

Principled Measurement and Evaluation of AI Systems Under Practical Constraints

Luke Guerdan

CMU-HCII-26-108

August 2026

School of Computer Science
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:

Kenneth Holstein, Carnegie Mellon University, Co-Chair
Zhiwei Steven Wu, Carnegie Mellon University, Co-Chair
Jeffrey Bigham, Carnegie Mellon University
Alexandra Chouldechova, Abridge

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.*

Copyright © 2026 Luke Guerdan

This work was supported by the UL Research Institutes through the Center for Advancing Safety of Machine Intelligence (CASMI) at Northwestern University, the Carnegie Mellon University Block Center for Technology and Society (Award No. 53680.1.5007718), the National Science Foundation Graduate Research Fellowship Program (Award No. DGE-1745016), and the National Institute of Standards and Technology and Carnegie Mellon University AI Measurement Science and Engineering Center (AIMSEC).

Abstract

As artificial intelligence (AI) systems are introduced in a range of socially consequential settings, measuring their capabilities, risks, and limitations has become crucial. Yet many properties of AI systems—such as the “helpfulness” of a chatbot response, or the “fairness” of a predictive algorithm—are unobservable latent constructs. Obtaining valid, reliable, and informative measurements of such constructs is challenging in practice, given both their contested nature and the often limited resources available to measure them effectively. In response, this thesis introduces the concept of a **measurement intervention**: a targeted approach designed to help organizations improve the validity of AI system performance measurements under practical constraints.

The first part of this thesis advances measurement interventions in the context of predictive modeling for algorithmic decision support (ADS). I propose a conceptual framework that characterizes threats to the validity of predictive models used in ADS. I then report findings from qualitative interviews with data scientists, which illustrate the nuanced trade-offs they make while balancing the validity of a predictive modeling formulation with competing desiderata, such as its resource requirements and portability across institutional contexts. Based on these findings, I devise two statistical approaches—“expert anchors” and “uncertainty cancellation”—that help teams more effectively repurpose existing organizational data for new measurement tasks.

The second part of this thesis advances measurement interventions for Generative AI (GenAI) evaluation. I first conduct a formative study with sixteen industry AI practitioners, to examine their current practices for assessing the validity of GenAI evaluations. I find that, while Subject Matter Experts (SMEs) often have a rich understanding of latent constructs being measured in GenAI outputs, AI practitioners currently lack mechanisms for effectively incorporating this expertise into the design of GenAI measurement instruments.

Based on these findings, I introduce RubricStudio: an interactive system that supports cross-functional teams of SMEs and AI practitioners in operationalizing validity principles while designing GenAI evaluations. RubricStudio automatically flags potential validity issues in evaluation results, and provides scaffolding designed to help SMEs and AI practitioners diagnose and resolve issues together through synchronous co-refinement sessions. I evaluate RubricStudio through a controlled between-subjects study, which shows that RubricStudio’s validity scaffolding helps teams make more targeted diagnoses of validity-related issues, and creates opportunities for SMEs to directly shape the course of the evaluation design process.

Finally, I propose two statistical measurement interventions, designed to help teams scale up GenAI evaluations via an LLM-as-a-judge—i.e., a second GenAI system that acts as an automated evaluator. The first intervention proposes a multi-label rating elicitation, aggregation, and performance measurement framework for validating LLM-as-a-judge systems on subjective measurement tasks, where more than one rating may be reasonably considered “correct.” The second leverages a doubly-robust estimation approach to yield informative evaluation results when human ratings are scarce but LLM-as-a-judge ratings cheap and abundant. Based on these overall findings, I conclude by reflecting on the opportunities and challenges for measurement interventions to advance the evaluation of future human-AI systems.

Acknowledgments

I would like to thank my advisors Ken Holstein and Steven Wu, who have been at the center of my journey over the past five years. As an advising team, Steven and Ken have always pushed me to pursue fundamental research questions, and to focus on producing the highest quality of work possible. Steven is a sharp theoretician and an even better theory mentor: he has an eye for a clean problem formulation, a clear theorem statement, and making theory work legible to people with different backgrounds. He has taught many of these skills through his deep hands-on mentorship in technical topics. Ken has an exceptional interdisciplinary theory of mind. He is able to situate arguments in ways that are likely to be compelling to readers of many backgrounds—this is a skill I aspire to emulate in my own interdisciplinary work. I have also long admired Ken’s mission to foreground human knowledge, agency, and expertise in the design of AI systems, and hope to keep these goals at the center throughout my career.

I would also like to thank the rest of my committee, including Jeff Bigham and Alex Chouldechova. Jeff has played a central role in shaping the conceptual development of the thesis, including how I integrate the principled and practical dimensions of measurement interventions. Jeff also pushed me to think about different ways of incorporating human expertise in the design and evaluation of AI/ML systems, beyond narrow conceptualizations of “human feedback” traditionally pursued by ML researchers. I have always looked up to Alex’s work as an example of how to bring statistical methods to bear on societally impactful problems. This is why I was so excited to work with Alex as part of an internship at MSR, where I embarked on a new chapter of work applying measurement principles to the evaluation of GenAI systems.

My other mentors from my internship at MSR NYC, including Hanna Wallach and Solon Barocas, also played an invaluable role in shaping my intellectual growth. Hanna’s *Measurement and Fairness* paper was the impetus for much of the measurement focus in this thesis. Working with her helped me approach measurement problems with conceptual rigor, and explain my work clearly. Solon’s sociological approach to research helped me zoom out, and consider how AI systems sit within our broader society. I always appreciated Solon’s ability to quickly boil down a complex topic into its fundamental questions; I hope to cultivate this skill in my career.

My research collaborators have also shaped many aspects of the research in this thesis. Amanda Coston taught me how to apply causal inference methods to the evaluation of algorithmic decision support tools. Devansh Saxena taught me much about how to design semi-structured interviews and navigate qualitative data analysis. Stevie Chancellor taught me how to write a compelling qualitative findings section. I would also like to thank the many other researchers I have been fortunate to collaborate with, including Razvan Amironesei, Fernando Diaz, Patrick Hall, Jessica Huynh, Anna Kawakami, Lydia Liu, Inioluwa Deborah Raji, Charvi Rastogi, Renee Shelby, Kimberly Truong, Ming Wang, Justin Whitehouse and Angela Zhou.

Generous funding, research participants, and support from administrative staff have also played a critical role in shaping my research. I am grateful to the NSF for awarding me a GRFP, and for additional funding from the Center for Advancing Safety of Machine Intelligence (CASMI) at Northwestern University, and the Carnegie Mellon University AI Measurement Science and Engineering Center (AIMSEC). I would also like to thank the many people who have

participated in my interview studies and the evaluation study of RubricStudio, without whom this research would not have been possible. And finally, I would like to thank Julia Chadwick, Karen Harlan and Kate Schaich for their terrific administrative support and encouraging attitude.

Throughout my PhD, my lab-mates have created a wonderful intellectual atmosphere through which to explore my research. From the CoALA lab, I would like to thank Wesley Deng, Alicia DeVrio, Frederic Gmeiner, and Tzu-Sheng Kuo. From Steven's lab, I would like to thank Hao-Fei Cheng, Keegan Harris, Terrance Liu, Daniel Ngo, Anusha Sinha, Logan Stapleton, Gokul Swamy and Pratiksha Thaker. These two communities together formed a warm and welcoming home across HCI and ML.

My PhD cohort and Pittsburgh friends have always supported me through the highs and lows of the PhD journey. From my PhD cohort, I'd like to thank Adinawa Adjagbodjou, Riku Arakawa, Hyunsung Cho, Lisa Egede, Catarina Fidalgo, Seyun Kim, Yunzhi Li, Cella Sum, Jordan Taylor, Kimi Wenzel, and Jiajia Wu. I would also like to thank my Pittsburgh friends, including Pranav Khadpe, Ezra Awumey, Samarth Wahal, Sanika Moharana, Sam Estep, Nathan DeVrio, Tejus Gupta, Katelyn Morrison, Rose Chang, Hugo Simao, Yi Fei Cheng, William Epperson, Andrew Kuznetsov, Venkatesh Sivaraman, Yi-Hao Peng, Howie Wang, Ji-Youn Jung and Shivani Kapania, for being a source of fun and delight during my PhD journey.

Additionally, I'd like to thank my long-time friends and family for keeping me grounded. To John, Madison, Lokesh, Justin, Sam, Jake, Josh, Andrew and Alex: you've remained a strong reminder of what is ultimately most important. To my Mom, thank you for always being there for me, and for enthusiastically moving the plants so I could put my fold-up desk by the window when working back at home. To my Dad, thank you for listening to me go into research rabbit holes on the phone, and for sending great beach photos at all times of day and year. To Laura, Harper, Natalie, Tate and Nick, thank you for creating a warm family environment for my visits back home in St. Louis. I will never forget our trips to the pool and contentious Uno games.

And, last but not least, I would like to thank Evie, for being such a wonderful partner and source of support throughout this last stretch of my dissertation. I will always remember the moments we've shared together baking bread, reading books, and going for long walks. I deeply appreciate the questions you ask about my work, and our conversations about topics big and small.

The many people who have supported me in this journey remind me of a quote from the sociologist William Bruce Cameron:

"not everything that can be counted counts, and not everything that counts can be counted."

Contents

List of Figures	x
List of Tables	xiv
1 Introduction	1
1.1 Predictive Models for Algorithmic Decision Support	3
1.2 Generative AI (GenAI) Systems	3
1.3 Contributions	5
1.3.1 Thesis Structure	6
2 A Conceptual Framework for Characterizing Threats to the Validity of Predictive ADS Tools	8
2.1 Motivation	9
2.2 Contribution	10
2.3 Related work	10
2.3.1 Human-AI decision-making	10
2.3.2 Modeling challenges in algorithmic decision support	11
2.3.3 Measurement and validity in algorithmic systems	11
2.4 Framework	12
2.4.1 Scope	12
2.4.2 Framework development	12
2.4.3 Causal diagram	13
2.4.4 Mapping algorithmic decision support tasks to sources of target variable bias	14
2.5 Model Development	18
2.5.1 Measurement model	18
2.5.2 Prediction model	20
2.5.3 Measurement model evaluation	20
2.6 Assessing Gaps and Opportunities for Experimental Research	22
2.6.1 Mapping existing experimental study designs to our causal diagram	23
2.6.2 Assessing the ecological validity of lab-based studies	23
2.6.3 Scaffolding a science of human-AI decision-making	24
2.7 Discussion	24
3 Examining how Data Scientists Navigate Measurement Decisions while Constructing Target Variables for Predictive Modeling Tasks	26
3.1 Motivation	27
3.2 Contribution	28
3.3 Background and Related Work	29

3.3.1	Bricolage and the Science of the Concrete	29
3.3.2	Tracing the Social Construction of Target Variables	30
3.3.3	Empirical Studies of Data Science Practice	31
3.3.4	Toolkits Supporting Target Variable Construction	32
3.3.5	Measurement and Validity in AI	33
3.4	Methods	34
3.4.1	Study Design	34
3.4.2	Participants	35
3.4.3	Data and Analysis	35
3.4.4	Positionally	36
3.5	Findings	36
3.5.1	Balancing Validity With Other Criteria	37
3.5.2	Strategies for (Re)formulating Prediction Tasks	39
3.5.3	Evaluating the Validity of a Problem Formulation	43
3.6	Discussion	46
3.6.1	Data Science as Bricolage	47
3.6.2	Design Implications: Reconciling Top-Down Measurement Interventions with Bricolage Practice	49
3.6.3	Design Implications: Developing a Scaffolding for Effective Bricolage Practice	51
3.6.4	Study Limitations	56
3.7	Conclusion	57
4	Improving the Validity of Predictive Modeling Formulations via Expert Anchors	58
4.1	Motivation	59
4.2	Contribution	60
4.3	Background and related work	60
4.3.1	AI functionality and validity concerns	60
4.3.2	Outcome measurement error	60
4.3.3	Counterfactual prediction	61
4.4	Preliminaries	62
4.5	Methodology	63
4.5.1	Unbiased risk minimization	63
4.5.2	Error parameter identification and estimation	66
4.6	Experiments	69
4.6.1	Models	69
4.6.2	Experiments on synthetic data	71
4.6.3	Semi-synthetic experiments on healthcare and employment data	72
4.7	Discussion	75
4.7.1	Decision points and complexities in measurement error modeling	75
4.7.2	Challenges of linking causal and statistical estimands	76
5	Supporting Informative Pre-Deployment Model Evaluations via Uncertainty Cancellation	77
5.1	Motivation	78
5.2	Related Work	80
5.3	Preliminaries	81
5.3.1	Problem Formulation	81
5.3.2	Partial Identification of Policy Performance	82

5.4	Regret Bound Identification	83
5.5	Mapping Causal Assumptions to Informative Regret Bounds	84
5.6	Regret Bound Estimation	85
5.6.1	Plug-in Estimator	86
5.6.2	Doubly Robust Estimator	86
5.7	Numerical Experiments	87
5.8	Real-World Application: Comparing Healthcare Enrollment Policies	88
5.9	Conclusion	90
6	A Statistical Framework for Improving the External Validity of Generative AI Evaluations	92
6.1	Motivation	93
6.2	Related Work	95
6.3	Preliminaries	96
6.4	Methodology	97
6.4.1	Baseline Approaches and Their Limitations	97
6.4.2	Doubly-Robust Estimator	98
6.5	Experiments	100
6.5.1	Dataset Generation Procedure	100
6.5.2	Setup Details	102
6.5.3	Results	102
6.6	Conclusion	104
7	A Statistical Framework for Validating LLM-as-a-Judge Systems under Rating Indeterminacy	105
7.1	Motivation	106
7.2	Related Work	107
7.3	Framework	109
7.3.1	Conceptualizing Judge System Performance Under Rating Indeterminacy	110
7.3.2	Operationalizing Judge System Performance: Measuring Human–Judge Agreement	110
7.3.3	Limitations of Standard Practice Relying on Forced-Choice Elicitation	111
7.3.4	Proposed Alternative: Measuring Human–Judge Agreement via Multi-Label Metrics	112
7.4	Experiments	112
7.4.1	Experiment Design	113
7.4.2	Results	114
7.5	Conclusion: Implications for Practice & Limitations	117
8	RubricStudio: An Interactive Toolkit for Operationalizing Construct Validity Principles in Cross-Functional Generative AI Evaluation Design	118
8.1	Introduction	119
8.2	Background & Related Work	121
8.2.1	Measurement and Evaluation of GenAI Systems	121
8.2.2	Interactive Systems for GenAI Evaluation Design	123
8.2.3	Tools and Processes Supporting Collaboration in GenAI Evaluation Design	124
8.3	Formative Study	125
8.3.1	Methods	125
8.3.2	Findings	126
8.3.3	Design Goals	128

8.4	RubricStudio	128
8.4.1	System Overview	129
8.4.2	Authoring and Managing Evaluation Results	129
8.4.3	Validity Scaffolding	131
8.4.4	Implementation	135
8.5	Summative Study	135
8.5.1	Task	135
8.5.2	Participants	136
8.5.3	Procedure	137
8.5.4	Data Analysis	137
8.6	Findings	138
8.6.1	Dyads with Validity Scaffolding engaged in deeper exploration, review, and diagnosis of issues (RQ1)	139
8.6.2	Validity scaffolding created more opportunities for SMEs to shape how issues were diagnosed and resolved (RQ2)	142
8.7	Discussion	146
8.7.1	Envisioning the Role of Automation in Evaluation Design Scaffolding	146
8.7.2	Opportunities to Support Collaborative Case Labeling	147
8.7.3	Opportunities to Guide Resource Investment Across Project Maturity Phases	148
8.7.4	Limitations	149
8.8	Conclusion	149
9	Conclusion	151
9.0.1	Future Direction: Breaking Out of the “Bag-of-Tasks” when Designing Measures of Human–AI System Performance	152
9.0.2	Future Direction: New Frontiers in Resource-Efficient Measurement	153
9.0.3	Closing Thoughts	154
	References	155

List of Figures

1.1	A visual illustration of chapters in this thesis. Chapters 2–3 establish a conceptual foundation for measurement interventions in the context of predictive ADS. Chapters 4–8 introduce measurement interventions targeting specific stages of the measurement process: conceptualization, operationalization, and application.	5
2.1	Our causal diagram for ADS, which represents a space of causal graphs, spanning different possible relationships between predictors, decisions, target variables, and their proxies in algorithmic decision support tasks.	13
2.2	Sub-graphs of the diagram in Figure 4.1 introducing statistical biases that impact the target variable Y^* .	15
2.3	A colored sub-space of our DAG, which shows the 66 of the 72 studies ($\approx 92\%$) in our review that examine a narrow sub-region of our proposed causal diagram.	22
3.1	An illustration of the relationship between the target variable and problem formulation.	27
3.2	An illustration of the target variable construction process as characterized by our study findings.	37
3.3	Data scientists evaluated the validity, predictive performance, portability, simplicity, and resource requirements of their problem formulation.	38
4.1	A causal diagram of treatment-conditional outcome measurement error.	62
4.2	An illustration of treatment-conditional OME in heart attack prediction.	63
4.3	An illustration of the oracle propensity and outcome functions in our synthetic experiment setup.	70
4.4	Accuracy of learned models on target potential outcomes as a function of training dataset sample size.	70
4.5	Bias in Average Treatment Effect estimates obtained from models on OHIE and JOBS data.	75
5.1	Illustration of uncertainty in comparing two policies in a toy setting. Our method leverages this to reduce policy comparison uncertainty.	78
5.2	Flow of causal assumptions within our framework.	84
5.3	The improvement in bounds offered by the δ -regret interval over the baseline interval.	86
5.4	A comparison between δ -regret and baseline regret interval end-points averaged over $N = 20$ trials and $N_s = 20,000$ samples.	87
5.5	Top: Coverage of accuracy regret interval estimates as a function of total sample size. Bottom: 95% bootstrap confidence intervals around upper and lower regret bounds over $N = 25$ trials.	87
5.6	Comparison of the δ -regret interval against the baseline for the NPV ($m_{a=0}$).	89

