic models transfer substantially better than the black box algorithms. In fact, isolating the difference across domains to be differences across lotteries exaggerates the relative value of economic models even relative to our original Figure 4.4 (which uses a definition of domains that combines several sources of variation). For consistency, Figure 4.6 reports confidence intervals for $r = 1$ (corresponding to training on one lottery and predicting on another), but we show in Appendix C.5.7 that the qualitative features of this figure hold also for $r = 3$ and $r = 5$.

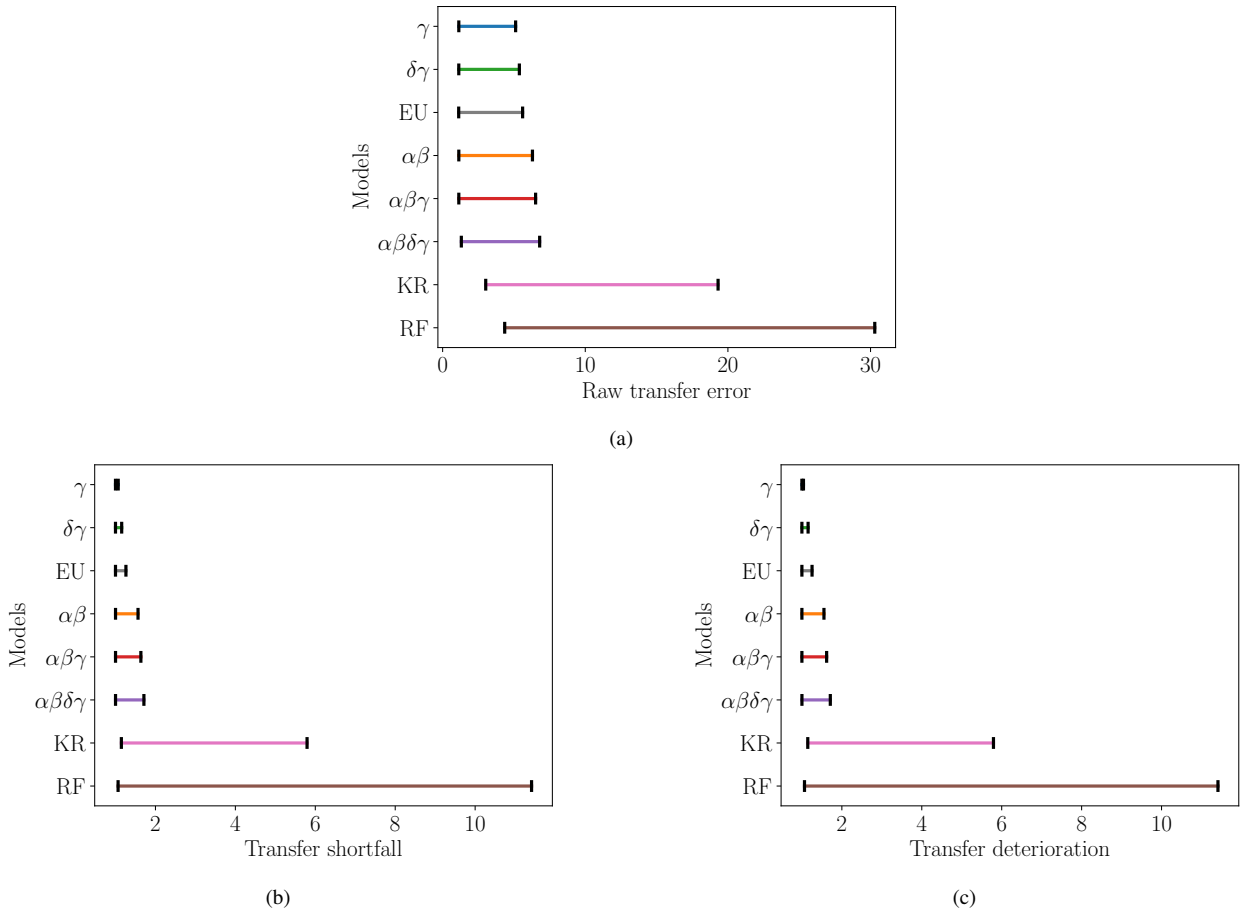


Figure 4.6: *This figure reports 83% (n=24, τ=0.95) forecast intervals when we transfer across lotteries in the l'Haridon and Vieider (2019) data.*

4.5.6 Two kinds of transfer problems

Section 4.5.5 shows that differences in transfer performance between economic models and black box algorithms cannot be explained solely by flexibility and overfitting.[23] This motivates distinguishing between different sources of cross-domain variation, using the fact that our framework allows the distribution P governing the training sample and the distribution P' governing the test

[23] In fact, the flexibility gap between the black boxes and economic models is not large: many conditional mean functions (for binary lotteries) can be well approximated by CPT for some choice of parameters values $\alpha, \beta, \delta, \gamma$ (Fudenberg et al., 2023).

sample to differ. At one extreme, P and P' may share a common marginal distribution on the feature space $\mathcal{X}$, but have very different conditional distributions $P_{Y|X}$ and $P'_{Y|X}$ (known as *model shift*). In our application, this would mean that the distribution over lotteries is the same, but the conditional distribution of reported certainty equivalents is different across domains. At another extreme, the conditional distributions $P_{Y|X}$ and $P'_{Y|X}$ might be the same, but the marginal distributions over the feature space could differ across domains, e.g., if different kinds of lotteries are used in different domains (known as *covariate shift*).

Our findings in Figure 4.5 suggest that black boxes do as well as economic models at transfer prediction when the marginal distribution over features P_X is held constant across samples. Intuitively, when the relevant feature vectors are the same in every sample, a black box algorithm can perform well by simply memorizing a prediction for each of these feature vectors. In contrast, when the set of lotteries varies across samples, then good transfer prediction necessarily involves extrapolation, and an algorithm that hasn't identified the right structure for relating behavior across lotteries will fail to generalize. Since economic models of risk preferences are intended to relate an individual's preferences across lotteries that permits extrapolation of this form, our empirical results suggest that they do so effectively.

For a simple, stylized, example of this contrast, consider three domains with degenerate distributions over observations. In domain 1, the distribution is degenerate at the lottery $(z_1, z_2, p) = (10, 0, 1/2)$ and certainty equivalent $y = 3$. In domain 2, the distribution is degenerate at the lottery $(z_1, z_2, p) = (10, 0, 1/2)$ and certainty equivalent $y = 4$. In domain 3, the distribution is degenerate at a new lottery $(z_1, z_2, p) = (20, 10, 1/10)$ and certainty equivalent $y = 11$. Suppose EU and a decision tree are both trained on a sample from domain 1. The CRRA parameter $\eta \approx 0.64$ perfectly fits the observation $(10, 0, 1/2; 3)$, as does the trivial decision tree that predicts $y = 3$ for all lotteries. The estimated EU model and decision tree are equivalent for predicting observations

in domain 2: both predict $y = 3$ and achieve a mean-squared error of 1. But their errors are very different on domain 3: the EU prediction for the new lottery is approximately 10.8 with a mean-squared error of approximately 0.05, while the decision tree's prediction is 3 with a mean-squared error of 64.

4.5.7 Predicting the relative transfer performance of black boxes and economic models

The preceding sections suggest that the relative transfer performance of black boxes and economic models is determined primarily by shifts in which lotteries are sampled, rather than shifts in behavior conditional on those lotteries. To further test this conjecture, we examine how well we can predict the ratio of the raw random forest transfer error to the raw CPT transfer error given information about the training and test lotteries but not about the distribution of certainty equivalents in either sample. If the relative performance of these methods depended importantly on behavioral shifts in the two domains—i.e., a change in the distribution of certainty equivalents for the same lotteries—then we would expect prediction of the relative performance based on lottery information alone to be poor. We find instead that lottery information has substantial predictive power for this ratio.

For each sample $S = \{(z_{1,i}, z_{2,i}, p_i; y_i)\}_{i=1}^m$, we consider the following features: the mean, standard deviation, max, and min value of z_1 among the lotteries in S; the mean, standard deviation, max, and min value of z_2 among the lotteries in S; the mean, standard deviation, max, and min value of p among the lotteries in S; the mean, standard deviation, max, and min value of $1 - p$ among the lotteries in S; the mean, standard deviation, max, and min of $pz_1 + (1-p)z_2$ among the lotteries in S; the size of S; and an indicator variable for whether $z_1, z_2 \geq 0$ for all lotteries in S.

We consider three possible feature sets: (a) *Training Only,* which includes all features derived from the training sample $\mathbf{M}_{\mathcal{T}}$; (b) *Test Only,* which includes all features derived from the test

sample S_d, (c) *Both,* which includes all features derived from the training sample $\mathbf{M}_{\mathcal{T}}$ and the test sample S_d. We evaluate two prediction methods: OLS and a random forest algorithm. Table 4.3 reports tenfold cross-validated errors for each of these feature sets and prediction methods. As a benchmark, we also consider the best possible constant prediction.

	Train Only	Test Only	Both
Constant	2.57	2.57	2.57
OLS	1.00	2.61	0.94
RF	0.98	2.52	0.76

Table 4.3: *Cross-Validated MSE*

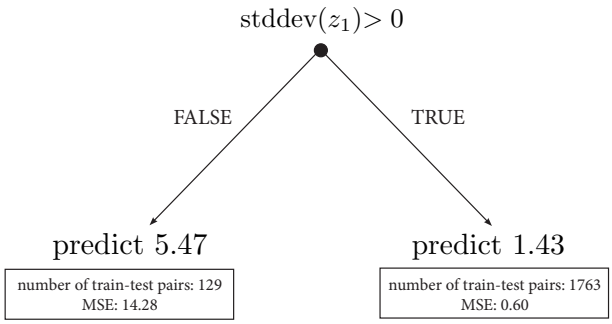


Figure 4.7: *Best 1-split decision tree based on training and test features.*

The best constant prediction achieves a mean-squared error of 2.57, which can be more than halved using features of the training set alone. Using features of both the training and test sets, the random forest algorithm reduces error to 30% of the constant model.

The random forest algorithm is too opaque to deliver insight into how it achieves these better predictions, but we can obtain some understanding by examining the best 1-split decision tree,

shown in Figure 4.7 below. This decision tree achieves a cross-validated MSE of 1.75, reducing the error of the constant model by 32%. It partitions the set of (train,test) domain pairs into two groups depending on whether the standard deviation of z_1 (the larger prize) in the training set of lotteries exceeds zero. There are three domains in which the prizes (z_1, z_2) are held constant across all training lotteries (although the probabilities vary). In the 129 transfer prediction tasks where one of these three domains is used for training, the decision tree predicts the ratio of the random forest error to the CPT error to be 5.47. For all other transfer prediction tasks, the decision tree predicts 1.43. This finding reinforces our intuition that the relative performance of the black boxes and economic models is driven in part by whether the training sample covers the relevant part of the feature space. When the training observations concentrate on an unrepresentative part of the feature space (such as all lotteries that share a common pair of prize outcomes), then the black boxes transfer much more poorly than economic models.

Our results also clarify a contrast between transfer performance and classical out-of-sample performance. In out-of-sample testing, the marginal distribution on $\mathcal{X}$ is the same for the training and test samples, so the set of training lotteries is likely to be representative of the set of test lotteries as long as the training sample is sufficiently large. When test and training samples are governed by distributions with different marginals on $\mathcal{X}$, the set of training lotteries can be unrepresentative of the set of test lotteries regardless of the number of training observations. Training on observations pooled across many domains alleviates the potential unrepresentativeness of the training data, but the number of domains needed will depend on properties of the distribution: An environment where each domain puts weight on exactly one lottery that is itself sampled i.i.d. may be difficult for black-box algorithms,[24] while an environment where the marginal distribution is degenerate on the same lottery in all domains may be easier. There is no analog in out-of-sample testing for the

[24] In this case, the number of domains black boxes need to achieve good transfer performance is likely comparable to the number of observations they need for good out-of-sample performance, which can be quite large.

role played by variation in the marginal distribution on $\mathcal{X}$ across domains. Moving beyond our specific application, we expect this variation to be an important determinant of the relative transfer performance of black box algorithms and economic models in general.

4.6 Conclusion

Our measures of transfer error quantify how well a model's performance on one domain extrapolates to other domains. We applied these measures to show that the predictions of expected utility theory and cumulative prospect theory outperform those of black box models on out-of-domain tests, even though the black boxes generally have lower out-of-sample prediction errors within a given domain. The relatively worse transfer performance of the black boxes seems to be because the black box algorithms have not identified structure that is commonly shared across domains, and thus cannot effectively extrapolate behavior from one set of features to another. Our finding that the economic models transfer better supports the intuition that economic models capture regularities that are general across a variety of domains.

One may view our theoretical results, and the statistical assumptions which justify them, from two perspectives. Taken literally, we consider a researcher who has access to data from a small set of domains, and an analyst who is interested in how well the researcher's procedure extrapolates. To provide guarantees we restrict the ways domains relate to each other; the simplest and strongest of these is that these is that the domains are exchangeable. We also assume that the analyst has access to data from a larger set of domains than the researcher does. One might wonder why the researcher uses a small set of domains for training, when the larger meta dataset is available to the analyst. In particular, the researcher might prefer to use the larger cross-domain dataset for training, leaving no "fresh" domains for performance evaluation. Although one could potentially weaken our data requirements to cover such cases by restricting the researcher's estimator, this

would rule out the black box prediction methods that are a focus of our analysis.

Alternatively, one can view our results as a tool for evaluating and comparing the performance of different methods for cross-domain prediction. The starting point for our analysis of a given method is the matrix collecting that method's cross-domain prediction errors for a set of observed domains. These matrices are rich objects, and it is not obvious how to compare them, but our theoretical analysis points to a natural set of summary statistics: quantiles of the observed error distribution. Focusing on these quantiles, or equivalently on our forecast intervals, provides a theoretically-motivated way to make comparisons across methods.

CHAPTER 5
CONCLUSION AND FUTURE DIRECTIONS

This dissertation set out to explore a central tension in the computational turn of economics: the same tools that expand our ability to measure and predict also introduce new questions about what those measurements mean, and whether those predictions generalize. The three papers collected here engage this tension from complementary angles.

Chapter 2 demonstrates that a latent economic concept, scientific novelty, can be operationalized with far greater precision than prior work allowed, but only once it is recognized as multidimensional. The spatial-temporal decomposition reveals that the economics profession has been implicitly conflating two distinct strategies for producing knowledge, strategies that carry fundamentally different consequences for scientific progress. Chapter 3 shows that an analogous decomposition applies to strategic behavior: the game-theoretic ideal and the cognitively accessible are not the same thing, and the gap between them, complexity, can be strategically managed in the interaction to gain advantages. Chapter 4 closes the loop by asking whether the predictive models that enable Chapters 2 and 3 can be trusted when applied outside their estimation environment. The answer is qualified: black-box algorithms are powerful within a domain, but that power does not transfer freely, and models grounded in economic theory offer a form of robustness that raw predictive performance obscures.

Taken together, these papers advance a methodological argument: computational tools in economics are most productive when they are disciplined by theory on the way in, and interrogated critically on the way out. Measurement without theory produces metrics whose economic meaning is uncertain. Prediction without attention to transferability produces models whose policy

relevance is fragile. The contribution of this dissertation is to demonstrate, across three substantive domains, what it looks like to hold both standards simultaneously.

5.1 Broader Frontiers

The three papers in this dissertation use computational tools primarily as instruments of measurement and inference. Yet the frontier of computational economics is moving rapidly in directions that this work does not fully engage, and that represent natural extensions of the questions raised here. Three developments seem particularly consequential.

Large Language Models as Simulators of Economic Agents. The most immediate frontier is the use of LLMs not merely to measure text, as in Chapter 2, but to simulate human behavior directly. A growing body of work has begun to explore whether LLMs can serve as credible stand-ins for human subjects in economic experiments, including but not limited to generating choices, stated preferences, and strategic responses that approximate those of real agents (Horton, 2023). This possibility is consequential for the kind of research conducted in Chapter 3: if LLMs can simulate players of varying skill levels, they could be used to run controlled experiments on the emergence of complexity that are infeasible with observational chess data alone, such as exogenously varying the skill gap between opponents or shutting down specific strategic incentives. More broadly, LLM-based simulation could substantially reduce the cost of generating behavioral data in settings such as auctions, bargaining, financial markets, where human subject experiments are expensive or ethically constrained.

The critical open question, however, is one of validity. Simulated agents must be calibrated against real behavior, and the conditions under which LLM responses faithfully approximate human economic decisions are not well understood. The transfer error framework developed in Chap-

ter 4 provides a natural lens for this problem: an LLM simulator is reliable precisely to the extent that its "in-domain" performance (replicating known experimental results) transfers to the "out-of-domain" setting of interest. Establishing this kind of disciplined validation is, in the author's view, among the most important methodological tasks facing the next generation of computational economists.

Generative Models and Synthetic Data. A related development is the use of generative AI to produce synthetic datasets that preserve the statistical properties of sensitive or proprietary data. The potential applications in economics are significant: administrative microdata, clinical records, and financial transactions are often inaccessible to researchers precisely because of privacy constraints. Synthetic data generation could in principle democratize access to rich observational datasets.

Yet the tension identified in Chapter 4 applies here with particular force. A generative model trained on data from one environment will encode the distributional features of that environment. Synthetic data produced from it will inherit those features, and any model trained on that synthetic data will face a transfer problem when applied to real data from a different context. Understanding the conditions under which synthetic training data produces models that generalize to real-world deployment is a problem that the forecast interval framework of Chapter 4 is well-positioned to address, and represents a natural extension of that paper's methodology.

Hypothesis Generation and the Automation of Discovery. Perhaps the most speculative but consequential frontier is the use of LLMs to generate scientific hypotheses (Ludwig and Mullainathan, 2024). The semantic novelty framework of Chapter 2 provides a way to locate any research papers, including an LLM-generated conjecture, within the existing intellectual landscape of a discipline. This raises an intriguing possibility: one could use the spatial-temporal decom-

position to audit the novelty profile of LLM-generated hypotheses, assessing whether they tend to cluster in well-explored regions of the knowledge space (high temporal, low spatial novelty) or whether they are capable of identifying genuinely underexplored intellectual territory.

If, as seems plausible, LLM-generated hypotheses are systematically biased toward the frontier of existing discourse—trained as they are on the most widely circulated text—then the spatial novelty measure could serve as a diagnostic for identifying the blind spots of automated discovery. More broadly, integrating the measurement tools of Chapter 2 with generative models offers a path toward a quantitative science of science that goes beyond citation analysis to map the actual topology of knowledge production and its automation.

5.2 A Final Remark

The computational turn in economics is still unfolding. The tools available to researchers today are qualitatively different from those of a decade ago, and the tools of a decade hence will likely be qualitatively different again. What will not change is the need to ask, of any new instrument, the questions this dissertation has tried to ask: What does it actually measure? Under what conditions does it behave as theory predicts? And when it is applied somewhere new, how much of its performance travels with it?

These are not obstacles to the use of computational tools in economics. They are the conditions under which those tools become genuinely scientific.

REFERENCES

Mohammed Abdellaoui, Peter Klibanoff, and Lætitia Placido. Experiments on compound risk in relation to simple risk and to ambiguity. *Management Science*, 61(6):1306–1322, 2015.

Johannes Abeler and Simon Jäger. Complex tax incentives. *American Economic Journal: Economic Policy*, 7(3):1–28, August 2015. doi: 10.1257/pol.20130137. URL https://www.aeaweb.org/articles?id=10.1257/pol.20130137.

Christopher Adjaho and Timothy Christensen. Externally valid treatment choice. *arXiv preprint arXiv:2205.05561*, 1, 2022.

Alonso Alfaro-Urena, Benjamin Faber, Cecile Gaubert, Isabela Manelici, and Jose P. Vasquez. Responsible sourcing? theory and evidence from costa rica. 2023. Working Paper.

Hunt Allcott. Site selection bias in program evaluation. *Quarterly Journal of Economics*, 130(3): 1117–1165, 2015.

Vital Anderhub, Werner Güth, Gneezy, and Sonsino. On the interaction of risk and time preferences: An experimental study. *German Economic Review*, 2(3):239–253, 2001.

Isaiah Andrews and Emily Oster. A simple approximation for evaluating external validity bias. *Economics Letters*, 178:58–62, 2019. ISSN 0165-1765. doi: https://doi.org/10.1016/j.econlet.2019.02.020. URL https://www.sciencedirect.com/science/article/pii/S0165176519300655.

Joshua Angrist, Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Feng Lu. Economic research evolves: Fields and styles. *American Economic Review*, 107(5):293–297, 2017.

Peter M Aronow and Donald KK Lee. Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika*, 100(1):235–240, 2013.

Sam Arts, Nicola Melluso, and Reinhilde Veugelers. Beyond citations: Measuring novel scientific ideas and their impact in publication text. *Review of Economics and Statistics*, pages 1–33, 2025.

Elliott Ash and Stephen Hansen. Text algorithms in economics. *Annual Review of Economics*, 15: 659–688, 2023.

Susan Athey. Beyond prediction: Using big data for policy problems. *Science*, 355:483–485, 2017.

Pierre Azoulay, Joshua S Graff Zivin, and Gustavo Manso. Incentives and creativity: evidence from the academic life sciences. *The RAND Journal of Economics*, 42(3):527–554, 2011.

Cuimin Ba, Aislinn Bohren, and Alex Imas. Over- and underreaction to information. Working

Paper, 2023.

Scott R Baker, Nicholas Bloom, and Steven J Davis. Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4):1593–1636, 2016.

Abhijit Banerjee, Arun G. Chandrasekhar, Suresh Dalpath, Esther Duflo, John Floretta, Matthew O. Jackson, Harini Kannan, Francine Loza, Anirudh Sankar, Anna Schrimpf, and Maheshwor Shrestha. Selecting the most effective nudge: Evidence from a large-scale experiment on immunization. *Econometrica*, 2024. Forthcoming.

Abhijit V. Banerjee, Sylvain Chassang, and Erik Snowberg. Decision theoretic approaches to experiment design and external validity. In Abhijit V. Banerjee and Esther Duflo, editors, *Handbook of Field Experiments*, volume 1, chapter 4, pages 141–174. North-Holland, Amsterdam, 2017. doi: 10.1016/bs.hefe.2016.08.005.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision*, September 2018.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: Pretrained language model for scientific text. In *EMNLP*, 2019.

Vidmantas Bentkus. On hoeffding's inequalities. *The Annals of Probability*, 32(2):1650–1673, 2004.

Daniel E Berlyne. Conflict, arousal, and curiosity. 1960.

B Douglas Bernheim and Charles Sprenger. On the empirical validity of cumulative prospect theory: Experimental evidence of rank-independent probability weighting. *Econometrica*, 88 (4):1363–1409, 2020.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Ranoua Bouchouicha and Ferdinand M Vieider. Accommodating stake effects under prospect theory. *Journal of Risk and Uncertainty*, 55(1):1–28, 2017.

Adrian Bruhin, Helga Fehr-Duda, and Thomas Epper. Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica*, 78(4):1375–1412, 2010.

Federico Bugni, Ivan Canay, Azeem Shaikh, and Max Tabord-Meehan. Inference for cluster randomized experiments with non-ignorable cluster sizes, 2023. Working Paper.

Colin F Camerer, Gideon Nave, and Alec Smith. Dynamic unstructured bargaining with private information: theory, experiment, and outcome prediction via machine learning. *Management*

Science, 65(4):1867–1890, 2019.

Andrew Caplin, Mark Dean, and Daniel Martin. Search and satisficing. *American Economic Review*, 101(7):2899–2922, December 2011. doi: 10.1257/aer.101.7.2899. URL https://www.aeaweb.org/articles?id=10.1257/aer.101.7.2899.

David Card and Alan B. Krueger. Time-series minimum-wage studies: A meta-analysis. *The American Economic Review*, 85(2):238–243, 1995. ISSN 00028282. URL http://www.jstor.org/stable/2117925.

Matias D Cattaneo, Luke Keele, Rocío Titiunik, and Gonzalo Vazquez-Bare. Extrapolating treatment effects in multi-cutoff regression discontinuity designs. *Journal of the American Statistical Association*, 116(536):1941–1952, 2021.

Abraham Charnes and William W Cooper. Programming with linear fractional functionals. *Naval Research logistics quarterly*, 9(3-4):181–186, 1962.

Sylvain Chassang and Sam Kapon. Designing randomized controlled trials with external validity in mind. Working Paper, 2022.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68, 2018.

Denis Chetverikov, Zhipeng Liao, and Victor Chernozhukov. On cross-validated lasso in high dimensions. *The Annals of Statistics*, 49(3):1300–1317, 2021.

P.-A. Chiappori, S. Levitt, and T. Groseclose. Testing mixed-strategy equilibria when players are heterogeneous: The case of penalty kicks in soccer. *American Economic Review*, 92(4):1138–1151, September 2002. doi: 10.1257/00028280260344678. URL https://www.aeaweb.org/articles?id=10.1257/0002828026034467 8.

Carl von Clausewitz. *Vom Kriege*. Dümmlers Verlag, Berlin, 1832. Original German edition.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*, 2020.

Peter V. Coveney, Edward R. Dougherty, and Roger R. Highfield. Big data need big theory too. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2080):20160153, 2016.

Vincent P Crawford, Miguel A Costa-Gomes, and Nagore Iriberri. Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51(1):5–62, 2013.

Mark Dean and Pietro Ortoleva. The empirical relationship between nonstandard economic behaviors. *Proceedings of the National Academy of Sciences*, 116(33):16262–16267, 2019.

Rajeev Dehejia, Cristian Pop-Eleches, and Cyrus Samii. From local to global: External validity in a fertility natural experiment. *Journal of Business & Economic Statistics*, 39(1):217–243, 2021. doi: 10.1080/07350015.2019.1639407. URL https://doi.org/10.1080/07350015.2019.1639407.

Stefano DellaVigna and Devin Pope. Stability of experimental results: Forecasts and evidence. Working Paper, 2019.

Fabio Dennstädt, Johannes Zink, Paul Martin Putora, Janna Hastings, and Nikola Cihoric. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. *Systematic reviews*, 13(1):158, 2024.

Yixi Ding, Jiaying Wu, Tongyao Zhu, Yanxia Qin, Qian Liu, and Min-Yen Kan. Ccsbench: Evaluating compositional controllability in llms for scientific document summarization. *arXiv preprint arXiv:2410.12601*, 2024.

Roman Egger and Joanne Yu. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7:886498, 2022.

Liran Einav, Amy Finkelstein, Iuliana Pascu, and Mark R. Cullen. How general are risk preferences? choices under uncertainty in different domains. *American Economic Review*, 102(6): 2606–38, May 2012. doi: 10.1257/aer.102.6.2606. URL https://www.aeaweb.org/articles?id=10.1257/aer.102.6.2606.

Keaton Ellis, Shachar Kariv, and Erkut Y. Ozbay. Predicting and understanding individual-level choice under risk. https://eml.berkeley.edu/~kariv/EKO_I.pdf, 2024a. Working Paper.

Keaton Ellis, Shachar Kariv, and Erkut Y. Ozbay. Predicting and understanding individual-level choice under uncertainty. https://eml.berkeley.edu/~kariv/EKO_II.pdf, 2024b. Working paper. Version dated 30 September 2024.

Matthew Embrey, Guillaume R. Fréchette, and Sevgi Yuksel. Cooperation in the finitely repeated prisoner's dilemma. *The Quarterly Journal of Economics*, 133(1):509–551, 2018. doi: 10.1093/qje/qjx033.

Nathalie Etchart-Vincent and Olivier l'Haridon. Monetary incentives in the loss domain and behavior toward risk: An experimental comparison of three reward schemes including real losses. *Journal of risk and uncertainty*, 42(1):61–83, 2011.

Christian Ewerhart. Chess-like games are dominance solvable in at most two steps. *Games and Economic Behavior*, 33, 2000. doi: doi:10.1006/game.1999.0763.

Yuyu Fan, David V Budescu, and Enrico Diecidue. Decisions with compound lotteries. *Decision*, 6(2):109, 2019.

Helga Fehr-Duda, Adrian Bruhin, Thomas Epper, and Renate Schubert. Rationality on the rise: Why relative risk aversion increases with stake size. *Journal of Risk and Uncertainty*, 40(2): 147–180, 2010.

Dana Foarta and Massimo Morelli. The common determinants of legislative and regulatory complexity. Working Paper, 2023.

Nicole M. Fortin, Andrew Hill, and Jeff Huang. Superstition in the housing market. *Economic Inquiry*, 52(3):974–993, 2014. doi: https://doi.org/10.1111/ecin.12066. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ecin.12066.

Jacob G Foster, Andrey Rzhetsky, and James A Evans. Tradition and innovation in scientists' research strategies. *American sociological review*, 80(5):875–908, 2015.

Jacob G Foster, Feng Shi, and James Evans. Surprise! measuring novelty as expectation violation. In *2019 Meeting American Sociological Association, New York, NY. OSF preprint: http://osf.io/preprints/socarxiv/2t46f*, 2021.

Drew Fudenberg and Gustav Karreskog Rehbinder. Predicting cooperation with learning models. *American Economic Journal: Microeconomics*, 16(1):1–32, 2024.

Drew Fudenberg and Annie Liang. Predicting and understanding initial play. *American Economic Review*, 109(12):4112–4141, 2019.

Drew Fudenberg, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. Measuring the completeness of economic models. *Journal of Political Economy*, 130(4):956–990, 2022a.

Drew Fudenberg, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. Measuring the completeness of economic models. *Journal of Political Economy*, 130(4):956–990, 2022b.

Drew Fudenberg, Wayne Gao, and Annie Liang. How flexible is that functional form? quantifying the restrictiveness of theories. *Review of Economics and Statistics,* forthcoming, 2023.

Russell J Funk and Jason Owen-Smith. A dynamic network measure of technological change. *Management science*, 63(3):791–817, 2017.

Xavier Gabaix and David Laibson. Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets*. *The Quarterly Journal of Economics*, 121(2):505–540, 05 2006. ISSN 0033-5533. doi: 10.1162/qjec.2006.121.2.505. URL https://doi.org/10.1162/qjec.2006.121.2.505.

Michael Gechter, Cyrus Samii, Rajeev Dehejia, and Cristian Pop-Eleches. Evaluating ex ante counterfactual predictions using ex post causal inference, 2019. Working Paper.

Matthew Gentzkow and Jesse M Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.

William M Goldstein and Hillel J Einhorn. Expression theory and the preference reversal phenomena. *Psychological review*, 94(2):236–254, 1987.

Jeremy Greenwood, Zvi Hercowitz, and Per Krusell. Long-run implications of investment-specific technological change. *The American Economic Review*, 87(3):342–362, 1997. ISSN 00028282. URL http://www.jstor.org/stable/2951349.

Rahul Kumar Gupta, Ritu Agarwalla, Bukya Hemanth Naik, Joythish Reddy Evuri, Apil Thapa, and Thoudam Doren Singh. Prediction of research trends using lda based topic modeling. *Global Transitions Proceedings*, 3(1):298–304, 2022.

Yoram Halevy. Ellsberg revisited: An experimental study. *Econometrica*, 75(2):503–536, 2007.

Binglan Han, Teo Susnjak, and Anuradha Mathrani. Automating systematic literature reviews with retrieval-augmented generation: a comprehensive overview. *Applied Sciences*, 14(19): 9103, 2024.

David W. Harless and Colin F. Camerer. The predictive utility of generalized expected utility theories. *Econometrica*, 62(6):1251–1289, 1994. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951749.

Jason S Hartford, James R Wright, and Kevin Leyton-Brown. Deep learning for predicting human strategic behavior. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016a. URL https://proceedings.neurips.cc/paper/2016/file/7eb3c8be3d411e8ebfab08eba5f49632-Paper.pdf.

Jason S Hartford, James R Wright, and Kevin Leyton-Brown. Deep learning for predicting human strategic behavior. *Advances in Neural Information Processing Systems*, pages 2424–2432, 2016b.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

John D. Hey and Chris Orme. Investigating generalizations of expected utility theory using experimental data. *Econometrica*, 62(6):1291–1326, 1994. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951750.

Toshihiko Hirasawa, Michihiro Kandori, and Akira Matsushita. Using big data and machine learning to uncover how players choose mixed strategies, 2022. Working Paper.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Jake M. Hofman, Duncan J. Watts, Susan Athey, Filiz Garip, Thomas L. Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J. Salganik, Simine Vazire, Allesandro Vespignani, and Tal Yarkoni. Integrating explanation and prediction in computational social science. *Nature*, 595:181–188, 2021.

Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, 117(17):9284–9291, 2020.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

V. Joseph Hotz, Guido W. Imbens, and Julie H. Mortimer. Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, 125(1-2):241–270, 2005.

Greg Howard. A check for rational inattention. *Journal of Political Economy Microeconomics*, 2023. doi: 10.1086/727556. URL `https://doi.org/10.1086/727556`.

Sung-Lin Hsieh, Shaowei Ke, Chen Zhao, and Zhaoran Wang. A logit neural-network utility model. Working Paper, 2023.

Taisuke Imai, Tom A Rutter, and Colin F Camerer. Meta-Analysis of Present-Bias Estimation using Convex Time Budgets. *The Economic Journal*, 131(636):1788–1814, 09 2020. ISSN 0013-0133. doi: 10.1093/ej/ueaa115. URL `https://doi.org/10.1093/ej/ueaa115`.

Uday Karmarkar. Subjectively weighted utility: A descriptive extension of the expected utility model. *Organizational Behavior & Human Performance*, 21(1):67–72, 1978.

Bryan Kelly, Asaf Manela, and Alan Moreira. Text selection. *Journal of Business & Economic Statistics*, 39(4):859–879, 2021.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–495, 2015.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.

Phillip Külpmann and Christoph Kuzmics. Comparing theories of one-shot play out of treatment. 2022.

Pamela K Lattimore, Joanna R Baker, and A Dryden Witte. The influence of probability on risky choice: A parametric examination. *Journal of Economic Behavior & Organization*, 17(3):315–436, 1992.

Erin Leahey, Jina Lee, and Russell J Funk. What types of novelty are most disruptive? *American Sociological Review*, 88(3):562–597, 2023.

You-Na Lee, John P Walsh, and Jian Wang. Creativity in scientific teams: Unpacking novelty and impact. *Research policy*, 44(3):684–697, 2015.

Mathieu Lefebvre, Ferdinand M Vieider, and Marie Claire Villeval. Incentive effects on risk attitude in small probability prospects. *Economics Letters*, 109(2):115–120, 2010.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5): 911–938, 2021.

Lihua Lei, Roshni Sahoo, and Stefan Wager. Policy learning under biased sample selection. *arXiv preprint arXiv:2304.11735*, 2023.

Christian Leibel and Lutz Bornmann. What do we know about the disruption index in scientometrics? an overview of the literature. *Scientometrics*, 129(1):601–639, 2024. doi: 10.1007/s11192-023-04873-5. URL https://doi.org/10.1007/s11192-023-04873-5.

Olivier l'Haridon and Ferdinand M Vieider. All over the map: A worldwide comparison of risk preferences. *Quantitative Economics*, 10(1):185–215, 2019.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Zihang Lin, Yian Yin, Lu Liu, and Dashun Wang. Sciscinet: A large-scale open data lake for the science of science research. *Scientific Data*, 10(1):315, 2023.

Jens Ludwig and Sendhil Mullainathan. Machine learning as a tool for hypothesis generation. 2023. Working Paper.

Jens Ludwig and Sendhil Mullainathan. Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827, 2024.

Zhuoran Luo, Wei Lu, Jiangen He, and Yuqi Wang. Combination of research questions and methods: A new measurement of scientific novelty. *Journal of Informetrics*, 16(2):101282, 2022.

Kensuke Maeba. Extrapolation of treatment effect estimates across contexts and policies: An application to cash transfer experiments, 2022. URL https://bpb-us-e1.wpmucdn.com/sites.northwestern.edu/dist/f/5630/files/2022/12/Extrapolation_draft.pdf. Working Paper.

Charles F. Manski. Econometrics for decision making: Building foundations sketched by haavelmo and wald. *Econometrica*, 89(6):2827–2853, 2021. doi: https://doi.org/10.3982/ECTA17985. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA17985.

Andreas Maurer. Concentration inequalities for functions of independent variables. *Random Structures & Algorithms*, 29(2):121–138, 2006.

Daniel McFadden. The measurement of urban travel demand. *Journal of Public Economics*, 3(4): 303–328, 1974.

Reid McIlroy-Young, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. Aligning superhuman ai with human behavior: Chess as a model system. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1677–1687, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403219. URL https://doi.org/10.1145/3394486.3403219.

Alisdair McKay, Emi Nakamura, and Jón Steinsson. The power of forward guidance revisited. *American Economic Review*, 106(10):3133–58, October 2016. doi: 10.1257/aer.20150063. URL https://www.aeaweb.org/articles?id=10.1257/aer.20150063.

Rachael Meager. Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91, 2019.

Rachael Meager. Aggregating distributional treatment effects: A bayesian hierarchical analysis of the microcredit literature. *American Economic Review*, 112(6):1818–47, June 2022. doi: 10.1257/aer.20181811. URL https://www.aeaweb.org/articles?id=10.1257/aer.20181811.

Konrad Menzel. Transfer estimates for causal effects across heterogeneous sites. *arXiv preprint arXiv:2305.01435*, 2023.

Shubhanshu Mishra and Vetle I Torvik. Quantifying conceptual novelty in the biomedical literature. In *D-Lib magazine: the magazine of the Digital Library Forum*, volume 22, pages 10–1045, 2016.

Magne Mogstad and Alexander Torgovitsky. Identification and extrapolation of causal effects with instrumental variables. *Annual Review of Economics*, 10:577–613, 2018.

Sendhil Mullainathan and Ziad Obermeyer. Does machine learning automate moral hazard and error? *American Economic Review*, 107(5):476–480, 2017.

Zahra Murad, Martin Sefton, and Chris Starmer. How do risk attitudes affect measured confidence? *Journal of Risk and Uncertainty*, 52(1):21–46, 2016.

Kevin Murphy. *Machine Learning: a Probabilistic Perspective*. MIT Press, 2012.

Paulo Natenzon. Random choice and learning. *Journal of Political Economy*, 127(1):419–457, 2019. doi: 10.1086/700762. URL https://doi.org/10.1086/700762.

OECD. Purchasing power parities, 2023. URL https://data.oecd.org/conversion/purchasing-power-parities-ppp.htm. Accessed on 10 November, 2022.

Ryan Oprea. What makes a rule complex? *American Economic Review*, 110(12):3913–51, December 2020. doi: 10.1257/aer.20191717. URL https://www.aeaweb.org/articles?id=10.1257/aer.20191717.

Florian Oswald. The effect of homeownership on the option value of regional migration. *Quantitative Economics*, 10(4):1453–1493, 2019. doi: https://doi.org/10.3982/QE872. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/QE872.

Tabea MG Pakull, Hendrik Damm, Ahmad Idrissi-Yaghir, Henning Schäfer, Peter A Horn, and Christoph M Friedrich. Wispermed at biolaysumm: Adapting autoregressive large language models for lay summarization of scientific articles. *arXiv preprint arXiv:2405.11950*, 2024.

Ignacio Palacios-Huerta. Professionals Play Minimax. *The Review of Economic Studies*, 70(2):395–415, 04 2003. ISSN 0034-6527. doi: 10.1111/1467-937X.00249. URL https://doi.org/10.1111/1467-937X.00249.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

Bruce Pandolfini. *Pandolfini's Chess Complete: The Most Comprehensive Guide to the Game, from History to Strategy*. Touchstone, 1992.

Parag Pathak and Peng Shi. Simulating alternative school choice options in boston - main report. 2013. Working Paper.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, ..., and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Joshua C. Peterson, David D. Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L. Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214, 2021. doi: 10.1126/science.abe2629. URL https://www.science.org/doi/abs/10.1126/science.abe2629.

Alexander Peysakhovich and Jeffrey Naecker. Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior & Organization*, 133:373–384, 2017. ISSN 0167-2681. doi: https://doi.org/10.1016/j.jebo.2016.08.017. URL https://www.sciencedirect.com/science/article/pii/S0167268116301846.

Ori Plonsky, Reut Apel, Eyal Ert, Moshe Tennenholtz, David Bourgin, Joshua Peterson, and ...and I. Erev. Predicting human decisions with behavioral theories and machine learning. *CoRR*, abs/1904.06866, 2019. URL `http://arxiv.org/abs/1904.06866`.

Hina Raja, Asim Munawar, Mohammad Delsoz, Mohammad Elahi, Yeganeh Madadi, Amr Hassan, Hashem Abu Serhan, Onur Inam, Luis Hermandez, Sang Tran, et al. Using large language models to automate category and trend analysis of scientific articles: an application in ophthalmology. *arXiv preprint arXiv:2308.16688*, 2023.

David M Ritzwoller, Joseph P Romano, and Azeem M Shaikh. Randomization inference: Theory and applications. *arXiv preprint arXiv:2406.09521*, 2024.

Paul R Rosenbaum. Sensitivity analysis in observational studies. *Encyclopedia of statistics in behavioral science*, 2005.

Evan Russek, Daniel Acosta-Kane, Bas van Opheusden, Marcelo G Mattar, and Tom Griffiths. Time spent thinking in online chess reflects the value of computation, Oct 2022. URL `osf.io/preprints/psyarxiv/8j9zx`.

Roshni Sahoo, Lihua Lei, and Stefan Wager. Learning from a biased sample. *arXiv preprint arXiv:2209.01754*, 2022.

Yuval Salant and Jorg Spenkuch. Complexity and satisficing: Theory with evidence from chess. Working Paper, 2023.

Suproteem K Sarkar and Keyon Vafa. Lookahead bias in pretrained language models. *Available at SSRN 4754678*, 2024.

Claude E Shannon. Xxii. programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 41(314):256–275, 1950.

Sotaro Shibayama, Deyun Yin, and Kuniko Matsumoto. Measuring novelty in science with word embedding. *PloS one*, 16(7):e0254034, 2021.

Matthias Sutter, Martin G Kocher, Daniela Glätzle-Rützler, and Stefan T Trautmann. Impatience and uncertainty: Experimental decisions predict adolescents' field behavior. *American Economic Review*, 103(1):510–31, 2013.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Elizabeth Tipton and Robert B Olsen. A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, 47(8):516–524, 2018.

Denis Trapido. How novelty in knowledge earns recognition: The role of consistent identities. *Research Policy*, 44(8):1488–1500, 2015.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

Adilson Vital Jr, Filipi N Silva, Osvaldo N Oliveira Jr, and Diego R Amancio. Predicting citation impact of research papers using gpt and other text embeddings. *Physica A: Statistical Mechanics and its Applications*, page 130789, 2025.

Eva Vivalt. How Much Can We Generalize From Impact Evaluations? *Journal of the European Economic Association*, 18(6):3045–3089, 09 2020. ISSN 1542-4766. doi: 10.1093/jeea/jvaa019. URL https://doi.org/10.1093/jeea/jvaa019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Mark Walker and John Wooders. Minimax play at wimbledon. *The American Economic Review*, 91(5):1521–1538, 2001. ISSN 00028282. URL http://www.jstor.org/stable/2677937.

Zhiping Paul Wang, Priyanka Bhandary, Yizhou Wang, and Jason H Moore. Using gpt-4 to write a scientific review article: a pilot evaluation study. *BioData Mining*, 17(1):16, 2024.

Lingfei Wu, Dashun Wang, and James A Evans. Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382, 2019.

Yan Yan, Shanwu Tian, and Jingjing Zhang. The impact of a paper's new combinations and new components on its citation. *Scientometrics*, 122(2):895–913, 2020.

Yi Zhao and Chengzhi Zhang. A review on the novelty measurements of academic papers. *Scientometrics*, 130(2):727–753, 2025.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. Working Paper, 2021.

COMPUTATIONAL METHODS IN ECONOMICS

Approved by:

Annie Liang
Economics and Computer Science
Northwestern University

Benjamin Golub
Economics and Computer Science
Northwestern University

Jessica Hullman
Computer Science
Northwestern University

Harry Pei
Economics
Northwestern University

Date Approved: May 8, 2026

APPENDIX A
APPENDIX FOR CHAPTER 2

A.1 Technical Details of Semantic Isolation Metrics

This appendix provides the formal definitions for the semantic isolation metrics discussed in Section 2.2.

A.1.1 Fundamental Definitions

Let S be the corpus of all papers. For each paper $i \in S$, we have a vector embedding e_i and a publishing year t_i. Let $S(t) = \{j \in S \mid t_j \leq t\}$ be the set of papers published in or before year t.

The core of our measurement is the set of pairwise distances between a focal paper i and and a specified comparison set of papers. For a focal paper i and a temporal offset c (in years), we define the comparison set as $S(t_i + c)$. The set of distances between paper i and all other papers $j \in S(t_i + c)$ (where $j \neq i$) is given by:

$$D_{\{i,c\}} = \{d(e_i, e_j) \mid j \in S(t_i + c),\ j \neq i\}$$

where $d(\cdot)$ denotes a measure of distance (e.g., Euclidean distance). The offset c defines the temporal window of the comparison:

- **Contemporaneous ($c = 0$):** $D_{\{i,0\}}$ is the set of distances between paper i and all other papers published up to and including its own publication year t_i.
- **Prospective ($c > 0$):** $D_{\{i,3\}}$ is the set of distances between paper i and all papers published

up to three years after its publication.

- **Retrospective ($c < 0$):** $D_{\{i,-3\}}$ is the set of distances between paper i and all papers that were published at least three years prior to its own publication.

A.1.2 Point-in-Time Isolation Metrics

The following four metrics quantify a paper's isolation at a specific moment in time, defined by the offset c. These metrics form the basis of our spatial novelty score. For clarity, a higher value for distance-based metrics indicates greater isolation, while a lower value for density-based metrics indicates greater isolation.

Metric 1: Neighborhood Density. The number of papers within a radius defined by a fraction γ of the mean distance from paper i:

$$N(i,c) = \left|\{d \in D_{\{i,c\}} \mid d \leq \gamma \cdot Mean\left(D_{\{i,c\}}\right)\}\right|$$

Metric 2: $k-$Nearest Neighbors ($k-$NN) Distance. The distance to the $k-$th closest paper in the comparison set:

$$d^k_{\{i,c\}} = k\text{-th smallest element of } D_{\{i,c\}}$$

A larger value of $d^k_{\{i,c\}}$ indicates that a paper's closest intellectual peers are semantically distant, implying higher isolation.

Metric 3: Average $k-$NN Distance. The average distance to the k nearest neighbors:

$$\bar{d}^k_{\{i,c\}} = \frac{1}{k}\sum_{j=1}^{k}\left(j\text{-th smallest element of } D_{\{i,c\}}\right)$$

This provides a more robust measure of local isolation than the single k-NN distance. A higher value indicates greater isolation.

Metric 4: Kernel Density. A Gaussian kernel density estimate of the local semantic environment:

$$density(i, c, k) = \sum_{d \in D_{\{i,c\}}} e^{-d^2/2\sigma^2}$$

The bandwidth σ is adapted for each paper by setting it to that paper's k-NN distance, $\sigma = d^k_{\{i,c\}}$. This provides a smooth, weighted measure of local density.

A.1.3 Dynamic Isolation Metrics

Dynamic metrics capture temporal novelty by measuring the change in a paper's semantic environment over time. For any point-in-time metric M defined above, we define its change between two time points, specified by offsets c_1 and c_2 where $c_1 < c_2$, as:

$$\Delta M(i, c_1, c_2) = M(i, c_1) - M(i, c_2)$$

The interpretation of the sign of ΔM depends on whether M is a distance or density metric. We primarily use two types of dynamic metrics:

- **Retrospective Change:** The change in isolation in the period leading up to a paper's publication. For example, using the k-NN distance metric, we can calculate $\Delta d^k(i, -3, 0) = d^k_{\{i,-3\}} - d^k_{\{i,0\}}$. A large positive value for this measure indicates that the k-NN distance decreased significantly over the three years prior to publication, implying the neighborhood became rapidly denser. This signals that the paper is entering a "hot" or consolidating research frontier.

- **Prospective Change:** The change in isolation in the period following the focal paper's publication, such as $\Delta d^k(i, 0, 3) = d^k_{\{i,0\}} - d^k_{\{i,3\}}$. While not used for prediction (as it uses future information), this serves as a validation measure. A large positive value indicates that a paper's neighborhood became much denser after it was published, suggesting it was a precursor to a new line of inquiry.

A.1.4 Baseline Semantic Indicators

To isolate paper-specific novelty from corpus-wide trends (e.g., the overall densification of the semantic space over time), we construct a set of baseline indicators. These serve as control variables for semantic isolation metrics.

- **Paper-level Mean Distance:** For each paper i, its mean distance to all prior work: $Mean(D_{\{i,0\}})$.
- **Corpus-level Annual Mean Distance:** For each year t, the average distance between all pairs of papers published in or before that year. The measure is $Mean_{i,j \in S(t), i \neq j} \left(d(e_i, e_j)\right)$.
- **Temporal Change in Corpus-level Distance:** The year-over-year change in the corpus-level annual mean distance. This controls for secular trends in the semantic space.

A.2 Database Construction and Feature Implementation

This appendix provides a detailed description of the data collection, processing, and feature engineering pipeline used in our analysis.

A.2.1 Dataset Construction

Data Sources and Scope. Our dataset is constructed from three primary sources: RePEc (Research Papers in Economics) for metadata, Google Scholar for citation counts and PDF access, and

SciSciNet (Lin et al., 2023) for bibliometric indicators. The scope of our corpus is defined by the list of economics journals compiled by Angrist et al. (2017), which includes a broad set of general interest and top field journals.

Data Collection and Integration. The data construction process involved three main steps. First, we programmatically collected article metadata (title, author, year, journal, volume) from the RePEc website for all journals in our target list. Second, we queried each article title on Google Scholar to retrieve citation counts (as of December 2024) and links to publicly available PDF versions. Third, we downloaded the full-text PDFs and matched each article to its corresponding entry in the SciSciNet database using title, author, and publication year to retrieve pre-computed disruption and atypicality scores. We performed extensive cleaning to retain only main research articles and ensure accurate matching across sources.

Sample Definition. As described in Section 2.3, the final dataset is partitioned to facilitate a causally-grounded analysis.

- **Knowledge Base (1980–1995):** This set of papers provides the historical semantic context against which the novelty of later papers is measured.
- **Focal Period (1996–2015):** This is the primary analysis sample, for which we measure novelty and subsequent impact.

After filtering and matching procedures, the final analysis dataset contains 52,766 papers from the focal period with complete metadata, text, and impact data. A list of the journals included is provided in Table A.1.

Journal Name	Journal Name
American Economic Review	American Economist
American Journal of Agricultural Economics	Annals of Economic and Social Measurement
Applied Economics	Bell Journal of Economics
Bell Journal of Economics and Management Science	Brookings Papers on Economic Activity
Canadian Journal of Economics	Carnegie-Rochester Conference Series on Public Policy
Econometric Theory	Econometrica
Economic Development and Cultural Change	Economic Inquiry
Economic Record	Economic Theory
Economica	Economics Letters
European Economic Review	Explorations in Economic History
Federal Reserve Bank of St. Louis Review	Games and Economic Behavior
Industrial and Labor Relations Review	International Economic Review
International Journal of Game Theory	International Journal of Industrial Organization
International Labour Review	Journal of Business & Economic Statistics
Journal of Econometrics	Journal of Economic Behavior & Organization
Journal of Economic Dynamics and Control	Journal of Economic History
Journal of Economic Issues	Journal of Economic Literature
Journal of Economic Perspectives	Journal of Economic Theory
Journal of Economics & Management Strategy	Journal of Environmental Economics and Management
Journal of Health Economics	Journal of Human Resources
Journal of Industrial Economics	Journal of Institutional and Theoretical Economics
Journal of International Economics	Journal of International Money and Finance
Journal of Labor Economics	Journal of Law and Economics
Journal of Law, Economics, & Organization	Journal of Legal Studies
Journal of Mathematical Economics	Journal of Monetary Economics
Journal of Money, Credit and Banking	Journal of Political Economy
Journal of Productivity Analysis	Journal of Public Economics
Journal of Regional Science	Journal of Regulatory Economics
Journal of Urban Economics	Journal of the European Economic Association
Kyklos	NBER Macroeconomics Annual
National Tax Journal	Oxford Bulletin of Economics and Statistics
Oxford Economic Papers (New Series)	Public Policy
Quarterly Journal of Economics	Quarterly Review of Economics and Business
Quarterly Review of Economics and Finance	RAND Journal of Economics
Review of Economic Dynamics	Review of Economic Studies
Review of Economics and Statistics	Review of Income and Wealth
Review of Radical Political Economics	Review of Social Economy
Social Choice and Welfare	Southern Economic Journal
The Economic Journal	The Energy Journal
Theory and Decision	World Development

Table A.1: *List of Economics Journals included in the Corpus. We use the list of journals from Angrist et al. (2017).*

A.2.2 Descriptive Statistics of Data

A.2.3 Semantic Content Extraction Pipeline

A central innovation of our study is the extraction of structured, multi-dimensional representations of each paper's intellectual contribution using Large Language Models (LLMs).

Text Processing. The pipeline begins with the raw PDF files. Each PDF is converted to plain text using an LLM-based Optical Character Recognition (OCR) model to handle complex layouts and mathematical notation. To manage computational costs while preserving essential content, we employ a two-step summarization process using Gemini Pro. First, the full text is condensed into an structured comprehensive summary (approx. 1500–2000 words). This summary is designed to

Field	Count	Mean cite (std)	Median cite	75-th cite	Disruptive rate
macroeconomics	5353	203.7 (540.85)	58.0	165.0	0.135
international	3951	251.8 (681.66)	75.0	224.0	0.171
financial	3953	207.8 (690.30)	59.0	171.0	0.112
economic history	1061	125.9 (234.46)	57.0	126.0	0.203
microeconomic theory	8079	109.4 (463.79)	37.0	93.0	0.089
industrial organization	3985	167.5 (378.25)	60.0	162.0	0.166
econometrics & quantitative	6331	237.6 (872.89)	66.0	175.0	0.128
public	3828	178.4 (398.10)	68.0	169.0	0.196
health	1824	164.4 (325.80)	78.0	175.0	0.230
development	2405	320.7 (715.11)	132.0	317.0	0.308
labor	4946	230.7 (489.91)	90.0	235.75	0.212
agricultural	928	151.1 (226.69)	74.5	181.0	0.347
urban & regional	933	246.8 (400.85)	119.0	259.0	0.193
environmental & resource	1520	175.9 (337.95)	75.0	174.25	0.251
behavioral	2543	298.9 (786.60)	95.0	257.5	0.091
other	1126	198.7 (579.87)	58.0	161.75	0.256
all	52766	201.1 (587.07)	65.0	176.0	0.163

Table A.2: *Descriptive Statistics by Economic Subfield. "Disruptive rate" is the fraction of papers with a disruption index greater than zero, following Wu et al. (2019).*

retain all core arguments, methods, and findings.

Multi-Dimensional Insight Extraction. Second, from this comprehensive summary, we prompt the LLM to extract five distinct aspects of the paper's contribution. This decomposition allows us to measure novelty along different intellectual dimensions. The five insight types are:

- **Paragraph**: A concise overview of the paper's primary contribution.
- **Topic**: The paper's subfield classification, research subject, and specific issues addressed
- **Question-Conclusion**: The core research questions posed and the principal findings or conclusions.

- **Methodology**: The data sources, empirical strategy, analytical models, and technical approaches used.
- **Theory**: The theoretical frameworks, conceptual models, and analytical innovations introduced.

The prompts for insights and comprehensive summary are in Appendix A.6.

A.2.4 Feature Engineering

The final stage of our pipeline involves generating the semantic isolation metrics from the structured text data.

Embedding and Distance Calculation. To ensure the robustness of our semantic representations, we generate embeddings for each of the five insight types using three distinct models: SciBERT, SPECTER, and Qwen. We then calculate pairwise distances between all papers in the corpus using both Euclidean distance and cosine similarity.

Systematic Feature Construction. The full set of 3,240 candidate features is generated by systematically combining choices along five axes:

1. **Insight Type (5 options):** Paragraph, Topic, Question-Conclusion, Methodology, Theory.
2. **Embedding Model (3 options):** SciBERT, SPECTER, Qwen.
3. **Distance Metric (2 options):** Euclidean, Cosine similarity.
4. **Isolation Metric Type:** Neighborhood Density, k-NN Distance, Average k-NN Distance, Kernel Density, and the Dynamic metrics.

5. **Hyperparameters:** The metrics are calculated across a range of parameters to capture isolation at different scales:

    - For Neighborhood Density: $\gamma \in \{0.5, 1.0, 1.5\}$.
    - For k-NN based metrics: $k \in \{3, 5, 10, 20, 30, 50\}$.
    - For Dynamic Change metrics: The change is computed over intervals defined by offsets $c_1, c_2 \in \{-5, -3, 0\}$ years relative to the publication date (e.g., $[-3, 0]$ for retrospective).

This systematic, combinatorial approach ensures a comprehensive measurement of a paper's semantic position and trajectory, forming the empirical foundation for our analysis.

A.3 More Results on Validating Semantic Isolation Metrics

This appendix provides supplementary tables and analyses that support the validation of our semantic isolation metrics presented in Section 2.4.

A.3.1 Full Results for Predictive Tasks

This section presents the complete set of results for the prediction tasks, comparing the performance of our Isolation metrics against all other feature groups.

Predicting Log Citation Counts. Table A.3 details the performance of each individual feature group in predicting log citation counts. Notably, the Isolation feature set demonstrates predictive power (measured by MSE reduction) that is highly competitive with standard textual representations like TF-IDF. Table A.4 shows that the dimensionally-reduced Isolation-PCA feature set (200 components) retains nearly all the predictive power of the full set, justifying its use in the

SHAP analysis in the main text and confirming that the signal is robust, not an artifact of high dimensionality.

feature set	MSE	comp
control	1.7529	-0.0285
	(0.00161)	
control+subfield	1.7044	0.0
	(0.00154)	
control+journal	1.3252	0.2225
	(0.00122)	
control+textual	1.4558	0.1459
	(0.00139)	
control+LDA	1.5698	0.0790
	(0.00149)	
control+TFIDF	1.3935	0.1824
	(0.00135)	
control+atypicality	1.7110	-0.0039
	(0.00156)	
control+isolation	1.5352	0.0993
	(0.00133)	
control+isolation-PCA	1.5728	0.0772
	(0.00137)	

Table A.3: *Individual feature groups predicting log citation counts.*

Predicting Disruptive Papers. The full results for the disruption prediction task are presented in Tables A.5 and A.6. As with citations, the Isolation metrics are a top-performing feature group, with precision and F1-scores comparable to TF-IDF. The Isolation-PCA features again provide a significant, additional improvement in predictive performance, reinforcing the findings from the main text.

feature set	MSE	comp
control+subfield	1.7044	0.0
	(0.00154)	
all features (excl. isolation)	1.1465	0.3273
	(0.00111)	
all features (incl. isolation)	1.1092	0.3492
	(0.00105)	
all features (incl. isolation-PCA)	1.1105	0.3485
	(0.00105)	

Table A.4: *Additional value of isolation-PCA in predicting log citations.*

A.3.2 Decomposition of Predictive Power by Insight Type

To understand which dimensions of a paper's content contribute most to our novelty measures, we decompose the Isolation feature set by the five LLM-generated insight types. Tables A.7 through A.10 present the results. For both citations and disruption, isolation metrics based on the Paragraph Summary and Question-Conclusion consistently yield the highest predictive performance. This reinforces the finding that novelty in a paper's central claims is a primary predictor of impact. Metrics based on Theory also perform strongly. While Methodology and Topic isolation show slightly weaker performance, they still provide significant predictive value, demonstrating that novelty along these dimensions is also an important, albeit secondary, signal of future influence.