5.7	Expected cost regret as a function of false positive to false negative cost ratio.	89
6.1	Comparison of our doubly-robust estimator with baselines on three datasets from our Persona Simulation Framework.	94
6.2	Our framework produces estimates for the target parameter θ_t using (i) complete rating tuples from the source distribution (blue, left), (ii) unlabeled samples from the target distribution (yellow, right), and (iii) persona ratings produced for both source and target samples (red, top).	95
6.3	Coverage by persona quality (top), covariate shift (center), and selection bias (bottom).	102
6.4	Average Bias (MAE), Coverage, and Interval Width across experimental conditions presented in Fig. 6.3.	103
6.5	Coverage of DR (Riesz) versus RePPI when varying covariate shift and selection bias with persona ratings from different LLM judges.	103
7.1	Examples of rating indeterminacy in (1) toxic language, (2) factuality, (3) helpfulness, and (4) relevance rating tasks.	106
7.2	Framework applied to “toxicity” task. Judge rankings change substantially when ranked by categorical agreement with forced-choice human ratings versus downstream task performance measured against multi-label ratings.	108
7.3	Our framework connects two key meta-evaluation goals: measuring general-purpose human–judge agreement metrics (left), and validating judge systems on specific downstream evaluation tasks (right).	110
7.4	An illustration of our rating model applied to an item in an underspecified Yes/No rating task.	111
7.5	Sub-optimality from selecting a judge system via human–judge agreement metrics vs. directly on the downstream task metric.	115
7.6	Estimates of task-specific sensitivity parameters ($\hat{\beta}_t^J$) recovered from four judge systems. These estimates vary considerably between judges and across tasks.	116
7.7	Illustrative examples of rank inversions among four pairs of metrics and three rating tasks reported in the main results (Figure 7.5).	117
8.1	An overview of the <i>validity scaffolding for co-refinement</i> provided by Rubric-Studio. A dyad begins by designing an initial rubric. They then test it by running an evaluation, reviewing the results, and diagnosing issues surfaced by automated validity warnings. A review workflow helps the dyad diagnose the root cause of an issue (e.g., a problem with a evaluation criterion, the LLM-as-a-judge, or the dataset) and address it through a targeted refinement action (e.g., an updated evaluation criterion). The dyad iteratively identifies and resolves evaluation design issues until they are satisfied with their results, then they export their final evaluation configuration.	120
8.2	The Designer page. While reviewing input-output pairs from the evaluation dataset using the Case Reviewer (A), the dyad authors their rubric in the Rubric Builder (B), where each property (C) is decomposed into binary evaluation criteria (D). The prompt editor (E) enables the dyad to edit the prompt that is used to apply each criterion to an input-output pair.	130

8.3	The Evaluations view. The Rubric Version & Results Tracker (A) records which LLM-as-a-judge models have been run against each rubric version, while the Evaluation Runner (B) launches a run and reports its estimated time, cost, and progress. Here, edits to three criteria trigger a re-run of only those criteria—75 ratings rather than the full rubric (DG3).	130
8.4	Examples of the two validity checks, each performed on a different evaluation run. The Pass Rate Check (A) flags criteria with high or low pass rates (here, No Giveaways at 24%); clicking one opens a Pass/Fail breakdown. The Structural Relationship Check (B) flags pairs of criteria within a property that have very high (potentially redundant) or very low (potentially unrelated) correlations. In both checks, clicking on the warning directs the user to a breakdown that scaffolds their review of cases.	132
8.5	The warning-review workflow. After a check fires, the dyad reviews information about the warning (A). They then review input-output pairs together to gather evidence (B) before entering a diagnosis (C). Finally, the dyad reviews recommended evaluation refinement actions (D) and marks the warning as resolved once they are satisfied it has been addressed. Handshake icons indicate reminders for the SME and the AI practitioner to combine their expertise as they work through the step together.	134
8.6	Assignment of the 12 dyads to a 2×2 between-subjects design crossing condition (Validity Scaffolding vs. Unstructured Collaboration) with domain. Three dyads were assigned to each cell. Two pilot dyads (D01, D02) are excluded.	136
8.7	Dyads in the validity scaffolding condition investigated fewer issues overall (left), but spent more time on each of them (right). This greater time investment concentrated on early-stage exploration, review, and diagnosis of issues. Left: Median number of issues investigated per-dyad (marker shapes indicate domain). Right: mean minutes per issue, broken down by investigation phase.	139
8.8	Left: Investigations in the Validity Scaffolding condition were primarily warning-initiated, while those in the Unstructured Collaboration condition were user-initiated or procedural. Investigations can have multiple initiation pathways; hence values do not sum to one. Center: dyads in the Validity Scaffolding condition reviewed more cases per issue. Right: dyads in the Validity Scaffolding condition spent longer reviewing each case. Dots denote per-dyad means and the horizontal bars denote the median over per-dyad means.	140
8.9	Diagnosis codes per issue by condition. Each dot denotes a dyad-level mean, with bars denoting the median across dyads. Left: Dyads in the Validity Scaffolding condition diagnosed more issues overall, and were more likely to attribute issues to the evaluation dataset or autorater. Right: Dyads in the Validity Scaffolding condition engaged in more careful diagnostic reasoning activities, including differential diagnosis and criteria differentiation codes.	141
8.10	Question topics by condition and initiator. Confirmatory questions (left) and open-ended questions (right), broken-down by topic category. Each bar is stacked to show the breakdown of questions initiated by the AI practitioner (AIP) and subject matter expert (SME). Bars represent the mean number of questions per issue across dyads (i.e., the mean of per-dyad means). In the Validity Scaffolding condition, the increase in confirmatory questions about labeling and interpretation was driven primarily by the AIP.	144

8.11 SME-influenced-analysis codes by condition and investigation phase. Each marker denotes a moment in which an SME shared a non-obvious insight that shaped how an issue was addressed. When an insight impacted multiple phases of an investigation (e.g., shaping both diagnosis and the resulting action), we assigned it to the earliest investigation phase in which the SME's observation occurred; dyads with multiple markers had multiple independent SME-influenced-analysis moments. Such moments arose in five of six dyads in the Validity Scaffolding condition, but in only one dyad in the Unstructured Collaboration condition.	145
---	-----

List of Tables

2.1	Taxonomy of measurement and prediction approaches. Top: methods proposed in ADS literature. Bottom: methods applied in machine learning, social sciences, and bio-statistics.	19
2.2	Experimental studies examining the under-studied sub-region provided in Figure 4.1.	22
3.1	Demographic overview of study participants: data scientists' experience, roles, and affiliations.	35
3.2	A description of criteria that data scientists considered while evaluating their problem formulation and illustrative examples of how participants balanced each with validity.	38
3.3	Data scientists engaged in both theory-driven and data-driven practices while evaluating validity.	45
4.1	Multiple combinations of min, max, and base rate anchor assumptions (shown via $\checkmark$) enable identification of α_t, β_t	67
4.2	Model accuracy (s.e.) across error parameter settings (α_0, β_0) at $N = 60k$ samples over 10 runs. Top-2 performance across each (α_0, β_0) setting shown in bold.	70
6.1	Examples of statistical parameters that can be estimated via our M-estimation framework.	97
8.1	Background of the 16 formative study participants. Collaborator roles are abbreviated: R = Research, PM = Product Management, DS = Data Science, Eng = Engineering, QA = Quality Assurance, Exec = Executives, Pol = Policy, UX = User Experience; "—" denotes a participant who worked independently. Validity Familiarity and Validity Application refer to participants' self-reported familiarity with, and frequency of applying, validity principles, respectively. . .	125
8.2	Diagnosis codes per issue by condition. Values are medians of per-dyad means; U and p are from Mann–Whitney tests comparing the six Validity Scaffolding dyads against the six Unstructured Collaboration dyads. $*p < .05$	141
8.3	Taxonomy of questions that participants in a dyad posed to one another. Question topics were developed using the same approach used to establish event codes: we coded transcripts from the first six sessions—which contained at least one dyad from each condition–dyad cell (Fig. 8.6)—then iteratively refined the topic categories. Each question was assigned to a single topic. . .	143

Chapter 1

Introduction

Artificial intelligence (AI) systems are increasingly deployed in socially consequential settings, where they influence key decisions about individuals' health, economic opportunities, and legal outcomes. These systems range from algorithmic decision support (ADS) tools that predict risk and recommend actions, to generative AI (GenAI) systems used in business and consumer applications. As these diverse AI technologies become embedded in organizational processes, it is critical that we are able to reliably measure their capabilities, risks, and limitations. Yet many key properties we care about in AI systems—such as the "fairness" of a predictive algorithm, or the "factual consistency" of a chatbot response—are unobservable, latent constructs that are only indirectly reflected in datasets used to measure AI system performance.

Validity describes the extent to which a measurement approach accurately reflects the latent construct it purports to measure (Drost, 2011; Campbell, 2017; Messick, 1995; Jacobs and Wallach, 2021). Obtaining valid, reliable, and informative measurements of AI system performance remains a significant challenge. These difficulties stem from both the contested nature of constructs themselves, as well as the practical constraints organizations face when designing operationalizations and collecting data to measure them effectively. For example, when evaluating a healthcare risk prediction model, we might aim to measure how well it identifies patients' "healthcare needs"—but this multifaceted construct is often reduced to simplistic proxy measures like "cost of medical care", which can be biased by historical disparities in access to patient care (Obermeyer et al., 2019). A credible measurement approach requires both a clear conceptual definition of "healthcare needs"—informed by domain knowledge—and a suitable dataset aligned with that definition. Yet ensuring that available data is consistent with our measurement goals is a formidable task; a hospital may record a narrow snapshot of a patient's journey through the healthcare system, or represent their data with incommensurate formats. Prior work has documented many cases in which flaws in the measurement process contributed to the deployment of systems that failed to meet their intended goals (Jacobs and Wallach, 2021; Stray et al., 2021; Kawakami et al., 2024; Goyal et al., 2022; Obermeyer et al., 2019; Mullainathan and Obermeyer, 2021; Coston et al., 2023a).

The quantitative social sciences have developed a rich set of conceptual and statistical tools to scaffold the measurement process (Drost, 2011; Campbell, 2017). However, applying these tools to evaluate AI systems remains challenging. In particular, measurement theory traditionally assumes a "top-down" process, starting with the theoretical conceptualization of a construct, followed by its operationalization, and then data collection (Van't Veer and Giner-Sorolla, 2016; Miller and Salkind, 2002; Everett, 2013). As a result, existing measurement scaffolding often presumes that the data available for measurement is tailor-made to an

organization’s goals. Yet this is rarely the case in practice. Instead, prior empirical studies (Barocas et al., 2017; Passi and Jackson, 2018; Kawakami et al., 2022) and my own findings (Chapter 3) highlight a *mismatch* that often arises between organizational measurement goals and the data resources that happen to be available. This forces organizations to navigate a tradeoff between the validity benefits and resource costs involved with data collection, among a range of other practical tradeoffs (Chapter 3). Yet because existing measurement tools and frameworks are designed for idealized lab-based settings, in which validity is the focal criteria for a measurement instrument (Messick, 1995), existing frameworks are poorly suited to help practitioners navigate such tradeoffs. Taken together, this mismatch between existing measurement scaffolding and practical constraints may cause organizations to de-prioritize validity principles in favor of easily quantifiable system performance metrics.

To address these challenges, this thesis introduces the concept of a **measurement intervention**: a targeted approach designed to improve the validity of AI system performance measurements under practical constraints. Measurement interventions acknowledge the competing demands involved with measurement, and provide targeted support to help teams navigate tradeoffs effectively. Measurement interventions can scaffold both the organizational *and* technical components of measurement, including AI practitioner training, data collection, model development and model evaluation. These interventions also target different *stages* of the measurement process, including:

- **Theoretical conceptualization** of a concept to be measured in an evaluation,
- **Operationalization** of a theoretical conceptualization into a measurement instrument,
- **Application** of a measurement instrument to obtain statistically valid performance estimates in a specific system deployment context of interest.

Taken together, these ideas motivate the central argument of this thesis:

Thesis Statement. By taking a principled approach to AI evaluation design, we can devise targeted interventions that help practitioners improve the validity of AI performance measurements under competing practical constraints.

Measurement and Evaluation. Following Lynch’s model of evaluation, measurement, and testing (Adom et al., 2020; Bachman, 1990; Lynch, 2001), this thesis conceptualizes measurement as one component of the broader AI evaluation process. Measurement involves “the assignment of quantifiable data by using one or more instruments, such as a test or rating scale” (Adom et al., 2020), while evaluation involves the interpretation of one or more measurements—potentially in combination with other evidence such as qualitative data—against a normative standard establishing what constitutes “good” or “desirable” performance (Adom et al., 2020). This thesis focuses on measurement interventions: strategies to improve the validity and informativeness of measurements under resource constraints. These improved measurements in turn have the potential to improve the broader evaluation process.

Thesis Scope. This thesis establishes the conceptual and statistical foundations of measurement interventions in the context of two widely used classes of AI systems: predictive models for algorithmic decision support (ADS) and large-scale generative AI (GenAI) systems.

1.1 Predictive Models for Algorithmic Decision Support

The first portion of this thesis explores how measurement interventions can advance the evaluation of predictive models used for Algorithmic Decision Support (ADS; referred to as predictive ADS throughout). Evaluating predictive ADS tools is critical given their widespread adoption in high-stakes sectors, such as healthcare, education, finance, and employment (Raji et al., 2022; Chouldechova et al., 2018; Mullainathan and Obermeyer, 2017). Yet the outcomes used to train and evaluate predictive ADS tools, such as a patient’s “cost of care”, are often an imperfect proxy for the constructs pertinent to an organizational decision-making process, such as a patient’s “healthcare needs” (Obermeyer et al., 2019). Evaluations based on imperfect proxy labels can yield misleading performance measurements, and misguided system deployment decisions (Coston et al., 2023a; Goyal et al., 2022).

To establish a conceptual foundation for measurement interventions in predictive ADS, I¹ first develop a theoretical framework that disentangles threats to the validity of target variables used to train and evaluate predictive ADS tools (Chapter 2). Based on the validity threats surfaced by this framework, I identify concrete modeling tools, evaluation protocols, and lab-based experimental research designs that can close gaps between how the research community has traditionally approached ADS evaluation, and the validity challenges that emerge in practice. Next, I conduct qualitative interviews to investigate how these validity threats manifest in real-world data science projects (Chapter 3). I specifically study the process of *target variable construction*—or data scientists’ practices of operationalizing a construct into a target variable for predictive modeling tasks. I find that data scientists engage in measurement by *making do* with the limited organizational data at hand. This process of making do requires data scientists to navigate tradeoffs between validity and competing priorities, such as a model’s predictive performance, operational simplicity, and resource requirements. While data scientists apply a range of adaptive strategies to navigate such tradeoffs, these strategies can inadvertently compromise validity when applied ineffectively. I therefore develop interventions to organizational and model development protocols that help data scientists *make do* more effectively by strengthening measurement validity while working within practical constraints.

To establish a statistical foundation for measurement interventions in predictive ADS, I devise two methods designed to help teams better *make do* with available data. The first, *expert anchors*, flexibly incorporates domain experts’ knowledge about unobserved constructs to improve the validity of predictive models that rely on proxy outcomes (Chapter 4). The second, *uncertainty cancellation*, is designed to help teams collect more informative measurements of the *relative* performance of two decision-making approaches (e.g., AI-assisted versus human-only decision-making) when real-world contextual factors are unobserved in evaluation data (Chapter 5). I validate both methods through theoretical analysis and semi-synthetic experiments using real-world data from the healthcare and employment sectors. These approaches illustrate how organizations can leverage measurement interventions to improve the validity of performance measurements, even when additional data collection remains practically infeasible.

1.2 Generative AI (GenAI) Systems

The second part of this thesis explores the potential for measurement interventions to improve the validity of Generative AI (GenAI) evaluations. Measurements of GenAI system performance are often collected by first establishing a theoretically grounded definition of the concept being

¹Throughout, I use the pronoun *I* to refer to projects I have led or co-led. All work was performed in collaboration with my advisors, Ken Holstein and Steven Wu, in addition to external collaborators without whom this work would not be possible. Where relevant, each chapter lists the collaborators involved with each project.

measured through a process of *conceptualization*. This definition is then *operationalized* into rating instructions that specify how outputs from the GenAI system being evaluated should be scored (Wallach et al., 2025; Shankar, Zamfirescu-Pereira, Hartmann, Parameswaran and Arawjo, 2024; Desmond et al., 2025). The choice of rating method (human or automated) used to score outputs is also an important operationalization consideration (Tan et al., 2024). Finally, once an operationalization has been established, it is *applied* over a specific population of system inputs and outputs to produce a statistical performance estimate. However, flaws in any of these conceptualization, operationalization, or application stages of measurement can threaten the validity of an evaluation (Szymanski, Ziems, Eicher-Miller, Li, Jiang and Metoyer, 2024; Wallach et al., 2025; He et al., 2025). Therefore, in Chapters 6-8, I devise measurement interventions designed to target each of these stages of measurement.

Application. In Chapter 6, I introduce a measurement intervention targeting the *application* stage of measurement. This chapter addresses a core challenge impacting many GenAI evaluation tasks: the *source population* of rated input-output pairs available at evaluation time can differ from the *target population* encountered upon system deployment. In many cases, it can be challenging to obtain human ratings from the target population at deployment time, due to limited human rater availability. While an LLM-as-a-judge system can rate input-output pairs from the target distribution in place of human raters, these judge ratings may be an *imperfect proxy* for human ratings. To address this challenge, I develop a doubly-robust estimation framework that combines human ratings, observed from the source distribution under sampling bias, with LLM-as-a-judge ratings to provide statistically valid performance estimates over the target population. This estimator remains valid so long as either the model predicting human ratings or the model correcting for sampling bias is sufficiently accurate. This estimator is validated via synthetic and semi-synthetic experiments that experimentally manipulate covariate shift, selection bias and persona quality.

Operationalization. Chapter 6 views ratings produced by an LLM-as-a-judge as black-box predictions, and uses these predictions to improve statistical inference. However, in practice, LLM-as-a-judge ratings often serve not just as predictions but as *per-item scores* for a concept being measured in GenAI outputs. The status quo approach for validating that LLM-as-a-judge systems provide accurate ratings is to measure human-judge agreement—i.e., the per-item agreement between human ratings and judge ratings on a small sample of input-output pairs drawn from an evaluation corpus. A key assumption of existing LLM-as-a-judge validation approaches is that there is a single “correct” rating for each item in the validation corpus. Yet this assumption is often unreasonable on subjective GenAI rating tasks, where ambiguity or vagueness in rating criteria can yield more than one “correct” rating—a condition I call *rating indeterminacy*. To address this challenge, I develop a framework for validating LLM-as-a-judge systems under rating indeterminacy (Chapter 7). This framework measures agreement between humans and LLM-as-a-judge systems via multi-label “response set” ratings rather than forced-choice ratings. Through experiments on eleven real-world rating tasks, I show that standard forced-choice validation approaches select judge systems that perform as much as 31% worse than those selected by my proposed multi-label validation framework.

Conceptualization. Finally, in Chapter 8, I propose a measurement intervention targeting the *conceptualization* and *operationalization* stages of measurement. While subject matter experts (SMEs) often have the richest understanding of concepts being measured in GenAI outputs, in existing evaluation design workflows, SMEs’ knowledge is often incorporated into the evaluation design process *after* evaluation criteria have already been created. In particular, my formative findings suggest that technical teams often design evaluation criteria internally, then instruct SMEs to provide ratings with respect to these existing criteria. These ratings are then used to verify the quality of an LLM-as-a-judge through measures of human-judge agreement (e.g.,

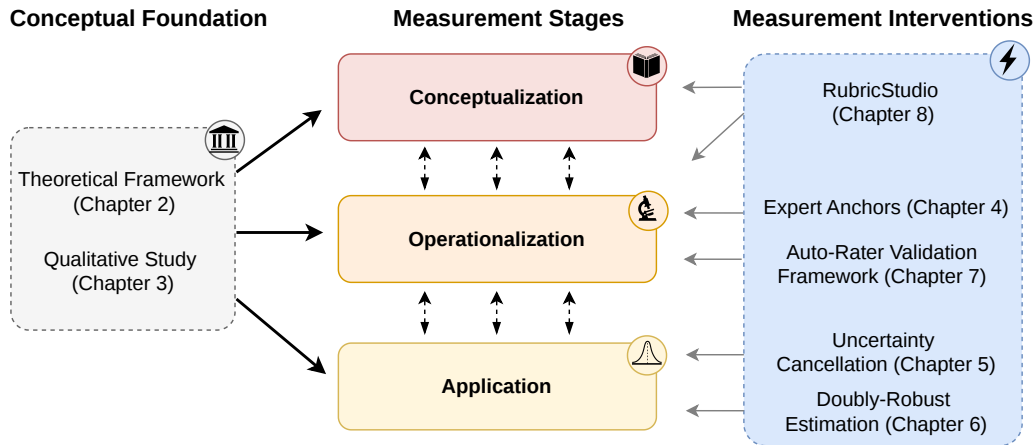


Figure 1.1: A visual illustration of chapters in this thesis. Chapters 2–3 establish a conceptual foundation for measurement interventions in the context of predictive ADS. Chapters 4–8 introduce measurement interventions targeting specific stages of the measurement process: conceptualization, operationalization, and application.

as in Chapter 7). However, because this practice leaves SMEs out of the *upstream* concept definition and refinement process, technical teams often later discover that their evaluation criteria are under-developed, requiring that valuable SME time be spent on collecting updated ratings.