Feature set	Precision	Comp-prec	F1	Comp-F1
control	0.2875	-0.0387	0.3885	-0.0455
	(0.00046)		(0.00049)	
control+subfield	0.3140	0.0	0.4151	0.0
	(0.00053)		(0.00057)	
control+journal	0.3398	0.0377	0.4384	0.0397
	(0.00055)		(0.00054)	
control+textual	0.3786	0.0942	0.4597	0.0762
	(0.00068)		(0.00062)	
control+LDA	0.3717	0.0841	0.4634	0.0825
	(0.00059)		(0.00056)	
control+TFIDF	0.4171	0.1503	0.4835	0.1169
	(0.00067)		(0.00052)	
control+atypical	0.3231	0.0133	0.4220	0.0118
	(0.00057)		(0.00058)	
control+isolation	0.4064	0.1347	0.4699	0.0937
	(0.00064)		(0.00059)	
control+isolation-PCA	0.4007	0.1263	0.4599	0.0764
	(0.00074)		(0.00063)	

Table A.5: *Individual feature groups predicting disruptive papers.*

Feature set	Precision	Comp-prec	F1	Comp-F1
control	0.2875	-0.0387	0.3885	-0.0455
	(0.00046)		(0.00049)	
control+subfield	0.3140	0.0	0.4151	0.0
	(0.00053)		(0.00057)	
all	0.4270	0.1647	0.4915	0.1305
	(0.0007)		(0.0006)	
all+isolation	0.4517	0.2007	0.5063	0.1559
	(0.00081)		(0.00067)	
all+isolation-PCA	0.4514	0.2003	0.5053	0.1541
	(0.00082)		(0.00066)	

Table A.6: *Additional value of isolation-PCA in predicting disruption.*

Feature set	MSE	Completeness
control+subfield	1.7044	0.0
	(0.00154)	
control+atypical	1.7110	-0.0039
	(0.00156)	
control+isolation	1.5352	0.0993
	(0.00133)	
control+isolation-paragraph	1.5914	0.0663
	(0.00145)	
control+isolation-methodology	1.6191	0.0500
	(0.00137)	
control+isolation-theory	1.6182	0.0506
	(0.00142)	
control+isolation-topic	1.6342	0.0412
	(0.00146)	
control+isolation-question-conclusion	1.6060	0.0577
	(0.00139)	

Table A.7: *Predicting log citations: performance of individual insight types.*

Feature set	MSE	Completeness
control+subfield	1.7044	0.0
	(0.00154)	
all features (excl. isolation)	1.1465	0.3273
	(0.00111)	
all features (incl. isolation)	1.1092	0.3492
	(0.00105)	
all features (incl. isolation-paragraph)	1.1065	0.3508
	(0.00109)	
all features (incl. isolation-methodology)	1.1359	0.3335
	(0.00112)	
all features (incl. isolation-theory)	1.1175	0.3444
	(0.00107)	
all features (incl. isolation-topic)	1.1313	0.3363
	(0.00105)	
all features (incl. isolation-ques-conc)	1.1118	0.3477
	(0.00106)	

Table A.8: *Predicting log citations: additional value of insight types*

feature	prec	comp-prec	f1	comp-f1
control	0.2875	-0.0387	0.3885	-0.0455
	(0.00046)		(0.00049)	
control+subfield	0.3140	0.0	0.4151	0.0
	(0.00053)		(0.00057)	
control+journal	0.3398	0.0377	0.4384	0.0397
	(0.00055)		(0.00054)	
control+textual	0.3786	0.0942	0.4597	0.0762
	(0.00068)		(0.00062)	
control+LDA	0.3717	0.0841	0.4634	0.0825
	(0.00059)		(0.00056)	
control+TFIDF	0.4171	0.1503	0.4835	0.1169
	(0.00067)		(0.00052)	
control+atypical	0.3231	0.0133	0.4220	0.0118
	(0.00057)		(0.00058)	
control+isolation	0.4064	0.1347	0.4699	0.0937
	(0.00064)		(0.00059)	
control+isolation-paragraph	0.3773	0.0923	0.4556	0.0691
	(0.00067)		(0.00063)	
control+isolation-methodology	0.3618	0.0697	0.4292	0.0241
	(0.00071)		(0.00065)	
control+isolation-theory	0.3610	0.0685	0.4360	0.0356
	(0.00059)		(0.00051)	
control+isolation-topic	0.3505	0.0533	0.4191	0.0068
	(0.00065)		(0.00055)	
control+isolation-question-conclusion	0.3662	0.0762	0.4416	0.0453
	(0.00058)		(0.00056)	

Table A.9: *Predicting disruption: performance of individual insight types*

Feature set	Precision	Comp-prec	F1	Comp-F1
control	0.2875	-0.0387	0.3885	-0.0455
	(0.00461)		(0.00486)	
control+subfield	0.3140	0.0	0.4151	0.0
	(0.0053)		(0.0057)	
all features (excl. isolation)	0.4270	0.1647	0.4915	0.1305
	(0.00701)		(0.00602)	
all features (incl. isolation)	0.4517	0.2007	0.5063	0.1559
	(0.00081)		(0.00067)	
all features (incl. isolation-paragraph)	0.4419	0.1864	0.5062	0.1557
	(0.00073)		(0.00062)	
all features (incl. isolation-method)	0.4338	0.1746	0.4932	0.1334
	(0.00072)		(0.0006)	
all features (incl. isolation-theory)	0.4380	0.1808	0.5003	0.1457
	(0.00073)		(0.00063)	
all features (incl. isolation-topic)	0.4357	0.1774	0.4961	0.1385
	(0.00075)		(0.00058)	
all features (incl. isolation-ques-conc)	0.4404	0.1843	0.5029	0.1500
	(0.00073)		(0.00061)	

Table A.10: *predicting disruption: additional value of insight types.*

A.3.3 Subgroup Analysis by Economic Subfield

To assess the generalizability of our findings, we conduct the prediction exercises separately for each of the 16 economic subfields in our sample. Tables A.11 and A.12 report the results for predicting log citations and disruption, respectively.

While the overall predictive power of isolation metrics holds across most fields, the relative importance of different insight types varies. Notably, isolation metrics derived from the Question-Conclusion summary are the strongest individual predictors in a majority of subfields for both citations and disruption. This suggests that novelty in a paper's core argument - the questions it asks and the answers it provides - is a particularly potent and generalizable predictor of scientific impact. We also observe that overall predictive accuracy tends to be lower in more established or theoretical fields (e.g., Microeconomic Theory, Macroeconomics), which also exhibit lower baseline rates of disruption.

field	control	+atypical	+iso	+method	+theory	+topic	+qu-co
macroeconomics	2.2014 (0.000)	2.1250 (0.035)	1.8342 (0.167)	1.9457 (0.116)	1.9157 (0.130)	1.9562 (0.111)	1.9141 (0.131)
	(0.00062)	(0.00065)	(0.00058)	(0.00055)	(0.00060)	(0.00057)	(0.00058)
international	2.1205 (0.000)	2.0368 (0.040)	1.7264 (0.186)	1.8271 (0.138)	1.8125 (0.145)	1.8404 (0.132)	1.7916 (0.155)
	(0.00065)	(0.00064)	(0.00055)	(0.00063)	(0.00059)	(0.00057)	(0.00058)
financial	2.1997 (0.000)	2.0934 (0.048)	1.8225 (0.171)	1.9334 (0.121)	1.8914 (0.140)	1.9251 (0.125)	1.8678 (0.151)
	(0.00066)	(0.00064)	(0.00056)	(0.00061)	(0.00057)	(0.00063)	(0.00054)
economic history	1.6489 (0.000)	1.5483 (0.061)	1.3266 (0.195)	1.3932 (0.155)	1.3978 (0.152)	1.3635 (0.173)	1.3631 (0.173)
	(0.00103)	(0.00099)	(0.00106)	(0.00093)	(0.00093)	(0.00099)	(0.00099)
microeconomics theory	1.7256 (0.000)	1.6635 (0.036)	1.4741 (0.146)	1.5519 (0.101)	1.5235 (0.117)	1.5330 (0.112)	1.5263 (0.116)
	(0.00040)	(0.00037)	(0.00036)	(0.00035)	(0.00038)	(0.00036)	(0.00035)
industrial organization	1.7983 (0.000)	1.7007 (0.054)	1.5167 (0.157)	1.5792 (0.122)	1.5727 (0.125)	1.6107 (0.104)	1.5620 (0.131)
	(0.00058)	(0.00050)	(0.00048)	(0.00051)	(0.00051)	(0.00051)	(0.00057)
econometrics & quantitative	2.1885 (0.000)	2.1258 (0.029)	1.7725 (0.190)	1.8670 (0.147)	1.8526 (0.153)	1.8881 (0.137)	1.8547 (0.153)
	(0.00059)	(0.00060)	(0.00049)	(0.00054)	(0.00049)	(0.00050)	(0.00049)
public	1.7914 (0.000)	1.7073 (0.047)	1.5004 (0.162)	1.5653 (0.126)	1.5647 (0.127)	1.5670 (0.125)	1.5614 (0.128)
	(0.00051)	(0.00060)	(0.00046)	(0.00051)	(0.00043)	(0.00049)	(0.00047)
health	1.6354 (0.000)	1.5511 (0.052)	1.3217 (0.192)	1.3811 (0.156)	1.3758 (0.159)	1.4088 (0.139)	1.3791 (0.157)
	(0.00081)	(0.00068)	(0.00064)	(0.00069)	(0.00068)	(0.00068)	(0.00061)
development	1.8143 (0.000)	1.7386 (0.042)	1.4770 (0.186)	1.5588 (0.141)	1.5665 (0.137)	1.5652 (0.137)	1.5172 (0.164)
	(0.00083)	(0.00073)	(0.00069)	(0.00066)	(0.00070)	(0.00069)	(0.00062)
labor	1.9207 (0.000)	1.8229 (0.051)	1.5170 (0.210)	1.6266 (0.153)	1.6060 (0.164)	1.6290 (0.152)	1.5770 (0.179)
	(0.00057)	(0.00049)	(0.00043)	(0.00045)	(0.00045)	(0.00046)	(0.00047)
agricultural	1.5922 (0.000)	1.4544 (0.087)	1.2468 (0.217)	1.3216 (0.170)	1.2831 (0.194)	1.3234 (0.169)	1.2595 (0.209)
	(0.00097)	(0.00086)	(0.00091)	(0.00091)	(0.00088)	(0.00091)	(0.00093)
urban & regional	1.8135 (0.000)	1.6648 (0.082)	1.4731 (0.188)	1.5187 (0.163)	1.5601 (0.140)	1.5722 (0.133)	1.5344 (0.154)
	(0.00112)	(0.00109)	(0.00098)	(0.00104)	(0.00105)	(0.00090)	(0.00097)
environmental & resource	1.7978 (0.000)	1.6711 (0.070)	1.4622 (0.187)	1.5374 (0.145)	1.5127 (0.159)	1.5355 (0.146)	1.4934 (0.169)
	(0.00098)	(0.00088)	(0.00097)	(0.00093)	(0.00093)	(0.00094)	(0.00093)
behavioral	2.1318 (0.000)	1.9431 (0.088)	1.6765 (0.214)	1.7759 (0.167)	1.7541 (0.177)	1.7583 (0.175)	1.7319 (0.188)
	(0.00076)	(0.00068)	(0.00065)	(0.00069)	(0.00070)	(0.00060)	(0.00066)
other	2.2479 (0.000)	2.0730 (0.078)	1.9137 (0.149)	2.0001 (0.110)	1.9539 (0.131)	1.9471 (0.134)	1.9072 (0.152)
	(0.00125)	(0.00103)	(0.00126)	(0.00125)	(0.00112)	(0.00118)	(0.00104)

Table A.11: *The bootstrapped MSE of different feature groups when predicting log citation count of papers in different subfields. The value in parenthesis below the MSE is standard error of the 100 time bootstrapping, and the value in parenthesis on the right of the MSE is the completeness, setting control group as worst case (same to the subfield baseline) and perfect prediction as best case.*

field	control	+atypical	+iso	+method	+theory	+topic	+qu-co
macroeconomics	0.2947 (-0.000)	0.3378 (0.061)	0.4840 (0.268)	0.4213 (0.179)	0.3934 (0.140)	0.4108 (0.165)	0.4705 (0.249)
	(0.00023)	(0.00028)	(0.00053)	(0.00046)	(0.00050)	(0.00064)	(0.00049)
international	0.3464 (-0.000)	0.4039 (0.088)	0.5302 (0.281)	0.4185 (0.110)	0.4520 (0.162)	0.4147 (0.105)	0.4728 (0.193)
	(0.00024)	(0.00025)	(0.00045)	(0.00038)	(0.00032)	(0.00039)	(0.00040)
financial	0.2446 (-0.000)	0.2874 (0.057)	0.3815 (0.181)	0.3561 (0.148)	0.3535 (0.144)	0.2927 (0.064)	0.3224 (0.103)
	(0.00033)	(0.00034)	(0.00081)	(0.00078)	(0.00066)	(0.00075)	(0.00073)
economic history	0.3672 (-0.000)	0.4097 (0.067)	0.5280 (0.254)	0.5060 (0.219)	0.4981 (0.207)	0.5049 (0.218)	0.5323 (0.261)
	(0.00051)	(0.00049)	(0.00065)	(0.00064)	(0.00067)	(0.00073)	(0.00067)
microeconomics theory	0.1616 (-0.000)	0.1863 (0.030)	0.3701 (0.249)	0.3181 (0.187)	0.3027 (0.168)	0.2944 (0.158)	0.3005 (0.166)
	(0.00017)	(0.00021)	(0.00060)	(0.00053)	(0.00046)	(0.00068)	(0.00051)
industrial organization	0.3036 (-0.000)	0.3566 (0.076)	0.5147 (0.303)	0.4252 (0.175)	0.4831 (0.258)	0.4404 (0.196)	0.4655 (0.233)
	(0.00024)	(0.00027)	(0.00049)	(0.00042)	(0.00047)	(0.00045)	(0.00040)
econometrics & quantitative	0.2535 (-0.000)	0.3042 (0.068)	0.4342 (0.242)	0.3552 (0.136)	0.3485 (0.127)	0.3180 (0.086)	0.3777 (0.166)
	(0.00017)	(0.00022)	(0.00050)	(0.00041)	(0.00044)	(0.00037)	(0.00041)
public	0.3531 (-0.000)	0.4332 (0.124)	0.5290 (0.272)	0.4821 (0.199)	0.4560 (0.159)	0.4508 (0.151)	0.4596 (0.165)
	(0.00022)	(0.00028)	(0.00035)	(0.00037)	(0.00036)	(0.00036)	(0.00036)
health	0.3415 (-0.000)	0.3711 (0.045)	0.5249 (0.279)	0.4406 (0.151)	0.4681 (0.192)	0.4233 (0.124)	0.4499 (0.165)
	(0.00033)	(0.00038)	(0.00045)	(0.00053)	(0.00052)	(0.00045)	(0.00038)
development	0.4689 (-0.000)	0.5594 (0.170)	0.6098 (0.265)	0.5783 (0.206)	0.5539 (0.160)	0.5704 (0.191)	0.5764 (0.202)
	(0.00025)	(0.00030)	(0.00030)	(0.00027)	(0.00029)	(0.00031)	(0.00027)
labor	0.3503 (-0.000)	0.4180 (0.104)	0.5145 (0.253)	0.4614 (0.171)	0.4570 (0.164)	0.4642 (0.175)	0.4746 (0.191)
	(0.00018)	(0.00020)	(0.00032)	(0.00030)	(0.00025)	(0.00029)	(0.00032)
agricultural	0.4809 (-0.000)	0.5426 (0.119)	0.6036 (0.236)	0.5717 (0.175)	0.5699 (0.171)	0.5737 (0.179)	0.5950 (0.220)
	(0.00042)	(0.00046)	(0.00045)	(0.00042)	(0.00047)	(0.00046)	(0.00050)
urban & regional	0.3392 (-0.000)	0.3531 (0.021)	0.4983 (0.241)	0.3997 (0.092)	0.4756 (0.206)	0.4446 (0.160)	0.4301 (0.137)
	(0.00056)	(0.00062)	(0.00092)	(0.00090)	(0.00075)	(0.00071)	(0.00083)
environmental & resource	0.4153 (-0.000)	0.4617 (0.079)	0.5594 (0.246)	0.5387 (0.211)	0.5066 (0.156)	0.4891 (0.126)	0.5161 (0.172)
	(0.00038)	(0.00043)	(0.00053)	(0.00053)	(0.00054)	(0.00049)	(0.00051)
behavioral	0.2504 (-0.000)	0.3453 (0.127)	0.3932 (0.190)	0.4118 (0.215)	0.3950 (0.193)	0.2874 (0.049)	0.3453 (0.127)
	(0.00052)	(0.00058)	(0.00122)	(0.00112)	(0.00124)	(0.00110)	(0.00096)
other	0.4122 (-0.000)	0.4810 (0.117)	0.5098 (0.166)	0.4774 (0.111)	0.4932 (0.138)	0.4632 (0.087)	0.4974 (0.145)
	(0.00039)	(0.00045)	(0.00058)	(0.00050)	(0.00053)	(0.00055)	(0.00051)

Table A.12: *The bootstrapped precision of different feature groups when predicting disruptive papers in different subfields. The value in parenthesis below the MSE is standard error of the 100 time bootstrapping, and the value in parenthesis on the right of the MSE is the completeness, setting control group as worst case (same to the subfield baseline) and perfect prediction as best case.*

A.3.4 Including Prospective Features

In this section we add perspective dynamic isolation metrics and check the predicting performance. Specifically, for dynamic metrics: The change is computed over intervals defined by offsets $c_1, c_2 \in \{-5, -3, 0, 3, 5\}$ years relative to the publication date (e.g., $[-3, 0]$ for retrospective, $[0, 3]$ for prospective). We call this new set of isolation features *isolation-wf*, meaning it is with future information.

Predicting Log Citation Counts . Table A.13 presents the standalone predictive power of the atypicality and isolation feature groups. When added to the control variables, the atypicality score offers no improvement over the baseline. In stark contrast, our isolation metrics achieve a completeness score of 17.9%, demonstrating substantial predictive power.

Feature set	MSE	Completeness
control+subfield	1.7044	0.0
	(0.00154)	
control+atypicality	1.7110	-0.0039
	(0.00156)	
control+isolation-wf	1.3999	0.1787
	(0.00133)	

Table A.13: *Predicting log citation counts: individual feature groups. Standard errors from 100 bootstrap iterations in parentheses. Completeness is the proportional reduction in MSE relative to the baseline.*

Table A.14 shows that adding atypicality to the baseline provides a modest 2.3% improvement in completeness. However, subsequently adding our isolation metrics increases completeness to 19.0% - an incremental gain of 16.7 percentage points. This strong additional effect provides compelling evidence that isolation and atypicality are complementary, measuring distinct aspects of novelty. Atypicality captures the novelty of a paper's inputs (unusual combinations of references), while our metrics capture the novelty of its output (the semantic content of its contribution).

Feature set	MSE	Completeness
control+subfield	1.7044	0.0
	(0.00154)	
control+subfield+atypicality	1.6646	0.0234
	(0.00154)	
control+subfield+atypicality+isolation-wf	1.3801	0.1903
	(0.00131)	
all features (excl. isolation)	1.1465	0.3273
	(0.00111)	
all features (incl. isolation)	1.0439	0.3875
	(0.00102)	

Table A.14: *Predicting log citation counts: additional value of isolation metrics.*

The bottom panel of Table A.14 shows that even when added to a model containing all other feature groups, our isolation metrics provide a statistically and economically significant improvement, increasing completeness by an additional 6.1 percentage points.

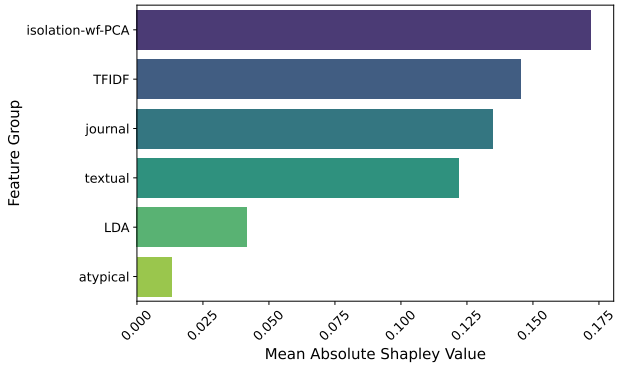


Figure A.1: *Feature importance (mean absolute SHAP) for predicting log citations.*

To assess the relative importance of each feature group, we compute SHAP (SHapley Additive exPlanations) values[1]. As shown in Figure A.1, the isolation metrics (reduced to 200 dimensions via PCA for comparability) contribute more to the model's predictions than any other feature

[1]We apply 5-fold cross validation, and combine the prediction on test folds to calculate the shap contribution.

group, including TFIDF features. This confirms that semantic isolation is not a peripheral signal but a central predictor of scientific impact.

Feature set	Precision	Comp-prec	F1	Comp-F1
control+subfield	0.3140 (0.00053)	0.0	0.4151 (0.00057)	0.0
control+atypicality	0.3231 (0.00057)	0.0133	0.4220 (0.00058)	0.0118
control+isolation	0.4115 (0.00072)	0.1421	0.4692 (0.00059)	0.0925

Table A.15: *Predicting disruptive papers: individual feature groups. Standard errors from 100 bootstrap iterations in parentheses. Comp-prec shows the completeness of error in precision.*

Predicting Disruptive Papers The results for disruption prediction, presented in Table A.15, mirror the patterns observed for citations. Our isolation metrics substantially outperform the atypicality score, achieving a 14.21% completeness score on precision compared to just 1.33% for atypicality. This consistency across different prediction tasks strengthens the evidence that isolation captures fundamental aspects of research novelty.

Feature set	Precision	Comp-prec	F1	Comp-F1
control+subfield	0.3140 (0.00053)	0.0	0.4151 (0.00057)	0.0
control+subfield+atypicality	0.3437 (0.00054)	0.0434	0.4424 (0.00051)	0.0466
control+subfield+atypicality+isolation	0.4242 (0.00074)	0.1606	0.4797 (0.00062)	0.1104
all features (excl. isolation)	0.4270 (0.00070)	0.1647	0.4915 (0.00060)	0.1305
all features (incl. isolation)	0.4551 (0.00075)	0.2057	0.5052 (0.00059)	0.1539

Table A.16: *Predicting disruptive papers: additional value of isolation metrics.*

Moreover, the metrics are again complementary, as shown in Table A.16. Adding isolation metrics to a model that already includes the baseline and atypicality features increases precision completeness from 4.34% to 16.06%. This additional effect confirms that our measure of semantic distance captures a dimension of novelty relevant for disruption that is distinct from the recombinative novelty captured by atypicality.

The bottom panel of Table A.16 shows that isolation metrics continue to provide significant predictive value even in the presence of all other features, increasing precision completeness from 16.47% to 20.57%. Although the 4.1 percentage point improvement is more modest than for citation prediction, the standard error shows that the improvement in precision is still significant. It demonstrates the consistent and non-redundant value of isolation across different impact dimensions.

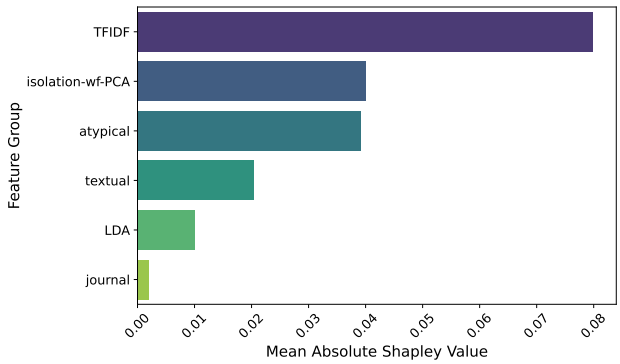


Figure A.2: *Feature importance (mean absolute SHAP) for predicting disruption.*

The SHAP analysis[2] in Figure A.2 further confirms that our isolation features are a key driver of the model's predictions. This confirms that isolation captures meaningful signals for identifying potentially disruptive research, further validating our approach across different dimensions of research impact.

[2]We apply 5-fold cross validation, and combine the prediction on test folds to calculate the shap contribution.

A.4 More results on Typology

This appendix provides technical details on the construction of our novelty scores and presents supplementary statistics on the distribution of papers within our two-dimensional framework.

A.4.1 From Isolation Metrics to Novelty Scores via PCA

We construct the spatial and temporal novelty scores by applying Principal Component Analysis (PCA) to the relevant subsets of our standardized semantic isolation metrics. The first principal component (PC1) of each set serves as our composite index.

- **Spatial Novelty Score:** PC1 of the standardized contemporaneous point-in-time isolation metrics and the baseline paper-level mean distance.
- **Temporal Novelty Score:** PC1 of the standardized retrospective dynamic isolation metrics.

To interpret the novelty scores, we examine the contributions (factor loadings) of the underlying features to the first principal component (PC1) for both the spatial and temporal analyses, as shown in Figure A.3.

For both novelty scores, the interpretation is anchored by the dominant, positive loadings of the distance-based metrics: k-NN distance (m2) and average k-NN distance (m3). This provides a clear and intuitive orientation for the final index. For the spatial novelty score, the strong positive contributions from m2 and m3 confirm that a higher value on PC1 corresponds to greater semantic isolation. For the temporal novelty score, their positive contributions confirm that a higher value on PC1 corresponds to a more rapidly densifying neighborhood (i.e., a larger positive change in isolation from $t = 3$ to $t = 0$), which is our measure of engagement with a "hot" research frontier.

The more complex loading patterns for the density-based metrics, particularly the mixed signs for neighborhood density (m1), reflect the imperfect correlations between the different proxies

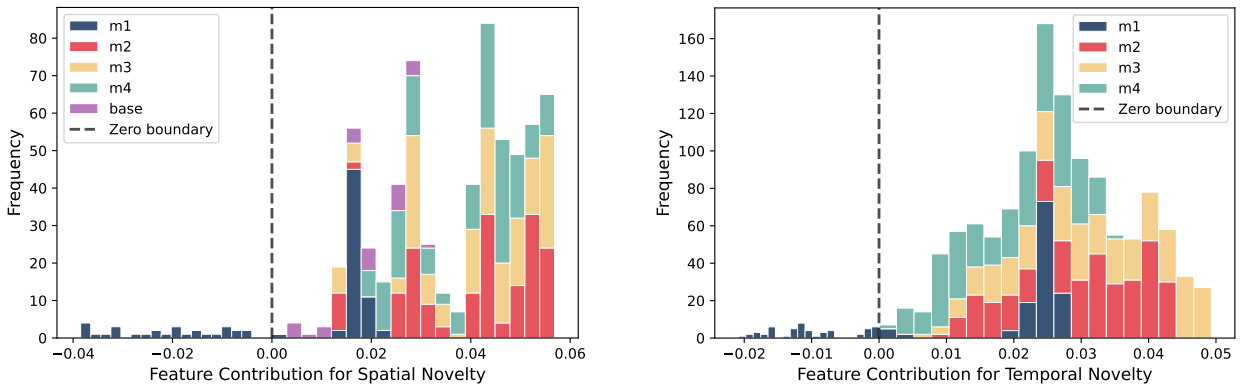


Figure A.3: *Distribution of feature contribution. Panel (a) is for spatial novelty, and panel (b) is for temporal novelty. For the metric name: m1 is neighborhood density; m2 is $k-$NN distance; m3 is $k-$NN average distance; and m4 is Gaussian density.*

for novelty. While all four metric types are designed to capture aspects of novelty, they are not perfectly collinear. PCA constructs the optimal linear combination that captures the maximum shared variance. In this process, a feature might receive a small or even negative loading if doing so helps to purify the principal component, making it a better index of the primary variation defined by the dominant m2 and m3 metrics. This is a standard outcome of PCA when synthesizing multiple, related indicators.

Therefore, the resulting PC1 is a statistically robust index that synthesizes information from all our isolation measures, with its overall direction and interpretation reliably governed by the strongest and most intuitive distance-based features.

An analysis of the top-20 most influential features for each score reveals that a diverse set of metrics contributes to the final indices. For both scores, metrics derived from various insight types (especially Question-Conclusion and Paragraph Summary), embedding models (SPECTER and Qwen), and distance measures are all highly represented. This diversity underscores the value of our comprehensive feature engineering approach and demonstrates that the resulting novelty

scores are robust composites rather than being driven by a single measurement choice.

A.4.2 Descriptive Statistics of the Novelty Framework

Figure A.4 shows the distribution of spatial novelty score and temporal novelty score.

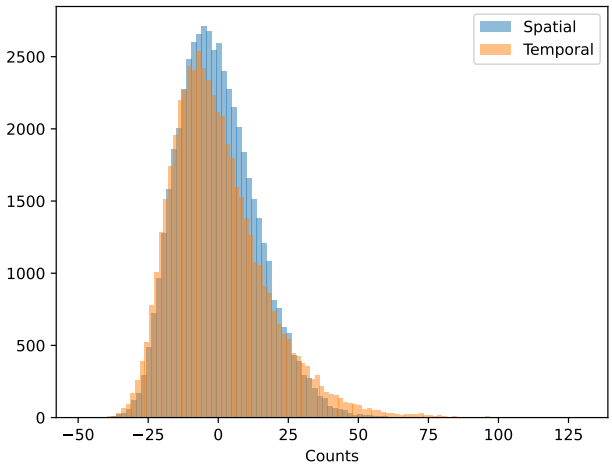


Figure A.4: *Distribution of novelty scores.*

Table A.17 presents the distribution of papers across the four novelty archetypes, defined by the median splits of the two scores. The prevalence of the off-diagonal archetypes - Trendy and Outlying neighborhoods - quantifies the structural trade-off discussed in the main text. Trailblazing neighborhoods, where papers successfully combine high spatial and temporal novelty, represent the least common archetype, comprising about one-fifth of the sample.

Table A.18 details the distribution of papers from each economic subfield across the four quadrants, reinforcing the field-mapping analysis in the main text. For example, Macroeconomics papers are heavily concentrated in the Consolidating or Trendy quadrants, while Economic History papers are mainly in Outlying quadrants. Figure A.5 visualizes the distributions for these two fields, illustrating that while their central tendencies differ, there is substantial overlap. This overlap confirms that our framework captures within-field variation in research strategies, rather than

High Temporal	Trendy (N=15468, Percent=29.3%)	Trailblazing (N=10915, Percent=20.7%)
Low Temporal	Consolidating (N=10915, Percent=20.7%)	Outlying (N=15468, Percent=29.3%)
	Low Spatial	High Spatial

Table A.17: *Distribution of papers across the four novelty archetypes.*

simply acting as a proxy for subfield identity.

field	Consolidating	Outlying	Trendy	Trailblazing
macroeconomics	0.31	0.149	0.474	0.067
international	0.257	0.185	0.42	0.139
financial	0.151	0.307	0.287	0.255
economic history	0.057	0.719	0.057	0.167
microeconomics theory	0.302	0.204	0.379	0.115
industrial organization	0.266	0.364	0.207	0.163
econometrics & quantitative	0.16	0.276	0.333	0.23
public	0.213	0.342	0.241	0.203
health	0.026	0.31	0.13	0.535
development	0.084	0.37	0.163	0.383
labor	0.241	0.307	0.274	0.178
agricultural	0.094	0.57	0.089	0.247
urban & regional	0.239	0.363	0.21	0.188
environmental & resource	0.103	0.474	0.126	0.297
behavioral	0.098	0.282	0.24	0.38
other	0.079	0.467	0.083	0.37
all	0.207	0.293	0.293	0.207

Table A.18: *Distribution of novelty archetypes within economic subfields. Each row shows the fraction of papers from a given subfield that fall into each of the four novelty archetypes.*

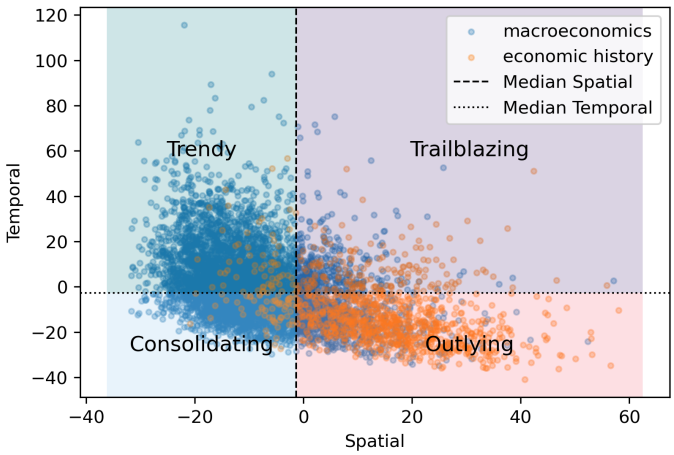


Figure A.5: *Distribution of macroeconomics and economic history papers, illustrating the distinct central tendencies but substantial overlap between the two fields.*

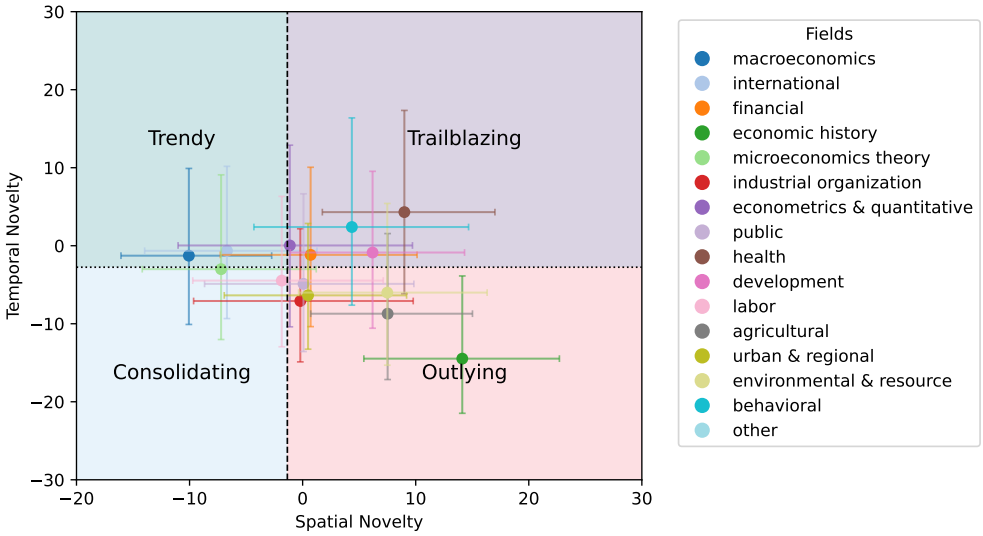


Figure A.6: *Median novelty profiles by economic subfield. Each point represents the median spatial and temporal novelty score for all papers in a given subfield. Error bars represent 25-75 inter-quartile range.*

A.5 More Results on Typology and Impacts

This appendix provides the detailed statistical results and robustness checks for the analysis presented in Section 2.6.

A.5.1 Robustness of High-Citation Analysis

This section provides the full results for the analysis of high-citation rates. Table A.19 presents the detailed statistics underlying Figure 2.8. The non-overlapping 95% confidence intervals confirm that the differences between novelty archetypes are statistically significant.

In Table A.19, we first show the detailed values for Figure 2.8 in Section 2.6.1. The high-cited rate and standard deviation are calculated from 100 resampling. We calculate the 95% confidence interval for the values, which show that the differences between novelty types are statistically significant.

Neighborhood Archetype	High-Cite Rate	SE	95% CI lower	95% CI upper
Consolidating	0.1808	0.00031	0.18023	0.18143
Outlying	0.1970	0.00027	0.19650	0.19755
Trendy	0.2914	0.00028	0.29080	0.29191
Trailblazing	0.3247	0.00032	0.32403	0.32527
all	0.2477	0.00001	0.24771	0.24775

Table A.19: *High-citation rates (Top 25%) by novelty type. Statistics are calculated from 100 bootstrap iterations. A paper is defined as highly cited if its citation count is in the top 25% of its publication-year-subfield cohort.*

Tables A.20 and A.21 test the robustness of this finding using stricter definitions of "highly cited" (top 10% and top 1%, respectively). The results confirm that the primary pattern holds and suggest that the relative advantage of high temporal novelty becomes even more pronounced for papers that achieve elite levels of impact. For the top 1% threshold, high-temporal-novelty papers (in Trendy or Trailblazing Neighborhoods) are nearly twice as likely to be highly cited as their low-temporal-novelty counterparts.

Neighborhood Archetype	High-Cite Rate	SE	95% CI lower	95% CI upper
Consolidating	0.0646	0.00021	0.06416	0.06499
Outlying	0.0736	0.00017	0.07322	0.07390
Trendy	0.1178	0.00020	0.11739	0.11815
Trailblazing	0.1345	0.00029	0.13394	0.13506
all	0.0973	0.00001	0.09725	0.09729

Table A.20: *High-citation rates (Top 10%) by novelty type.*

Neighborhood Archetype	High-Cite Rate	SE	95% CI lower	95% CI upper
Consolidating	0.0054	0.00006	0.00526	0.00549
Outlying	0.0074	0.00005	0.00733	0.00753
Trendy	0.0108	0.00005	0.01071	0.01091
Trailblazing	0.0137	0.00007	0.01359	0.01387
all	0.0093	0.00001	0.00929	0.00931

Table A.21: *High-citation rates (Top 1%) by novelty type.*

A.5.2 Analysis of Citation Count Distributions

Beyond the binary classification of high impact, we examine the relationship between novelty archetypes and the full distribution of citation counts. Table A.22 reports the mean, median, and standard error (SE) for citations within each quadrant.

The results confirm the primary role of temporal novelty: papers in Trendy or Trailblazing Neighborhoods have mean citation counts more than double those of papers in Consolidating or Outlying neighborhoods. However, the analysis of median citations and the SE reveals an important secondary role for spatial novelty. Among high-temporal-novelty papers, papers in Trailblazing neighborhoods have a substantially higher median citation count than papers in Trendy neighborhoods (105.3 vs. 85.7). This pattern, combined with the lower SE for papers in Trailblazing neighborhoods, indicates that while both archetypes can produce blockbuster hits, high spatial novelty leads to a more consistently high level of impact and a less skewed citation distribution.

High Temporal	Trendy (Mean=269.6, Median=85.7, SE=48.54)	Trailblazing (Mean=270.9, Median=105.3, SE=41.40)
Low Temporal	Consolidating (Mean=121.4, Median=40.3, SE=26.79)	Outlying (Mean=138.0, Median=49.0, SE=30.98)
	Low Spatial	High Spatial

Table A.22: *Citation distribution statistics by novelty type. Statistics are averaged over 100 bootstrap iterations. SE is the standard error of the mean (Std. Dev. / sqrt Mean). For the full sample, Mean=200.6, Median=65.1, CV=2.87.*

A.5.3 Robustness of Disruption Analysis

This section provides the full results for the analysis of disruptive impact. Table A.23 presents the statistics underlying Figure 2.9. The confidence intervals confirm that papers high in spatial novelty (in Outlying or Trailblazing neighborhoods) are significantly more likely to be disruptive than their low-spatial-novelty counterparts.

Neighborhood Archetype	Disruption Rate	SE	95% CI lower	95% CI upper
Consolidating	0.1184	0.00034	0.11778	0.11910
Outlying	0.2245	0.00032	0.22385	0.22513
Trendy	0.1053	0.00024	0.10484	0.10578
Trailblazing	0.2021	0.00037	0.20135	0.20279
all	0.1630	0.00017	0.16264	0.16331

Table A.23: *Disruption rates by novelty type (full sample).*

Tables A.24 and A.25 test the robustness of this finding by examining disruption rates exclusively within the subsamples of highly cited papers. Across all specifications, papers with high spatial novelty maintain a significantly higher disruption rate, confirming that this relationship is not confounded by a paper's overall level of attention.

Neighborhood Archetype	Disruption Rate	SE	95% CI lower	95% CI upper
Consolidating	0.1038	0.00065	0.10254	0.10507
Outlying	0.2012	0.00073	0.19974	0.20259
Trendy	0.1062	0.00051	0.10519	0.10719
Trailblazing	0.1694	0.00052	0.16839	0.17045
all	0.1451	0.00029	0.14453	0.14569

Table A.24: *Disruption rates by novelty type (top 25% cited paper only).*

Neighborhood Archetype	Disruption Rate	SE	95% CI lower	95% CI upper
Consolidating	0.1327	0.00136	0.13000	0.13533
Outlying	0.2271	0.00122	0.22467	0.22947
Trendy	0.1219	0.00071	0.12054	0.12334
Trailblazing	0.1819	0.00082	0.18028	0.18349
all	0.1638	0.00048	0.16291	0.16478

Table A.25: *Disruption rates by novelty type (top 10% cited paper only).*

A.5.4 Robustness of the Advantage of Trailblazing Archetype

This section provides robustness checks for the advantage of papers in Trailblazing neighborhoods in achieving dual impact. Table A.26 provides the detailed statistics for the main analysis (top 25% cited + disruptive). Table A.27 demonstrates that the synergistic advantage of papers in Trailblazing neighborhoods persists when using a stricter high-citation threshold (top 10%). In both cases, papers in Trailblazing neighborhoods are significantly more likely than any other archetype to be both highly cited and disruptive.

Neighborhood Archetype	Dual Impact Rate	SE	95% CI Lower	95% CI Upper
Consolidating	0.0188	0.00012	0.01854	0.01900
Outlying	0.0396	0.00016	0.03932	0.03995
Trendy	0.0309	0.00015	0.03064	0.03123
Trailblazing	0.0550	0.00017	0.05466	0.05534
all	0.0359	0.00007	0.03580	0.03609

Table A.26: *Probability of being both highly cited (top 25%) and disruptive.*

We then consider papers that are both top 10% high-cited and disruptive and show the result in in Table A.25. We can find that papers in Trailblazing neighborhoods still have advantage over other types.

Neighborhood Archetype	Dual Impact Rate	SE	95% CI Lower	95% CI Upper
Consolidating	0.0086	0.00009	0.00839	0.00874
Outlying	0.0167	0.00009	0.01652	0.01688
Trendy	0.0144	0.00009	0.01419	0.01453
Trailblazing	0.0245	0.00012	0.02423	0.02469
all	0.0159	0.00005	0.01585	0.01603

Table A.27: *Probability of being both highly cited (top 10%) and disruptive.*

A.5.5 Robustness to Alternative Novelty Threshold

To ensure our findings are not sensitive to the median split used to define the novelty archetypes, we replicate the analysis using a more restrictive 75th percentile threshold. As shown in Table A.28, papers in Trailblazing neighborhoods becomes considerably rarer (4.0% of the sample), underscoring the difficulty of simultaneously achieving high levels on both novelty dimensions. Despite this shift in distribution, the qualitative results remain unchanged. Tables A.29 and A.30 show that temporal novelty remains the primary predictor of high citations, while spatial novelty remains the primary predictor of disruption.

High Temporal	Trendy (N=11086, Percent=21.0%)	Trailblazing (N=2106, Percent=4.0%)
Low Temporal	Consolidating (N=28488, Percent=54.0%)	Outlying (N=11086, Percent=21.0%)
	Low Spatial	High Spatial

Table A.28: *Distribution of papers across the four quadrants, setting high as 75-th percentile.*

Neighborhood Archetype	High-Cite Rate	SE	95% CI lower	95% CI upper
Consolidating	0.2136	0.00018	0.21322	0.21395
Outlying	0.2188	0.00038	0.21809	0.21957
Trendy	0.3401	0.00039	0.33932	0.34085
Trailblazing	0.3756	0.00097	0.37365	0.37746
all	0.2477	0.00001	0.24771	0.24775

Table A.29: *High-Citation rates (Top 25%) using 75th percentile threshold.*

We also check the disruptive rate using this typology, and show the result in Table A.30. We can see that types with high spatial novelty (in Outlying or Trailblazing neighborhoods) have much higher high-cited rate compared to types with low spatial novelty.

Neighborhood Archetype	Disruption Rate	SE	95% CI lower	95% CI upper
Consolidating	0.1442	0.00024	0.14370	0.14463
Outlying	0.2508	0.00040	0.25004	0.25161
Trendy	0.1150	0.00031	0.11441	0.11561
Trailblazing	0.2075	0.00091	0.20576	0.20933
all	0.1630	0.00017	0.16264	0.16331

Table A.30: *Disruption rate using 75th percentile threshold (full sample).*

A.5.6 Field-Level Analysis

To ensure our findings are not driven by compositional effects across disciplines, we replicate the analysis by constructing the novelty scores and examining impact rates within each subfield. Tables A.31 and A.32 present the results. Despite variation in the magnitude of effects, the core qualitative patterns hold across nearly all subfields: papers high in temporal novelty (in Trendy or Trailblazing neighborhoods) consistently have higher citation rates, while papers high in spatial novelty (in Outlying or Trailblazing neighborhoods) consistently have higher disruption rates. This confirms that our framework captures general pathways to scientific influence.

field	Consolidating	Outlying	Trendy	Trailblazing
macroeconomics	0.196	0.194	0.291	0.322
	(0.0012)	(0.0009)	(0.0009)	(0.0013)
international	0.149	0.175	0.324	0.343
	(0.0012)	(0.0010)	(0.0013)	(0.0014)
financial	0.179	0.195	0.317	0.293
	(0.0013)	(0.0011)	(0.0010)	(0.0016)
economic history	0.178	0.184	0.319	0.273
	(0.0029)	(0.0018)	(0.0017)	(0.0032)
microeconomics theory	0.186	0.227	0.238	0.362
	(0.0009)	(0.0007)	(0.0007)	(0.0011)
industrial organization	0.162	0.227	0.248	0.361
	(0.0012)	(0.0011)	(0.0010)	(0.0014)
econometrics & quantitative	0.160	0.197	0.286	0.361
	(0.0011)	(0.0008)	(0.0007)	(0.0012)
public	0.187	0.196	0.278	0.339
	(0.0012)	(0.0010)	(0.0010)	(0.0015)
health	0.218	0.161	0.319	0.289
	(0.0019)	(0.0016)	(0.0016)	(0.0019)
development	0.212	0.175	0.312	0.291
	(0.0015)	(0.0013)	(0.0014)	(0.0016)
labor	0.183	0.177	0.308	0.333
	(0.0011)	(0.0008)	(0.0010)	(0.0013)
agricultural	0.191	0.169	0.322	0.283
	(0.0027)	(0.0019)	(0.0021)	(0.0035)
urban & regional	0.233	0.152	0.312	0.275
	(0.0027)	(0.0022)	(0.0026)	(0.0029)
environmental & resource	0.163	0.208	0.283	0.327
	(0.0020)	(0.0019)	(0.0018)	(0.0023)
behavioral	0.185	0.222	0.272	0.309
	(0.0014)	(0.0013)	(0.0015)	(0.0020)
other	0.136	0.186	0.323	0.328
	(0.0024)	(0.0020)	(0.0025)	(0.0027)

Table A.31: *High-citation rates (top 25%) by novelty archetype, by subfield. Table entries report the mean high-citation rate with bootstrapped standard errors in parentheses. Novelty archetypes are defined within each subfield.*

field	Consolidating	Outlying	Trendy	Trailblazing
macroeconomics	0.104	0.192	0.081	0.158
	(0.0010)	(0.0010)	(0.0006)	(0.0013)
international	0.137	0.239	0.109	0.196
	(0.0011)	(0.0012)	(0.0009)	(0.0013)
financial	0.073	0.165	0.075	0.121
	(0.0010)	(0.0011)	(0.0008)	(0.0013)
economic history	0.181	0.231	0.210	0.139
	(0.0032)	(0.0023)	(0.0023)	(0.0031)
microeconomics theory	0.052	0.135	0.055	0.105
	(0.0006)	(0.0007)	(0.0004)	(0.0008)
industrial organization	0.100	0.203	0.139	0.219
	(0.0009)	(0.0015)	(0.0011)	(0.0013)
econometrics & quantitative	0.105	0.186	0.068	0.158
	(0.0009)	(0.0009)	(0.0006)	(0.0011)
public	0.151	0.244	0.127	0.270
	(0.0011)	(0.0014)	(0.0009)	(0.0015)
health	0.162	0.336	0.151	0.268
	(0.0018)	(0.0024)	(0.0015)	(0.0020)
development	0.253	0.376	0.262	0.332
	(0.0019)	(0.0019)	(0.0020)	(0.0029)
labor	0.144	0.257	0.198	0.234
	(0.0010)	(0.0011)	(0.0011)	(0.0014)
agricultural	0.353	0.308	0.379	0.367
	(0.0033)	(0.0028)	(0.0033)	(0.0041)
urban & regional	0.200	0.269	0.107	0.203
	(0.0028)	(0.0030)	(0.0019)	(0.0028)
environmental & resource	0.159	0.282	0.221	0.346
	(0.0019)	(0.0021)	(0.0021)	(0.0029)
behavioral	0.049	0.143	0.058	0.104
	(0.0009)	(0.0013)	(0.0009)	(0.0013)
other	0.313	0.269	0.163	0.288
	(0.0029)	(0.0025)	(0.0022)	(0.0024)

Table A.32: *Disruption rates by novelty archetype, by subfield. Table entries report the mean disruption rate with bootstrapped standard errors in parentheses. Novelty archetypes are defined within each subfield.*

A.6 Prompts

Listing A.1: *Prompt for comprehensive summary.*

```
"""
Generate an in-depth 1000-2000 word structured summary of this economics research paper that simultaneously reconstructs its
     content and provides critical analysis. Adhere to these specifications:
### 1. Extended Structural Reconstruction (800-1200 words)
For each section, include:
#### Introduction
- Research context & motivations (2-3 paragraphs)
- Explicitly stated hypotheses/research questions
- Literature review synthesis: Map key citations to research gaps
#### Theoretical Framework
- Foundational Theories: Outline the core economic theories and models that underpin the research.
- Assumptions and Constructs: Detail the theoretical assumptions made and define key constructs.
- Conceptual Integration: Explain how these theories inform the research questions, methodology, and analysis.
- Critical Perspective: Evaluate the strengths and limitations of the theoretical approach in addressing the identified research
     gaps.
#### Methodology
- Data sources: Providers, timeframes, inclusion/exclusion criteria
- Analytical framework: Equations/diagrams with plain-English explanations
- Technical procedures: Step-by-step workflow
#### Experiments/Analysis
- Implementation chronology
- Control variables/benchmarks
- Sensitivity testing documentation
#### Results
- Primary/secondary findings with statistical significance
- Non-obvious patterns in data
- Failed hypotheses (if any)
#### Discussion
- Mechanisms explaining results
- Comparative analysis with 3-5 key cited works
- Policy recommendations with implementation pathways

### 2. Comprehensive Evaluation (500-800 words)
Develop dedicated critique sections:
#### Innovation Audit
- Novelty: Conceptual | Methodological | Empirical
- Patentable prior art comparison
- Disruptiveness score (1-5) with rationale
#### Methodological Rigor
- STROBE/RDD checklist compliance
- Data limitations
- Counterfactual analysis completeness
#### Communication Effectiveness
- Readability metrics: Flesch-Kincaid grade level | Jargon density
```

```
- Argumentation coherence mapping
- Visual aid effectiveness assessment
#### Scalability Potential
- Replication constraints analysis
- Generalizability (Contexts/Data/Methods)
- Cost-benefit simulation

### 3. Enhancement Protocol (200-300 words)
Provide actionable improvements for:
- Theory-building extensions
- Methodological upgrades
- Policy translation frameworks
### Format Requirements:
- Use MLA citation style for all references
- Include 4-6 verbatim quotes exemplifying key strengths/weaknesses
- Present technical content through Definition Boxes
- Highlight critical insights in **bold**
- Maintain a graduate-level academic tone
- Deliver output as a hybrid document integrating summary and peer-review elements

Here is the paper:
{Paper_Content}
"""
```

Listing A.2: *Prompt for Paragraph insight.*

```
"""
You are an expert in economics with experience in reading and summarizing academic research papers.
I have a detailed summary of an economics research paper, and I would like you to write a shorter, more concise summary while
      retaining all the key information.

**Goal:**
Write a summary of several paragraphs that clearly conveys the main contributions, innovations, and findings of the paper.
The summary should be clear and informative enough for a researcher to understand the core of the paper without reading it.

**Instructions:**
- Keep the summary concise but comprehensive.
- Clearly explain the research question, methodology, data, and findings.
- Highlight the paper's main contributions and innovations.
- Avoid unnecessary details or lengthy explanations.
- Maintain a formal and academic tone.

Here is the detailed summary of the paper:
{paper_content}

Please write the shorter summary below:
"""
```

Listing A.3: *Prompt for Topic insight.*

```
"""
Please extract the research topic of the given text in one sentence strictly following this format:
"The economic subfield is [subfield name], the main research subject is [subject], it focuses on [specific issue], and it uses [
     approach]."
Your response should be at most 40 words.

Here is the input text:
{paper_content}

Please write the response below:
"""
```

Listing A.4: *Prompt for Question-Conclusion insight.*

```
"""
Please summarize the research question and conclusions of the given text according to the following structure (≤100 words):

1. Core Research Question: [What the paper seeks to answer, framed as a clear question]
2. Approach and Hypothesis: [For empirical studies, describe the hypothesis; for theoretical papers, state the key assumption]
3. Key Findings: [Discoveries, including a comparison with expectations or existing literature]
4. Novelty and Implications: [What makes this conclusion different or impactful]

Here is the input text:
{paper_content}

Please write the summary below:
"""
```

Listing A.5: *Prompt for Methodology insight.*

```
"""
Please summarize the methodology of the given text in a structured format (≤80 words):

- Data Source: [Dataset name, time span, sample size]
- Core Model: [Mathematical expression, e.g., Y = βX + ϵ]
- Analytical Techniques: [e.g., IV estimation, text mining process]
- Assumptions: [Key economic assumptions, if applicable]
- Methodological Innovation: [How this method improves upon or differs from existing methods]

Here is the input text:
{paper_content}

Please write the summary below:
"""
```

Listing A.6: *Prompt for Theory insight.*

```
"""
Please summarize the theoretical contributions of the given text according to the following structure (≤100 words):

1. Fundamental Theory: [The main theoretical framework it builds upon]
2. Extensions: [Modified assumptions, newly introduced mechanisms]
3. Theoretical Advances: [Innovations in key formulas or theorem proofs]
4. Scope of Explanation: [Extensions of the new theory to applicable scenarios]
5. Empirical Relevance: [How the theoretical insights connect with empirical findings, if applicable]

Here is the input text:
{paper_content}

Please write the summary below:
"""
```

APPENDIX B

APPENDIX FOR CHAPTER 3

B.1 Additional empirical results

B.1.1 Additional Models and Algorithms

We now augment Table 3.1 to report the performance of alternative models and algorithms.

A more flexible interpretation of the standard game-theoretic model is that it partitions positions into three strategically identical classes of positions (depending on their tablebase value), but does not necessarily pin down the prediction at these classes to be the tablebase value. This model can be written as the class of functions

$$f_{\hat{y}_W,\hat{y}_D,\hat{y}_B}(p_i) = \begin{cases} \hat{y}_W & \text{if } v_i^{TB} = 1 \\ \hat{y}_D & \text{if } v_i^{TB} = 1/2 \\ \hat{y}_B & \text{if } v_i^{TB} = 0 \end{cases}$$

indexed to free parameters $(\hat{y}_W, \hat{y}_D, \hat{y}_B) \in [0,1]^3$, which we estimate on the training data. We report the fivefold cross-validated mean-squared error of this rule.

We additionally consider a variant of our machine learning algorithm, which is restricted to use only features that do not have knowledge of the tablebase value of the position. The performance of this algorithm tells us how informative the non-tablebase statistics of the position are. (Since tablebase is a highly complex function of the position, we do not expect that our hand-coded features can reproduce it, and do not find that they do.)

Table B.1 reports the cross-validated mean-squared errors of each of these approaches. We find

that Flexible Tablebase improves substantially on Tablebase, but does not attain the performance of our original ML algorithm. If however we remove tablebase as a feature, the performance of the ML algorithm falls below the performance of Tablebase, again suggesting that tablebase is a highly predictive single statistic of the position.

	Error	Comp
Perfect Prediction	0	1
ML Algorithm	0.0035 (1.1×10^{-5})	93%
Flexible Tablebase	0.0079 (2.3×10^{-5})	84%
Tablebase	0.0138 (3.3×10^{-5})	72%
ML Algorithm without Tablebase	0.0152 (5.0×10^{-5})	69%
Worst-Case Benchmark	0.0494 (12.2×10^{-5})	0

Table B.1: *Performance of the different prediction methods.*

B.1.2 Alternative Best Case Benchmark

In Tables 3.1 and B.1, we use perfect prediction as the best case benchmark against which to evaluate the performance of Tablebase. We now replace this idealized benchmark with the performance of a closely related algorithm. For each position p_i, we split the observations for that position into five equally sized sets of games. In each round of cross-validation, select one of the subsets to use as a test set, and pool the remaining observations to form a training set. The prediction for p_i in the test set is the average game outcome for this position in the training data.[1] Given a sufficiently large number of observations per position, this rule learns the mean outcome for each position, and

[1]The cross-validation we use to evaluate this rules is different from the cross-validation that we use to evaluate the others.

thus minimizes mean-squared error among all prediction rules based on the position alone. The performance of this rule turns out to be sufficiently close to perfect prediction that the reported completeness of the methods would not change substantially had we used this rule instead. Below we reproduced Table 3.1 with this benchmark instead.

	Error	Comp
Best-Case Benchmark (Emp)	0.0012	1
	(0.1×10^{-5})	
ML Algorithm	0.0035	95%
	(1.1×10^{-5})	
Tablebase	0.0138	74%
	(3.3×10^{-5})	
Worst-Case Benchmark	0.0494	0
	(12.2×10^{-5})	

Table B.2: *Performance of the different prediction methods. We use the empirical prediction as the best-case benchmark.*

B.1.3 Performance of Tablebase by Rating Group

Table B.3 reproduces Table B.1, reporting results for each rating group introduced in Section 3.2.2.