To address this gap, I develop *RubricStudio*, an interactive system that helps AI practitioners and SMEs collaborate to design and validate rubrics for LLM-as-a-judge evaluations. *RubricStudio* runs automated validity checks that surface potential problems with an evaluation, and provides diagnostic scaffolding designed to help AI practitioners and SMEs identify and address validity issues through synchronous co-refinement sessions. The system supports conceptualization by helping dyads decompose concepts like “sycophancy” or “helpfulness” into hierarchical rubrics of binary evaluation criteria. The system also supports operationalization by helping dyads compare the quality of ratings produced by different models used as an LLM-as-a-judge. Through a controlled between-subjects study with 12 AI practitioner–SME dyads, I find that the validity scaffolding in *RubricStudio* led dyads to engage in more deliberate diagnostic reasoning, and drew SMEs into the process of diagnosing and acting upon validity issues—interpretive work from which our formative study suggests they are often excluded in current evaluation design workflows.

1.3 Contributions

The contributions in this thesis are summarized below and developed in full in the chapters that follow.

Measurement Interventions for Predictive ADS.

- A *causal framework* that identifies the sources and implications of *target variable bias* in ADS by characterizing the data generating process that gives rise to predictive model training datasets. This framework disentangles which sources of target variable bias (e.g., measurement error, intervention effects, selective labels) are of concern in different types of ADS tasks, and identifies opportunities for future modeling and empirical human subjects research to account for target variable bias more carefully (Chapter 2).

- Findings from an interview study with fifteen data scientists working in education and healthcare sectors, which illustrates how they construct target variables for prediction tasks. This work argues that target variable construction is a *bricolage practice*, in which practitioners apply creative and adaptive problem (re)formulation strategies to make do with their imperfect data under competing practical constraints (Chapter 3).
- A statistical learning method designed to correct for outcome measurement error, treatment effects, and selection bias in parallel while learning a predictive model from observational data. This approach leverages *expert anchors*, which are flexible assumptions that encode subject matter experts' knowledge into a measurement-error correction, to learn a reliable predictive model from imperfect proxy labels (Chapter 4).
- *Uncertainty cancellation*, a partial-identification technique that yields informative bounds on the *relative* performance of two decision-making policies under unmeasured confounding by isolating comparison-relevant uncertainty. This approach is validated theoretically, via synthetic data experiments, and with an illustrative pre-deployment evaluation of a predictive model used to inform healthcare enrollment decisions (Chapter 5).

Measurement Interventions for GenAI Evaluation.

- A formalization of GenAI system quality estimation that explicitly accounts for both covariate shift and selection bias in available human ratings. This formulation is paired with a *doubly-robust estimator* that combines human ratings with LLM-as-a-judge “persona” ratings to yield valid estimates so long as either a model predicting human ratings or a model correcting for sampling bias is of sufficient quality (Chapter 6).
- The first framework for validating LLM-as-a-judge systems under *rating indeterminacy*, in which evaluation items can have multiple “correct” ratings. Leveraging this framework, it is shown both theoretically and empirically that standard validation approaches can lead to the selection of highly sub-optimal judge systems. This motivates concrete recommendations for more principled LLM-as-a-judge validation practices (Chapter 7).
- *RubricStudio*, an interactive system that helps AI practitioners and subject matter experts collaborate to design and validate rubrics for LLM-as-a-judge evaluations. A controlled between-subjects study with twelve dyads shows that the validity scaffolding provided by RubricStudio led dyads to reason more deliberately about validity issues, and created opportunities for SMEs to shape the course of the evaluation design process (Chapter 8).

1.3.1 Thesis Structure

Figure 1.1 outlines the structure of this thesis. In Chapters 2–3, I establish a conceptual foundation for measurement interventions in the context of predictive ADS. Building on this foundation, Chapters 4–8 introduce specific measurement interventions spanning conceptualization, operationalization, and application. Chapters 4 and 5 introduce measurement interventions supporting the design and evaluation of predictive ADS tools, while Chapters 6–8 introduce measurement interventions supporting GenAI evaluation. Finally, Chapter 9 concludes the thesis with a reflection on the promises and challenges of measurement interventions and directions for future research. Relevant background and related work is provided within each chapter.

Part I

Measurement Interventions for Predictive ADS Tools

Chapter 2

A Conceptual Framework for Characterizing Threats to the Validity of Predictive ADS Tools

This chapter is drawn from the paper:

Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein.
Ground(less) Truth: A Causal Framework for Proxy Labels in Human-Algorithm Decision-Making. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

2.1 Motivation

A growing body of research aims to combine predictive machine learning (ML) models with human judgment to improve decision-making processes. In the machine learning community, researchers have proposed improvements to *ML models* to better address gaps in human judgment (e.g., (Madras et al., 2018; Wilder et al., 2020; Tan et al., 2018; Hilgard et al., 2021)). In the human-computer interaction community, behavioral interventions have been developed to help *humans* better incorporate model outputs into their decision-making (e.g., (Bansal et al., 2021; Bućinca et al., 2021; Liu et al., 2021; Bucinca et al., 2020; Carton et al., 2020; Bussone et al., 2015; Lai et al., 2021)). However, current evaluations of both model-level and human behavioral interventions typically assess the quality of human decisions, algorithmic predictions, and hybrid combinations of the two by comparing their accuracy on “ground-truth” labels that are readily available in existing data. This practice assumes that the labels targeted by predictive models serve as a reliable measure of the underlying goals and objectives of human decision-makers.

Yet in real-world deployments of algorithmic decision support (ADS) tools, labels are often imperfect proxies for the target outcomes of interest to human experts. While making decisions, content moderators frequently assess the “toxicity” of online comments (Halfaker and Geiger, 2020) while physicians often consider the “cardiovascular disease risk” of patients (Anderson et al., 1991). “Toxicity” and “cardiovascular disease risk” are examples of latent constructs which are unobserved in data. Because observed labels (e.g., toxicity annotations and diagnostic test results) serve as indirect measurements of these phenomena (Jacobs and Wallach, 2021), they can be subject to *measurement error*. Additionally, humans often select among multiple possible actions (e.g., medical treatments, social welfare interventions) in hopes of improving a downstream outcome of interest. Because an outcome is only observed for the selected action, labeled data does not contain the counterfactual outcome that *would* occur had a different option been chosen instead. This introduces a set of additional challenges, including selective labels (Lakkaraju et al., 2017), intervention effects (Coston, Mishler, Kennedy and Chouldechova, 2020), and selection bias (Rambachan and Roth, 2019), which interact with measurement error in nuanced ways depending on the nature of the specific decision-support task.

We refer to this collection of challenges – which can be characterized as sources of statistical bias impacting labels – as *target variable bias* (TVB).¹ Following common terminology in statistics, we use the term “bias” to describe systematic differences between the target outcome of interest to human experts and its imperfect operationalization in available data. Thus, while TVB describes a broad conceptual difference between outcomes of interest and their observed proxies, this difference can be formally studied under existing statistical frameworks.

Target variable bias has been widely documented in real-world deployments of algorithmic systems (Kleinberg et al., 2018; Fogliato, Xiang, Lipton, Nagin and Chouldechova, 2021; Bao et al., 2021; Butcher et al., 2022; Kawakami et al., 2022; Cheng et al., 2022; Obermeyer et al., 2019; Mullainathan and Obermeyer, 2017, 2019; Chalfin et al., 2016). Predictive models impacted by target variable bias have contributed to unwarranted firing of teachers (Chalfin et al., 2016), perpetuated disparities in access to medical resources (Obermeyer et al., 2019), and raised concerns among social workers investigating allegations of child abuse and neglect (Kawakami et al., 2022; Cheng et al., 2022). Surprisingly, existing modeling efforts and human subjects experiments in the human-AI decision-making literature have largely overlooked this

¹The term Target Variable Bias was introduced in (Chouldechova et al., 2018; Fogliato et al., 2020). We use this as an umbrella term describing sources of statistical bias known to impact proxy labels in decision support tasks.

challenge. **Left unaddressed, this disconnect could undermine the ultimate goal of human-AI decision-making research: to develop algorithmic systems that meaningfully improve decision-making in real-world contexts.**

2.2 Contribution

Therefore, in this work, we bridge the divide between challenges encountered in real-world deployments of predictive models and current human-AI decision-making research practices by (i) raising awareness of target variable bias, (ii) identifying gaps in previously published modeling approaches and human subjects experiments, and (iii) providing guidelines for improved research practices going forward. In particular, we develop a *causal framework* which identifies the sources and implications of target variable bias in human-AI decision-making by examining the data generating process which gives rise to predictive model training datasets. Our framework enables us to distill ADS tasks studied in prior literature into their underlying structural components, and identify which sources of TVB (e.g., measurement error, intervention effects, selective labels) are of concern in a specific task. Using our framework, we identify opportunities to better address target variable bias through two lines of human-AI decision-making research:

- **Model development.** We develop a *measurement* and *prediction* decomposition that articulates target variable modeling assumptions. We use our decomposition to create a taxonomy of model-level improvements proposed in previous literature. We also propose a set of recommended measurement model evaluation strategies.
- **Experimental human subjects studies.** We use our framework to re-examine the design of prior human subjects experiments studying human-AI decision-making. Our analysis identifies systematic gaps in our current understanding of human-AI decision-making due to target variable bias.

2.3 Related work

We begin by introducing the body of human-AI decision-making research our framework is designed to inform. We then summarize modeling challenges and broader validity concerns that draw current research practices (i.e., modeling assumptions, experimental study designs, and measures of decision quality) into question.

2.3.1 Human-AI decision-making

Recent machine learning research proposes techniques designed to complement the limitations of human judgement. Drawing from a long line of work showing that actuarial risk assessments can outperform expert judgement in many prediction tasks (Grove et al., 2000; Dawes et al., 1989), methods have been proposed that learn to complement humans by adaptively routing decision instances (Madras et al., 2018; Gao et al., 2021), leveraging heterogeneity in human and machine decision performance (Tan et al., 2018; Wilder et al., 2020; Donahue et al., 2022; Charusaie et al., 2022), leveraging consistency in expert decisions (De-Arteaga et al., 2021), and adapting to (Hilgard et al., 2021) and training (Mozannar et al., 2022) human mental representations of model outputs. Yet these techniques operate on a set of simplifying assumptions about the world, which may or may not hold in a given deployment context. We provide a framework for articulating modeling assumptions, and show that many common

assumptions made by prior work involving proxy labels are unlikely to hold in practice. Recent research has also studied opportunities for human-AI complementary in algorithm-assisted *human* decision-making (Lai and Tan, 2019; Green and Chen, 2019; Bućinca et al., 2021; Gajos and Mamykina, 2022; Cai et al., 2021; Lai et al., 2020). This work investigates the potential for tools such as training protocols (Cai et al., 2021; Lai et al., 2020; Cai et al., 2019), explanations (Liu et al., 2021; Bussone et al., 2015), and other behavioral interventions (Bućinca et al., 2021; Gajos and Mamykina, 2022), to improve how humans make use of model outputs. While many online experimental studies have focused on interventions to improve predictive performance, little work to date has experimentally studied other key factors that are present in real-world deployment contexts, such as asymmetric access to information (Holstein et al., 2023; Hemmer et al., 2022), measurement error (Gordon et al., 2022), and omitted payoffs (Green and Chen, 2021).

2.3.2 Modeling challenges in algorithmic decision support

Prior work has surfaced a litany of challenges impacting predictive models designed for algorithmic decision support (ADS), including unobservables (Kleinberg et al., 2018), selective labels (Lakkaraju et al., 2017), selection bias (Singh et al., 2021; De-Arteaga et al., 2018), and intervention effects (Coston, Mishler, Kennedy and Chouldechova, 2020). Additional work has examined the quality of proxy labels in decision support tasks. For example, Obermeyer et al. (2019) surfaced “*label choice bias*”, in which racial disparities in access to health resources were introduced by poor label selection decisions. “*Omitted payoffs bias*” describes factors of interest to humans that are incompletely reflected by predictive models targeting available labels (De-Arteaga et al., 2021; Chalfin et al., 2016; Kleinberg et al., 2018). While this bias describes challenges specific to prediction (e.g., model unobservables, measurement error (Kleinberg et al., 2018)), this term also applies when humans care about a broader set of decision-making factors beyond predictive risk (Chalfin et al., 2016; Green and Chen, 2019). In this work, we use the lens of measurement and validity to examine systematic differences between target outcomes of interest to humans and proxy labels observed in data (Jacobs and Wallach, 2021). In adopting this lens, we draw upon a rich set of existing knowledge and methodologies from adjacent disciplines (e.g., psychology, political science, sociology) designed to evaluate how latent phenomena of interest to humans are quantified in data (Roberts, 1985).

2.3.3 Measurement and validity in algorithmic systems

Recent work has raised broader concerns regarding whether algorithmic systems successfully achieve their purported function (Coston et al., 2023b; Jacobs and Wallach, 2021; Bao et al., 2021). Synthesizing concepts from measurement theory in the quantitative social sciences, Jacobs and Wallach (2021) argue that “*algorithmic fairness*” is a latent construct that is imperfectly operationalized by statistical fairness measures. Bao et al. (2021) examine statistical biases present in criminal justice datasets (e.g., ProPublica’s COMPAS Dataset (Angwin et al., 2016)) used in fairness benchmarks of algorithmic Risk Assessment Instruments (RAIs). This analysis identifies several biases in the outcome variable Y target by models, which we further characterize in this work. Coston et al. (2023b) highlight validity concerns impacting RAIs, including many discussed in § 2.3.2. Recent work has also surfaced validity issues in content moderation (Gordon et al., 2022) and recommender systems (Milli et al., 2021; Stray et al., 2022).

Despite this growing awareness, we currently lack a holistic understanding of validity threats to prediction targets in human-AI decision-making. Addressing this gap is critical for preventing

algorithmic harms in real-world deployment contexts. Therefore, in this work, we use causal diagrams to examine the relationship between measurement error and additional modeling challenges (i.e., § 2.3.2) that can impact the validity of prediction targets in real-world decision support settings. To our knowledge, our work offers the first holistic examination of how measurement error, unobservables, selection bias, intervention effects, and confounding interact to impact target variable validity in real-world ADS deployments.

2.4 Framework

We now describe our framework scope (§ 2.4.1) and development process (§ 2.4.2) before introducing our causal diagram (§ 2.4.3). We then use our framework to map algorithmic decision support tasks to relevant sources of target variable bias (§ 2.4.4).

2.4.1 Scope

Our framework applies to settings in which a supervised learning model is introduced to augment human decision-making by predicting (i) a future event (e.g., medical (Obermeyer et al., 2019), criminal justice (Dieterich et al., 2016), child welfare (Chouldechova et al., 2018), or real estate (Holstein et al., 2023; Poursabzi-Sangdeh et al., 2021) related outcomes); (ii) a subjective human annotation (e.g., perceived content toxicity (Gordon et al., 2022)); or (iii) factual information (e.g., food nutrition (Bućinca et al., 2021)). In these settings, model predictions are combined with human decision-making, either by showing model predictions to a human (i.e., algorithm-in-the-loop (Green and Chen, 2019)), who makes the final decision, or via a hybrid flow of agency (e.g., deferral-based learning (Madras et al., 2018), learning with bandit feedback (Gao et al., 2021)). Given our focus on prediction-based decision tasks, we do not directly examine decision-support settings involving unsupervised learning (e.g., clustering), tasks relying upon generative models (e.g., text or image generation), or sequential settings with time dependency (e.g., reinforcement learning) in this work.

2.4.2 Framework development

Understanding which statistical biases are of concern in a given ADS task requires *examining the historical data generating process* that gave rise to the model training dataset. Causal diagrams, which are graphs that show causal relationships between nodes via connected edges (Pearl, 1995), are tools specifically designed for this purpose. A causal diagram shows each variable being modeled as a *node*. Causal connections between nodes are shown via *edges*. If the direction of a causal pathway is known, this is shown via a directed arrow from the parent to child node. An undirected edge is used to connect nodes when the causal direction is unknown or varies across settings described by the diagram (Pearl, 1995). Our framework introduces a causal diagram to examine challenges impacting the labels available in data. Therefore, we specifically consider variables (i.e., *nodes*) and relationships (i.e., *edges*) that directly relate to the target variable; we *abstract away* other important factors, such as the training (Cai et al., 2021; Lai et al., 2020; Cai et al., 2019), decision-making process (Green and Chen, 2021), and workflow (Green and Chen, 2019) of the human decision-makers using the predictive model. While prior work has examined these factors in detail (Cai et al., 2021; Lai et al., 2020; Cai et al., 2019; Green and Chen, 2021, 2019), our framework foregrounds factors most salient for understanding target variable bias. In § 2.6.3, we outline how our approach can be extended to systematically examine a broader set of components beyond target variables in human-AI decision-making research.

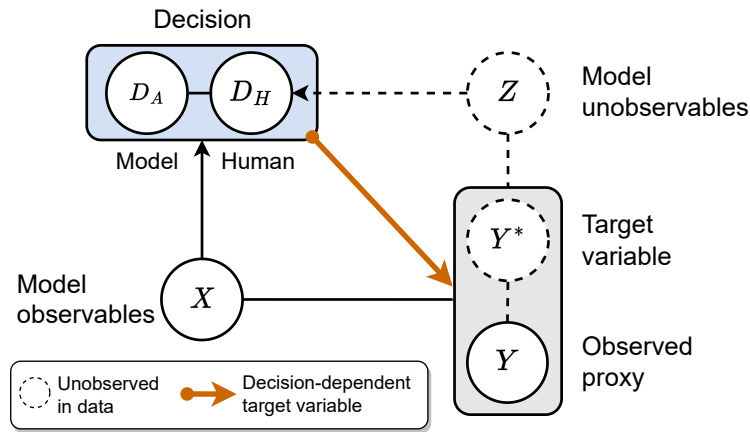


Figure 2.1: Our causal diagram represents a space of causal graphs, spanning different possible relationships between predictors, decisions, target variables, and their proxies in algorithmic decision support tasks. Edges with directionality that can vary across ADS settings are indicated via undirected edges. Observed variables are shown with solid lines, while unobserved variables are shown in dotted lines. An arrow pointing to a shaded box is shorthand for separate arrows pointing from the source to nodes contained within the box.

Our causal diagram was developed and refined through an iterative series of discussions among the authors and external researchers spanning a range of disciplines. Based on a review of real-world case studies, we synthesized candidate causal diagrams that could adequately characterize the target variable of interest across settings, and then stress-tested these diagrams by attempting to identify counterexamples. Through our discussions with external researchers, we also cross-referenced our framework with existing terminology and methods developed in adjacent disciplines, such as medical diagnostic testing, educational assessment, behavioral health, and statistics.

2.4.3 Causal diagram

Diagram structure

Figure 4.1 shows our proposed causal diagram, which represents a **space of directed acyclic graphs (DAGs)** describing the relationship between predictors, decisions, target variables, and their proxies in ADS tasks.

Predictors. X describes *covariates* used to generate model predictions. Covariates are often drawn from administrative data sources (e.g., medical records, lending history) available to an organization for model development. In ADS settings, humans can also make use of *unobserved contextual information* Z while making decisions. For example, a physician might consider real-time medical test results (e.g., electrocardiograms (Mullainathan and Obermeyer, 2019)) unavailable to a model, while a social worker might weigh contextual factors described via phone calls while deciding whether to recommend investigation of child maltreatment allegations (Kawakami et al. (2022)). In some cases, human decision-makers can also be unaware of a subset of covariates (e.g., due to organizational policy or prohibitively large datasets) (Holstein et al., 2023). Figure 4.1 refers to X and Z as *model observables* and *model unobservables*, respectively, based on whether the predictors are available to a model.

Decisions. The blue shaded box in Figure 4.1 shows the joint human-algorithm decision D . We decompose this node into separate variables for human decisions D_H and algorithm

predictions D_A . Prior to deployment of an algorithm, decisions result solely from human judgement (D_H). In some cases, post-deployment decisions result from humans incorporating predictions into their decision-making (i.e., algorithm-in-the-loop (Green and Chen, 2019)). In other cases, the joint decisions result from a learned combination of D_H and D_A (Gao et al., 2021; Madras et al., 2018; Wilder et al., 2020; Tan et al., 2018; Hilgard et al., 2021; Charusaie et al., 2022; Donahue et al., 2022).

Target variables. The node Y^* describes the unobserved target variable of interest to human decision-makers. For example, a model might be introduced to weigh the risk of unobserved constructs such as “medical need”, “recidivism”, “creditworthiness”, or “job performance.” Y describes the *observed proxy* that is targeted by a model in place of Y^* . For example, a model might predict “cost of medical care” (Obermeyer et al., 2019), “re-arrest” (Fogliato et al., 2020), “loan default”, or “supervisor performance reviews” in place of the targets listed previously. The grey box in Figure 4.1 represents a *measurement model* mapping the unobserved construct to the observed proxy targeted by a predictive model (see § 2.4.4).

Edges. We now describe the space of possible relationships connecting nodes in ADS tasks. Covariates and model unobservables both contribute to human decisions (D_H), while algorithmic predictions (D_A) are only influenced by covariates (X). For example, a physician might make use of medical records (X) and real-time test results (Z), while an algorithm only has access to medical records (X). We show these relationships via directed arrows $X \rightarrow D$ and $Z \rightarrow D_H$. Decisions (D) can also influence the target *and* proxy outcomes (Y^* , Y). For example, enrollment in a medical treatment program can increase medical costs (Y) while also improving patient health (Y^*). We show this relationship via the directed arrow $D \rightarrow Y, Y^*$.

The direction of causality between covariates (X), unobservables (Z), and prediction targets (Y and Y^*) can vary across ADS domains. In Figure 4.1, we convey this ambiguity via undirected edges. Causal diagrams for prediction tasks often show covariates (X) and unobservables (Z) contributing to downstream outcomes (Y, Y^*) via a domain-specific causal pathway (Barocas et al., 2019). In our diagram, this flow of information would be communicated via directed edges from X to (Y, Y^*), and from Z to (Y, Y^*). However, in some cases, the causal pathway can be *reversed* (Hardt and Kim, 2022). For example, this is possible if a patient’s unobserved disease status (Y^*) contributes to their medical history (X) or real-time test results (Z). Therefore, bidirectional edges shown in Figure 4.1 map to directed edges with directionality that varies depending on the domain. ²

2.4.4 Mapping algorithmic decision support tasks to sources of target variable bias

We now leverage our causal framework to identify sources of target variable bias that can impact predictive models in algorithmic decision support tasks. We begin by introducing two distinct *regimes of ADS tasks* described by our generalized diagram shown in Figure 4.1; those with: (1) decision-dependent target variables, and (2) decision-independent target variables. **ADS tasks with decision-dependent target variables are subject to more sources of TVB than those with decision-independent target variables.** While it is possible to define many other specific regimes of our generalized diagram (e.g., different directions of causality between (Y, Y^*) and X or Z , or different flows of agency between D_A and D_H), we introduce

²In order for the causal diagram to remain valid (i.e., a directed *acyclic* graph), one of the edges connecting nodes must remain disconnected in these settings (i.e., when target variables are decision-dependent and $Y^* \rightarrow X, Y \rightarrow X, Y^* \rightarrow Z$, or $Y \rightarrow Z$). While this requirement is consistent with the scope of our framework, which considers non-sequential settings, feedback loops are an important factor to consider in sequential settings Ensign et al. (2018).

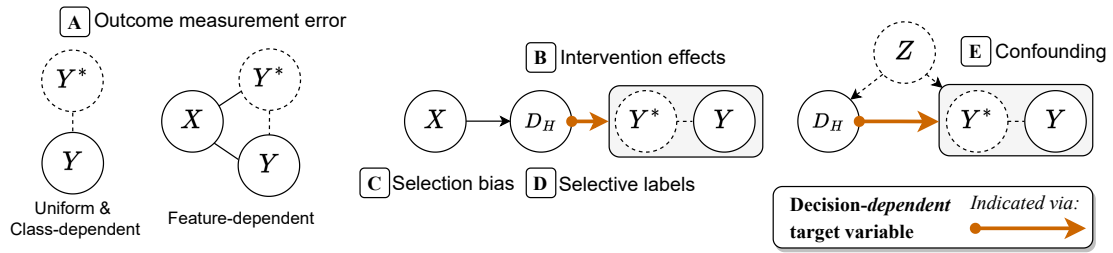


Figure 2.2: Sub-graphs of the diagram in Figure 4.1 introducing statistical biases that impact the target variable Y^* . *Outcome measurement error* (A) can occur in settings with both decision-dependent and independent target variables. In decision-dependent settings, *intervention effects* (B), *selection bias* (C), *selective labels* (D), and *confounding* (E) are also of concern.

the distinction between decision-dependent and independent target variables here because it is useful for identifying task-specific sources of TVB.

ADS tasks with **decision-dependent target variables** occur when *the decision informed by an algorithm also impacts the downstream outcomes Y and Y^** . Real-world ADS deployments often involve prediction tasks with decision-dependent target variables. For example, re-arrest is only observed among defendants released on bail, while child welfare screening decisions can influence the risk of adverse care outcomes (Coston, Mishler, Kennedy and Chouldechova, 2020). More generally, decisions informed by algorithms often constitute *risk mitigating interventions* (e.g., medical treatments, educational programs) or *opportunities* (e.g., loans, new candidate hires) that change the likelihood of the target outcome (e.g., disease prognosis, educational attainment). Settings with decision-dependent target-variables include the orange arrow from D to Y and Y^* shown in Figure 4.1.³

In contrast, the target variable is *not* influenced by the proposed decision in ADS tasks with **decision-independent target variables**. ADS are frequently deployed in the real world with the goal of informing decisions that can change the predicted outcomes. However, lab-based experimental studies of human-AI decision-making often conduct evaluations via ADS tasks with decision-independent target variables. For instance, studies have examined models that predict factual content (e.g., food nutrition (Bućinca et al., 2021)) and perceptual information (e.g., counts of objects (Park et al., 2019), geometric shapes (Zhang et al., 2022)). These tasks are decision-independent because the prediction target (i.e., food nutrition, geometric shape) is *not* influenced by the prediction made by a human and/or model.⁴ ADS tasks in the decision-independent regime do not contain the arrow from D to Y and Y^* in Figure 4.1.

We now introduce five sources of target variable bias relevant in ADS tasks. **Outcome measurement error is of concern in both decision-dependent and decision-independent regimes, while intervention effects, selective labels, selection bias, and confounding bias are only relevant in decision-dependent regimes.**

³ This regime maps directly to the “predictive optimization” setting recently studied by ? and the discussion of predictive model validity provided by Coston et al. (2023b).

⁴ ADS tasks in which labels are assigned via human annotations also fall within the decision-independent regime. In these tasks, the ratings of human annotators (Y) serve as a proxy for the broader construct of interest (Y^*) of interest in the model deployment setting (e.g., comment “toxicity” or “hate speech” (Gordon et al., 2022)).

Outcome measurement error

Human experts often make decisions involving unobserved, latent constructs such as “*recidivism risk*” and “*job performance*.” These latent constructs are not directly observable in the world, but can be operationalized via a *measurement model* (Hand, 2004; Jacobs and Wallach, 2021). Adopting a label observed in data as a proxy for an unobserved latent construct serves as a *de facto* measurement model. For instance, in criminal justice settings, defendant re-arrest is commonly adopted as a proxy for recidivism risk (Fogliato, Xiang, Lipton, Nagin and Chouldechova, 2021; Bao et al., 2021), while in commercial hiring settings, manager reviews are frequently adopted as a proxy for future job performance. Outcome measurement error (Figure 2.2.A) occurs when there is a systematic difference between the target variable of interest to experts and policymakers (Y^*) and its operationalization by a proxy (Y). This challenge has been extensively documented in judicial (Fogliato et al., 2020; Butcher et al., 2022), child welfare (Kawakami et al., 2022; Cheng et al., 2022), and hiring (Chalfin et al., 2016) ADS domains.

Because proxy labels impacted by measurement error offer an incomplete reflection of the actual goals of human decision-makers, they serve as an incomplete measure of human-AI decision quality. Therefore, before adopting a proxy as a measure of human-AI decision quality, it is critical to assess whether it serves as a satisfactory approximation of the target variable of interest to humans. Measurement theory in the quantitative social sciences provides tools to conduct this assessment by weighing the *construct validity* and *reliability* of observed labels (Hand, 2004; Jacobs and Wallach, 2021) (see § 2.5.3). In practice, measurement error in proxies is often studied via *measurement error models*. These models make *assumptions on the relationship* between the target outcome (Y^*) and its proxy (Y). Outcome measurement error is of concern in decision-dependent and independent regimes because observed labels can be subject to construct validity and reliability concerns in both settings.

Intervention effects

In many ADS tasks, decisions serve as *risk mitigating interventions* intended to improve the chances of a favorable policy-relevant outcome (Barabas et al., 2018; Kube et al., 2019; Coston, Mishler, Kennedy and Chouldechova, 2020). As a result, past human decisions D_H influence the probability of the target outcome Y^* and its proxy Y (Figure 2.2.B). However, many existing predictive techniques mistakenly assume that decisions D and outcomes Y , Y^* are statistically independent (Barabas et al., 2018; Kube et al., 2019; Coston, Mishler, Kennedy and Chouldechova, 2020). This practice can be traced back to formulation of ADS as a prediction-policy problem (Kleinberg et al., 2015), in which models are trained to maximize predictive performance with respect to observed outcomes without considering causal effects from D to Y and Y^* . Yet, we argue that accounting for the causal connection between decisions and outcomes is of central interest in many ADS tasks. For instance, consider two distinct policy problems that arise in tasks with decision-dependent target variables:

- **Selective Intervention (SI):** In this policy setting, organizations provide resources to individuals who are at *high baseline risk under no intervention*. For example, developers of the Allegheny Family Screening Tool (AFST) introduced the tool with the goal of assessing “*latent risk*” of maltreatment prior to county child welfare interventions (Vaithianathan et al., 2019). Similarly, predictive models have been introduced in educational settings to identify students at-risk of failing given no tutoring resources (Smith et al., 2012). This task requires causal inference because it involves inferring what *would occur* if an individual does not receive the proposed intervention.

- **Selective Opportunity (SO):** In this policy setting, an organization grants an opportunity (e.g., a new loan, or pre-trial release on bail) to decision subjects while trying to minimize risk of an adverse outcome (e.g., loan default, recidivism) given an individual receives the opportunity. This prediction task requires causal inference because it involves predicting what would occur under the *hypothetical scenario* that an individual receives the opportunity under consideration.

Naively structuring an ADS task as a prediction policy problem in SI and SO settings can lead to misleading assessments of model performance. For example, [Coston, Mishler, Kennedy and Chouldechova \(2020\)](#) demonstrate that predicting observational modeling and evaluation in SI settings systematically underestimates the risk for high-risk individuals who would respond most favorably to the intervention. The underlying modeling and evaluation challenge introduced by intervention effects stems from the fact that downstream outcomes are only observed under one of the possible decisions ([Pearl, 2009](#); [Perdomo et al., 2020](#); [Johansson et al., 2020](#)). This source of bias is only relevant in the decision-dependent regime because intervention effects are introduced by the connection $D \rightarrow Y, Y^*$.

Selective labels

Another challenge introduced by the connection $D \rightarrow Y, Y^*$ in the decision-dependent regime is *selective labels* (Figure 2.2.D). This bias has been widely discussed in connection to pre-trial risk assessments, where recidivism-related proxy outcomes (e.g., re-arrest, failure to appear) are only observed among defendants released on bail ([Kleinberg et al., 2018](#); [Fogliato, Xiang, Lipton, Nagin and Chouldechova, 2021](#); [Bao et al., 2021](#); [Lakkaraju et al., 2017](#)). Selective labels also occur in child welfare settings, in which some outcomes (e.g., placement in foster care) are only observed among cases screened-in for investigation ([De-Arteaga et al., 2018](#)). Selective labels maps directly to *selective intervention* and *selective opportunity* policy problems because we never observe how an individual *would have* benefited from a missed opportunity (SO), or how an intervention *would have* impacted an individual who historically received no additional resources. Selective labels pose the greatest challenge when selection bias was also present in the data generating process.

Selection bias

This bias, which occurs when covariates (X) or model unobservables (Z) influenced past decisions (D) (Figure 2.2.C), complicates selective labels and intervention effects. Because a previous decision-making policy may have been more likely to intervene (SI) or grant opportunities (SO) to some sub-populations, these groups may be systematically over- or under-represented in historical outcome data. As a result, ADS models trained on historical data will not perform equally well on all sub-populations during deployment ([Berk, 1983](#)). This effect has been well-documented in recidivism prediction settings, in which models predicting re-arrest outcomes have worse performance among sub-populations historically denied bail ([Kallus and Zhou, 2018b](#)). While selection bias can cause challenges in any setting in which data is collected non-randomly ([Heckman, 1979](#)), this challenge is compounded in decision-dependent outcome tasks because the connection $X \rightarrow D_H \rightarrow Y^*, Y$ causes selection effects to cascade to selective observation of outcomes Y and Y^* . The connection between selection bias and other downstream issues (e.g., intervention effects, selective labels) underscores the importance of considering the full data generating process while diagnosing sources of bias impacting proxy labels.

Confounding bias

This bias occurs when unmeasured variables influence both the treatment and outcome variable (Pearl, 1995). Confounding impacts ADS tasks when unobservables influenced past decisions and downstream outcomes (Figure 2.2.E) (Pearl, 1995). When confounding impacts ADS models, it is not possible to fully mitigate treatment effects and selective labels via traditional causal inference techniques (Pearl, 2009). Yet, confounding is *not* introduced by model unobservables Z in decision-independent tasks because there is no arrow from decisions D to outcomes Y and Y^* . In these tasks, unobservables may serve as an opportunity for complementarity between humans and models arising from asymmetric access to information (Holstein et al., 2023; Hemmer et al., 2022). Therefore, by mapping an ADS task to its underlying causal diagram and identifying the appropriate task regime, it is possible to identify whether model unobservables pose a treat or opportunity for an ADS deployment.

2.5 Model Development

We now provide a framework for specifying target variable assumptions during predictive model development. We argue that predictive modeling for ADS involves two distinct steps: *measurement* and *prediction*. During the measurement step, tool developers construct a *measurement model* that operationalizes the target variable of interest Y^* using readily available datasets. During the second step, tool designers train a *prediction model* that targets the *proxy outcome* returned by the measurement model. We now discuss each of these modeling steps in detail.

2.5.1 Measurement model

During the measurement step, the unobserved outcome of interest (Y^*) is approximated using historical data from the causal diagram in Figure 4.1. This step involves establishing a *measurement hypothesis* ($\hat{Y}^*$) using observed information: covariates X , past decisions D , and one or more outcome proxies Y . In some settings, a subset of unobservables are available during model development, but unavailable in during deployment. Such runtime confounders $Z_r \subseteq Z$ can occur when protected attributes (e.g., race, gender) are available during development, but not in deployment for legal reasons (Diana et al., 2022; Coston, Kennedy and Chouldechova, 2020). Given information X , Z_r , D , Y recorded in existing data, we can construct a measurement model approximating the target variable Y^* :

$$\hat{Y}^* = F_m[X, Z_r, D, Y] \quad (2.1)$$

Unlike statistical models commonly used in machine learning contexts, a measurement model cannot be learned from past data because the target outcome Y^* is unobserved. Instead, F_m relies on *measurement assumptions* concerning the relationship between the unobserved outcome of interest and recorded information available for modeling. Therefore, it is not possible to assess the quality of $\hat{Y}^*$ by comparing against held-out data, as is common in prediction settings. Instead, evaluating measurement models requires a multifaceted approach, including assessments of construct validity, synthetic experiments, sensitivity analyses, and other evaluation strategies described in § 2.5.3.

Work	Measurement (F_m)	Prediction (F_p)	Assumptions	Bias Mitigated
Gao et al. (2021) Madras et al. (2018) Wilder et al. (2020) Tan et al. (2018) Hilgard et al. (2021)	$\hat{Y}^* = F_m[Y]$	$\hat{Y} = \hat{F}_p[X, D_H]$ Human decisions D_H available at runtime	Proxy and target variables are equivalent $Y^* = Y$	None
De-Arteaga et al. (2021)	$\hat{Y}^* = F_m[X, D, Y]$, where $\hat{Y}^* = D$ <i>expert consistency</i> instances and Y otherwise		Expert consistency assumption	Measurement error, Selection bias
Lakkaraju et al. (2017) Coston, Mishler, Kennedy and Chouldechova (2020)	$\hat{Y}^* = F_m[Y]$	$\hat{Y} = \hat{F}_p[X]$ Human decisions D_H unavailable at runtime	Heterogeneous acceptance rates	Selection bias
Coston, Kennedy and Chouldechova (2020)	$\hat{Y}^* = F_m[Y_d]$, where Y_d is a potential outcome		Causal identifiability conditions	Intervention effects
Wang et al. (2021)	$\hat{Y}^* = F_m[Y_d, Z_r]$, where Y_d is a potential outcome		Causal identifiability conditions	Intervention effects, Confounding
	$\hat{Y}^* = F_m[Y]$, where Y error is group-dependent		Confident learning assumptions (see (Northcutt et al., 2021))	Measurement error
Label noise Menon et al. (2015)	$\hat{Y}^* = F_m[Y]$, where Y class-conditional or positive and unlabeled	ERM with surrogate loss (see (Natarajan et al., 2013))	Weak separability	Measurement error
Latent Class Analysis McCutcheon (1987)	$\hat{Y}^* = F_m[Y]$, where $Y = \{Y^1, \dots, Y^K\}$ are independent factors	3-step LCA with covariates (see (Vermunt, 2010))	$Y^i \perp Y^j \mid Y^*$	Measurement error
Hui-Walter Framework Hui and Walter (1980)	$\hat{Y}^* = F_m[Y]$, where $Y = \{Y^1, \dots, Y^K\}$ are diagnostic tests	N/A	Test Se/Sp identifiability assumptions	Measurement error

Table 2.1: Taxonomy of measurement and prediction approaches. Top: methods proposed in ADS literature. Bottom: methods applied in machine learning, social sciences, and bio-statistics.