All of the qualitative findings from Section 3.2.2 hold when we partition the data in this way, but we can also note new comparative statics. The performance of the ML algorithm (with tablebase) is similar across rating groups, suggesting that the predictability of play is comparable. But interestingly play is predicted by different features. The completeness of Tablebase and Flexible Tablebase increase substantially as players become more highly-rated, reflecting that tablebase value is a better prediction of play between higher-rated players. Interestingly, ML without Tablebase *outperforms* both Tablebase and also Flexible Tablebase for the low-rated group, suggesting that other (more apparent) features of the position are more predictive of play than tablebase. These

	Low-rated		Medium-rated		High-rated	
	Error	Comp	Error	Comp	Error	Comp
Best-Case Benchmark	0	1	0	1	0	1
ML	0.0040 (3.6×10^{-5})	93%	0.0.0038 (2.2×10^{-5})	91%	0.0030 (1.9×10^{-5})	95%
ML without Tablebase	0.0088 (8.0×10^{-5})	84%	0.0134 (7.7×10^{-5})	70%	0.0195 (9.7×10^{-5})	66%
Flexible Tablebase	0.0113 (7.7×10^{-5})	79%	0.0090 (4.4×10^{-5})	80%	0.0076 (3.9×10^{-5})	87%
Tablebase	0.0347 (22.1×10^{-5})	36%	0.0163 (6.9×10^{-5})	64%	0.0101 (4.8×10^{-5})	83%
MA (Worst Case Benchmark)	0.0543 (35.1×10^{-5})	0	0.0449 (19.4×10^{-5})	0	0.0580 (21.9×10^{-5})	0

Table B.3: *Performance of the different prediction methods, in different rating groups.*

findings are consistent with lower-rated players being more ignorant about the tablebase value of a position, in which case their choices can't correlate too strongly with this statistic.

B.2 Proofs for Results in Main Text

B.2.1 Proof of Proposition 1

Fix a Markov strategy profile. The payoff-relevant state is x_t (we will drop time subscripts). Under this strategy profile, there is a value v_s and v_c to the moving player of starting from each state.

In each position $x \in \{c, s\}$, given insight I_x^t, in the complex position, the expected value of playing action m_t is

$$\begin{cases} \alpha_c I_x^t (1 - v_c) & \text{if} \quad m_t = C \\ \alpha (1 - v_s) & \text{if} \quad m_t = S \end{cases}$$

In both states, the cutoff rule takes the form: play C if

$$I_x^t > \rho := \frac{1 - v_s}{1 - v_c} \tag{B.1}$$

and play L if $I_x < \rho$, with ties broken arbitrarily. The fact that best responses are as claimed follows from these observations.

B.2.2 Proof of Proposition 2

Fix any Markov-perfect equilibrium. Now we write down recursive equations for value using the definitions in 3.3.2.3 above:

$$v_c = g_c\left(1 - v_c\right) + \alpha_c\left(1 - p_c\right)\left(1 - v_s\right) \tag{B.2}$$

$$v_s = g_s\left(1 - v_c\right) + \alpha_s\left(1 - p_s\right)\left(1 - v_s\right). \tag{B.3}$$

Solving (B.2) and (B.3) and plugging into (B.1) yields an equation for the cutoff specified in B.1

$$\rho = \frac{1 - v_s}{1 - v_c} = \frac{1 - g_s + g_c}{1 + \alpha_s(1 - p_s) - \alpha_c(1 - p_c)}. \tag{B.4}$$

Notice that the right-hand side is $\mathcal{R}(\rho)$ defined in 3.1. What we have said implies that this is a necessary condition for an equilibrium.

Lemma 1. *All solutions to (3.1) are interior. For generic (α, α_s), there is at least one ρ satisfying $\rho = \mathcal{R}(\rho)$ and $\mathcal{R}'(\rho) < 1$.*

Proof. Note that

$$\mathcal{R}(0) = 1 - \alpha_s \mathbb{E}[I_s] + \alpha_c \mathbb{E}[I_h] > 0$$

while $\mathcal{R}(1) < 1$. Since I_s and I_h have differentiable densities, $\mathcal{R}$ is continuous, so by the intermediate value theorem it crosses the 45-degree line.

Let E be the set of points ρ where $\mathcal{R}(\rho) = \rho$. By our assumption that the densities are differentiable, $\mathcal{R}$ is differentiable everywhere. For generic (α, α_s), the set E is of dimension zero and none

of the solutions of $\mathcal{R}(\rho) = \rho$ involve a tangency between $\mathcal{R}(\rho)$ and the 45-degree line. Consider such an (α, α_s) pair. It is clear that $\mathcal{R}'(\rho) \leq 1$ at some $\rho \in E$, otherwise $\mathcal{R}$ could not cross the 45-degree line. But no tangencies implies that $\mathcal{R}'(\rho) < 1$. □

Given such a fixed point, we must show that it corresponds to an equilibrium. Given a cutoff ρ satisfying $\mathcal{R}(\rho) = \rho$, set v_c and v_s as follows, with all the endogenous variables on the right-hand side being functions of the cutoff ρ:

$$v_c = -\frac{2g_c - p_s g_c + \alpha_c - p_c\alpha_c - g_s\alpha_c + p_c g_s \alpha_c}{-2 + p_s - 2g_c + p_s g_c + g_s\alpha_c - p_c g_s \alpha_c},$$
$$v_s = -\frac{1 - p_s + g_s + g_c - p_s g_c - g_s\alpha_c + p_c g_s \alpha_c}{-2 + p_s - 2g_c + p_s g_c + g_s\alpha_c - p_c g_s \alpha_c}.$$

It can be directly verified that these solve (B.2), (B.3), and (B.4), so they are "physically consistent" values and playing according to the threshold strategy is a best response. This completes the proof of the proposition.

B.2.3 Proof of Proposition 4

The claims in the first paragraph follow from the implicit function theorem.

Let z parameterize a perturbation. First consider a perturbation that increases α_s. Let $\rho(z)$ be a solution to $\rho = \mathcal{R}_z(\rho)$, defined locally round the starting z. At a stable equilibrium, where $\mathcal{R}_z(\rho)$ crosses the 45-degree line from above, $\frac{d}{dz}\rho(z) < 0$ is equivalent to $\frac{\partial}{\partial z}\mathcal{R}_z(\rho) < 0$.

To facilitate writing the derivative, define $A_x = \int_\rho^1 r \, \mathrm{d}F_x(r)$, where $x \in \{s, c\}$.

Then we have

$$\frac{\partial}{\partial z}\mathcal{R}_z(\rho) = -\frac{(1 - p_c)(1 + g_c) + A_s(1 - \alpha_c(1 - p_c))}{(\alpha_s(1 - p_s) - \alpha_c(1 - p_c) + 1)^2}.$$

Since this is always negative, we conclude that $\frac{d}{d\alpha_s}\rho < 0$. Let B_s denote the numerator of the expression above, $B_s = (1 - p_c)(1 + g_c) + A_s(1 - \alpha_c(1 - p_c))$.

The reasoning is analogous for a perturbation that increases α_c. In that case we compute

$$\frac{\partial}{\partial z}\mathcal{R}_z(\rho) = \frac{(1 - p_c)(1 - g_s) + A_c(1 + \alpha_s(1 - p_s))}{(\alpha_s(1 - p_s) - \alpha_c(1 - p_c) + 1)^2}.$$

Note this derivative is always positive. Let B_c denote the numerator, $B_c = (1 - p_c)(1 - g_s) + A_c(1 + \alpha_s(1 - p_s))$.

Now consider a perturbation that brings about a small increase $\dot{\alpha_c}$ in the accuracy of play in complex positions, and a small increase $\dot{\alpha_s}$ in the accuracy of play in simple positions. For a decrease in the threshold (cancelling common factors that do not affect the sign) it must be that

$$\dot{\alpha_s}B_s \geq \dot{\alpha_c}B_c.$$

Defining $\gamma = B_c/B_s$, this gives the condition (3.2). Note that for $\rho \approx 1$, both values A_x are approximately 0, as are the values g_x and p_x. Then $\gamma \approx 1$.

B.3 Feature List

Below we use "tb value" as shorthand for the tablebase value. We also use "difficulty" to mean the fraction of feasible moves that maintain the tablebase value.

Table B.4: *Tablebase related features.*

No.	Description
1	the number of moves that maintain the tb value
2	the number of moves that change the tb value
3	the fraction of moves that maintain the tb value
4	the difficulty of the next position, averaged across all feasible moves of the player-to-move
5	the difficulty of the next position, averaged across all moves of the player-to-move that maintain the tb value
6	the max difficulty of the next position, among all feasible moves of the player
7	the max difficulty of the next position, averaged across all moves of the player-to-move that maintain the tb value
8	the average number of feasible moves for opponent, after the player playing a move that maximizes the difficulty of the next position
9	the material value of the winning player
10	depth to mate
11	tablebase value
12-14	stockfish evaluation of depth 10/16/20
15-17	stockfish win/draw/loss of depth 10/16/20

18-20stockfish evaluating time of depth 10/16/20

Additionally, for each of features 1-8, we include interaction effects with whether the player to move is White or Black.

Table B.5: *Non-tablebase related features.*

No.	Description
1	the player to move (White or Black)
2	the number of feasible moves
3	the maximum number of feasible moves in the next position, averaged across all feasible moves in the current position
4	the minimum number of feasible moves in the next position, averaged across all feasible moves in the current position
5	the number of moves that lead to a check
6	the material value for White
7	the material value for Black
8	the absolute difference of the material values for White and Black
9	the number of pieces that White has which Black does not (with each piece treated distinctly)
10	the number of pieces that Black has which White does not (with each piece treated distinctly)
11	the sum of the two previous features
12	the total point value of pieces that White has which Black does not
13	the total point value of pieces that Black has which White does not
14	the sum of the two previous features
15	$\mathbb{1}$(only pieces remaining are kings and pawns)
16	the number of moves in the shortest path to promotion of a pawn for the player to move (set to a large number if every pawn is blocked on its file)

17	the number of moves in the shortest path to promotion of a pawn for the opponent (set to a large number if every pawn is blocked on its file)
18	the absolute difference of the two previous features
19	the area of the smallest rectangle enclosing all pieces
20	$\mathbb{1}$(the player to move is in check)
21	$\mathbb{1}$(the player to move is in check and is White)
22	$\mathbb{1}$(the player to move is in check and is Black)
23	$\mathbb{1}$(there is a feasible move leading to stalemate)
24	the number of unique pieces that can move

B.4 Detailed Results underlying Structural Estimates

In this section we provide intermediate results in our structural estimation. Table B.6 reports the structurally estimated numbers underlying Figure 3.6.

x	Low-rated	Medium-rated	High-rated
$(s, \{S, C\})$	0.028	0.0109	0.0081
	(0.00088)	(0.00033)	(0.00021)
$(c, \{S, C\})$	0.0316	0.0162	0.0102
	(0.00079)	(0.00040)	(0.00024)

Table B.6: *The estimated insight discount $\hat{\delta}_x$ for each rating group following the position states where players have a choice. Bootstrapped standard errors are presented in parentheses following each estimate.*

We now present the empirical estimates of the inputs used to produce these estimates, following the discussion in Section 3.4.2.

B.4.1 Blunder rates

Table B.7 reports the empirical estimates $\hat{\boldsymbol{b}}$ of the blunder vector in each rating group. These are computed simply as the empirical rate of transition from the indicated position state to a lower tablebase value for the moving player.

	Low-rated	Medium-rated	High-rated
$(s, \{S\})$	0.0511	0.0495	0.0256
	(0.0005)	(0.0002)	(0.0001)
$(s, \{C\})$	0.0475	0.0472	0.0542
	(0.0038)	(0.0015)	(0.0012)
$(s, \{S, C\})$	0.006	0.0106	0.0095
	(0.0003)	(0.0002)	(0.0001)
$(c, \{S\})$	0.1306	0.057	0.0361
	(0.0036)	(0.0007)	(0.0007)
$(c, \{C\})$	0.081	0.0762	0.0439
	(0.0009)	(0.0003)	(0.0002)
$(c, \{S, C\})$	0.0323	0.022	0.0128
	(0.0016)	(0.0006)	(0.0003)

Table B.7: *The estimated blunder vector $\hat{\boldsymbol{b}}$ with bootstraped standard errors.*

B.4.2 Transition matrices

Table B.8 reports the empirical estimates $\hat{\boldsymbol{T}}$ of the transition matrix among position states in each rating group. These are computed simply as the empirical rate of transition from the indicated position state to each other one. Note that we do *not* condition on the moving player not blundering in computing these estimates. Thus, the sum of each row is one minus the blunder probability from that state.

B.4.3 Equilibrium values of position states

Equipped with the estimates above, we can compute the values of the position states for the moving player.

	$(S,\{S\})$	$(S,\{C\})$	$(S,\{S,C\})$	$(C,\{S\})$	$(C,\{C\})$	$(C,\{S,C\})$
$(S,\{S\})$	0.8468	0.0	0.1021	0.0	0.0	0.0
	(0.0009)	(0.0001)	(0.0008)	(0.0)	(0.0)	(0.0)
$(S,\{C\})$	0.0	0.0	0.0	0.1145	0.8368	0.0012
	(0.0)	(0.0)	(0.0)	(0.0138)	(0.0149)	(0.0033)
$(S,\{S,C\})$	0.294	0.0	0.053	0.0954	0.4153	0.1365
	(0.0033)	(0.0)	(0.0015)	(0.0026)	(0.0036)	(0.0022)
$(C,\{S\})$	0.3739	0.0	0.4955	0.0	0.0	0.0
	(0.009)	(0.0)	(0.0084)	(0.0)	(0.0)	(0.0)
$(C,\{C\})$	0.0	0.0	0.0	0.0357	0.8516	0.0317
	(0.0)	(0.0)	(0.0)	(0.0006)	(0.0013)	(0.0007)
$(C,\{S,C\})$	0.4101	0.0146	0.2193	0.005	0.3044	0.0143
	(0.0082)	(0.0013)	(0.0069)	(0.0005)	(0.0085)	(0.0019)

(a) *Low-rated*

	$(S,\{S\})$	$(S,\{C\})$	$(S,\{S,C\})$	$(C,\{S\})$	$(C,\{C\})$	$(C,\{S,C\})$
$(S,\{S\})$	0.8291	0.0065	0.115	0.0	0.0	0.0
	(0.0004)	(0.0001)	(0.0004)	(0.0)	(0.0)	(0.0)
$(S,\{C\})$	0.0	0.0	0.0	0.472	0.3854	0.0955
	(0.0)	(0.0)	(0.0)	(0.0044)	(0.0037)	(0.0038)
$(S,\{S,C\})$	0.3073	0.0092	0.0504	0.205	0.2816	0.1359
	(0.0013)	(0.0003)	(0.0005)	(0.0011)	(0.0011)	(0.0008)
$(C,\{S\})$	0.6816	0.0091	0.2523	0.0	0.0	0.0
	(0.0026)	(0.0006)	(0.0025)	(0.0)	(0.0)	(0.0)
$(C,\{C\})$	0.0	0.0	0.0	0.0272	0.8717	0.0249
	(0.0)	(0.0)	(0.0)	(0.0002)	(0.0005)	(0.0002)
$(C,\{S,C\})$	0.6281	0.007	0.1921	0.0194	0.1078	0.0237
	(0.0038)	(0.0013)	(0.0025)	(0.0014)	(0.0023)	(0.0013)

(b) *Medium-rated*

	$(S,\{S\})$	$(S,\{C\})$	$(S,\{S,C\})$	$(C,\{S\})$	$(C,\{C\})$	$(C,\{S,C\})$
$(S,\{S\})$	0.8699	0.0055	0.099	0.0	0.0	0.0
	(0.0002)	(0.0001)	(0.0002)	(0.0)	(0.0)	(0.0)
$(S,\{C\})$	0.0	0.0	0.0	0.3816	0.3042	0.2601
	(0.0)	(0.0)	(0.0)	(0.0029)	(0.0029)	(0.0025)
$(S,\{S,C\})$	0.252	0.0104	0.0683	0.1728	0.3357	0.1512
	(0.0007)	(0.0002)	(0.0003)	(0.0006)	(0.0008)	(0.0006)
$(C,\{S\})$	0.6303	0.0227	0.3109	0.0	0.0	0.0
	(0.0018)	(0.0008)	(0.0018)	(0.0)	(0.0)	(0.0)
$(C,\{C\})$	0.0	0.0	0.0	0.0293	0.9052	0.0216
	(0.0)	(0.0)	(0.0)	(0.0002)	(0.0004)	(0.0002)
$(C,\{S,C\})$	0.4849	0.0474	0.1539	0.0718	0.1968	0.0323
	(0.0026)	(0.0012)	(0.0015)	(0.0013)	(0.0021)	(0.0012)

(c) *High-rated*

Table B.8: *The estimated transition matrices $\hat{\boldsymbol{T}}$ with bootstraped standard errors.*

	Low-rated	Medium-rated	High-rated
$(s, \{S\})$	0.4853	0.4859	0.493
	(0.0001)	(0.0000)	(0.0000)
$(s, \{C\})$	0.5012	0.4942	0.4816
	(0.0021)	(0.0008)	(0.0006)
$(s, \{S, C\})$	0.5164	0.5088	0.5031
	(0.0003)	(0.0001)	(0.0001)
$(c, \{S\})$	0.4321	0.4789	0.4858
	(0.0019)	(0.0004)	(0.0004)
$(c, \{C\})$	0.4796	0.48	0.4887
	(0.0003)	(0.0001)	(0.0001)
$(c, \{S, C\})$	0.493	0.4988	0.5006
	(0.0009)	(0.0003)	(0.0002)

Table B.9: *Estimated equilibrium value vector $\hat{\boldsymbol{v}}$ with bootstraped standard errors in parentheses.*

APPENDIX C

APPENDIX FOR CHAPTER 4

C.1 Proofs

C.1.1 Notation

Throughout let $\mathcal{N} \equiv \{1, \ldots, n\}$. The set $\mathbb{T}_{r,n}$ consists of all vectors of length r with distinct values in $\mathcal{N}$. For any $(d_1, \ldots, d_{r+1}) \in \mathbb{T}_{r+1,n+1}$, let $f(d_1, \ldots, d_{r+1}) = e_{(d_1,\ldots,d_r),d_{r+1}}$ denote the transfer error from training samples $S_{d_1}, \ldots, S_{d_r}$ to test sample $S_{d_{r+1}}$.

C.1.2 Proofs of Proposition 6 and Proposition 7

These are special cases of Theorem 1 with $\Gamma = 1$. To avoid repetition, we will only prove Theorem 1.

C.1.3 Proof of Theorem 1 and Corollary 1

We start by proving a simple lemma.

Lemma 2. *Let $H = (1-\pi)F + \pi G$ be a mixture of two distributions F and G. Further let Z_H be a draw from H and Z_F a draw from F. Then, for any $0 \leq \tau_1 < \tau_2 \leq 1$,*

$$(1-\pi)(\tau_2 - \tau_1) \leq \mathbb{P}\left(Z_H \in [\underline{Q}_{\tau_1}(F), \overline{Q}_{\tau_2}(F)]\right) \leq (1-\pi)\mathbb{P}\left(Z_F \in [\underline{Q}_{\tau_1}(F), \overline{Q}_{\tau_2}(F)]\right) + \pi$$

Proof. Let Z_G be a draw from G, and let W be a binary random variable, independent of Z_F and

Z_G, with $\mathbb{E}[W] = \pi$. Then $Z_H \stackrel{d}{=} (1-W)Z_F + WZ_G$ and

$$\begin{aligned}
&\mathbb{P}\left(Z_H \in [\underline{Q}_{\tau_1}(F), \overline{Q}_{\tau_2}(F)]\right) \\
&= (1-\pi)\mathbb{P}\left(Z_F \in [\underline{Q}_{\tau_1}(F), \overline{Q}_{\tau_2}(F)]\right) + \pi\mathbb{P}\left(Z_G \in [\underline{Q}_{\tau_1}(F), \overline{Q}_{\tau_2}(F)]\right) \\
&\in (1-\pi)\mathbb{P}\left(Z_F \in [\underline{Q}_{\tau_1}(F), \overline{Q}_{\tau_2}(F)]\right) + [0, \pi].
\end{aligned}$$

By definition of upper and lower quantiles, $\mathbb{P}\left(Z_F \in [\underline{Q}_{\tau_1}(F), \overline{Q}_{\tau_2}(F)]\right) \geq \tau_2 - \tau_1$. The result then follows. □

Proof of Theorem 1. Throughout the proof we condition on the unordered samples $\{S_1, \ldots, S_{n+1}\}$ and denote by $\{S_{(1)}, \ldots, S_{(n+1)}\}$ any typical realization. Let $\mathcal{F}$ denote the sigma-field generated by the unordered set $\{S_{(1)}, \ldots, S_{(n+1)}\}$. Under the assumed data-generating process,

$$e_{(d_1,\ldots,d_r),n+1} \mid \mathcal{F} \stackrel{d}{=} f(\boldsymbol{\pi}^w(d_1), \ldots, \boldsymbol{\pi}^w(d_r), \boldsymbol{\pi}^w(n+1)), \quad \forall (d_1, \ldots, d_r) \in \mathbb{T}_{r,n}.$$

where $\boldsymbol{\pi}^w$ is a random permutation on $\{1, \ldots, n+1\}$ distributed according to

$$\mathbb{P}(\boldsymbol{\pi}^w = \pi) = \frac{1}{n!} \frac{\omega(S_{\pi(n+1)})}{\sum_{j=1}^{n+1} \omega(S_j)}. \tag{C.1}$$

On the other hand, $\mathbf{T} \stackrel{d}{=} (\boldsymbol{\pi}^n(1), \ldots, \boldsymbol{\pi}^n(r))$ where $\boldsymbol{\pi}^n$ denotes a uniform random permutation on $\{1, \ldots, n\}$, so $e_{\mathbf{T},n+1} \mid \mathcal{F} \stackrel{d}{=} f(\boldsymbol{\pi}^w \circ \boldsymbol{\pi}^n(1), \ldots, \boldsymbol{\pi}^w \circ \boldsymbol{\pi}^n(r), \boldsymbol{\pi}^w(n+1))$. By (C.1), we have

$$e_{\mathbf{T},n+1} \mid \mathcal{F} \stackrel{d}{=} f(\boldsymbol{\pi}^w(1), \ldots, \boldsymbol{\pi}^w(r), \boldsymbol{\pi}^w(n+1)). \tag{C.2}$$

For any $(d_1, \ldots, d_r, k) \in \mathbb{T}_{r+1,n+1}$, let

$$W'_k = \mathbb{P}((\boldsymbol{\pi}^w(1), \ldots, \boldsymbol{\pi}^w(r), \boldsymbol{\pi}^w(n+1)) = (d_1, \ldots, d_r, k)) = \frac{(n-r)!}{n!} \frac{\omega(S_k)}{\sum_{j=1}^{n+1} \omega(S_j)}.$$

Thus,

$$e_{\mathbf{T},n+1} \mid \mathcal{F} \sim \sum_{(d_1,\ldots,d_r,k) \in \mathbb{T}_{r+1,n+1}} W'_k \cdot \delta_{f(d_1,\ldots,d_r,k)}. \tag{C.3}$$

Because $\mathbb{P}(Z \leq \bar{Q}_\tau(F)) \geq \tau$,

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{Q}_\tau \left(\sum_{(d_1,\ldots,d_r,k) \in \mathbb{T}_{r+1,n+1}} W'_k \cdot \delta_{f(d_1,\ldots,d_r,k)} \right) \mid \mathcal{F} \right) \geq \tau. \tag{C.4}$$

Let

$$\Omega_{n+1} = \sum_{(d_1,\ldots,d_r,k) \in \mathbb{T}_{r+1,n+1} \setminus \mathbb{T}_{r+1,n}} W'_k. \tag{C.5}$$

By definition, the element $n+1$ must belong to every tuple $(d_1, \ldots, d_r, k) \in \mathbb{T}_{r+1,n+1} \setminus \mathbb{T}_{r+1,n}$. Thus, W'_{n+1} shows up $|\mathbb{T}_{r+1,n+1} \setminus \mathbb{T}_{r+1,n}|/(r+1)$ times in the sum (C.5). By symmetry, each of the other W_k's shows up $|\mathbb{T}_{r+1,n+1} \setminus \mathbb{T}_{r+1,n}|r/(r+1)n$ times. Since

$$|\mathbb{T}_{r+1,n+1} \setminus \mathbb{T}_{r+1,n}| = |\mathbb{T}_{r+1,n+1}| - |\mathbb{T}_{r+1,n}| = \frac{(n+1)!}{(n-r)!} - \frac{n!}{(n-r-1)!} = (r+1)\frac{n!}{(n-r)!},$$

we obtain that

$$\begin{aligned} \Omega_{n+1} &= \frac{n!}{(n-r)!} W'_{n+1} + \frac{r(n-1)!}{(n-r)!} \sum_{k=1}^{n} W'_k \\ &= \frac{r}{n} + \frac{(n-1)!}{(n-r-1)!} W'_{n+1} = \frac{r}{n} + \frac{n-r}{n} \frac{\omega(S_{n+1})}{\sum_{j=1}^{n+1} \omega(S_j)}. \end{aligned} \tag{C.6}$$

where the second to last equality uses $\sum_{k=1}^{n+1} W'_k = \frac{(n-r)!}{n!}$. By (C.6) and (4.9),

$$W_k = \frac{W'_k}{1-\Omega_{n+1}}.$$

Thus, the distribution in (C.3) can be written as the following mixture distribution

$$\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}} W'_k \cdot \delta_{f(d_1,\ldots,d_r,k)} = (1-\Omega_{n+1})F^{\omega}_{\mathbf{M}} + \Omega_{n+1}\cdot G$$

where $G = \sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}\setminus\mathbb{T}_{r+1,n}} (W'_k/\Omega_{n+1})\cdot\delta_{f(d_1,\ldots,d_r,k)}$. By Lemma 2 with $\tau_1 = 0, \tau_2 = \tau$, and $F = F^{\omega}_{\mathbf{M}}$, we have

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \le \bar{Q}_\tau(F^{\omega}_{\mathbf{M}}) \mid \mathcal{F}\right) \ge \tau(1-\Omega_{n+1}).$$

Moreover, when $f(d_1,\ldots,d_r,k)$ are mutually distinct,

$$\mathbb{P}\left(Z \le \bar{Q}_\tau(F^{\omega}_{\mathbf{M}})\right) \le \tau + \max_k W_k,$$

where Z is the draw from the distribution inside $\bar{Q}_\tau$ and

$$\max_k W_k = \frac{(n-r-1)!}{(n-1)!}\frac{\max_{k\le n}\omega(S_k)}{\sum_{j=1}^{n}\omega(S_j)}.$$

Thus, Lemma 2 implies

$$\begin{aligned}
&\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{Q}_\tau(F^\omega_{\mathbf{M}}) \mid \mathcal{F}\right) \\
&\leq \left(\tau + \frac{(n-r-1)!}{(n-1)!}\frac{\max_{k\leq n}\omega(S_k)}{\sum_{j=1}^n \omega(S_j)}\right)(1-\Omega_{n+1}) + \Omega_{n+1} \\
&= 1 - \left(1 - \tau - \frac{(n-r-1)!}{(n-1)!}\frac{\max_{k\leq n}\omega(S_k)}{\sum_{j=1}^n \omega(S_j)}\right)(1-\Omega_{n+1}).
\end{aligned}$$

The result then follows by the iterated law of expectation and (C.6), which implies

$$1 - \Omega_{n+1} = \frac{n-r}{n}\frac{\sum_{j=1}^n \omega(S_j)}{\sum_{j=1}^{n+1}\omega(S_j)}.$$

The two-sided guarantee can be similarly obtained by Lemma 2 with $\tau_1 = 1-\tau$ and $\tau_2 = \tau$.

To prove Corollary 1, we note that

$$\frac{\omega(S_{n+1})}{\sum_{j=1}^{n+1}\omega(S_j)} \leq \frac{\Gamma}{n\Gamma^{-1}+\Gamma} = \frac{1}{n\Gamma^{-2}+1}.$$

Thus,

$$1 - \Omega_{n+1} = \frac{n-r}{n}\left(1 - \frac{\omega(S_{n+1})}{\sum_{j=1}^{n+1}\omega(S_j)}\right) \geq \frac{n-r}{n}\frac{n\Gamma^{-2}}{n\Gamma^{-2}+1} = \frac{n-r}{n+\Gamma^2}.$$

□

C.2 Other Transfer Problems

Although we have focused on specifications of transfer error that evaluate how well a model transfers from one domain to another, our results apply for the substantially broader class of random variables given in 4.2.2. We discuss below other interesting specifications of $e_{\mathcal{T},d^*}$ and what they

might measure.

C.2.1 Parameter Transfer

When a model has interpretable parameters, we may be interested in whether the parameter values estimated on the training data will be a good proxy for the best-fitting parameters in the target sample.

Example 6 (Effectiveness of a Job Training Program). An economist has estimated the effectiveness of a job training program using a data set from one location (as in Hotz et al. (2005)). How similar would the estimate be if the program were implemented at another location?

Example 7 (Loss Aversion). An economist observes on a data set of choice over lotteries that "losses loom larger than gains," specifically that the loss aversion parameter in Prospect Theory has a value larger than 1. If the economist were to elicit choices over a different set of lotteries, would this qualitative conclusion continue to hold?

Consider any model that can be defined as a set $\mathcal{F}_\Theta = \{f_\theta\}_{\theta\in\Theta}$ of prediction rules $f_\theta : \mathcal{X} \to \mathcal{Y}$, which depend continuously on a parameter θ in a compact parameter space Θ. Given any training data $S_{\mathbf{T}}$, let $\hat{\theta}(S_{\mathbf{T}}) = \operatorname{arginf}_{\theta\in\Theta} \sum_{d\in\mathbf{T}} \frac{|S_d|}{\sum_{d\in\mathbf{T}}|S_d|} e(f_\theta, S_d)$ be the parameter value that minimizes a weighted sum of the errors across the samples in the training data, and let $f_{\hat{\theta}(S_{\mathbf{T}})}$ denote the corresponding prediction rule.[1] To assess parameter variation, first fix a distance metric $d(\theta, \theta')$ (e.g., Euclidean distance) to assess how different two parameter vectors θ and θ' are. Then the transfer error

$$e_{\mathbf{T},n+1} = d\left(\hat{\theta}(S_{\mathbf{T}}), \hat{\theta}(S_{n+1})\right)$$

measures how far the estimated parameters on the training data are from the best-fitting parameters on the target sample.

[1] If there are ties, break them arbitrarily

We can also assess how well a qualitative prediction that is based on the estimated parameters will transfer to the target sample (e.g., a prediction that some coefficient is positive). Let A denote any event that can be described as a function of the parameter θ. Then

$$e_{\mathbf{T},n+1} = \begin{cases} 1 & \text{if } \mathbb{1}\left(\hat{\theta}(S_{\mathbf{T}}) \in A\right) = \mathbb{1}\left(\hat{\theta}(S_{n+1}) \in A\right) \\ 0 & \text{otherwise} \end{cases}$$

is a transfer error which tells us whether the prediction about A based on the training samples also holds in the target sample.

C.2.2 Other Estimation Procedures

In the examples above, a model is trained on r training samples and used to predict properties of a target sample. Our results apply also for other training procedures. To avoid introducing extensive notation, we describe these procedures informally.

Example 8 (Transfer Learning). In *transfer learning* problems in computer science (see e.g., Pan and Yang (2010)), some observations from the target sample are available in addition to the training samples $S_{\mathbf{T}}$. The model or algorithm is trained on these observations jointly, with some specification of how to weight the target sample observations relative to the other training data. The performance of a model estimated in this way is another transfer error.

Example 9 (Transfer of Specific Parameters). While some economic parameters are viewed as constant across domains, other parameters may be viewed as domain-specific. For example, spatial models of trade often have structural parameters (e.g., the elasticity of demand substitution between goods produced in different countries) whose values are set using estimates from another paper, and "fundamentals" (e.g., productivity in each country), which are re-estimated on each sample (see for example Alfaro-Urena et al., 2023). The performance of a model that is estimated and

evaluated in this way is a transfer error.

Example 10 (Using Cross-Validation to Tune Parameters). Our framework can also accommodate training procedures in which cross-validation is used to tune select model parameters. For example, black box algorithms often have a complexity parameter (e.g., the penalization parameter in LASSO or the depth of decision trees in a random forest algorithm). One way of choosing the size of this parameter is based on out-of-sample fit (Hastie et al., 2009; Chetverikov et al., 2021). In our setting, this means holding out one of the training samples to use for testing, training the algorithm on the remaining $r - 1$ training samples, and evaluating fit on the remaining test sample. The chosen complexity parameter is the one that yields the lowest average error across the r possible choices of the test sample. Fixing this value for the complexity parameter, the algorithm is then fit to the entire training data. The performance of such an algorithm on the target sample is a transfer error.

Example 11 (Counterfactual Predictions). One way that economic models are used is to form predictions for outcomes under policy changes that have yet to be implemented. For instance, McFadden (1974) predicted the demand impacts of the then-new BART rapid transit system in the San Franciso Bay Area, and Pathak and Shi (2013) predicted demand for schools under changes to the Boston school choice system. One can generalize our framework to cover the case where each sample S_d is instead a pair of two observations, $S_d = (S_d^0, S_d^1)$. The pre-intervention samples $(S_1^0, \ldots, S_{n+1}^0)$ are drawn i.i.d. from one distribution, while the post-intervention samples $(S_1^1, \ldots, S_{n+1}^1)$ are drawn i.i.d. from another. In this more general setting, a transfer error is any function of the training pairs $\{(S_d^0, S_d^1)\}_{d \in \mathbf{T}}$, the target pair (S_{n+1}^0, S_{n+1}^1), and potentially an independent noise variable.[2]

[2]Our theoretical results generalize completely for transfer errors defined in this way; the main limitation is the difficulty of obtaining sufficiently many pre- and post-intervention pairs. We mention this potential application in the case that such data does eventually become available.

C.3 Extensions and further results

Our main results focus on forecasting realized transfer errors, which is useful when we want to know the range of plausible errors in transferring a given model to a new domain. We now complement those results with procedures for inference focused on population quantities: Section C.3.2 provides confidence intervals for quantiles of the transfer error distribution, and Section C.3.3 provides a confidence interval for the expected transfer error. Since these quantities can be perfectly recovered given data from an infinite number of domains, we expect the lengths of these intervals to vanish as the number of observed domains grows large, unlike the forecast intervals from Section 4.3.2.

C.3.1 Preliminary Lemma

We start by establishing a bound that will be useful in the subsequent construction of confidence intervals. Let

$$U = \frac{(n-k)!}{n!} \sum_{(i_1,\ldots,i_k)\in\mathbb{T}_{k,n}} \phi(Z_{i_1},\ldots,Z_{i_k})$$

be an arbitrary U-statistic of degree k with a bounded (and potentially asymmetric) kernel ϕ that takes values in $[0,1]$.

Definition 10. For every $n, k \in \mathbb{Z}_+$ and $x, y \in \mathbb{R}$, define

$$B_{n,k}(x;y) \equiv \min\left\{ b^1_{n,k}(x;y), b^2_{n,k}(x;y), b^3_{n,k}(x;y) \right\}$$

where

$$b^1_{n,k}(x;y) \equiv \exp\left\{-\lfloor n/k \rfloor \left(x \wedge y \log\left(\frac{x \wedge y}{y}\right) + (1 - x \wedge y)\log\left(\frac{1 - x \wedge y}{1 - y}\right)\right)\right\}$$
$$b^2_{n,k}(x;y) \equiv e \cdot \mathbb{P}\left(\text{Binom}(\lfloor n/k \rfloor; y) \leq \lceil \lfloor n/k \rfloor \cdot x \rceil\right)$$
$$b^3_{n,k}(x;y) \equiv \exp\left\{\min_{\lambda > 0} \frac{n\lambda}{k}\left(x - \frac{\lambda}{\lambda + kG(\lambda)} y\right)\right\}$$

Lemma 3. *If $\phi(Z_1, \ldots, Z_k) \in [0,1]$ almost surely, then $P(U \leq x) \leq B_{n,k}(x; \mathbb{E}(U))$ for every $x \in [0,1]$.*

C.3.2 Quantiles of transfer error

Let F denote the CDF of $e_{\mathbf{T},n+1}$, which we assume is continuous. This section builds a confidence interval for the β-th quantile of F, denoted q_β.

For arbitrary $q \in \mathbb{R}$ and realized metadata $\mathbf{M} = \{S_1, \ldots, S_n\}$, define

$$\varphi(q, \mathbf{M}) = \frac{(n-r-1)!}{n!} \sum_{(d_1,\ldots,d_{r+1}) \in \mathbb{T}_{r+1,n}} \mathbb{I}(e_{(d_1,\ldots,d_r),d_{r+1}}) \leq q)$$

where $\mathbb{I}(\cdot)$ is the indicator function, recalling that $e_{(d_1,\ldots,d_r),d_{r+1}}$ denotes the observed transfer error from samples $(S_{d_1}, \ldots, S_{d_r})$ to sample $S_{d_{r+1}}$. This is the fraction of observed transfer errors in the metadata (from r training samples to one test sample) that are less than q. Then $U_\beta \equiv \varphi(q_\beta, \mathbf{M})$ is a U-statistic where by definition, $\mathbb{E}[U_\beta] = \beta$. Lemma 3 then implies

$$\mathbb{P}(U_\beta \leq x) \leq B_{n,r+1}(x, \beta) \qquad \mathbb{P}(U_\beta \geq x) = \mathbb{P}(1 - U_\beta \leq 1 - x) \leq B_{n,r+1}(1-x, 1-\beta). \quad \text{(C.7)}$$

Definition 11. For any quantile $\beta \in (0,1)$ and confidence level $1 - \alpha \in (0,1)$, let $\hat{u}^+_\beta(\alpha) =$

$\inf\{u : B_{n,r+1}(u;\beta) \geq \alpha\}$ and $\hat{u}_\beta^-(\alpha) = \sup\{u : B_{n,r+1}(1-u;1-\beta) \geq \alpha\}$. Further define $\hat{q}_\beta^{\mathrm{L}}(\alpha) \equiv \min\left\{q : \varphi(q,\mathbf{M}) \geq \hat{u}_\beta^+(\alpha)\right\}$ and $\hat{q}_\beta^{\mathrm{U}}(\alpha) \equiv \max\left\{q : \varphi(q,\mathbf{M}) \leq \hat{u}_r^-(\alpha)\right\}$.

Since $B_{n,r+1}(u;\cdot)$ is right-continuous in u, it follows from (C.7) that $\mathbb{P}(U_\beta < \hat{u}_\beta^+(\alpha)) \leq \alpha$ and $\mathbb{P}(U_\beta > \hat{u}_\beta^-(\alpha)) \leq \alpha$. Since $\varphi(q,\mathbf{M})$ is monotonically increasing in q, the event $\{U_\beta < \hat{u}_\beta^+(\alpha)\}$ is equivalent to $\{q_\beta < \hat{q}_\beta^{\mathrm{L}}(\alpha)\}$, while $\{U_\beta > \hat{u}_\beta^-(\alpha)\}$ is equivalent to $\{q_\beta > \hat{q}_\beta^{\mathrm{U}}(\alpha)\}$. This yields:

Proposition 8. *For any quantile* $\beta \in (0,1)$ *and confidence level* $1-\alpha \in (0,1)$,
$P(q_\beta \leq \hat{q}_\beta^U(\alpha)) \geq 1-\alpha$ *and* $\mathbb{P}\left(q_\beta \in \left[\hat{q}_\beta^L(\alpha/2), \hat{q}_\beta^U(\alpha/2)\right]\right) \geq 1-\alpha$.

Figure C.1 applies Proposition 8 to construct two-sided 81% confidence interval for the median raw transfer error, median transfer shortfall, and median transfer deterioration. As in Figure 4.4, these confidence intervals are substantially wider for the black box algorithms, and have higher upper bounds.

C.3.3 Expected transfer error

This section constructs confidence intervals for the expected transfer error, $\mu \equiv \mathbb{E}(e_{\mathbf{T},n+1})$, under the assumption that transfer errors are uniformly bounded (in which case it is without loss to set $e_{\mathbf{T},n+1} \in [0,1]$). Define the U-statistic

$$U = \frac{(n-r-1)!}{n!} \sum_{(d_1,\ldots,d_{r+1}) \in \mathbb{T}_{r+1,n}} e_{(d_1,\ldots,d_r),d_{r+1}}.$$

Because $\mathbb{E}[U] = \mu$, Lemma 3 implies that $\mathbb{P}(U \leq x) \leq B_{n,r+1}(x,\mu)$ and $\mathbb{P}(U \geq x) \leq B_{n,r+1}(1-x,1-\mu)$ for all $x \in \mathbb{R}$.

Definition 12. For any confidence guarantee $1-\alpha \in (0,1)$, let $\hat{\mu}^+(\alpha) = \sup\{\mu : B_{n,r+1}(U;\mu) \geq \alpha\}$ and $\hat{\mu}^-(\alpha) = \inf\{\mu : B_{n,r+1}(1-U;1-\mu) \geq \alpha\}$.

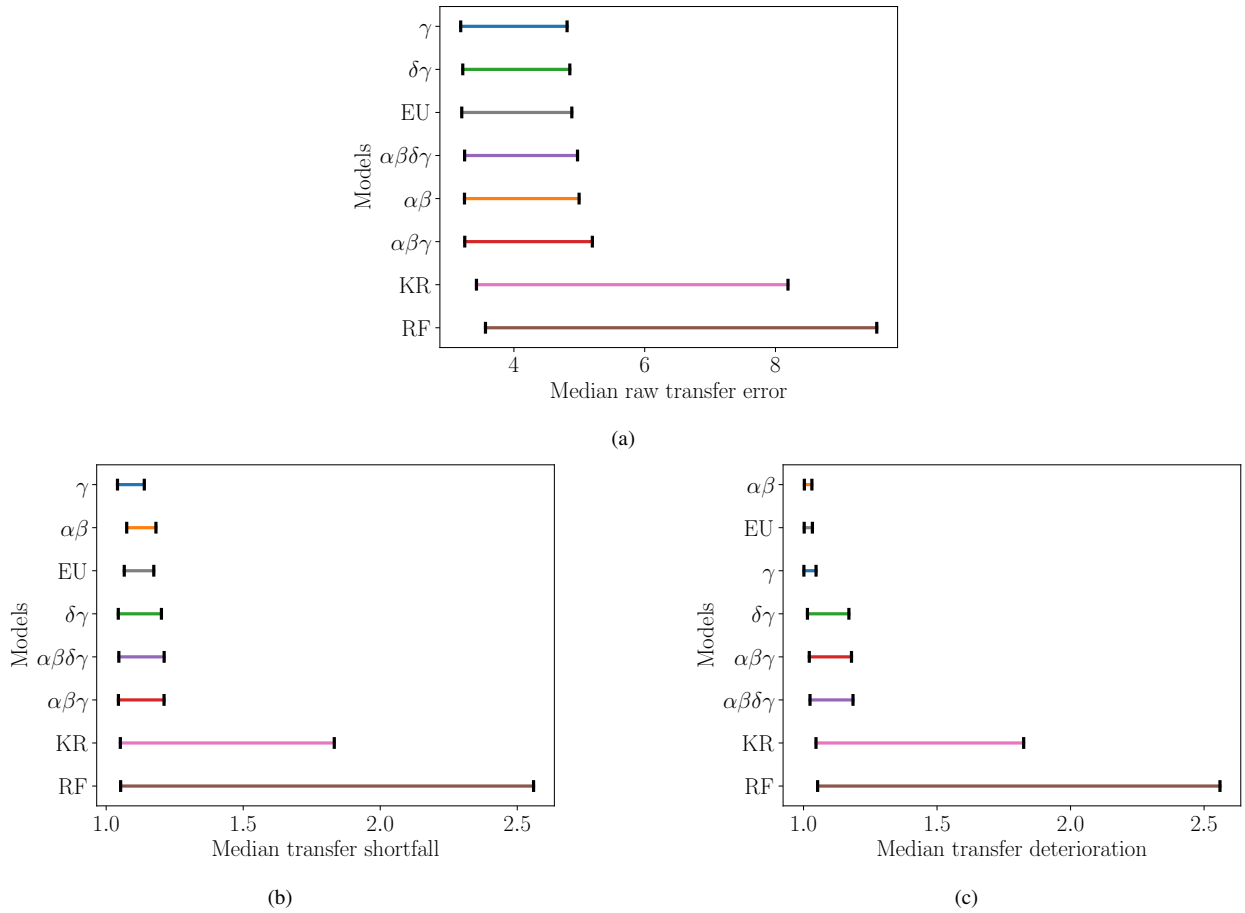


Figure C.1: *81% confidence intervals for the median of (a) raw transfer error, (b) transfer shortfall, and (c) transfer deterioration.*

It follows from (C.7) that $\mathbb{P}(U < \hat{u}^+(\alpha)) \leq \alpha$ and $\mathbb{P}(U > \hat{u}^-(\alpha)) \leq \alpha$, which implies:

Proposition 9. *If* $e_{\mathcal{T},d} \in [0,1]$ *almost surely, then* $\mathbb{P}\left(\mu \leq \hat{\mu}^+(\alpha)\right) \geq 1-\alpha$ *and* $\mathbb{P}\left(\mu \in [\hat{\mu}^-(\alpha/2), \hat{\mu}^+(\alpha/2)]\right) \geq 1-\alpha$.

Figure C.2 applies this result to construct two-sided 81% confidence intervals for the expected transfer errors we considered in Section 4.5.4. Since transfer shortfall and transfer deterioration are not bounded, we report instead confidence intervals for the expectation of their inverses $\frac{\min_{m\in\mathcal{M}} e\left(f^m_{S_{n+1}}, S_{n+1}\right)}{e(f_{S_\mathbf{T}}, S_{n+1})}$ and $\frac{e(f_{S_{n+1}}, S_{n+1})}{e(f_{S_\mathbf{T}}, S_{n+1})}$; lower values for these measures correspond to worse trans-

fer performance. We again find that the confidence intervals for the black box algorithms are qualitatively worse than those for the economic models.

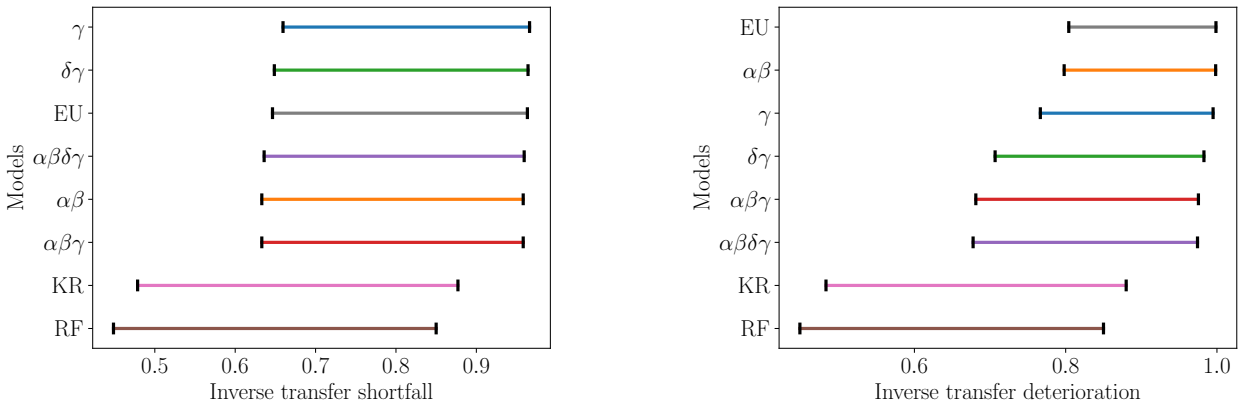


Figure C.2: *81% confidence intervals for (a) expected inverse transfer shortfall, (b) expected inverse transfer deterioration.*

C.3.4 Proof of Lemma 3

Hoeffding (1963) shows that $P(U \leq x) \leq b^1_{n,k}(x, \mathbb{E}(U))$, and Bates et al. (2021) shows that $P(U \leq x) \leq b^2_{n,k}(x, \mathbb{E}(U))$. We now show that if $x \in [0, 1]$ then $P(U \leq x) \leq b^3_{n,k}(x, \mathbb{E}(U))$. To do this, we use a series of intermediate results to extend a result of Bates et al. (2021) on U-statistics of degree 2 with bounded kernels to U-statistics with bounded kernels for any order $k \geq 2$.

Let $Z_1, \ldots, Z_n$ be i.i.d. random variables and $\phi : \mathbb{R}^k \to [0, 1]$ be a bounded function. Then a U-statistic of degree k is defined as

$$U = \frac{(n-k)!}{n!} \sum_{i_1,\ldots,i_k} \phi(Z_{i_1}, \ldots, Z_{i_k}), \tag{C.8}$$

where $\sum_{i_1,\ldots,i_k}$ denotes the sum over all k-tuples in $\mathcal{N}$ with mutually distinct elements. The average of Z_i is a special case of (C.8) with $k = 1$ and $\phi(z) = z$.

Let $m = \lfloor n/k \rfloor$ and $\boldsymbol{\pi}^n : \mathcal{N} \mapsto \mathcal{N}$ be a uniformly random permutation. For each permutation π, define

$$W_\pi = \frac{1}{m} \sum_{j=1}^{m} \phi\left(Z_{\pi((j-1)k+1)}, \ldots, Z_{\pi(jk)}\right).$$

Note that the summands in W_π are independent given π. Then $U = \mathbb{E}_{\boldsymbol{\pi}^n}[W_{\boldsymbol{\pi}^n}]$, where $\mathbb{E}_{\boldsymbol{\pi}^n}$ denotes the expectation with respect to $\boldsymbol{\pi}^n$ when conditioning on $Z_1, \ldots, Z_n$. By Jensen's inequality, for any convex function ψ, $\mathbb{E}[\psi(U)] = \mathbb{E}[\psi(\mathbb{E}_{\boldsymbol{\pi}^n}[W_{\boldsymbol{\pi}^n}])] \leq \mathbb{E}[\mathbb{E}_{\boldsymbol{\pi}^n}\psi(W_{\boldsymbol{\pi}^n})] = \mathbb{E}_{\boldsymbol{\pi}^n}[\mathbb{E}\psi(W_{\boldsymbol{\pi}^n})]$. Since W_π has identical distributions for all π,

$$\mathbb{E}[\psi(U)] \leq \mathbb{E}[\psi(W_{\mathbf{id}})] \tag{C.9}$$

where $\mathbf{id}$ is the permutation that maps each element to itself.

Recalling that Hoeffding's inequality is derived from the moment-generating function $\psi(z) = e^{\lambda z}$ (Hoeffding, 1963), and the Bentkus inequality is derived from the piecewise linear function $\psi(z) = (z - t)_+$ (Bentkus, 2004), the following tail inequalities for U-statistics are a direct consequence of (C.9).

Proposition 10. *Let U be a U-statistic of order k with a bounded kernel $\phi \in [0, 1]$ in the form of* (C.8) *and $m = \lfloor n/k \rfloor$. Then*

1. *(Hoeffding inequality for U-statistics, Section 5 of Hoeffding 1963)*

$$\mathbb{P}(U \leq x) \leq \exp\left\{-m h_1\left(x \wedge \mathbb{E}[U]; \mathbb{E}[U]\right)\right\},$$

where

$$h_1(y; \mu) = y \log\left(\frac{y}{\mu}\right) + (1 - y) \log\left(\frac{1 - y}{1 - \mu}\right).$$

2. (Bentkus inequality for U-statistics, modified from Bentkus 2004)

$$\mathbb{P}\left(U \leq x\right) \leq e\mathbb{P}\left(\mathrm{Bin}\left(m; \mathbb{E}[U]\right) \leq \lceil mx \rceil\right).$$

Other concentration inequalities can be derived from the leave-one-out property. Write $U(Z_1, \ldots, Z_n)$ for U and let $U_i = \inf_{z_i} U(Z_1, \ldots, Z_{i-1}, z_i, Z_{i+1}, \ldots, Z_n)$. Note that U_i is independent of Z_i. Since $\phi(\cdot) \geq 0$, we have $0 \leq U - U_i \leq \frac{(n-k)!}{n!} \sum_{j=1}^{k} \sum_{i_1,\ldots,i_k,i_j=i} \phi(Z_{i_1}, \ldots, Z_{i_k})$ so $\frac{n}{k}(U - U_i) \leq 1$ and

$$\begin{aligned}
\sum_{i=1}^{n}(U - U_i)^2 &\leq \frac{((n-k)!)^2}{(n!)^2} \sum_{i=1}^{n} \left(\sum_{j=1}^{k} \sum_{i_1,\ldots,i_k,i_j=i} \phi(Z_{i_1}, \ldots, Z_{i_k}) \right)^2 \\
&\overset{(i)}{\leq} \frac{k(n-k)!}{n \cdot n!} \sum_{j=1}^{k} \sum_{i=1}^{n} \sum_{i_1,\ldots,i_k,i_j=i} \phi(Z_{i_1}, \ldots, Z_{i_k})^2 \\
&\overset{(ii)}{\leq} \frac{k(n-k)!}{n \cdot n!} \sum_{j=1}^{k} \sum_{i=1}^{n} \sum_{i_1,\ldots,i_k,i_j=i} \phi(Z_{i_1}, \ldots, Z_{i_k}) \\
&= \frac{k^2}{n} U,
\end{aligned}$$

where (i) applies the Cauchy-Schwarz inequality and (ii) uses the fact that $\phi(\cdot) \leq 1$. If we let $W = (n/k)U$ and $W_i = (n/k)U_i$, then $W - W_i \leq 1, \quad \sum_{i=1}^{n}(W - W_i)^2 \leq kW$. This implies that W as a function of $Z_1, \ldots, Z_n$ satisfies the assumptions for the claim (34) in Theorem 13 of Maurer (2006) with constant $a = k$.[3]

Proposition 11 (Theorem 13, Maurer 2006)**.** *Let* $G(\lambda) = (e^{\lambda} - \lambda - 1)/\lambda$. *Then for any* $\lambda > 0$,

$$\log \mathbb{E}[e^{\lambda(\mathbb{E}[W] - W)}] \leq \frac{k\lambda G(\lambda)}{\lambda + kG(\lambda)} \mathbb{E}[W].$$

[3]Theorem 13 of Maurer (2006) states a weaker result that $\log \mathbb{E}[e^{\lambda(\mathbb{E}[W]-W)}] \leq \frac{k\mathbb{E}[W]}{2}\lambda^2$. The stronger version stated here can be found in the second last display in the proof of Theorem 13 of Maurer (2006).

This further implies that for any $x \in (0, \mathbb{E}[U])$*,*

$$\mathbb{P}\left(U \leq x\right) \leq \exp\left\{\min_{\lambda>0} \frac{n\lambda}{k}\left(x - \frac{\lambda}{\lambda + kG(\lambda)}\mathbb{E}[U]\right)\right\}.$$

Putting Proposition 10 and Proposition 11 together yields Lemma 3.

C.4 Supplementary Material to Section 4.4

C.4.1 A more general distribution shift model

In this subsection we discuss a more general distribution shift model that allows $S_1, \ldots, S_n$ to have non-identical distributions. Let p denote the joint density of $S_1, \ldots, S_{n+1}$ (with respect to a dominating measure). Let $\boldsymbol{\pi}^p$ be a random permutation on $\{1, \ldots, n+1\}$ such that, for any realization $\{s_1, \ldots, s_{n+1}\}$ of $\{S_1, \ldots, S_{n+1}\}$,

$$\begin{aligned}
&\mathbb{P}\left(\boldsymbol{\pi}^p(1) = d_1, \ldots, \boldsymbol{\pi}^p(n+1) = d_{n+1} \mid \{S_{(1)}, \ldots, S_{(n+1)}\} = \{s_1, \ldots, s_{n+1}\}\right) \\
&= \frac{p(s_{d_1}, \ldots, s_{d_{n+1}})}{\sum_{(d_1', \ldots, d_{n+1}') \in \mathbb{T}_{n+1,n+1}} p(s_{d_1'}, \ldots, s_{d_{n+1}'})}.
\end{aligned}$$

Then,

$$\begin{aligned}
&(s_{\boldsymbol{\pi}^p(1)}, \ldots, s_{\boldsymbol{\pi}^p(n)}) \mid \{S_{(1)}, \ldots, S_{(n+1)}\} = \{s_1, \ldots, s_{n+1}\} \\
&\overset{d}{=} (S_1, \ldots, S_{n+1}) \mid \{S_{(1)}, \ldots, S_{(n+1)}\} = \{s_1, \ldots, s_{n+1}\}.
\end{aligned}$$

Again let $\mathcal{F}$ denote the sigma-field generated by the unordered set $\{S_{(1)}, \ldots, S_{(n+1)}\}$.

Definition 13. For any $\Gamma \geq 1$, let $\mathcal{P}(\Gamma; r)$ be the class of distributions on $(S_1, \ldots, S_{n+1})$ with

$$\frac{(n+1)!}{(n-r)!}\mathbb{P}\left(\boldsymbol{\pi}^p(1) = d_1, \ldots, \boldsymbol{\pi}^p(r) = d_r, \boldsymbol{\pi}^p(n+1) = k \mid \mathcal{F}\right) \in \left[\Gamma^{-1}, \Gamma\right] \text{ almost surely,}$$

for any $(d_1, \ldots, d_r, k) \in \mathbb{T}_{n+1}$.

Above, $(n-r)!/(n+1)!$ is the probability under a uniform permutation and thus the LHS can be interpreted as the density ratio between $\boldsymbol{\pi}^p$ and a uniform permutation, which measures the deviation from exchangeability.

By (C.1), when $\nu \in \mathcal{W}(\Gamma)$, the joint density p satisfies

$$\begin{aligned}&\frac{(n+1)!}{(n-r)!}\mathbb{P}(\boldsymbol{\pi}^p(1) = d_1, \ldots, \boldsymbol{\pi}^p(r) = d_r, \boldsymbol{\pi}^p(n+1) = k)\\ &= \frac{(n+1)\omega(S_k)}{\sum_{j=1}^{n+1}\omega(S_j)} \leq \frac{(n+1)\Gamma}{n\Gamma^{-1}+\Gamma} = \frac{(n+1)\Gamma^2}{n+\Gamma^2}.\end{aligned}$$

Thus,

$$p \in \mathcal{P}\left(\frac{(n+1)\Gamma^2}{n+\Gamma^2}; r\right) \subset \mathcal{P}\left(\Gamma^2; r\right).$$

Next, we derive forecast intervals akin to Corollary 1.