All predictive models in ADS introduce a measurement model. However, this model is often *implicitly defined* and makes tacit assumptions on the relationship between available data sources (X , Z_r , D , Y) and the target variable (Y^*). Table 2.1 provides a detailed list of the measurement models assumed by existing ADS approaches. This table reifies often-implicit measurement assumptions adopted by prior work. In the bottom three rows, we apply our taxonomy to workhorse methods used in machine learning (Menon et al., 2015; Natarajan et al., 2013), quantitative social sciences (McCutcheon, 1987), and bio-statistics (Hui and Walter, 1980) literature. The *Bias Mitigated* column of Table 2.1 refers to the source of TVB addressed by the modeling technique. For instance, we mark “None” for Wilder et al. (2020) and Madras et al. (2018) because these approaches are not designed to mitigate any TVB sources listed in § 2.4.4.

2.5.2 Prediction model

After establishing a measurement model to estimate Y^* given (X, Z_r, D, Y) , tool designers then train a *prediction model* for use in decision-support settings. This prediction model takes observed covariates (X) and predicts the measurement hypothesis ($\hat{Y}^*$) established during the preceding measurement step. Because Z_r and Y are unavailable during deployment, these are not included in the prediction model. Most often, prediction models do not assume human decisions D are available at runtime (i.e., algorithm-in-the-loop (Green and Chen, 2019)). However, in some more nuanced decision-making workflows, models may also assume that human decisions are available at run-time as an additional input (i.e., (Madras et al., 2018; Wilder et al., 2020; Gao et al., 2021; Tan et al., 2018)). Given X and optionally D available at runtime, the prediction model estimates the *measurement hypothesis* $\hat{Y}^*$:

$$\hat{Y} = \hat{F}_p[X, D] \quad (2.2)$$

Whereas a measurement model is *constructed via measurement assumptions*, the prediction model $\hat{F}_p$ is a *learned mapping* from X (and in some cases D) to the measurement hypothesis $\hat{Y}^*$. Therefore, it is appropriate to evaluate generalization of $\hat{F}_p$ to held-out data via the standard slate of evaluation metrics (e.g., accuracy, AU-ROC, or statistical fairness measures). Critically, this evaluation is conducted with respect to the measurement hypothesis established during the measurement step ($\hat{Y}^*$) rather than the target outcome (Y^*) directly. Thus, showing strong performance of $\hat{F}_p$ is *not sufficient to claim a model generates valid predictions for the target outcome* Y^* .

2.5.3 Measurement model evaluation

Measurement model evaluation requires a holistic, multifaceted approach leveraging converging sources of evidence. Informed by methods used in statistics, quantitative social sciences, and learning sciences, we provide a recommended set of approaches for validating measurement models in ADS tasks.

Construct reliability and validity

Measurement theory offers a comprehensive set of criteria for assessing the quality of a measurement model. *Construct reliability* describes the degree to which a latent phenomena is consistently reflected by a measurement model (e.g. 2.1) over time. Threats to construct reliability have been well documented in settings in which target variables are assigned via subjective human annotations. In these settings, assignment of target outcomes can vary substantially based on rater identity (Denton et al., 2021), context (Pavlopoulos et al., 2020),

and specification of the annotation protocol (Poletto et al., 2021). Construct validity describes the extent to which a measurement model adequately captures an unobserved phenomenon of interest. Thus, while construct reliability is roughly analogous to the notion of statistical variance in F_m , construct validity is analogous to statistical *bias* in F_m (Jacobs and Wallach, 2021). We refer the reader to (Coston et al., 2023b) for a detailed discussion of sub-components of construct reliability and validity that pertain to risk assessment development and evaluation.

Outcome cross-validation

In many ADS domains, multiple proxies are available that are believed to be related to the target outcome of interest. In the criminal justice domain, courts often track multiple recidivism-related outcomes (e.g., 2-year general and violent recidivism, failure to appear). In the child welfare domain, government agencies may track substantiation of abuse allegations, acceptance for welfare services, agency re-referral, placement in foster care, and hospitalization (Vaithianathan et al., 2019). When multiple reference outcomes are available, *outcome cross-validation* can be used to train a model to predict one proxy, then evaluate this model on a slate of *additional reference variables* that domain experts expect may be reasonable proxies for the outcome of interest. If targeting a proxy *also* results in strong performance across other reference variables, this provides evidence suggesting that a proxy may serve as a suitable measurement model. Outcome cross-validation has been independently used by analyses of proxy outcomes in learning analytics, criminal justice (Kleinberg et al., 2018), child welfare (De-Arteaga et al., 2021), and healthcare (Obermeyer et al., 2019). Special cases of outcome cross-validation map to sub-components of construct validity. For example, a model demonstrates *predictive validity* if its predictions correlate with a reference outcome known to be related to the construct of interest (Hand, 2004). A model demonstrates *discriminant validity* if its predictions are *not* correlated with a conceptually distinct outcome.

Sensitivity analyses

Sensitivity analyses enable assessing the degree of *measurement model misspecification* permissible before evaluation of a *prediction model* is invalidated. This technique has traditionally been applied in causal inference settings to estimate the magnitude of unobserved confounding necessary to invalidate a treatment effect estimate (Rosenbaum, 2005; Díaz and van der Laan, 2013). More recently, sensitivity analyses have been developed for predictive model evaluation. For instance, Fogliato et al. (2020) proposed a sensitivity analyses framework that examines the degree of outcome measurement error permissible before fairness-related analyses are invalidated. Future work in ADS would benefit from sensitivity analysis frameworks that examine multiple sources of target variable bias in parallel.

Synthetic evaluation

A limitation of leveraging real-world datasets for measurement model validation is that one never knows the actual relationship between Y and Y^* in naturalistic data. Model-level evaluations in ADS typically circumvent this issue via *synthetic evaluations* which test whether proposed approaches are robust to experimentally manipulated bias (De-Arteaga et al., 2021; Coston, Mishler, Kennedy and Chouldechova, 2020; Menon et al., 2015; Wang et al., 2021). Yet, synthetic evaluations require assuming a specific measurement error model. If the data generating process adopted by a synthetic evaluation does not reflect real-world conditions, this can lead to overconfidence in model performance in more realistic settings. This concern is

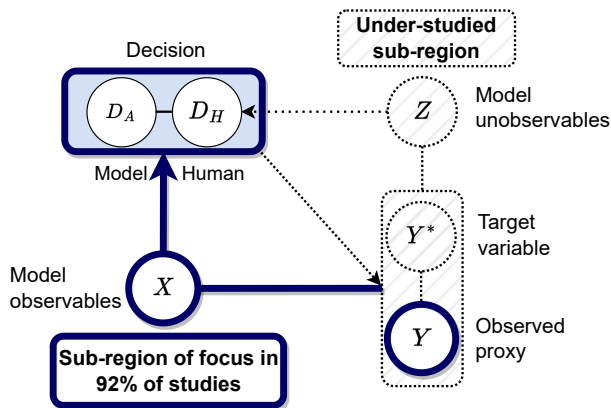


Figure 2.3: 66 of the 72 studies ($\approx 92\%$) in our review examine a narrow sub-region of our proposed causal diagram.

Work	Setting	Sub-region of causal diagram
Hemmer et al. (2022) Holstein et al. (2023)	House price prediction	Unobservables: $D \leftarrow Z \rightarrow Y$
Gordon et al. (2022)	Toxicity detection	Measurement error: $Y^* \rightarrow Y$
Peng et al. (2019)	Hiring	Selection bias: $X \rightarrow D$
Green and Chen (2021) Fogliato, Chouldechova and Lipton (2021)	Judicial	Omitted payoffs: $Q \rightarrow D_H$

Table 2.2: Experimental studies examining the under-studied sub-region provided in Figure 4.1.

salient because synthetic evaluations are often designed with bespoke data generating processes intended to highlight the specific challenge being addressed by the technique.

The Oracle Test

Chouldechova et al. (2018) propose a conceptual tool called the “Oracle Test”, which can surface unforeseen sources of target variable bias. This thought experiment supposes that we have access to an oracle model that can predict a proxy with perfect accuracy. The key question posed by this test is: “*What concerns remain given access to such an oracle?*”. Because we have a “perfect” prediction model (e.q. 2.2), remaining concerns are often related to measurement and validity (e.q. 2.1). For example, Chouldechova et al. (2018) surface concerns related to measurement error when they apply the Oracle Test to examine RAls designed for ADS models deployed in the child welfare domain. Green and Chen (2021) also leverage the Oracle Test by arguing that improvements to predictive accuracy do not equate to improved public policy outcomes when competing factors in addition to risk (i.e., defendant liberty) are overlooked.

2.6 Assessing Gaps and Opportunities for Experimental Research

In this section, we leverage our framework to assess the extent to which existing lab-based studies consider sources of target variable bias (§ 2.6.1). Our analysis finds systematic gaps in our current understanding of human-AI decision-making in light of TVB. We then show how our framework can be used by researchers to assess threats to the ecological validity and generalizability of lab-based studies (§ 2.6.2). We conclude by discussing opportunities to use our methodology to explore a broader space of open challenges in human-AI decision-making research (§ 2.6.3). We provide a resource that helps researchers apply our causal framework to examine the design and ecological validity of several experimental human-AI decision-making studies.

2.6.1 Mapping existing experimental study designs to our causal diagram

To assess the extent to which existing studies examine factors related to target variable bias in their study design, we revisit a comprehensive literature review conducted by Lai et al. (2021) through the lens of our causal diagram (Figure 4.1). Lai et al. (2021) review over one hundred experimental studies of algorithm-assisted decision-making published in premiere venues between 2018 and 2021. Our follow-up analysis extends this review to studies published in 2022 at the same set of venues, in addition to recent pre-prints. We further limit selection criteria applied by Lai et al. (2021) to studies examining prediction-based decision-making settings (i.e., scope outlined in § 2.4.1). Thus, we exclude studies included in the initial review with a focus on NLP-related tasks.

Our analysis finds that 66 out of 72 ($\approx 92\%$) studies satisfying our criteria conduct experimental evaluations focusing on a narrow sub-graph of our causal diagram. These studies investigate a modification to the joint decision-making process (i.e., the blue D_H and D_A region) using observed attributes X and an outcome proxy Y (Figure 2.3). Such studies assume that (1) the target variable and proxy are equivalent (i.e., no measurement error), (2) all predictors are observed by both the algorithm and the human (i.e., no model unobservables), and (3) decisions and outcomes are unrelated (i.e., no intervention effects).

Six of the remaining studies we review examine different sub-regions of the causal diagram described in Figure 4.1. Table 2.2 groups these studies by the sub-region under study, including unobservables (Hemmer et al., 2022; Holstein et al., 2023), measurement error (Gordon et al., 2022), selection bias (Peng et al., 2019), and omitted payoffs (Green and Chen, 2021). While these studies offer early insight into how target variable bias can impact algorithm-assisted human decision-making, our empirical understanding of these challenges remains limited compared to the joint human-AI decision region investigated by $\approx 92\%$ of studies. Critically, no work in our review experimentally manipulated factors related to *intervention effects* or examined *multiple intersecting sources of bias* in parallel. Given the prevalence of compounding challenges in real-world settings, **this gap opens a broad space of open questions and future opportunities for human-AI decision-making research.**

2.6.2 Assessing the ecological validity of lab-based studies

The gap we identify between real-world challenges and lab-based studies (i.e., Figure 2.3) carries implications for the ecological validity of experimental studies. Threats to ecological validity may be most acute when findings from a controlled study conducted under simplified conditions are *generalized* to real-world ADS deployments in which multiple sources of target variable bias are present. In these settings, measurement error and intervention effects could impact whether findings gathered via controlled experiments also apply in more complex real-world conditions.

Fortunately, our causal diagram provides a tool for assessing whether findings from a lab based study are likely to generalize to a given real-world ADS tool deployment. The first step in this process involves mapping the ADS task to its corresponding regime identified in § 2.4.4. Next, based on domain expertise, one can identify whether different sources of bias are likely to be relevant in the given real-world deployment. For instance, a model deployed to allocate tutoring resources (i.e., a decision-dependent task) may need to account for measurement error in learning outcomes and intervention effects from historical tutoring decisions. In contrast, a model deployed for a perceptual assessment task (e.g., predicting current forest cover from satellite imagery (Wang and Yin, 2021); a task with decision-independent outcomes) may not need to address these concerns. After identifying the appropriate ADS regime and relevant sources of bias, one can assess whether an experimental study is likely to generalize

to this setting by examining whether the study used a similar prediction task (e.g., also tested decision-dependent or decision-independent outcomes).

To demonstrate how causal diagrams can be used to assess ecological validity of lab-based studies, consider a previous lab-based assessment conducted by [Park et al. \(2019\)](#). This study – which is sampled from the 66 studies covered by the blue sub-region of the causal diagram provided in Figure 2.3 – examines whether introducing a delay between when humans view observed features X and algorithmic recommendations D_A improves their performance on a perceptual jellybean counting task. Because the true quantity of jellybeans does not depend on the decision under consideration, this study involves a task from the decision-independent outcome regime. Further, the influence of human-only observed attributes Z and measurement error is limited in this task. Therefore, findings from this work may most readily generalize to real-world decision-making settings with limited interference from measurement error, model unobservables, and intervention effects.

2.6.3 Scaffolding a science of human-AI decision-making

Our work leverages causal diagrams to characterize sources of bias impacting target variables. However, beyond this focus, causal diagrams also offer a powerful scaffolding for studying other aspects of human-AI decision-making, such as the joint human-AI decision-making process (i.e., the D node in our framework). For example, [Green and Chen \(2021\)](#) specify a causal diagram that models how judges weigh risk against other competing factors (e.g., culpability, value of defendant freedom) during pre-trial release decisions. The authors then experimentally verify a *hypothesized* edge in this causal diagram via a controlled online study. Through a series of such studies, it may be possible to develop a more generalized *theory* of AI-assisted human decision-making across decision support tasks. This process of specifying, testing, and refining causal models is central to existing empirical disciplines, including psychology and sociology ([Pearl, 1995](#)).

2.7 Discussion

Our work surfaces a disconnect between the challenges that arise in *real-world deployments* of algorithmic systems versus *current research practices* (i.e., experimental study designs, modeling assumptions, measures of human-AI decision quality) adopted in the human-AI decision-making literature. Left unaddressed, current gaps in this literature can amount to substantive downstream harms. For instance, while prior studies of real-world ADS tool deployments have surfaced patterns of *apparent* human under-reliance arising from imperfect prediction targets ([Kawakami et al., 2022](#); [Cheng et al., 2022](#); [Stevenson and Doleac, 2022](#)), no experimental human subjects studies to date have examined how to disentangle warranted skepticism in a misaligned model versus *unwarranted* under-reliance due to algorithm aversion. Absent such knowledge, organizations may continue to pressure domain experts to rely upon flawed predictive models ([Kawakami et al., 2022](#)), which have been shown to misallocate of medical resources ([Obermeyer et al., 2019](#)) and perpetuate historical patterns of bias ([Fogliato, Xiang, Lipton, Nagin and Chouldechova, 2021](#); [Akpınar et al., 2021](#); [Bao et al., 2021](#)).

Our work provides a critical first step for addressing this disconnect by clarifying the relationship between measurement error, intervention effects, unobserved confounding, selective labels, and selection bias via intuitive causal diagrams. Going forward, we hope that this framework will support more comprehensive assessment of modeling techniques (§ 2.5) and empirical human subjects studies (§ 2.6) designed to facilitate human-AI decision-making.

However, further work is needed to gain a comprehensive understanding of the sources and implications of target variable bias in human-AI decision-making research.

In particular, future research should develop holistic measures of decision-quality that reflect factors beyond statistical performance computed via a single outcome proxy. These measures should reflect both *process-oriented* considerations (i.e., how multiple decision-relevant factors are weighted (Green and Chen, 2019), and adherence to procedural, interpersonal, and informational justice) in addition to *outcome-oriented* considerations (i.e., whether a decision led to a beneficial outcome). Where possible, outcome-related measures should draw upon *multiple decision-relevant proxies* to better account for limitations of adopting any single proxy in isolation. While this practice is standard in disciplines such as learning sciences, diagnostic medical testing, and psychology, to date, human-AI decision-making research has primarily adopted outcome-oriented measures that hinge upon on a single potentially flawed proxy.

Our work also motivates exciting new lines of human-AI decision-making research. For instance, our review of prior modeling approaches finds that, while many techniques have been designed to address a subset of model reliability challenges (Table 2.1), few examine how various sources of target variable bias compound in real-world deployment scenarios. Additionally, our review of experimental human subjects research provides a set of tools for (i) identifying open empirical questions (i.e., Figure 2.3), (ii) designing studies with robust ecological validity, and (iii) synthesizing findings from multiple experimental studies into a complete scientific understanding of human-AI decision-making. We hope that our work will raise awareness of target variable bias in the human-AI decision-making research community and spur efforts to better align research practices with the complex challenges encountered in real-world ADS deployments.

Chapter 3

Examining how Data Scientists Navigate Measurement Decisions while Constructing Target Variables for Predictive Modeling Tasks

This chapter is drawn from the paper:

Luke Guerdan*, Devansh Saxena*, Stevie Chancellor[†], Zhiwei Steven Wu[†], and Kenneth Holstein[†]. Measurement as Bricolage: Examining How Data Scientists Construct Target Variables for Predictive Modeling Tasks. In *Proceedings of the ACM on Human-Computer Interaction* (CSCW), 2025. (*: co-first author, [†]: co-senior author)

3.1 Motivation

“Problem-solvers who do not proceed from top-down design but are arranging and rearranging a set of well-known materials can be said to be practicing bricolage. They tend to try one thing, step back, reconsider, and try another” (Turkle, p. 51).

Data scientists grapple with subtle but important considerations when translating organizational goals into a formal modeling task. For example, consider a data science team developing a model to help physicians decide which patients have sufficient “healthcare need” for enrollment in a high-risk care management program (e.g., (Obermeyer et al., 2019)). While developing a model, data scientists may work with clinicians to understand how to translate risk predictions to actionable interventions — for example, by selecting an appropriate risk cut-off for enrollment recommendations. Data scientists may also carefully consider whether the patient populations represented in their training data adequately match the characteristics of patients encountered when the model is deployed. The process of *problem formulation* describes how data scientists navigate such translational concerns as they map a high-level goal into a tractable computational problem (Passi and Barocas, 2019).