Theorem 2. *Suppose the joint distribution of* $(S_1, \ldots, S_{n+1})$ *lies in* $\mathcal{P}(\Gamma; r)$*. Define*

$$q_\Gamma(\tau) = 1 - \frac{1-\tau}{\Gamma} \cdot \frac{(n+1) - (r+1)\Gamma}{n-r} + \frac{(n-r-1)!}{n!}.$$

Then

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{e}^{\mathbf{M}}_{q_\Gamma(\tau)}\right) \geq \tau\left(1 - \frac{(r+1)\Gamma}{n+1}\right),$$

and

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \in \left[\bar{e}^{\mathbf{M}}_{1-q_\Gamma(\tau)}, \bar{e}^{\mathbf{M}}_{q_\Gamma(\tau)}\right]\right) \geq (2\tau - 1)\left(1 - \frac{(r+1)\Gamma}{n+1}\right).$$

Proof. For notational convenience, for any $(d_1, \ldots, d_r, k) \in \mathbb{T}_{r+1,n+1}$, let

$$A_{d_1,\ldots,d_r,k} = \mathbb{P}\left(\boldsymbol{\pi}^p(1) = d_1, \ldots, \boldsymbol{\pi}^p(r) = d_r, \boldsymbol{\pi}^p(n+1) = k \mid \mathcal{F}\right).$$

Again, we condition on the unordered samples $S_1, \ldots, S_{n+1}$ and denote by $S_{(1)}, \ldots, S_{(n+1)}$ a typical realization. By similar arguments used to show (C.3), we have

$$e_{\mathbf{T},n+1} \mid \mathcal{F} \sim \sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}} A_{d_1,\ldots,d_r,k} \cdot \delta_{f(d_1,\ldots,d_r,k)}. \tag{C.10}$$

Thus,

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{Q}_\tau\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}} A_{d_1,\ldots,d_r,k} \cdot \delta_{f(d_1,\ldots,d_r,k)}\right) \mid \mathcal{F}\right) \geq \tau.$$

Let

$$\Omega_{n+1} = \sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}\setminus\mathbb{T}_{r+1,n}} A_{d_1,\ldots,d_r,k}.$$

Then we can rewrite the distribution in (C.10) as a mixture distribution

$$\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}} A_{d_1,\ldots,d_r,k} \cdot \delta_{f(d_1,\ldots,d_r,k)} = (1 - \Omega_{n+1}) \cdot F + \Omega_{n+1} \cdot G,$$

where

$$F = \sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n}} \frac{A_{d_1,\ldots,d_r,k}}{1-\Omega_{n+1}} \cdot \delta_{f(d_1,\ldots,d_r,k)}, \quad G = \sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}\setminus\mathbb{T}_{r+1,n}} \frac{A_{d_1,\ldots,d_r,k}}{\Omega_{n+1}} \cdot \delta_{f(d_1,\ldots,d_r,k)}.$$

By Lemma 2 with $\tau_1 = 0, \tau_2 = \tau$, we have that

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{Q}_\tau\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n}} \frac{A_{d_1,\ldots,d_r,k}}{1-\Omega_{n+1}} \cdot \delta_{f(d_1,\ldots,d_r,k)}\right) \mid \mathcal{F}\right) \geq \tau(1-\Omega_{n+1}). \tag{C.11}$$

By definition,

$$A_{d_1,\ldots,d_r,k} \in \frac{(n-r)!}{(n+1)!} \cdot \left[\Gamma^{-1}, \Gamma\right]. \tag{C.12}$$

Moreover,

$$\Omega_{n+1} \in [\Gamma^{-1}, \Gamma] \cdot \frac{(n-r)!}{(n+1)!} \cdot (|\mathbb{T}_{r+1,n+1}| - |\mathbb{T}_{r+1,n}|) = \frac{(r+1)}{n+1} \cdot [\Gamma^{-1}, \Gamma]. \tag{C.13}$$

Sort the elements in $\{f(d_1,\ldots,d_r,k) : (d_1,\ldots,d_r,k) \in \mathbb{T}_{r+1,n}\}$ as

$$f_{(1)} \leq f_{(2)} \leq \ldots \leq f_{(|\mathbb{T}_{r+1,n}|)},$$

where $f_{(j)} = f(d_1^{(j)},\ldots,d_r^{(j)},k^{(j)})$. Write $A_{(j)}$ for $A_{d_1^{(j)},\ldots,d_r^{(j)},k^{(j)}}$ and N for $|\mathbb{T}_{r+1,n}| = n!/(n-r-1)!$. Then

$$\bar{Q}_\tau\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n}} \frac{A_{d_1,\ldots,d_r,k}}{1-\Omega_{n+1}} \cdot \delta_{f(d_1,\ldots,d_r,k)}\right) = f_{(N-L_\tau)},$$

where

$$L_\tau = \min\left\{L : \frac{1}{1-\Omega_{n+1}} \sum_{i=N-L}^{N} A_{(i)} > 1-\tau\right\}.$$

By (C.12) and (C.13),

$$\begin{aligned}
&\frac{(n-r)!}{(n+1)!}\Gamma \cdot \frac{L_\tau + 1}{1 - \frac{r+1}{n+1}\Gamma} > 1 - \tau \\
&\iff \frac{n-r}{n+1}\frac{\Gamma}{1 - \frac{r+1}{n+1}\Gamma}\frac{L_\tau + 1}{N} > 1 - \tau \\
&\iff \frac{L_\tau}{N} > \frac{(1-\tau)\{(n+1) - (r+1)\Gamma\}}{(n-r)\Gamma} - \frac{1}{N} \\
&\iff \frac{N - L_\tau}{N} < 1 - \frac{1-\tau}{\Gamma} \cdot \frac{(n+1) - (r+1)\Gamma}{n-r} + \frac{1}{N} = q_\Gamma(\tau)
\end{aligned}$$

Thus,

$$\begin{aligned}
&\bar{Q}_\tau \left(\sum_{(d_1,\ldots,d_r,k) \in \mathbb{T}_{r+1,n}} \frac{A_{d_1,\ldots,d_r,k}}{1 - \Omega_{n+1}} \cdot \delta_{f(d_1,\ldots,d_r,k)} \right) \\
&\leq \bar{Q}_{q_\Gamma(\tau)} \left(\frac{(n-r-1)!}{n!} \sum_{(d_1,\ldots,d_r,k) \in \mathbb{T}_{r+1,n}} \delta_{f(d_1,\ldots,d_r,k)} \right) = \bar{e}^{\mathbf{M}}_{q_\Gamma(\tau)},
\end{aligned}$$

By (C.11) and (C.13),

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{e}^{\mathbf{M}}_{q_\Gamma(\tau)} \mid \mathcal{F} \right) \geq \tau(1 - \Omega_{n+1}) \geq \tau \left(1 - \frac{r+1}{n+1}\Gamma \right).$$

Thus, the result for the one-sided interval is proved.

The result for the two-sided interval can be proved similarly by Lemma 2 with $\tau_1 = 1 - \tau, \tau_2 = \tau$ and by using the symmetry. □

C.4.2 Algorithm for evaluating worst-case dominance

We provide an algorithm that computes $\bar{e}_\tau(\Gamma)$ with a single τ in $O(rn^{r+1}\log n)$ time and computes $\bar{e}_\tau(\Gamma)$ for all $\tau \in (0,1)$ in $O(rn^{r+1}\log n + n^{r+2})$ time. First, sort the elements in $\{f(d_1,\ldots,d_{r+1}) :$

$(d_1, \ldots, d_{r+1}) \in \mathbb{T}_{r+1,n}\}$ as

$$f_{(1)} \leq f_{(2)} \leq \ldots \leq f_{(|\mathbb{T}_{r+1,n}|)},$$

where

$$f_{(j)} = f(d^{(j)}), \quad d^{(j)} = (d_1^{(j)}, \ldots, d_{r+1}^{(j)}) \in \mathbb{T}_{r+1,n}.$$

Let $\psi^{(j)} \in \{0,1\}^n$ with

$$\psi_i^{(j)} = I\left(d_{r+1}^{(j)} = i\right).$$

Further define the cumulative sum of $\psi^{(j)}$ as

$$\Psi^{(j)} = \sum_{\ell=1}^{j} \psi_i^{(\ell)}.$$

Let $w = (\omega(S_1), \ldots, \omega(S_n))^T$ and $\mathbf{1}_n = (1, 1, \ldots, 1)^T$. By (4.9), for each j,

$$f_{(j)} \geq \bar{e}_\tau^{\mathbf{M},\omega} \Longleftrightarrow \frac{(n-r-1)!}{(n-1)!} \frac{w^T \Psi^{(j)}}{w^T \mathbf{1}_n} \geq \tau.$$

Therefore,

$$\bar{e}_\tau^{\mathbf{M},\omega} = f_{J_\tau^\omega}, \quad \text{where } J_\tau^\omega = \min\left\{j : \frac{w^T \Psi^{(j)}}{w^T \mathbf{1}_n} \geq \tau \frac{(n-1)!}{(n-r-1)!}\right\}.$$

By definition, the set of w generated by all $\omega \in \mathcal{W}(\Gamma)$ is $[\Gamma^{-1}, \Gamma]^n$. Thus,

$$\bar{e}_\tau^{\mathbf{M}}(\Gamma) = f_{J_\tau(\Gamma)}, \quad \text{where } J_\tau(\Gamma) = \min\left\{j : \min_{w \in [\Gamma^{-1}, \Gamma]^n} \frac{w^T \Psi^{(j)}}{w^T \mathbf{1}_n} \geq \tau \frac{(n-1)!}{(n-r-1)!}\right\}. \tag{C.14}$$

Via some algebra, we can further simplify the expression of $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$.

Theorem 3. *Let $\bar{\Psi}_k^{(j)}$ be the average of the k-smallest coordinates of $\Psi^{(j)}$ and*

$$Q_j(\Gamma) = \frac{j}{n} + \min_{k\in\{1,\dots,n\}} \frac{\bar{\Psi}_k^{(j)} - \frac{j}{n}}{1 + \frac{n}{k(\Gamma^2-1)}}.$$

Then $Q_j(\Gamma)$ is strictly increasing in both j and Γ. Moreover, $\bar{e}_\tau^{\mathbf{M}}(\Gamma) = f_{(J_\tau(\Gamma))}$, where

$$J_\tau(\Gamma) = \min\left\{ j \geq \tau\frac{n!}{(n-r-1)!} : Q_j(\Gamma) \geq \tau\frac{(n-1)!}{(n-r-1)!} \right\}.$$

Proof. First, we prove that

$$\min_{w\in[\Gamma^{-1},\Gamma]^n} \frac{w^T\Psi^{(j)}}{w^T\mathbf{1}_n} = \min_{w\in\{\Gamma^{-1},\Gamma\}^n} \frac{w^T\Psi^{(j)}}{w^T\mathbf{1}_n}. \tag{C.15}$$

Let $g_j(w) = w^T\Psi^{(j)}/w^T\mathbf{1}_n$. Then g_j is continuous and bounded on the closed set $[\Gamma^{-1},\Gamma]^n$ and thus the minimum can be achieved. Let

$$w^{(j)}(\Gamma) = \operatorname*{argmin}_{w: g_j(w)=\min_{w\in[\Gamma^{-1},\Gamma]^n} g_j(w)} \sum_{i=1}^n \min\left\{|w_i-\Gamma|, |w_i-\Gamma^{-1}|\right\}.$$

Suppose there exists $i\in\mathcal{N}$ such that $w_i^{(j)}(\Gamma)\in(\Gamma^{-1},\Gamma)$. Then

$$g_j(w_i, w_{-i}) = \frac{\Psi_i^{(j)}w_i + \Psi_{-i}^{(j)T}w_{-i}}{w_i + \mathbf{1}_{n-1}^T w_{-i}} = \Psi_i^{(j)} + \frac{\Psi_{-i}^{(j)T}w_{-i} - \Psi_i^{(j)}\cdot\mathbf{1}_{n-1}^T w_{-i}}{w_i + \mathbf{1}_{n-1}^T w_{-i}},$$

where $\Psi_{-i}^{(j)}$ and w_{-i} are the leave-i-th-entry subvectors of $\Psi^{(j)}$ and w. Clearly, g_j is a monotone function of w_i for any given w_{-i}. Since $w^{(j)}(\Gamma)$ is a minimizer and $w_i^{(j)}(\Gamma)\in(\Gamma^{-1},\Gamma)$, we must have $\Psi_{-i}^{(j)T}w_{-i} - \Psi_i^{(j)}\cdot\mathbf{1}_{n-1}^T w_{-i} = 0$. Define $\widetilde{w}^{(j)}(\Gamma)$ with

$$\widetilde{w}_i^{(j)}(\Gamma) = \Gamma, \quad \widetilde{w}_{-i}^{(j)}(\Gamma) = w_{-i}^{(j)}(\Gamma).$$

Then

$$g_j(\widetilde{w}^{(j)}(\Gamma)) = g_j(w^{(j)}(\Gamma)) = \min_{w\in[\Gamma^{-1},\Gamma]^n} g_j(w),$$

while

$$\sum_{i=1}^{n} \min\left\{|\widetilde{w}_i^{(j)}(\Gamma) - \Gamma|, |\widetilde{w}_i^{(j)}(\Gamma) - \Gamma^{-1}|\right\} < \sum_{i=1}^{n} \min\left\{|w_i^{(j)}(\Gamma) - \Gamma|, |w_i^{(j)}(\Gamma) - \Gamma^{-1}|\right\}.$$

This contradicts the definition of $w^{(j)}(\Gamma)$, so $w^{(j)}(\Gamma) \in \{\Gamma^{-1}, \Gamma\}^n$, which completes the proof of (C.15).

For any $w \in \{\Gamma^{-1}, \Gamma\}^n$ with $|\{i : w_i = \Gamma\}| = k$, the Fréchet-Hoeffding inequality implies that Γ's are allocated to the k smallest entries of $\Psi^{(j)}$. Thus,

$$\min_{w\in\{\Gamma^{-1},\Gamma\}^n} \frac{w^T\Psi^{(j)}}{w^T\mathbf{1}_n} = \min_{k\in\mathcal{N}\cup\{0\}} \frac{\Gamma k\bar{\Psi}_k^{(j)} + \Gamma^{-1}(\mathbf{1}_n^T\Psi_i^{(j)} - k\bar{\Psi}_k^{(j)})}{\Gamma k + \Gamma^{-1}(n-k)}.$$

By definition, $\mathbf{1}_n^T\Psi_i^{(j)} = j$. Then for each k, the above expression can be simplified as

$$\begin{aligned}
&\frac{\Gamma k\bar{\Psi}_k^{(j)} + \Gamma^{-1}(\mathbf{1}_n^T\Psi_i^{(j)} - k\bar{\Psi}_k^{(j)})}{\Gamma k + \Gamma^{-1}(n-k)} = \frac{\Gamma k\bar{\Psi}_k^{(j)} + \Gamma^{-1}(j - k\bar{\Psi}_k^{(j)})}{\Gamma k + \Gamma^{-1}(n-k)} \\
&= \frac{(\Gamma - \Gamma^{-1})k\bar{\Psi}_k^{(j)} + \Gamma^{-1}j}{(\Gamma - \Gamma^{-1})k + \Gamma^{-1}n} = \frac{j}{n} + \frac{(\Gamma - \Gamma^{-1})k\left(\bar{\Psi}_k^{(j)} - \frac{j}{n}\right)}{(\Gamma - \Gamma^{-1})k + \Gamma^{-1}n} \\
&= \frac{j}{n} + \frac{\bar{\Psi}_k^{(j)} - \frac{j}{n}}{1 + \frac{n}{k(\Gamma^2-1)}}.
\end{aligned}$$

The above expression is j/n for both $k = n$ and $k = 0$, so we can remove 0 from the minimum, and thus

$$\min_{w\in[\Gamma^{-1},\Gamma]^n} \frac{w^T\Psi^{(j)}}{w^T\mathbf{1}_n} = Q_j(\Gamma).$$

By (C.14),

$$J_\tau(\Gamma) = \min\left\{ j : Q_j(\Gamma) \geq \tau \frac{(n-1)!}{(n-r-1)!} \right\}.$$

Finally, we can restrict to $j \geq \tau n!/(n-r-1)!$ because

$$Q_j(\Gamma) = \frac{j}{n} + \min_{k\in\{1,\ldots,n\}} \frac{\bar{\Psi}_k^{(j)} - \frac{j}{n}}{1 + \frac{n}{k(\Gamma^2-1)}} \leq \frac{j}{n} + \frac{\bar{\Psi}_n^{(j)} - \frac{n}{n}}{1 + \frac{n}{n(\Gamma^2-1)}} = \frac{j}{n}.$$

□

Since $Q_j(\Gamma)$ is increasing in j, $J_\tau(\Gamma)$ can be found via binary search with iteration complexity $O(\log n^{r+1}) = O(r \log n)$. Each iteration costs at most $O(n)$ operations to sort the entries of $\Psi^{(j)}$ based on the ordered version of $\Psi^{(j-1)}$, since there is only entry updated, and $O(n)$ additional operations to compute $Q_j(\Gamma)$. Thus, the overall computational overhead after obtaining $(f_{(1)}, \ldots, f_{(|\mathbb{T}_{r+1,n}|)})$ is just $O(rn \log n)$, which is much smaller than the cost of sorting f-values $O(n^{r+1} \log n^{r+1}) = O(rn^{r+1} \log n)$.

In some cases, we want to compute $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$ for all $\tau \in [0,1]$ at once. The following result links $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$ to an induced distribution on the f's.

Corollary 2. *For any $\Gamma \geq 1$, let μ_Γ be a weighted measure with*

$$\mu_\Gamma = \sum_{j=1}^{|\mathbb{T}_{r+1,n}|} \frac{(n-r-1)!}{(n-1)!} (Q_j(\Gamma) - Q_{j-1}(\Gamma)) \cdot \delta_{f_{(j)}},$$

where $Q_0(\Gamma) = 0$. Then $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$ is the τ-th quantile of μ_Γ.

Since the ordering takes $O(rn \log n)$ time and computing each $Q_j(\Gamma)$ takes $O(n)$ time, the total computational cost to compute $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$ for all $\tau \in [0,1]$ is $O(rn^{r+1} \log n + n^{r+2})$.

C.4.3 Algorithm for evaluating everywhere dominance

Let $f_{(j),1}$ and $f_{(j),2}$ be the j-th largest transfer errors for method 1 and 2, respectively. Similarly, the count vectors for two methods are denoted by $\Psi^{(j),1}$ and $\Psi^{(j),2}$. Then method 1 does NOT everywhere dominate method 2 at the τ-th quantile if and only if there exists $j_1, j_2 \in \{1, \ldots, |\mathbb{T}_{r+1,n}|\}$ and $w \in [0, \infty)^n$ such that

$$f_{(j_1),1} > f_{(j_2),2}, \quad \frac{(n-r-1)!}{(n-1)!} \frac{w^T \Psi^{(j_1-1),1}}{w^T \mathbf{1}_n} < \tau \leq \frac{(n-r-1)!}{(n-1)!} \frac{w^T \Psi^{(j_2),2}}{w^T \mathbf{1}_n}. \tag{C.16}$$

Above $\Psi^{(0),1} = (0, 0, \ldots, 0)^T$.

To avoid pairwise comparisons, which incur $O(n^{2(r+1)})$ computation, we can check (C.16) by only focusing on $j_1 = m(j), j_2 = j$ where

$$m(j) = \min\{j' : f_{(j'),1} > f_{(j),2}\}.$$

It is easy to see that (C.16) holds for some pair $(j_1, j_2) \in \{1, \ldots, |\mathbb{T}_{r+1,n}|\}^2$ if and only if it holds for $(m(j), j)$ for some $j \in \{1, \ldots, |\mathbb{T}_{r+1,n}|\}$. For any given j, (C.16) reduces to

$$\frac{(n-r-1)!}{(n-1)!} \frac{w^T \Psi^{(m(j)-1),1}}{w^T \mathbf{1}_n} < \tau \leq \frac{(n-r-1)!}{(n-1)!} \frac{w^T \Psi^{(j),2}}{w^T \mathbf{1}_n}, \quad w \in [0, \infty)^n.$$

This is equivalent to solving the following linear fractional programming problem and then checking if the objective is below τ:

$$\min \frac{w^T a^{(j)}}{w^T \mathbf{1}_n}, \quad \text{s.t.,} \frac{w^T b^{(j)}}{w^T \mathbf{1}_n} \geq \tau, \quad w \in [0, \infty)^n,$$

where

$$a^{(j)} = \Psi^{(m(j)-1),1} \cdot \frac{(n-r-1)!}{(n-1)!}, \quad b^{(j)} = \Psi^{(j),2} \cdot \frac{(n-r-1)!}{(n-1)!}.$$

We can apply the Charnes-Cooper transformation (Charnes and Cooper, 1962) by introducing $v = w/w^T\mathbf{1}_n$ to transform it into a linear programming problem:

$$\min v^T a^{(j)}, \quad \text{s.t., } v^T b^{(j)} \geq \tau, v^T \mathbf{1}_n = 1, v \in [0,\infty)^n. \tag{C.17}$$

Solving these $O(n^{r+1})$ LP problems can be accelerated by the following two observations:

1. Using the same argument as in the last step of the proof of Theorem 3, we can restrict

$$j \geq \tau \frac{n!}{(n-r-1)!}.$$

2. When $a_i^{(j)} \geq b_i^{(j)}$ for every $i \in \mathcal{N}$, then the objective of (C.17) can never be below τ.

C.5 Supplementary material for Section 4.5

C.5.1 Description of data

We briefly describe the individual samples in our meta-data.

Table C.1: *Description of Data.*

Source of Data	# Obs	# Subj	# Lottery	Country	Gains Only
Abdellaoui et al. (2015)	801	89	3	France	Y
Fan et al. (2019)	4750	125	19	US	Y
Bouchouicha and Vieider (2017)	3162	94	66	UK	N
Sutter et al. (2013)	661	661	4	Austria	Y
Etchart-Vincent and l'Haridon (2011)	3036	46	20	France	N

Fehr-Duda et al. (2010)	8560	153	56	China	N
Lefebvre et al. (2010)	72	72	2	France	Y
Halevy (2007)	366	122	2	Canada	Y
Anderhub et al. (2001)	183	61	1	Israel	Y
Murad et al. (2016)	2131	86	25	UK	Y
Dean and Ortoleva (2019)	1032	179	3	US	Y
Bernheim and Sprenger (2020)	1071	153	7	US	Y
Bruhin et al. (2010)	8906	179	50	Switzerland	N
Bruhin et al. (2010)	4669	118	40	Switzerland	N
l'Haridon and Vieider (2019)	1708	61	28	Australia	N
l'Haridon and Vieider (2019)	2548	95	28	Belgium	N
l'Haridon and Vieider (2019)	2350	84	28	Brazil	N
l'Haridon and Vieider (2019)	2240	80	28	Cambodia	N
l'Haridon and Vieider (2019)	2687	96	28	Chile	N
l'Haridon and Vieider (2019)	5711	204	28	China	N
l'Haridon and Vieider (2019)	3072	128	24	Colombia	N
l'Haridon and Vieider (2019)	2968	106	28	Costa Rica	N
l'Haridon and Vieider (2019)	2770	99	28	Czech Republic	N
l'Haridon and Vieider (2019)	3906	140	28	Ethiopia	N
l'Haridon and Vieider (2019)	2604	93	28	France	N
l'Haridon and Vieider (2019)	3639	130	28	Germany	N
l'Haridon and Vieider (2019)	2352	84	28	Guatemala	N
l'Haridon and Vieider (2019)	2492	89	28	India	N
l'Haridon and Vieider (2019)	2352	84	28	Japan	N
l'Haridon and Vieider (2019)	2716	97	28	Kyrgyzstan	N
l'Haridon and Vieider (2019)	1791	64	28	Malaysia	N
l'Haridon and Vieider (2019)	3360	120	28	Nicaragua	N
l'Haridon and Vieider (2019)	5638	202	28	Nigeria	N
l'Haridon and Vieider (2019)	2660	95	28	Peru	N

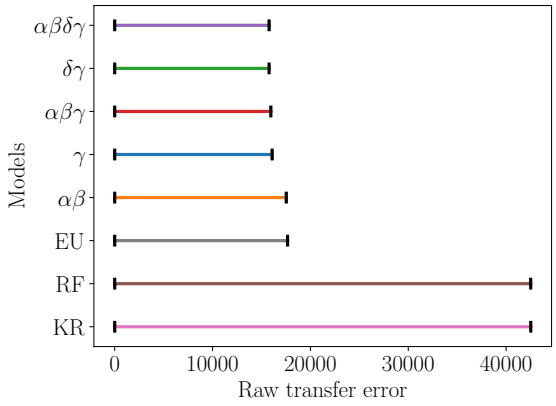
l'Haridon and Vieider (2019)	2491	89	28	Poland	N
l'Haridon and Vieider (2019)	1959	70	28	Russia	N
l'Haridon and Vieider (2019)	1819	65	28	Saudi Arabia	N
l'Haridon and Vieider (2019)	1988	71	28	South Africa	N
l'Haridon and Vieider (2019)	2240	80	28	Spain	N
l'Haridon and Vieider (2019)	2212	79	28	Thailand	N
l'Haridon and Vieider (2019)	2070	74	28	Tunisia	N
l'Haridon and Vieider (2019)	2240	80	28	UK	N
l'Haridon and Vieider (2019)	2701	97	28	US	N
l'Haridon and Vieider (2019)	2436	87	28	Vietnam	N

C.5.2 Papers as domains

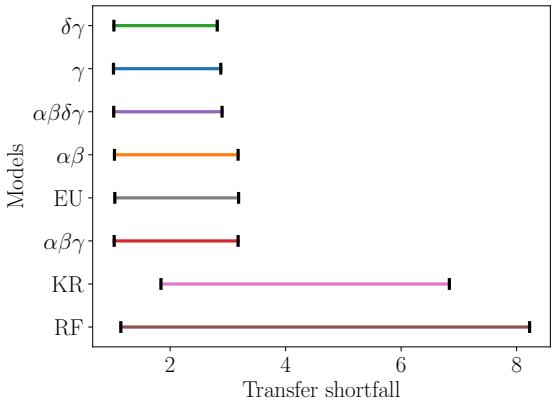
We now consider an alternative definition of domains, with each of the 14 papers representing a different domain. This changes the content of the i.i.d. assumption imposed in Section 4.3.2, where we now assume that samples are i.i.d. across papers, but may be dependent across subject pools within the same paper. We repeat our main analysis and report 78% two-sided forecast intervals in Figure C.3. These intervals are qualitatively similar to those reported in Figure 4.4.

C.5.3 Supplementary tables and figures for main analysis

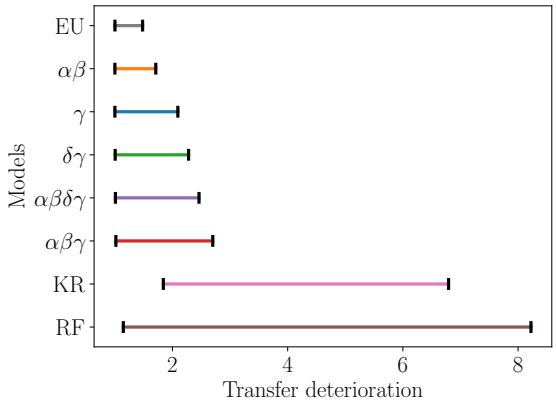
Table C.2 reports the forecast intervals that are depicted in Figure 4.4.



(a) *Raw transfer error*



(b) *Transfer shortfall*



(c) *Transfer deterioration*

Figure C.3: *78% (n=14, $\tau = 0.95$) forecast intervals for each of the three measures, treating each paper as a separate domain.*

Model	Raw Transfer Error	Transfer Shortfall	Transfer Deterioration
CPT variants			
γ	[2.50,15.83]	[1.03,2.54]	[1.00,1.47]
α, β	[2.56,16.13]	[1.04,2.35]	[1.00,1.30]
δ, γ	[2.48,17.19]	[1.02,2.47]	[1.00,1.53]
α, β, γ	[2.47,15.91]	[1.02,2.60]	[1.00,1.85]
$\alpha, \beta, \delta, \gamma$	[2.46,15.99]	[1.02,2.62]	[1.00,1.82]
EU models			
EU	[2.56,16.41]	[1.04,2.14]	[1.00,1.30]
ML algorithms			
Random Forest	[2.71,31.39]	[1.02,6.42]	[1.02,6.42]
Kernel Regression	[2.75,33.62]	[1.02,5.33]	[1.01,5.29]

Table C.2: *86% (n=44, $\tau = 0.95$) forecast intervals*

C.5.4 Alternative forecast intervals

In this section, we report alternative forecast intervals for our three measures. Table C.3 constructs 96% two-sided forecast intervals (setting $\tau = 1$),[4] and Table C.4 reports 91% one-sided forecast intervals (setting $\tau = 0.95$). All of the forecast intervals are qualitatively similar to the 86% two-sided forecast intervals reported in the main text.

Finally, Figure C.4 plots the τ-th percentile of the pooled transfer errors as τ varies. The figure shows that the qualitative conclusions we have drawn about the relative performance of black boxes

[4]The lower bounds of these intervals are the minimum transfer error (among the pooled transfer errors) and the upper bounds are the maximum transfer error.

Model	Raw Transfer Error	Transfer Shortfall	Transfer Deterioration
CPT main variants			
γ	[0.81,23104.96]	[1.01,7.31]	[1.00,7.22]
α, β	[0.71,19999.41]	[1.00,5.28]	[1.00,5.27]
δ, γ	[0.71,23052.76]	[1.00,7.25]	[1.00,7.18]
α, β, γ	[0.71,28122.26]	[1.00,5.65]	[1.00,5.60]
$\alpha, \beta, \delta, \gamma$	[0.71,27959.10]	[1.00,6.01]	[1.00,5.95]
EU models			
EU	[0.72,22787.99]	[1.00,4.44]	[1.00,1.75]
ML algorithms			
Random Forest	[0.96,42520.49]	[1.01,33.17]	[1.01,33.17]
Kernel Regression	[1.01,42519.23]	[1.01,6.835]	[1.00,6.79]

Table C.3: *96% (n=44, $\tau = 1$) two-sided forecast intervals*

and economic models are not specific to any choice of τ.[5] In fact, in Panels (a) and (c), the black box curves lie everywhere above the CPT and EU curves, so both the lower and upper bounds of the black boxes' forecast intervals are higher than those of the economic models for every choice of τ.

C.5.5 Forecast intervals for the ratio of raw RF and CPT transfer errors

Let $e_{\mathcal{T},d}$ be the ratio of the raw random forest transfer error to the raw CPT transfer error (i.e., using the specification in (4.1)), henceforth the *transfer error ratio*.

Panel (a) of Figure C.5 reports 86% two-sided forecast intervals for the raw transfer error ratio for each CPT specification. The lower bound for each CPT model is approximately 0.9, while the upper bound is as large as 4.5. Panel (b) of the figure is a histogram of raw transfer error ratios for the 4-parameter CPT model when the training domains $\mathcal{T}$ and the target domains d are drawn uniformly at random from the set of domains in the meta-data. This distribution has a large cluster

[5]To improve readability, we remove extreme numbers by truncating $\tau \in [5, 95]$, and show results only for the $\alpha\beta\gamma\delta$ specification of the CPT model.

Model	Raw Transfer Error	Transfer Shortfall	Transfer Deterioration
CPT main variants			
γ	[0,15.83]	[1,2.54]	[1,1.47]
α, β	[0,16.13]	[1,2.35]	[1,1.30]
δ, γ	[0,17.19]	[1,2.47]	[1,1.53]
α, β, γ	[0,15.91]	[1,2.60]	[1,1.85]
$\alpha, \beta, \delta, \gamma$	[0,15.99]	[1,2.62]	[1,1.82]
EU models			
EU	[0,16.41]	[1,2.14]	[1,1.30]
ML algorithms			
Random Forest	[0,31.39]	[1,6.42]	[1,6.42]
Kernel Regression	[0,33.62]	[1,5.33]	[1,5.29]

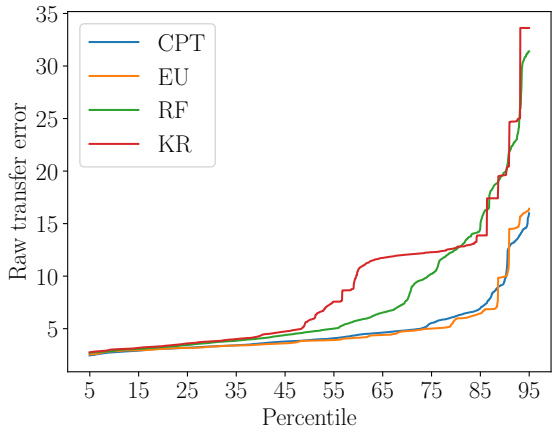
Table C.4: *91% (n=44, $\tau = 0.95$) one-sided forecast intervals*

of ratios around 1 (i.e., raw CPT transfer errors are similar to the raw random forest errors) and a long right tail of ratios achieving a max value of 32.8 (i.e., the random forest error can be up to 32 times as large as the CPT error). The cumulative distribution function of $e_{\mathcal{T},d}$, reported in Panel (c) of Figure C.5, shows that the random forest algorithm outperforms CPT in approximately 35% of $(\mathcal{T}, d)$ pairs, although CPT rarely has a much worse raw transfer error than the random forest and is sometimes much better.

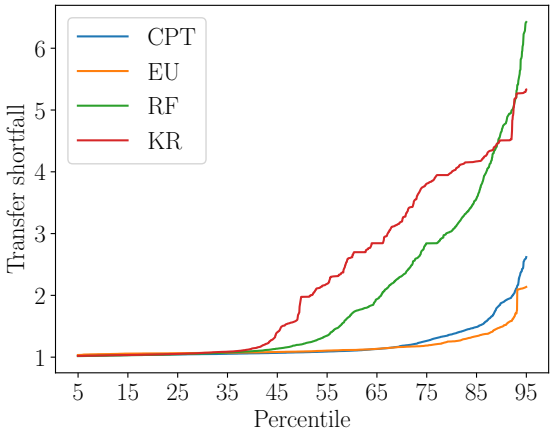
C.5.6 Alternative Choice of r

Here we consider an alternative choice for the number of training domains, setting $r = 3$ instead of $r = 1$.

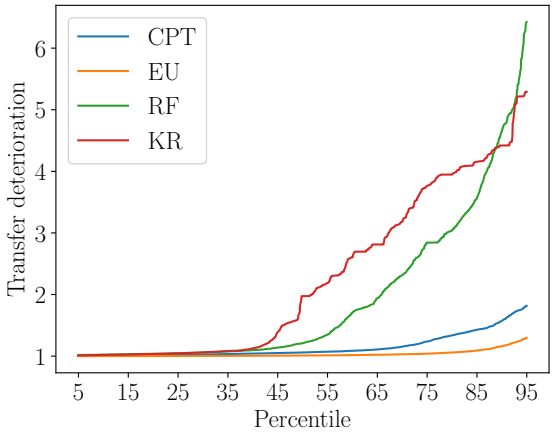
This corresponds to randomly choosing 3 of the 44 domains to be the training domains, finding the best prediction rule for this pooled data, and using the estimated prediction rule to predict the remaining 41 samples. For this analysis we use domain cross-validation to select tuning parameters for the black box algorithms, as described in Example 10.



(a) *Raw transfer error*



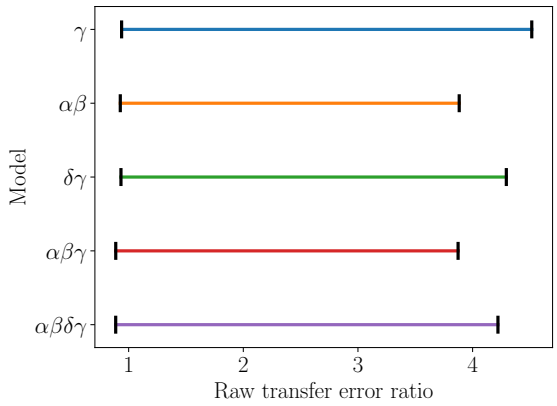
(b) *Transfer shortfall*



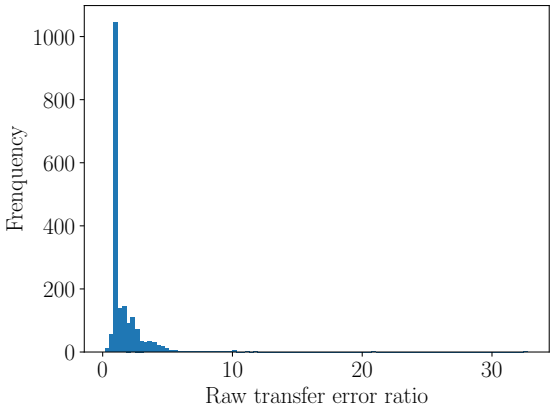
(c) *Transfer deterioration*

Figure C.4: *Error percentiles from 5 to 95 (truncated for readability).*

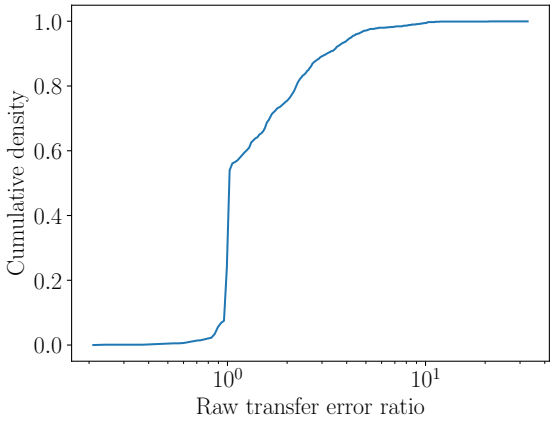
Figure C.6 is the analog of Figure 4.4. Again we choose $\tau = 0.95$, thus constructing forecast intervals whose lower bounds are the 5% percentile of pooled transfer errors, and whose upper bounds are the 95% percentile of pooled transfer errors. Applying Proposition 6, these are 82% forecast intervals. The most notable change is that the random forest forecast interval shrinks considerably, which suggests that the raw transfer error of the random forest algorithm becomes less variable when it is trained on more domains. Otherwise, all of the qualitative statements in the main text for $r = 1$ continue to hold. In particular, as with $r = 1$, we find that the forecast intervals



(a) *86% (n=44, $\tau = 0.95$) Forecast intervals*



(b) *Density*



(c) *CDF*

Figure C.5: *Forecast intervals, density, and cdf for the ratio of the raw random forest transfer error to the raw CPT transfer error.*

for all three of our measures have higher lower and upper bounds for the black box algorithms than for the CPT specifications.

C.5.7 Supplementary Material to Section 4.5.5

Here we consider an alternative choice of for the number of training samples, setting $r = 3$ and $r = 5$ instead $r = 1$. Recalling that each sample includes the observations associated with a unique lottery, this corresponds to randomly choosing three (or five) of the 24 lotteries for training,

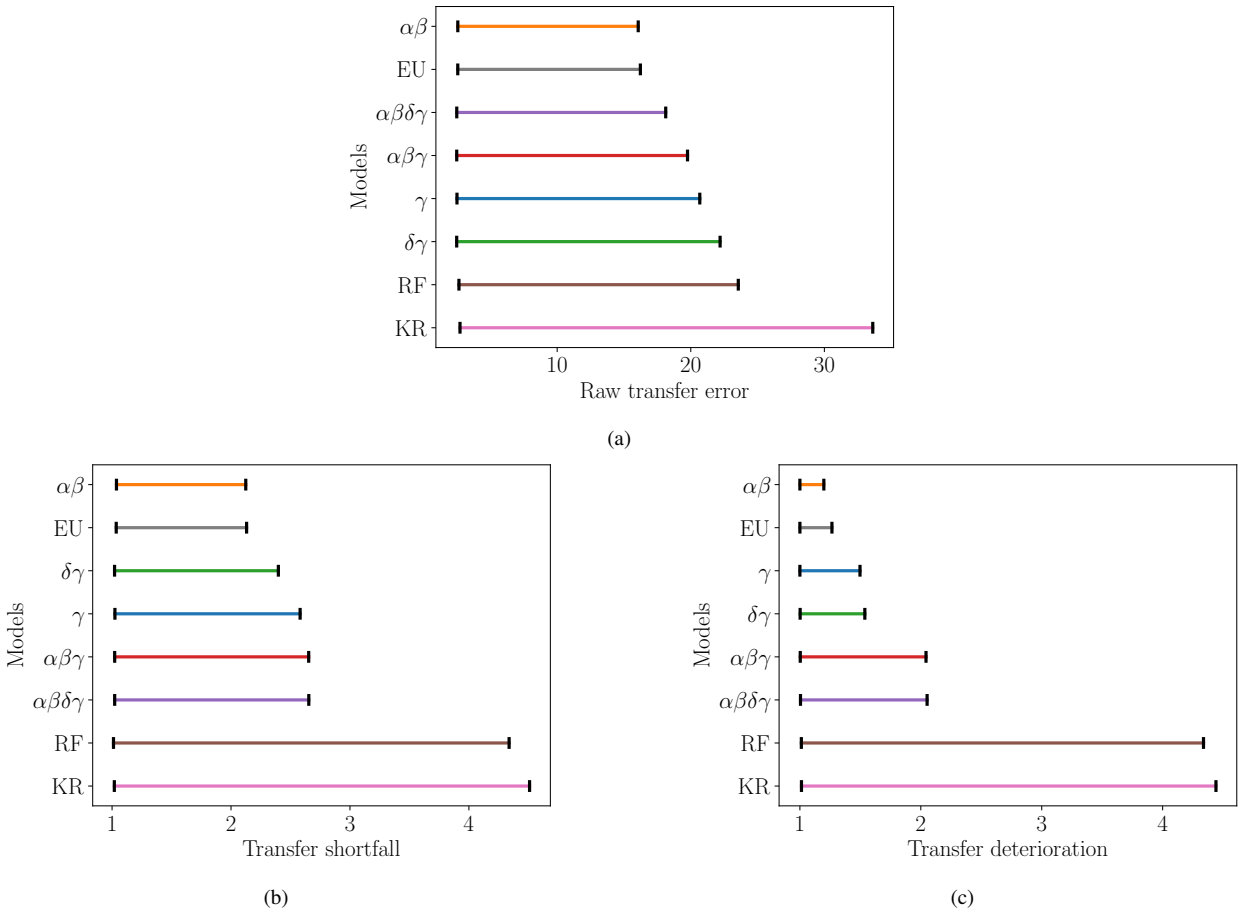


Figure C.6: *82% (n=44, $\tau = 0.95$) forecast intervals for (a) raw transfer error, (b) transfer shortfall, and (c) transfer deterioration, with the choice of $r = 3$.*

finding the best prediction rule for this pooled data, and using the estimated prediction rule to predict certainty equivalents for the remaining 21 (or 19) lotteries. We use domain cross-validation to select tuning parameters for the black box algorithms, as described in Example 10.

Figure C.7 and Figure C.8 are the analog of Figure 4.6, with $r = 3$ and $r = 5$ respectively. We again choose $\tau = 0.95$, thus constructing forecast intervals whose lower bounds are the 5% percentile of pooled transfer errors, and whose upper bounds are the 95% percentile of pooed transfer errors. Applying Proposition 6, these are 76% for $r = 3$ and 68% for $r = 5$ forecast

intervals. The most notable change is that the forecast intervals shrink for all of the prediction methods, which suggests that the raw transfer error becomes less variable when it is trained on more lotteries. Otherwise, all of the qualitative statements in the main text for $r = 1$ continue to hold, and in particular the economic models continue to transfer better than the black box algorithms do.

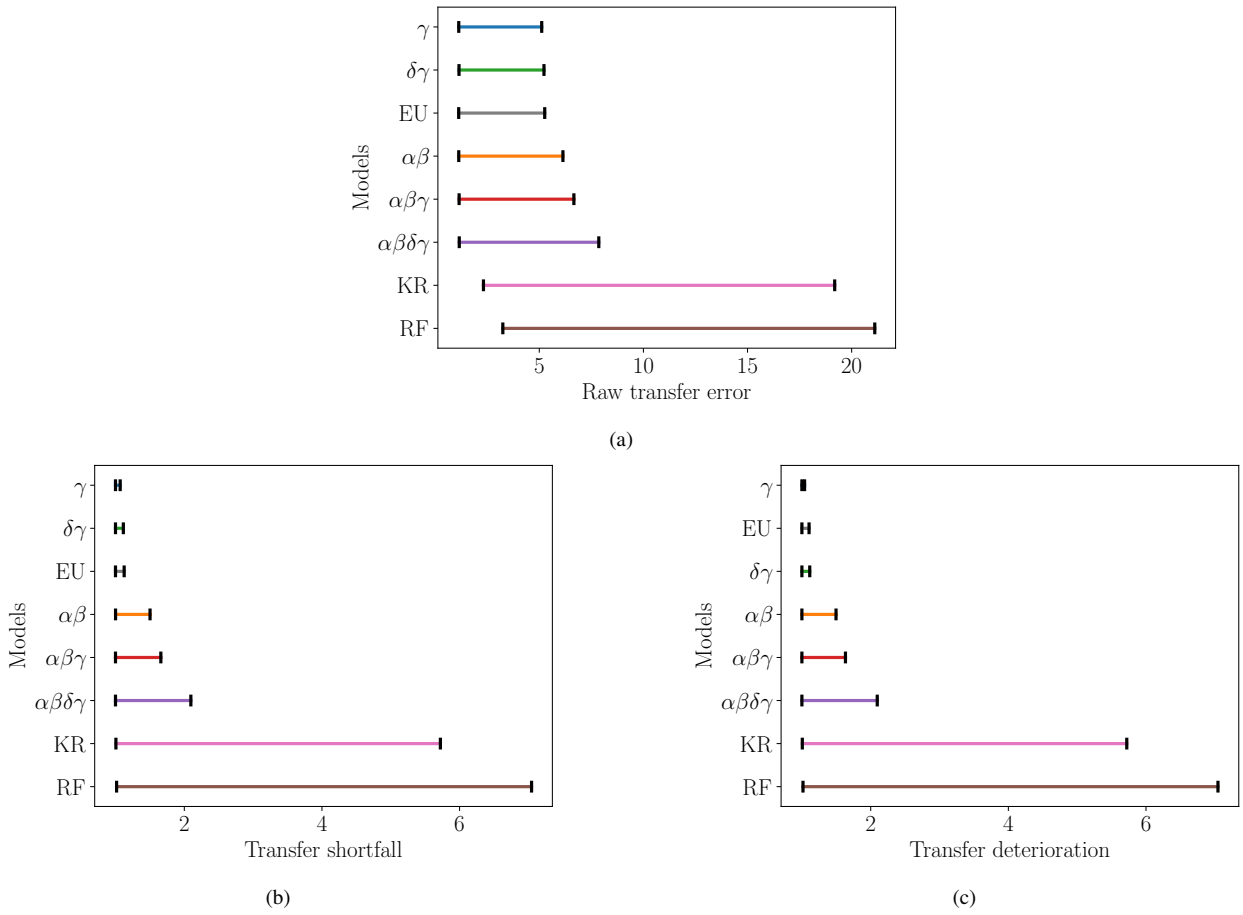


Figure C.7: *76% (n=24, τ=0.95) forecast intervals using common lotteries in l'Haridon and Vieider (2019), with the choice of $r = 3$.*

C.5.8 More details on worst-case dominance

Figures C.9 and C.10 compare the worst case upper bound of the forecast intervals for CPT and RF for our three transfer measures as either Γ or τ varies. In each case the dominance relation is

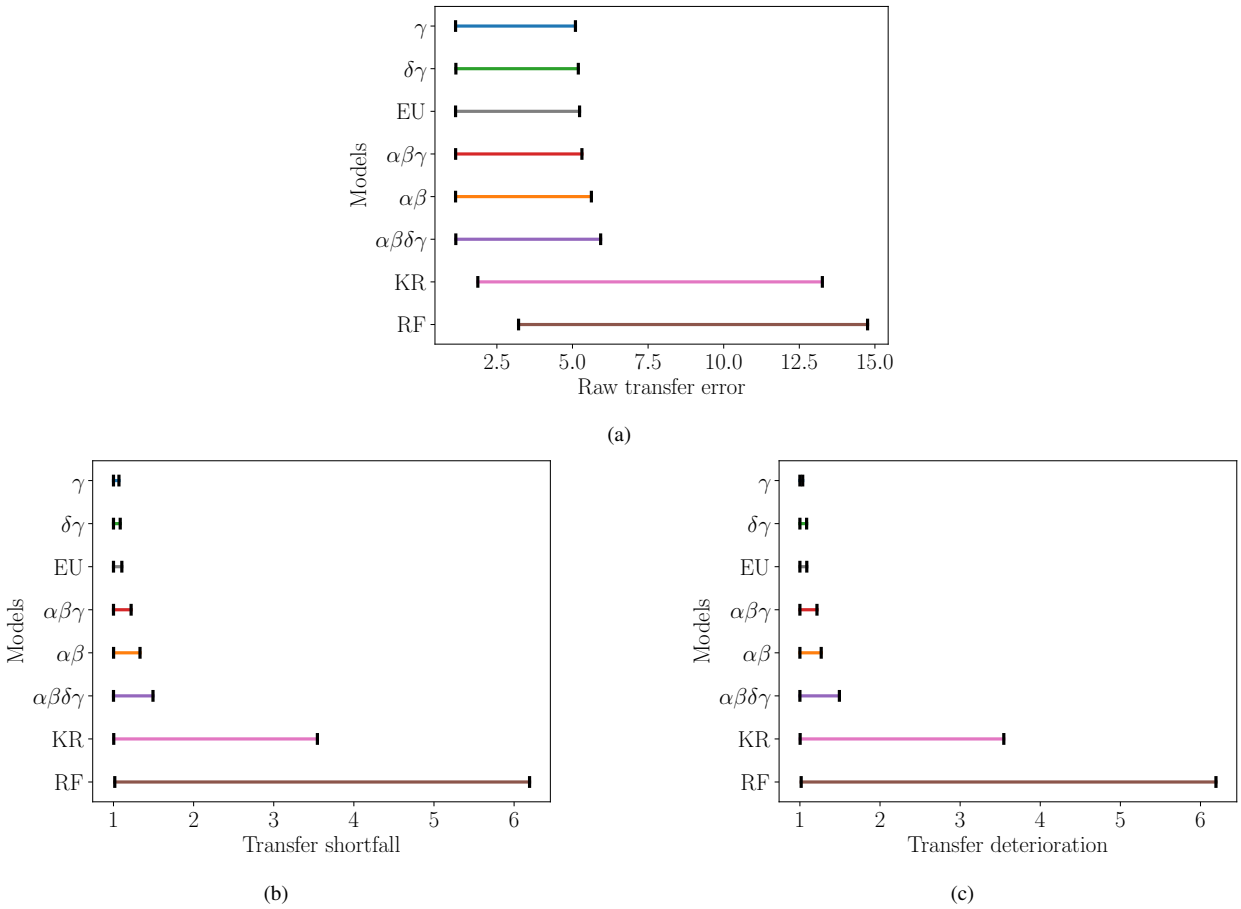


Figure C.8: *68% (n=24, τ=0.95) forecast intervals using common lotteries in l'Haridon and Vieider (2019), with the choice of* $r = 5$.

clear.

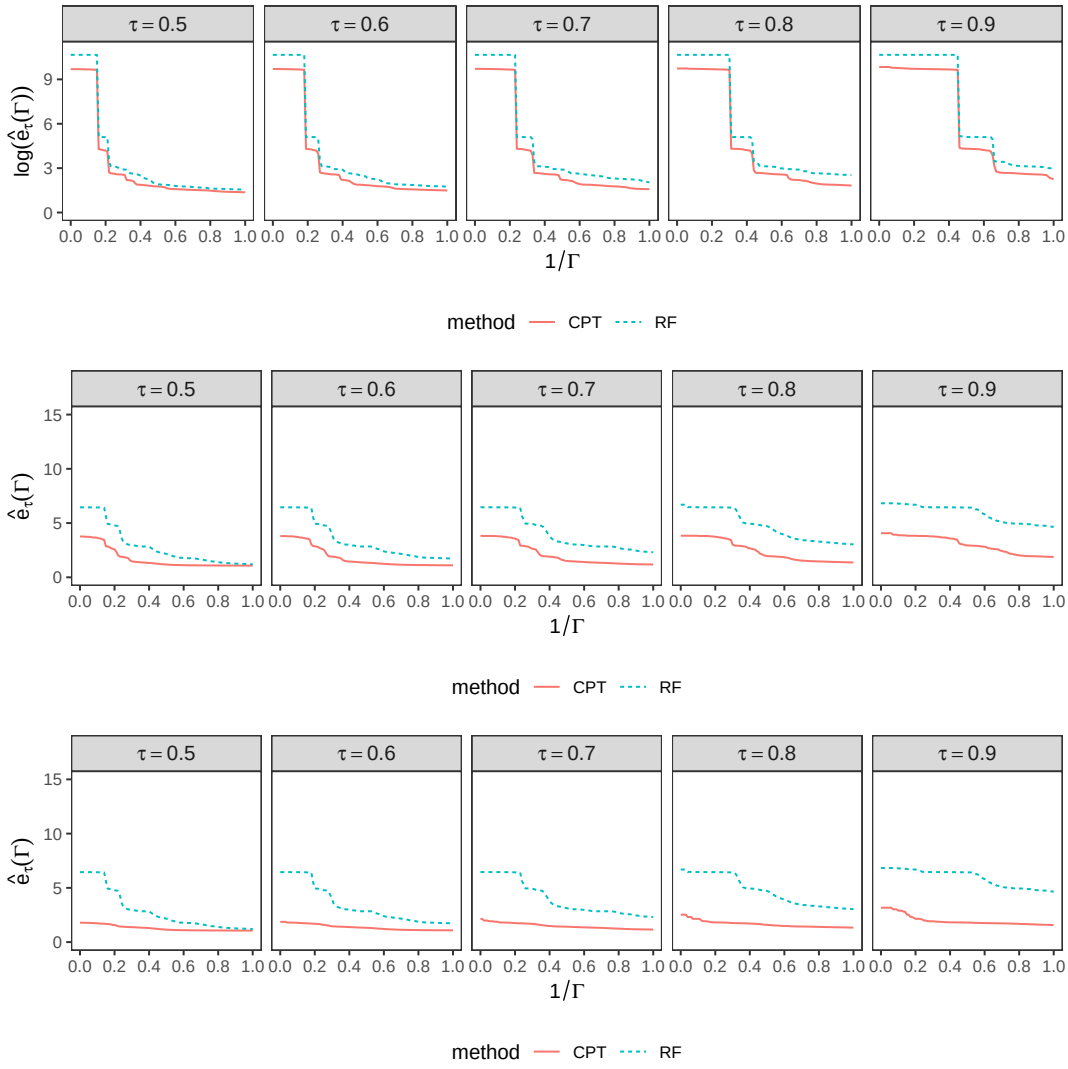


Figure C.9: *The worst case upper prediction bound* $\hat{e}_\tau(\Gamma)$ *(as defined in (4.10)) for (a) raw transfer error, (b) transfer shortfall, and (c) transfer deterioration of CPT and RF as a function of* $\Gamma \in [1, \infty)$*, discretized at* $100/i (i = 0, 1, \ldots, 100)$*, at different quantiles* $\tau \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

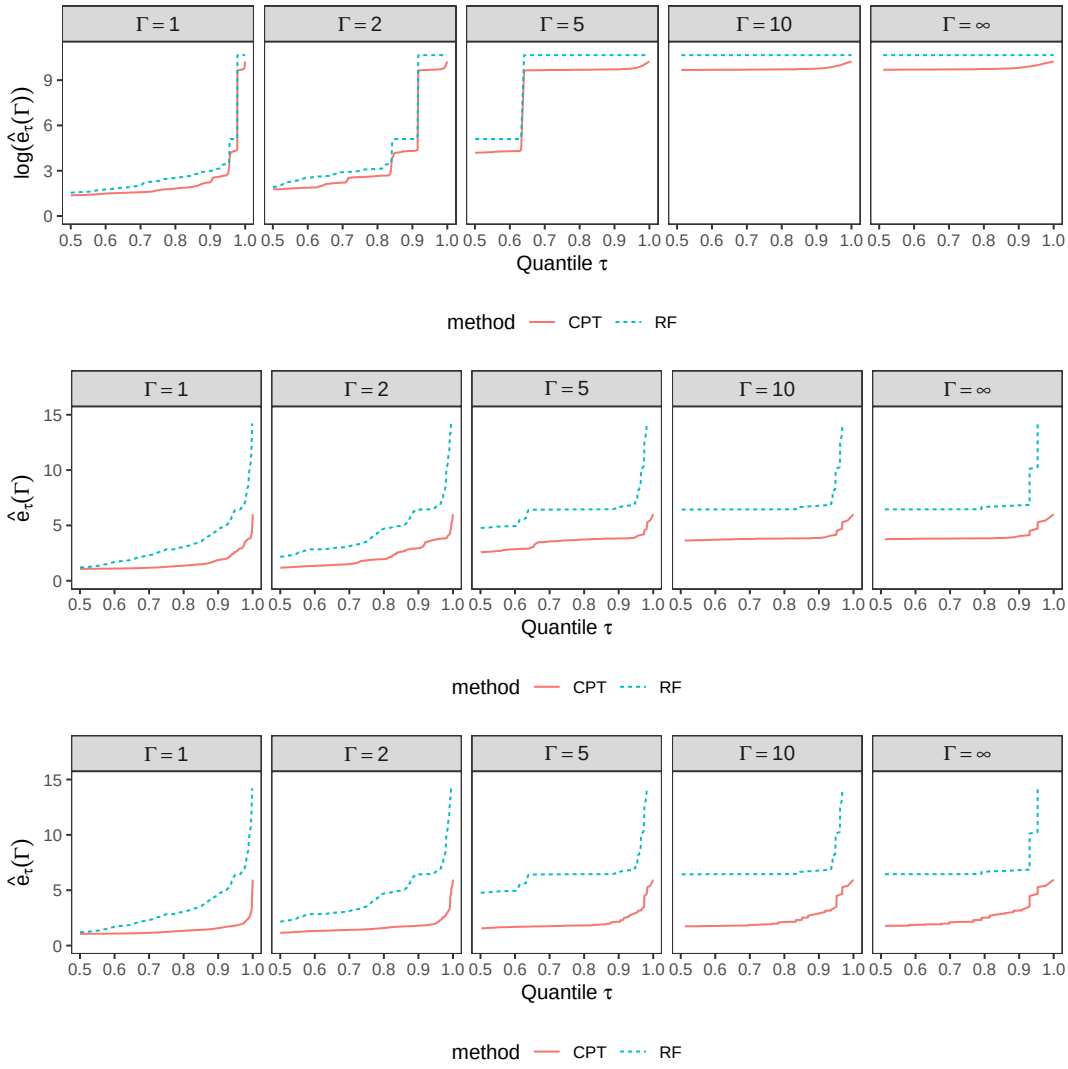


Figure C.10: *The worst case upper prediction bound* $\hat{e}_\tau(\Gamma)$ *(as defined in (4.10)) for (a) raw transfer error, (b) transfer shortfall, and (c) transfer deterioration of CPT and RF as a function of* $\tau \in [0.5, 1]$ *without discretization for* $\Gamma \in \{1, 2, 5, 10, \infty\}$.