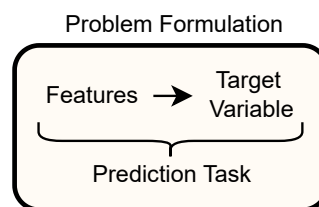


Figure 3.1: An illustration of the relationship between the target variable and problem formulation. A *prediction task* describes a mapping from *features* to a *target variable*. A *problem formulation* connects a prediction task to high-level modeling goals, including factors such as how risk scores are converted to intervention recommendations and the target population at deployment time.

One critical aspect of problem formulation is the process of *target variable construction* (Figure 3.1). Many outcomes of interest to data scientists — such as the “authenticity” of student writing or the “healthcare need” of patients — are latent, theoretical constructs that are not directly observable in data. To develop a model targeting these constructs, data scientists must construct a *proxy variable* based on outcomes that are readily available. For example, a data scientist might choose to predict patients’ cost of medical care or their number of recently diagnosed chronic health conditions as a proxy for their “healthcare need” (Obermeyer et al., 2019). Traditionally, measurement theory assumes that analysts follow a top-down workflow while mapping unobserved constructs to measurements in data (Jacobs and Wallach, 2021). For example, a learning scientist might identify a latent construct they would like to measure among students (e.g., “math proficiency”) and design an operationalization (e.g., a test) *before* collecting data (e.g., student test scores). Disciplines in the quantitative social sciences have developed a rich set of guidelines (e.g., (Van’t Veer and Giner-Sorolla, 2016; Miller and Salkind, 2002)) and modeling tools (e.g., (Everett, 2013; Singh, 2007)) to help analysts implement this top-down measurement workflow.

However, the traditional top-down conceptualization of measurement is in tension with our understanding of data science as a *bottom up* process constrained by *existing data* (Passi and Barocas, 2019; Kross and Guo, 2021). Passi and Barocas (2019) observe that problem formulation in data science is an iterative process, where data scientists progressively

reconcile their high-level objectives with low-level data and modeling constraints. As a result, while a traditional top-down measurement approach assumes that an analyst has the resources to collect data tailor-made to the goals of a study, data scientists are often forced to re-purpose the limited data that happens to be available. This disconnect creates a significant challenge: though many flawed data science projects can be traced back to target variable construction (e.g., (Turque, 2012; Amrein-Beardsley, 2014; Mullainathan and Obermeyer, 2021; Kawakami et al., 2022; Cheng et al., 2022)), existing tools, frameworks, and pedagogical resources reflect a top-down planning approach towards measurement, rarely accounting for the bottom-up, resource-constrained reality of data science practice.

3.2 Contribution

To understand data scientists' current practices and challenges in constructing target variables—and to identify opportunities to better support target variable construction in practice—in this work, we conducted semi-structured interviews with fifteen data scientists from education (N=8) and healthcare (N=7). We asked data scientists to walk us through how they navigate target variable construction in prior modeling projects. Participants also engaged in a target variable construction task, allowing us to observe their reasoning in a new modeling context. Our goal in these interviews was to understand how data scientists intertwine two conceptually distinct activities, measurement of latent constructs and prediction of target variables, during problem formulation (Mussnug, 2022). While prior empirical studies of data science practice have often centered prediction (e.g., (Alspaugh et al., 2018; Wongsuphasawat et al., 2019; Kandel et al., 2012)), we broaden our scope of inquiry to investigate the interplay between measurement validity and predictive utility in problem formulation.

We find that data scientists construct target variables by **leveraging creative and pragmatic approaches to make do with the limited data they have at hand**. Data scientists craft their formulation through a process of iterative *dialogue* with their data, tweaking and refining their target variable until they conclude that it provides a sufficient operationalization for the unobserved construct of interest. To understand data scientists' target variable construction practices, we draw upon the concept of *bricolage*, originally introduced by Claude Lévi-Strauss to contrast the rigid top-down methodology of Western science with the more tangible “science of the concrete” practiced in non-Western societies (Levi-Strauss, 1966). Rather than implementing abstract concepts through top-down design, the bricoleur solves problems by arranging and re-arranging well-known materials in a bottom-up process. Like bricoleurs, data scientists in our study often acknowledged imperfection in their improvisational approach to measurement, yet they also identified innovative solutions that would be difficult to anticipate through a process of top-down design.

Based on our findings, **we synthesize a new process model that better reflects data scientists' bricolage approach towards measurement** (Figure 3.2). Data scientists balance the validity of their formulation with other criteria such as predictive performance, simplicity, portability, and resource requirements. Data scientists narrow in on a suitable formulation by drawing upon a rich repertoire of problem (re)formulation strategies. Data scientists apply a strategy, take a step back to assess their progress along each criterion, and then apply a new strategy until all evaluation criteria are satisfied. Data scientists sometimes discontinue their projects if they conclude that no strategies are sufficient given the resources at hand.

Our characterization of target variable construction as bricolage **offers an important counterbalance to nascent views on measurement of AI systems**. While recent work correctly identifies measurement problems as the source of many AI failures (Coston et al., 2023a; Kawakami et al., 2024; Guerdan, Coston, Wu and Holstein, 2023; Jacobs and Wallach,

2021; Raji et al., 2022; Liu, Blodgett, Cheung, Liao, Olteanu and Xiao, 2024; Raji et al., 2021; Alaa et al., 2025), proposed solutions often assume or reinforce a top-down, idealized measurement process. Our findings suggest that in order for established measurement principles to have impact in practice, there is a need for thoughtful reconciliation with the resource-constrained, bottom-up reality of data science work. By documenting how measurement and prediction intertwine in data scientists' bricolage practices, we contribute a more grounded understanding of how measurement actually occurs in the wild, providing a foundation for developing more effective interventions. While discussing implications of our findings, we **present new opportunities for future CSCW, HCI, and ML research to develop tools that better support data scientists' bricolage practices.**

In summary, our contributions to the CSCW literature are:

- A novel characterization of measurement in data science as a *bricolage practice*, showing how data scientists negotiate between measurement validity and predictive utility by creatively combining existing data under resource constraints.
- Empirical findings from interviews with fifteen data scientists in healthcare and education, synthesized into a process model that reveals how practitioners weigh competing criteria (§ 3.5.1), apply five strategic approaches to reformulation (§ 3.5.2), and evaluate validity through both theory-driven and data-driven practices (§ 3.5.3).
- A framework for reconciling traditional measurement approaches with bricolage practices, including concrete opportunities for CSCW, HCI, and ML research to develop tools that scaffold and enhance target variable construction without sacrificing the flexibility that makes bricolage effective (§ 8.7).

3.3 Background and Related Work

In this section, we begin by introducing Claude Lévi-Strauss' theory of Bricolage. We then describe existing Science and Technology Studies (STS), HCI, and CSCW research that has investigated questions related to categorization, measurement, and target variable construction in data science practice. We conclude by providing an overview of key concepts from measurement theory in the quantitative social sciences which we build upon in this work.

3.3.1 Bricolage and the Science of the Concrete

Bricolage, defined by Claude Lévi-Strauss as “making the most of available resources” (Levi-Strauss, 1966), describes how actors (i.e., *bricoleurs*) construct artifacts in creative and adaptive ways. Bricolage is central to what Lévi-Strauss calls *the science of the concrete*, which contrasts with modern scientific thought by focusing on the materials directly at hand rather than abstract principles. Bricolage contains three elements: repertoire, dialogue, outcome (Duymedjian and Rüling, 2010). *Repertoire* describes the resources—physical artifacts, ideas, or skills—that the bricoleur brings to bear on a problem. Whereas Lévi-Strauss viewed the traditional engineer or scientist as acquiring resources tailor-made to specification, the bricoleur slowly assembles various “odds and ends” that might be useful in the future. Then, when faced with a new problem, the bricoleur engages in *dialogue* with these materials—a process of arranging and re-arranging materials in her repertoire to find a workable solution. While engaging in dialogue, the bricoleur adopts old objects for new purposes and combines concepts with materials in novel ways. Finally, *outcome* describes the artifact resulting from the bricolage process. While bricolage entails a tolerance for imperfection, Lévi-Strauss argues that bricoleurs sometimes identify “brilliant unforeseen results” (Levi-Strauss, 1966) as part of their process.

Prior research has used Lévi-Strauss' framework to describe many forms of knowledge work as bricolage (e.g., entrepreneurship (Mateus and Sarkar, 2024; Duymedjian and Rüling, 2010), design (Büscher et al., 2001; Louridas, 1999; Vallgård and Fernaeus, 2015), management (Duymedjian and Rüling, 2010), and programming (Turtle, 2011)). For instance, Baker and Nelson (2005) use bricolage to explain how entrepreneurs build business ventures under resource constraints. They observe that entrepreneurial success hinges on the ability of firms to combine their limited resources in creative, adaptive ways.

In her book *Life on the Screen*, Sherry Turtle uses bricolage to describe how programmers write computer code. Through the mid-1980s, the canonical coding approach taught in classrooms was *structured programming*—a linear, top-down process that proceeds by systematically mapping abstract plans into procedural instructions. Yet, Turtle observed that programmers—novices and experts alike—often engaged in a more unstructured *dialogue* with their code while solving real-world problems. Drawing an analogy between the programmer and bricoleur, she explains:

“In the context of programming, the bricoleur’s work is marked by a desire to play with lines of computer code, to move them around almost as if they were material things—notes on a score, elements of a collage, words on a page” (Turtle, p. 51-52)

Turtle describes this style of tinkering as important for developing an intuition for the complexity of computer programs. Like a painter who steps back to inspect a canvas between brush strokes, programmers try one approach, re-evaluate, and then try another. Turtle observes that this process often leads programmers to more elegant solutions than could be imagined through the more traditional process of structured planning.

In this work, we draw a novel connection between *data science work* and the practice of bricolage. Much like an entrepreneur, programmer, or painter, data scientists construct target variables by combining *existing* resources in creative, adaptive ways. Instead of instantiating abstract plans through top-down design—akin to the structured programming approach taught in classrooms—data scientists engage in more direct *dialogue* with their data and modeling tools, making do with the materials at hand. And analogous to past research investigating how to support such exploratory and improvisational coding practices Kery and Myers (2017), this has implications for the design of practical tools to support target variable construction in practice. We highlight connections between target variable construction and bricolage throughout the findings (§ 8.6) and describe implications for tool design in the discussion (§ 8.7).

3.3.2 Tracing the Social Construction of Target Variables

Target variable construction also relates to a rich body of literature in Science and Technology Studies (STS) and philosophy of science that investigates the social processes that give rise to categories in data (Latour and Woolgar, 2013; Bowker, 2000; Pine and Liboiron, 2015; Harvey et al., 2024; Porter, 1996; Crosby, 1997; Hacking, 1999; Mol, 2002; Suchman, 1993). Latour and Woolgar (2013) provide an account of the social construction of scientific facts within laboratories, highlighting that facts are not merely discovered but actively produced through a complex interplay of social and technical factors. Likewise, Bowker (2000) examines how systems of classification, such as the International Classification of Diseases (ICD), shape and are shaped by the social and political contexts in which they are embedded. Building on Bowker's work, recent studies have also explored the construction of racial and ethnic

categories within AI models (e.g., (Abdu et al., 2023; Mickel, 2024)) and charted measurement assumptions underpinning motion capture technologies (Harvey et al., 2024).

Most related to our study, Muller et al. (2021) interviewed fifteen data scientists to investigate the *data labeling* and *annotation* practices of data science teams. Muller et al. (2021) identified three approaches that teams used while defining labels: *principled*, *iterative*, and *improvisational*. Teams engaging in principled design followed a top-down planning approach, systematically specifying annotation guidelines before data labeling. Teams leveraging iterative design took a bottom-up approach, incorporating opportunities for annotation protocol refinement throughout the labeling process. Finally, teams engaging in improvisational design viewed data labeling as an open-ended, exploitative process. Our study investigates a distinct question of how data science teams construct target variables for *predictive modeling tasks*. Rather than studying how data science teams design “ground truth” labels anew, we examine how data scientists connect their *existing* data to their high-level modeling objectives as part of the problem formulation process.

3.3.3 Empirical Studies of Data Science Practice

Our work also builds upon empirical studies investigating how data scientists conduct their day-to-day work. Many studies have examined activities in the technical “inner loop” (Kross and Guo, 2021) of data science practice, such as exploratory data analysis (e.g., (Alspaugh et al., 2018; Wongsuphasawat et al., 2019)), feature engineering (e.g., (Muller et al., 2019)), and visualization (e.g., (Kandel et al., 2012)). Other studies have investigated broader “outer loop” activities performed by data scientists, including collaborating with other stakeholders (Kross and Guo, 2021; Passi and Jackson, 2018; Nahar et al., 2022; Mao et al., 2019; Kim et al., 2016; Zhang et al., 2020; Hou and Wang, 2017) and addressing ethical concerns (Holstein et al., 2019; Deng et al., 2022). CSCW research has described inner and outer loop activities as iterative and interconnected (Kross and Guo, 2021; Passi and Barocas, 2019). In particular, based on interviews with ten professionals spanning industry and academia, Kross and Guo (2021) observed that data scientists iteratively “framed the problem” (i.e., an outer loop activity) and returned to their data to gain an understanding of its affordances (i.e., inner loop activities).

However, we currently have an incomplete understanding of *how* data scientists construct target variables while formulating prediction tasks. To date, Passi and Barocas (2019) provide the most detailed account of how problem formulation unfolds in real-world modeling contexts. Over the course of an eighteen-month ethnographic study of a corporate data science team, Passi and Barocas (2019) observed that data scientists repeatedly reformulated their prediction task as they discovered new information about data availability constraints and the predictability of alternative outcome variables. Based on their findings, Passi and Barocas (2019) describe problem formulation as a *negotiated translation* between high-level objectives and a concrete prediction task. Our work builds upon this finding to identify the more general evaluation desiderata and (re)formulation strategies driving the problem formulation process. We develop a process model through analysis of our findings and use this model to identify targeted opportunities to better support data scientists’ problem formulation practices going forward. For example, when viewed through the lens of our findings, the team studied by Passi and Barocas (2019) used *one of five* (re)formulation strategies (i.e., *swapping*) to change outcome variables when the first proved infeasible due to data availability constraints. We present design opportunities to help data scientists more effectively apply swapping (§ 3.5.2) and other (re)formulation strategies.

3.3.4 Toolkits Supporting Target Variable Construction

The limitations of data science toolkits and analysis methods have been documented across empirical studies examining data science practices (Jun, Birchfield, De Moura, Heer and Just, 2022), software repositories (Hoess et al., 2025), and ML-based sciences (Kapoor and Narayanan, 2023). Jun et al. reviewed 20 commonly-used data analysis tools and found that the ecosystem of tools supports vastly different model implementations, even when using the same underlying mathematical equation (Jun, Birchfield, De Moura, Heer and Just, 2022). Although these tools support nuanced model implementations, their low-level technical abstractions make it challenging for analysts to engage in the high-level conceptual reasoning required for hypothesis formalization (Jun, Birchfield, De Moura, Heer and Just, 2022). As a result, interpretations of data, modeling assumptions, and implementation details can significantly influence outcomes and impact the validity of results (Hoess et al., 2025; Jun, Birchfield, De Moura, Heer and Just, 2022; Jun, Seo, Heer and Just, 2022; Kapoor and Narayanan, 2023). Research examining software engineering practices in code repositories, such as GitHub, further demonstrates how the choice of analysis tools can lead to substantial differences in simple metrics (e.g., commit counts and developer contributions). This further poses threats to validity of findings derived from one analysis tool (Hoess et al., 2025). Given this gap between technical abstractions and conceptual reasoning—and the embedded assumptions inherent in machine learning methods like feature engineering—it is not surprising that ML-based sciences continue to face a reproducibility crisis (Kapoor and Narayanan, 2023).

To mitigate these issues, the HCI, ML, and FAccT research communities have developed a range of toolkits to intervene at points along data scientists' inner and outer loop workflows. For instance, researchers have developed systems facilitating data collection and configuration (e.g., (Bhattacharya et al., 2024; Guo et al., 2011; Kandel et al., 2011; Rattenbury et al., 2017)), exploratory data analysis (e.g., (Wongsuphasawat et al., 2015; Kery et al., 2018)), model training (e.g., (Pang et al., 2020; Dingen et al., 2018)), caching intermediate states during data science experiments (e.g., (Schubotz et al., 2022)), and performance assessment (e.g., (Cabrera, Fu, Bertucci, Holstein, Talwalkar, Hong and Perer, 2023; Yan et al., 2021)). Many tools have also been developed to help data scientists explain their models (e.g., (Ribeiro et al., 2018, 2016; Lundberg, 2017)) or assess fairness-related concerns (e.g., (Tramer et al., 2017; Bellamy et al., 2019; Weerts et al., 2023; Saleiro et al., 2018; Liu, Chen, Shen and Choo, 2024; Cabrera et al., 2019; Ahn and Lin, 2019; Wexler et al., 2019; Dingen et al., 2018)). In the FAccT community, researchers have also developed responsible AI checklists and guidelines to help data scientists assess and document components of their AI system (e.g., dataset information (Geburu et al., 2021; Pushkarna et al., 2022), intended use cases (Mitchell et al., 2019)). To our knowledge, Gala et al. (2024) propose the only system specifically designed to support the target variable construction process. Instead, **existing toolkits are often designed to support data scientists' practices only after a target variable has been specified.** This presents a pressing gap in data science support at a critical point in their workflows.

However, prior HCI research suggests that such toolkits may prove ineffective if they are not designed around the *existing* practices of data scientists. For instance, past efforts to improve programming tools without a human-centered assessment of programmers' needs yielded ineffective languages and systems (Myers et al., 2016; Hazzan and Tomayko, 2004; Seffah et al., 2005; Ko et al., 2011). Similarly, the utility of ML fairness toolkits was found to be limited in practice as they were designed for isolated use by technical roles, while actual fairness work often relied on deep cross-functional collaborations (Deng et al., 2022). Our work aims to help the research community avoid similar challenges while developing tools to support data scientists' problem formulation processes. In particular, we identify opportunities for the HCI, CSCW, and ML research communities to develop tools that help data scientists more carefully

weigh trade-offs across multiple evaluation criteria (§ 3.6.3), select (§ 3.6.3) and apply (§ 3.6.3) strategies, and evaluate the validity of their problem formulation more effectively (§ 3.6.3).

3.3.5 Measurement and Validity in AI

A growing line of research traces AI system failures to *validity* pitfalls (Jacobs and Wallach, 2021; Milli et al., 2021; Stray et al., 2021; Guerdan, Coston, Wu and Holstein, 2023; Goyal et al., 2022; Obermeyer et al., 2019; Mullainathan and Obermeyer, 2021; Coston et al., 2023a). In the quantitative social sciences, the validity of a measurement instrument is established via a multifaceted assessment along several dimensions. *Construct validity* describes the extent to which a measurement instrument wholly and fully reflects the theoretical construct it purports to measure (Jacobs and Wallach, 2021). In a predictive modeling, construct validity is most often discussed in connection to whether the target variable predicted by a model reflects a theoretical construct of interest to model developers (Guerdan, Coston, Wu and Holstein, 2023; Coston et al., 2023a).¹ *Internal* and *external* validity are also important dimensions in experimental research. *Internal validity* describes the extent to which an experimental study can confidently establish a causal relationship between independent and dependent variables (Gergle and Tan, 2014). In predictive modeling, it relates to the existence of a defensible causal relationship between features and a target variable (Coston et al., 2023a). *External validity*, on the other hand, measures how well an analysis can be generalized beyond the specific conditions of the study. This dimension has been linked to the extent to which a predictive model's in-distribution performance generalizes across real-world deployment contexts (Coston et al., 2023a).

While these validity dimensions articulate helpful conceptual desiderata for a predictive model, they provide a limited picture of *how* practitioners actually navigate measurement decisions in their day-to-day work. Mussnug (2022) argue that current practice is characterized by a “predictive reframing”, in which data scientists cast a measurement problem (inferring existing but unobserved quantities) as a prediction task. This reframing shifts the epistemic aim from measuring a concept to predicting a given measurement of that concept. For example, when data scientists attempt to assess poverty levels using satellite imagery, they often frame their task as “predicting poverty” rather than “measuring poverty.” This practice centers the evaluation of a machine learning model around narrow predictive performance characteristics while sidestepping deeper questions of validity.

We expand upon Mussnug (2022)'s argument by characterizing how data scientists weigh validity and predictive performance during problem formulation. In support of Mussnug (2022)'s argument, we find that many prediction tasks pursued by data scientists are indeed measurement problems. For example, some participants in our study used supervised learning models to measure concepts recorded at future points in time — e.g., “10-year cardiovascular risk.” Others used supervised learning models as measurement instruments while measuring a concept at a current point of time — e.g., to score the “authenticity” of a student's writing. In both cases, data scientists heavily anchored on predictive performance while evaluating their problem formulation. However, counter to Mussnug (2022)'s argument, we find that data scientists *do* also engage with validity concepts during problem formulation, albeit with a more pragmatic, situated approach that lacks formal terminology. We uncover how data scientists evaluate validity in our findings (§ 8.6), and unpack implications for data science pedagogy (§ 3.6.2) and tooling support (§ 3.6.3) in the discussion.

¹Jacobs and Wallach (2021) provide a detailed review of sub-dimensions of construct validity as they pertain to measuring fairness properties of algorithmic systems.

3.4 Methods

To examine how data scientists construct target variables for predictive modeling tasks, we interviewed 15 data scientists from the education (N=8) and healthcare (N=7) sectors who had prior experience with developing and evaluating predictive models. We conducted one-hour semi-structured interviews to understand participants' current practices and perceived challenges for measurement as they engaged in problem formulation.

3.4.1 Study Design

Directed Storytelling

During the first interview segment, we used a directed storytelling approach [Evenson \(2016\)](#), inviting data scientists to reflect on a time in the past when they developed and evaluated a predictive model. We centered questions in this segment around data scientists' *target variable construction process*, a component of problem formulation which prior literature has shown is both complex and consequential ([Obermeyer et al., 2019](#)).

To ground our discussions, we first asked participants to share a few concrete experiences where they had developed and/or evaluated predictive models as part of their role. Building on these concrete experiences, the interviewer then narrowed in on one of the projects mentioned and asked participants the following follow-up questions, in a semi-structured fashion:

1. Were there ever times when you felt uncertain about your choice of target variable?
2. Could you walk me through how your team went about picking a target variable?
3. How did you know whether your team selected the "right" target variable?

We also asked participants a range of additional follow-up questions when these were not already covered by participants during their earlier responses, such as measurement guidelines and best practices they consulted throughout their project. On average, the first segment lasted 30-40 minutes of the interview.

Vignette-Based Problem Formulation Task

During the second segment, we asked participants to reason about a hypothetical problem formulation as we walked them through a vignette. We aimed to better understand how participants reason about measurement decisions on-the-fly, while potentially surfacing challenges and opportunities they might not have recalled while reflecting on their prior experiences. We asked participants with a healthcare background to imagine they were developing a model to predict patients with high future "medical acuity," and we asked participants with an education background to imagine they were developing a model to predict which students were "academically at-risk".² We provided participants a set of features and several outcome variables and instructed them to imagine that these were the data available for model development.

During the vignette, we asked participants to reflect on which outcomes might serve as a "better or worse choice" of prediction target given the stated modeling goals. After learning how participants initially approached this task, we then presented a series of model evaluation plots indicating the performance of predictive models targeting each outcome. After showing each plot, we paused and asked participants to again reflect on which outcomes might serve

²P1 was shown the healthcare vignette due to data availability constraints encountered at early phases of the interview schedule. During the interview, P1 signaled places where their interpretation of the vignette was informed by their domain experience. We took this into consideration during analysis.

PID	Domain	Postgrad Experience	Current Primary Role	Current Affiliation
P1	Education	15 years	Board Member	Industry
P2		7 years	Senior Analyst, Ph.D. Student	Industry/Academic
P3		5 years	Postdoc	Academic
P4		2 years	Ph.D. Student	Academic
P5		2 years	Ph.D. Student	Academic
P6		11 years	VP of Data Science	Industry
P7		10 years	Data Scientist	Industry
P8		11 years	Senior Research Scientist	Industry
P9	Healthcare	7 years	Senior Research Scientist	Industry
P10		6 years	Research Associate	Academic
P11		4 years	Senior Data Scientist	Industry
P12		11 years	Senior Data Science Analyst	Industry
P13		8 years	Senior Data Scientist	Industry
P14		8 years	PostDoc	Academic
P15		10 years	Assistant Professor	Academic

Table 3.1: Demographic overview of study participants: data scientists' experience, roles, and affiliations.

as better or worse choices for the stated modeling goals. We also asked participants to share any additional information that they would ideally like to have, to better inform their thought process. As participants engaged with the task over subsequent interviews, we refined the specificity of the scenario description and accompanying evaluation plots to further probe issues that surfaced in prior interviews. This second segment lasted 20-30 minutes on average.

3.4.2 Participants

Participants were recruited via purposive sampling techniques (Heckathorn, 2011) from authors' professional networks and social media (Table 3.1). To ensure that participants had a baseline level of domain expertise, we required that data scientists had domain-specific graduate-level training related to healthcare or education (e.g., biostatistics, learning sciences) or at least one year of full-time industry experience in a relevant domain to be eligible to participate. We further required that participants had prior experience developing and evaluating predictive models as part of their role. Our participants had a range of post-graduate experience at the time of the interview, spanning from 3-6 months to 15 years. At the time of the interview, participants had a variety of primary roles (e.g., Data Scientist, Assistant Professor, Board Member, PostDoc) from both academic and industry affiliations. As a result, our sample consists of data scientists with a baseline level of domain knowledge and a range of technical experience. This study was approved by the University Institutional Review Board and all participants were compensated with a \$35 Amazon gift card for completing the study.

3.4.3 Data and Analysis

One of the authors conducted interviews remotely via Zoom between September and December 2023. In total, we recorded and transcribed 14 hours and 30 minutes of audio after collecting verbal and written consent from all participants. Two authors independently performed open

coding on transcripts and met regularly to discuss and reconcile interpretation differences. After interviewing fifteen participants, we observed a common set of themes recurring across interviews and took this as an indicator that we had reached saturation (Saunders et al., 2018). We adopted a reflexive thematic analysis approach (Braun and Clarke, 2006; Clarke and Braun, 2017) to group codes into successively higher themes, informed by discussions among all of the authors. Our process gave rise to sixteen high-level themes, organized into three categories: (1) *factors* driving the target variable construction process, which we discuss in §3.5.1; (2) data scientists' *model development* and *evaluation* practices, which we discuss in §3.5.2 and §3.5.3; and (3) data scientists' *orientation* towards the problem formulation process, which we discuss throughout the findings, when drawing connections to the concept of bricolage.

3.4.4 Positionally

Our research approach is informed by our backgrounds spanning multiple disciplines relevant to the study of measurement and validity in data science practice. Our research team includes individuals with training in machine learning, human–computer interaction, cognitive psychology, and design. This interdisciplinary background helped our team understand the technical activities performed by data scientists, while also situating the practices we documented in interviews within their broader sociotechnical context. For instance, several team members have professional academic and industry experience working as data scientists. Several others have training in educational analytics and data science for behavioral health. This background helped our team recruit participants and build rapport during interviews. At the same time, our team's exposure to formal validity concepts through the quantitative social sciences helped our team situate the practices we documented during interviews in formal validity theory. While designing our protocol, conducting interviews, and analyzing data, this interdisciplinary background directed our focus towards key tensions between the messy reality of data science practice, which we personally encountered in prior projects, and formal validity concepts, which we engaged with in academic research. Several team members with experience in design had exposure to Claude Lévi-Strauss' concept of *bricolage*, which we eventually adopted as a central metaphor for describing our findings. In sum, the pragmatic orientation towards data science practice embodied in our findings emerged from a combination of academic social science training and real-world data science experience.

3.5 Findings

Our findings illustrate that target variable construction is a bricolage practice, in which data scientists design measurement instruments (i.e., target variables) for unobservable constructs by creatively making do with their existing data. While measurement has traditionally been conceptualized as a *monologue* (Turkle, 2011) — a one-way, top-down translation from theoretical constructs to measurements in data — we find that data scientists instead construct target variables through a rich process of *dialogue* with their data, driven by a need to make do with resource constraints.

The specific process of bricolage unfolds as depicted in Figure 3.2. Data scientists begin by formulating an initial prediction task based on available data. They then iteratively inspect their formulation along several dimensions: “Can the target variable be predicted with ‘reasonable’ accuracy?”, “Will the necessary data be available across deployment contexts?” When issues arise along one or more criteria, data scientists apply various (re)formulation strategies to address these gaps. These strategies serve as the “brushstrokes” of data science work. Like a painter with a canvas, data scientists apply a strategy, step back to survey their progress along

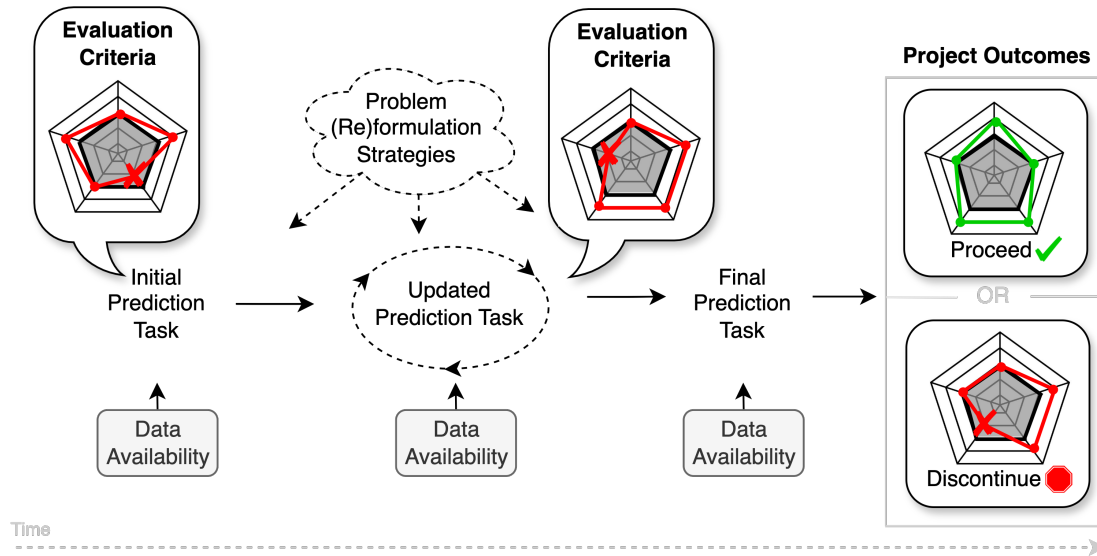


Figure 3.2: An illustration of the target variable construction process presented in our findings. Data scientists specify an initial target variable based on their available data, then iteratively refine their prediction task by applying (re)formulation strategies. Data scientists proceed with a formulation if it satisfies all criteria, or discontinue their project if strategies are exhausted.

multiple criteria, then consider a new strategy in response to revealed flaws. The resulting problem formulation, while rarely ideal, reflects what the bricoleur considers “fit for purpose”, by acknowledging imperfection while identifying creative and practical solutions that might never emerge through rigid top-down design. Data scientists eventually proceed with a formulation if they believe it satisfies their criteria across multiple dimensions, or discontinue their project if all strategies have been exhausted.

In the following subsections, we detail this bricolage process by describing how data scientists balance validity with other competing criteria (§ 3.5.1), the (re)formulation strategies they apply to navigate this balancing act (§ 3.5.2), and the theory-driven and data-driven approaches they use to evaluate the validity of their formulation (§ 3.5.3).

3.5.1 Balancing Validity With Other Criteria

While engaging in target variable construction, data scientists were forced to make do with the data they already had available for the task at hand. This process of making do forced data scientists to navigate a rich space of tradeoffs between validity and other important criteria, such as predictive performance, simplicity, resource requirements, and the ease by which it could be translated to different geographic or institutional contexts (i.e., portability) (Figure 3.3). In Table 3.2, we provide an overview of illustrative tradeoffs that data scientists made between validity and other criteria. We provide a more detailed description of how the process of making tradeoffs unfolded in the sections below.

Balancing Validity with Simplicity

While constructing prediction tasks, several data scientists navigated a tension between the *validity* and *simplicity* of their formulation. In particular, data scientists found complex formulations, such as those combining multiple outcome variables, hard to explain to other stakeholders and maintain in production. For example, P1 was a director of a data science

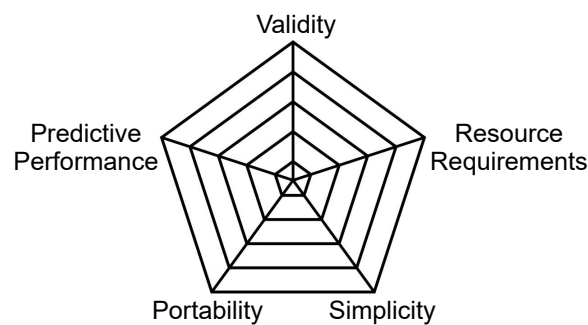


Figure 3.3: Data scientists evaluated the validity, predictive performance, portability, simplicity, and resource requirements of their problem formulation.

Criterion	Description	Illustrative Tradeoff
Validity	The extent to which a model assesses the specific concept or construct that the data scientist intends to measure (Coston et al., 2023a; Drost, 2011).	N/A
Predictive Performance	The extent to which a model's predictions align with observed outcomes (e.g., accuracy or AU-ROC).	Predicting cardiovascular rather than renal diabetes health outcomes (P10). See §3.5.2.
Resource Requirements	The organizational resources required to implement and maintain the formulation.	Making do with limited test scores to avoid imposing additional testing (P6). See § 3.5.1.
Portability	The ease by which a task definition can be transferred across geographic or institutional contexts.	Using coarse veterinary health information to support model inference across clinical contexts (P13).
Simplicity	The complexity of a task definition or model's structure and interpretation.	Switching from a complex multi-outcome model due to challenges in explanation (P2). See § 3.5.2.

Table 3.2: A description of criteria that data scientists considered while evaluating their problem formulation and illustrative examples of how participants balanced each with validity.

team at a large education company. Because the essay scoring tool designed by their team was used by schools throughout the country, P1 described the *simplicity* of their formulation as critical for ensuring the system was robust and maintainable. P1 *had* considered adopting a sophisticated multi-target formulation to improve validity. Yet, their team later abandoned the approach over practical concerns, explaining that the sophisticated models “are a real hassle in practice. The extra [...] bells and whistles [...] made it flakier code, something that we didn’t

necessarily want to deploy at scale, compared to a relatively straightforward single output variable model. Further, P1 later shared that they felt it was possible to proceed with using a simple single output variable model if they made several hyperparameter adjustments and user interface tweaks. In this experience, P1 was willing to settle for a specification that had minor validity drawbacks because they felt it would be simple and easy to maintain at scale.

Balancing Validity with Resource Requirements

Data scientists also evaluated the validity of their formulation against its resource requirements. For example, P6 was a data scientist developing an online math tutoring platform. P6 explained that the time and cost involved with acquiring “gold standard” student test scores to evaluate their models was a frequent challenge. School districts often had data sharing restrictions. Opportunities to test students were limited. However, P6 explained that even if they *could* test students more often, they would not recommend that approach. Testing students is “*just removing opportunity. Removing time from the students’ school calendar, which is already [...] crowded.*” Instead of collecting data, P6 devised a creative workaround that achieved their project’s goals using the data they *already* had available.³ While P6 acknowledged that their workaround was imperfect, they were fine with proceeding so long as they had “*at least some confidence*” in its validity. In particular, while reflecting on potential drawbacks of their modeling decisions, P6 shared that “*this [application] isn’t high stakes*” when compared to other potential modeling contexts (e.g., university admissions). However, had the model been deployed in other contexts, they explained that they might apply a different level of stringency to the validity of their approach.

Elasticity of Evaluation Criteria

We observed that data scientists applied varying levels of flexibility to different evaluation criteria, a characteristic we term *elasticity*. Participants frequently demonstrated elasticity with validity standards, tolerating imperfections when doing so reduced the complexity or resource requirements of their formulation. Data scientists actively *calibrated* this tolerance based on the success of compensatory adjustments (e.g., user interface tweaks) and the perceived “stakes” of their modeling decisions. Interestingly, participants maintained more rigid standards for other criteria. For instance, several participants cited hard cutoffs for acceptable predictive performance — e.g., P8 applied an AU-ROC threshold of 0.7 to determine whether to proceed with a project. This rigidity in predictive performance evaluation likely stems from the ease with which it can be distilled into a single, quantifiable metric, unlike validity, which requires a more holistic assessment involving multiple sources of evidence (Jacobs and Wallach, 2021). While tradeoffs are an inherent component of bricolage, motivated by adaptive response to resource constraints, the process of *navigating* tradeoffs remains poorly supported. We describe opportunities to help data scientists navigate tradeoffs in Section 3.6.3.

3.5.2 Strategies for (Re)formulating Prediction Tasks

Given their limited and imperfect data resources, how do data scientists engage in bricolage to craft a formulation that satisfies their objectives? Much like a painter refining a painting through successive brushstrokes, or a programmer compiling and re-compiling a system after adjusting lines of code (Turkle, 2011), data scientists iteratively construct target variables through *(re)formulation strategies*. These strategies form the backbone of the dialogue between the data scientist and their formulation, enabling them to respond adaptively to constraints encountered

³We detail the strategy P6 and their team used in Section 3.5.2.

while working with existing data. At times, data scientists apply these strategies instinctively, without a second thought. At other times, they shared eureka moments they encountered after discovering a strategy that leveraged unexpected data sources or combined information in novel ways. In the following sections, we detail five key strategies—piggybacking, composing, swapping, bridging, and refining—that data scientists used to navigate the constraints of their limited repertoire.

Piggybacking

During the initial stages of the problem formulation process, data scientists often constructed target variables through **piggybacking**: *adopting or extending a problem formulation that was previously used in a similar context in the past*. Participants justified piggybacking by citing precedents from related projects. For example, while developing a model to inform university admissions decisions, P5 suggested predicting whether students' college GPA exceeded a 3.0 because other data science teams used it for similar admissions pipelines in the past. Piggybacking provided a rationale for P5's formulation decisions and offered a point of comparison with prior work. However, they were also wary that piggybacking could introduce unintended validity consequences.

One participant encountered a validity flaw in their formulation after piggybacking. P1 was developing an automated essay scoring tool on a six-point scale, where six is the best. P1 initially piggybacked by adopting the most prominent evaluation metric in the literature (Quadratic Weighted Kappa) but quickly realized it led to jarring results.

"The model simply never gave that top score. The more accurate [model] predicted the two's, three's, four's, five's. There are students out there that simply don't accept that. They'll write 30, 40, 50, 60 attempts at an essay, trying to get [...] a perfect score."

Piggybacking off of other modeling contexts did not work for P1's dataset - optimizing for Quadratic Weighted Kappa yielded a model that almost never gave a perfect score. This led P1 to eventually adopt a performance metric better suited to their setting, which balanced predictive accuracy with some chance that students could get a perfect score. P1's experience illustrates that theoretical analysis of established precedent (Quadratic Weighted Kappa) was insufficient for making a well-informed piggybacking decision. Instead, P1's experience illustrates the "science of the concrete" characteristic of bricolage, where tangible interactions with materials are important for understanding how to assemble elements in the repertoire into a functional artifact.

Composing

Several data scientists constructed prediction tasks by **composing**: *assembling multiple outcome variables into a single prediction target*. Often, data scientists used composing after realizing that a single outcome variable did not capture all dimensions of a construct they were trying to measure. For example, P8 used composing to predict the "academic work ethic" of college freshmen from their high school performance. While high school GPA was an established predictor of collegiate success, P8 felt that GPA alone would not offer enough specificity to differentiate work ethic from other success factors. P8 reasoned about how they might *compose* standardized test scores and GPA in a single measure for "academic work ethic" by combining the two:

"If you assume that SAT or ACT scores are [...] indicators of academic knowledge, [...] this is where the person's at. If there's a difference between where the person should be and where they're at, we thought it was reasonable to conceptualize that as work ethic."

P8 designed a regression model that predicted GPA given test scores, creating a model that accounted for the gap between them to capture "work ethic." A negative residual indicated that students had lower grades than predicted by their test scores, thereby signaling work ethic concerns. While teaching an introductory freshman class, P8 was surprised to find that students predicted to have poor work ethic were more willing than others to copy homework. Her model's predictions also closely aligned with students' long-term persistence outcomes (e.g., retention and graduation). These observations provided evidence for the *construct validity* of their model's operationalization of "work ethic."

Yet some participants found that *composing* had simplicity limitations. For example, P2 tried to use a multi-outcome latent variable to predict student "engagement." However, when P2 presented the model to software engineers and product managers, they distrusted its predictions when the model contradicted their prior beliefs about new product features. P2 explained that:

"If I show them a latent model, and it says, 'No, it doesn't look like you improved engagement' then there's gonna be a lot of questions about that model. What is this actually measuring? How do you explain it?"

P2 found it easier to justify their model when it targeted a single outcome based on what software engineers and product managers viewed as "hard data." As a result, while P2 acknowledged validity benefits with latent variable models, they did not use them in industry practice. This tension between validity and simplicity highlights how *bricolage* solutions often bear the mark of compromise, as data scientists negotiate between theoretical ideals and practical constraints.

Swapping

As data scientists worked with their formulation, they sometimes reassessed their initial choice of outcome variable. The **swapping** strategy describes *changing the outcome variable when the initial one fails to meet evaluation criteria*. For example, P10 used swapping while developing a model to predict diabetes-related "health complications." P10 initially operationalized "health complications" using a cardiovascular outcome. Yet after discovering that cardiovascular outcomes were hard to predict, they later swapped to a measure of kidney function:

"At some point, after [...] seeing that this [cardiovascular] condition is quite hard to predict, I thought that switching to something that is strongly related to the biomarkers I had available would make more sense. [...] Seeing the data, getting to work with data, you might need to change your outcomes."

However, P10's use of renal outcomes was controversial because the renal biomarkers used as features were closely related to the outcome definition. As a result, it was not clear that the new model captured a true predictive signal in their data. This concern relates to *internal validity*, or "the existence of a defensible causal relationship between features and the target label" (Coston et al., 2023a). However, the goal of this project was to demonstrate a new prediction method, so P10 tolerated internal validity limitations. P10 explained that they would place greater weight on these internal validity concerns if they were deploying their

model in a clinical context. P10's experience embodies the responsive dialogue central to bricolage, where "seeing the data, getting to work with data" led to a reconsideration of the initial approach based on the artifact that could be constructed with the materials at hand.

Bridging

While working with their data, two participants identified "gold standard" outcomes with established validity that were rarely available and infeasible to directly predict. Data scientists addressed this challenge by **bridging**: *targeting a low-cost, readily available outcome as a surrogate for the more expensive gold standard measure*. While bridging, participants evaluated their surrogate model using the gold standard outcome to construct a "bridge" between the surrogate model and the unobservable construct of interest.

For example, P6 used bridging to evaluate the "readability improvement" of content changes they made to an online learning platform. While P6 wanted to assess whether their changes helped struggling readers, they only had access to "gold standard" test scores from a small fraction of schools. P6 circumvented this issue by using students' performance on early lessons in the software as a surrogate for their reading ability:

"We had this idea that came from another paper that [...] this early lesson [...] was a good proxy for reading ability. So we built some [...] models that took raw interaction data and [we] found a neural network model that does a reasonable job of predicting student's [test] performance from this first lesson."

P6 validated their surrogate model by checking its performance against the test scores they *did* have available. Once they were confident that the model had "reasonable" validity, they used it to evaluate whether content changes helped students predicted to be poor readers. They found that the content updates offered a large readability improvement and deployed them on the platform, declaring the project "a really big success." P6 acknowledged that both the gold standard tests and the formulation were imperfect. However, given the analytics focus of the project, P6 had some tolerance for validity imperfections in their bridging formulation. This tolerance for imperfection, balanced against practical gains, exemplifies the bricoleur's pragmatic approach to problem-solving under resource constraints.

Refining

While constructing a target variable, participants also iteratively tweaked and refined their formulation by making small changes. This strategy involved **refining**: *modifying the target variable definition in response to a defect along one or more criteria*. In contrast to the larger change in outcomes involved with swapping, participants used refining to make more granular adjustments to their target variable.

For example, P13 used *refining* to narrow in on definitions for animal diseases in her dataset. While human healthcare had a mature ICD-9 disease classification scheme, P13 explained that animal health conditions remained vague and poorly specified. Clinicians were also unsure of how to map their understanding of pet diseases into a **label definition**: or a *systematic set of classification rules in data*. P13 addressed this issue by refining the label definition with clinicians.

"Clinicians [...] can't think of how do I define the dog has diabetes? I would have the vet look at it and [ask if] there [are] symptoms and signs [...] I couldn't get right. That takes many times of iteration to find out."

At each stage of refinement, P13 and the vet used written clinical notes about the pet to identify animals who were labeled incorrectly under the current definition. P13 then revised the label definition to correctly categorize the misclassified instances (e.g., adding a new rule-based white blood cell count). These iterations continued 20 to 30 times until the label definition sufficiently captured the disease. P13 emphasized a need for practicality during the process: they was fine with 10% of diseases being mislabeled if the major dimensions of the disease definition were covered. P13's acceptance of imperfection while ensuring "major dimensions" were covered typifies the bricoleur's focus on creating solutions that are "fit for purpose" rather than ideal.

When Reformulation Fails.

After applying strategies, many data scientists eventually crafted a prediction task that satisfied their criteria. However, one discontinued a project after identifying a flaw that they could not patch through reformulation. P12 was developing a model to flag patients in the general care clinic who were at risk of Celiac disease. An initial model predicted test results with good accuracy. However, after carefully inspecting their formulation, P12 realized that patients with test results already appeared sick to physicians. Training a model on patients who *appeared sick* then deploying the model to *the full patient population* threatened the model's external validity. P12's team considered a variety of workarounds to patch this gap. However, when no solutions appeared feasible, P12 discontinued the project. While most data scientists applied strategies in response to visible defects in their formulation (e.g., predictive performance concerns), the flaw identified by P12 was *invisible*. In particular, identifying this issue required scrutinizing the connection between the dataset available for model development (i.e., patients who *appeared sick* to doctors) and the population at deployment time (i.e., the general population entering the clinic). We describe other approaches that data scientists used to shed light on invisible validity defects in Section 3.5.3.

3.5.3 Evaluating the Validity of a Problem Formulation

While applying strategies, data scientists often took a step back to evaluate the validity of their problem formulation. Drawing on domain knowledge, they engaged in theory-driven practices to probe conceptual relationships between their target variable and theoretical constructs related to their problem formulation. However, data scientists also treated their problem formulation as a material object to be scrutinized by inspecting concrete predictions generated on specific data instances. Lévi-Strauss calls this more tangible approach "the science of the concrete" — whereby the bricoleur understands their artifact through direct manipulation of material elements as opposed to abstract, conceptual reasoning. We now describe these theory-driven and data-driven practices in detail, as summarized in Table 3.3.

Theory-Driven Evaluation Practices.

While engaging in theory-driven practices, data scientists used their domain knowledge to reason about a problem formulation *conceptually*, without scrutinizing a specific data sample. These theory-driven practices were informed by participants' prior modeling experience and an understanding of the sociotechnical context (e.g., hospitals, classrooms) in which their models operate. Though few participants had formal measurement theory training, many theory-driven practices connect to established construct validity sub-criteria.

While engaging with the problem formulation vignette, several participants identified spurious factors that might affect an outcome variable while remaining unrelated to the construct of interest. When considering "length of hospital stay" as a measure for "health

acuity,” P11 noted that “*there’s so many other things that could contribute to somebody being in the emergency department for 12 hours that might not have anything to do with how sick they are*”, citing staffing levels and emergency department crowding as examples. In validity theory, this reasoning connects to *discriminant validity*, or the extent to which a problem formulation measures aspects of other unrelated constructs (Coston et al., 2023a).

Several participants also used multiple outcomes to develop a more complete understanding of how a construct should be conceptualized. P2 wanted to understand whether students who used hints were “gaming” the lesson – i.e., using hints to finish questions without seriously engaging with the material. Accuracy did not provide enough information to identify gaming from hint use because students could answer a question incorrectly, even after using a hint and applying effort. However, P2 found that students took longer to respond when they were engaged with the lesson. As a result, a combination of *accuracy*, *response time*, and *hint use* was necessary to triangulate gaming behavior.⁴ Reasoning across multiple outcomes connects to an assessment of content validity, or “the extent to which an operationalization wholly and fully captures the substantive nature of the construct purported to be measured” (Jacobs and Wallach, 2021).

Finally, participants described their choice of outcome variable as closely linked to the intervention being informed by a predictive model. For example, P6 needed to decide which outcome time horizon (i.e., question lesson, week, semester, or year) was most appropriate for measuring the affect of platform content changes on student learning. P6 shared that “*to get in the right neighborhood, where would we reasonably expect to move the needle?*” If the intervention included many content changes, they leaned towards an end-of-year student performance outcome. However, if a change to the platform was more granular, they would start at a more fine grained question or lesson level. This assessment of the relationship between interventions and outcomes connects to *consequential validity*, or “the real-world consequences of using the measurements obtained from a measurement model” (Jacobs and Wallach, 2021).

Data-Driven Evaluation Practices

While engaging in data-driven practices, participants checked whether their theoretical understanding of a construct was borne out by the measurements they observed on specific data instances. Gaps between theoretical expectations and observations signaled a potential flaw in their formulation to be scrutinized. By identifying, probing, and later reconciling these gaps through a strategy, data scientists slowly built trust in their formulation’s validity. In some cases, participants checked how a target variable categorized specific data instances (e.g., “ground truth” labels constructed by applying a label definition). In others, they checked the downstream predictions generated by a model trained on a target variable. Each of these practices depended heavily upon the specific data sample used to generate measurements. As such, our understanding of these activities emerged through the directed storytelling segment of interviews, as participants recounted specific times they manipulated a dataset in prior projects.

Several participants applied *spot checks and heuristics* to test whether measurements satisfied basic theoretical properties they expected. For example, while evaluating new test items, P6 plotted students’ scores as a function of their number of question attempts. A “*nice, monotonically increasing curve*” with a “*logarithmic flavor*” provided evidence that the item encoded knowledge of a single skill. P9 checked that the base rate of a disease derived from diagnostic codes matched the expected population prevalence. A large mismatch was a signal

⁴Participants who reasoned across multiple outcomes sometimes suggested composing. However, in some cases, participants used multiple outcomes to evaluate a time-lagged prediction model without explicitly leveraging multiple outcomes in the predictive model when composing wasn’t feasible (e.g., due to simplicity concerns).

Type	Practice	Description	Example
Theory-Driven	Identifying spurious causes of outcomes.	Identifying factors influencing the outcome unrelated to the construct of interest.	<i>"There's so many other things that could contribute to somebody being in the emergency department for 12 hours that might not have anything to do with how sick they are." (P11)</i>
	Reasoning across multiple outcomes.	Leveraging multiple outcomes to gain a more holistic understanding of a construct.	<i>"If you have high response times, and they're using [learning] supports, that's a good indication that they're trying, even if they don't get the next problem correct." (P2)</i>
	Linking interventions with outcomes.	Reasoning about potential causal relationships between an intervention and outcomes.	<i>"To get in the neighborhood of the right outcome measure, it is often like, well, what does this intervention really look like? [...] Where would we reasonably expect to move the needle?" (P6)</i>
Data-Driven	Validating predictions in real-world deployments.	Checking the extent to which model predictions align with expectations in real-world conditions.	<i>"We did an evaluation study where the clinicians would actually review cases while the information that our models predicted would be relevant was highlighted for them." (P15)</i>
	Poking holes in label definitions.	Attempting to identify instances that are misclassified under a current label definition.	<i>"[It was] a lot of literally, in some cases, printing off student essays and handing them to employees that were former teachers and saying, why did this person give this label?" (P1)</i>
	Applying quick checks and heuristics.	Using pragmatic checks to quickly gauge whether the task setup matches the data sample.	<i>"There's several models that existed in the literature that I attempted to transfer to my work. I checked it [...] using finger in the wind sort of [approaches]. Like, what's the distribution of my data? Is it closer to 7% or is it closer to 20%?" (P7)</i>

Table 3.3: Data scientists engaged in both theory-driven and data-driven practices while evaluating validity.

that piggybacking off of existing disease definitions was infeasible. Participants described these heuristics as imperfect; a necessary but insufficient condition for demonstrating validity. These checks and heuristics are an example of evaluating *face validity*, or the degree to which an operationalization measures what it purports to measure (Coston et al., 2023a).

Participants tried to *poke holes in label definitions* by identifying cases a label definition miscategorized the instance in comparison to theoretical expectations. While *refining*, P13 tried to poke holes by identifying specific cases where the expert judgment of the clinician diverged from the categorization assigned by the label definition. In education, P1 tried to poke holes in essay scoring rubrics by printing off physical copies and asking employees who were former teachers to explain “*why did this person give this label?*” Employees’ response served to build P1’s intuitive sense of *how* the rubric was used to assign instance-level scores, and identify potential gaps. By “*staying close to the data*”, P1 had the trust needed in the labeling process to justify specific essay scores to students, teachers, other external stakeholders. P1 described this trust as critical should a measurement decision later require formal legal justification — e.g., in the context of university admissions. For both P1 and P13, poking holes required working in tandem with domain experts, who had the knowledge required to verify whether the instance *really was* positive or negative for the construct of interest. The practice of poking holes and patching them connects to an assessment of *content validity*.

Finally, participants checked whether observed model behavior reflected their expectations in real-world deployment contexts. P15 checked the validity of a model trained to predict the “clinical relevance” of information in electronic medical records by conducting a field study that tested whether predictions saved clinicians time. The time savings shown by predictions provided evidence that the predicted fields really were “clinically relevant.” P15 and many other participants found both theory-driven and data-driven evaluation practices critical for establishing a holistic understanding of their formulation’s validity.

3.6 Discussion

Today, significant focus across the HCI, CSCW, ML, and FAccT communities has turned towards evaluating the extent to which algorithmic systems achieve their purported aims. Given this attention, researchers and practitioners increasingly draw upon validity concepts established in the quantitative social sciences to describe desirable properties of algorithmic systems. For instance, a growing body of work examines how validity sub-criteria — such as content, convergent, and discriminant validity — might be applied to evaluate algorithmic systems (Coston et al., 2023a; Kawakami et al., 2024; Guerdan, Coston, Wu and Holstein, 2023; Jacobs and Wallach, 2021; Raji et al., 2022; Liu, Blodgett, Cheung, Liao, Olteanu and Xiao, 2024; Raji et al., 2021; Alaa et al., 2025). Yet our findings uncover a disconnect between the traditional, top-down measurement process assumed by validity theory and the ways data scientists navigate measurement decisions in practice. In particular, data scientists typically must make do with limited resources by re-appropriating existing data for new measurement tasks. These constraints force data scientists to balance validity with other important criteria, such as predictive performance, simplicity, and resource requirements.

Based on these findings, we argue that **the challenge is not to *replace* these bricolage practices with top-down planning, but rather to develop forms of scaffolding that can enhance the rigor of bricolage while preserving its creativity and adaptability.** Our findings provide a foundation for the research community to develop this scaffolding, by identifying concrete opportunities to help data scientists balance validity with other criteria, identify and apply problem re-formulation strategies, and assess validity using both top-down and bottom-up evaluation practices. We now reflect on data science as a bricolage practice,

before providing targeted recommendations for supporting data scientists' problem formulation practices going forward.

3.6.1 Data Science as Bricolage

References to data science often invoke engineering metaphors: data scientists “engineer” features, “build” models, and “architect” computational solutions. Yet our findings suggest that the practice of target variable construction more closely resembles bricolage than engineering, a mode of knowledge production whereby the bricoleur makes do with the resources at hand as opposed to acquiring materials for a predefined plan. Our findings highlight how this process of making do unfolds in real-world projects, as data scientists re-appropriate data fields originally collected for entirely different purposes to accomplish their measurement goals. This practice of bricolage describes *how* data scientists orient themselves towards measurement problems, as well as nature and quality of their solutions.

The Data Science Repertoire

The bricoleur operates with an inventory of odds and ends called a “repertoire.” Data scientists’ repertoire includes both material resources (e.g., datasets, computational tools) and immaterial resources (e.g., domain theories, methods) brought to bear on a modeling task. The repertoire shapes the space of possibilities that bricoleurs imagine while confronting new problems. As [Louridas \(1999\)](#) observes while drawing connections between bricolage and design, “the bricoleur asks his collection, whereas the engineer, like the scientist, asks the universe.” P12 eloquently translated this to data science, lamenting that “*ultimately, you can’t create data that don’t exist.*” The constrained nature of data scientists’ repertoire forces them to re-examine their limited materials from new perspectives.

In particular, materials in data scientists’ repertoire function as what Lévi-Strauss calls “signs”, or concrete objects that represent abstract concepts, but remain constrained by their imagined uses. P6’s creative re-purposing of raw user interactions in an online learning platform as a proxy for “reading ability” exemplifies the use of a sign, by unexpectedly transforming a mundane data field into an operationalization of a construct. Yet, as [Levi-Strauss \(1966\)](#) notes, “the possibilities remain always limited by the particular history of each piece, and by what is predetermined in it due to the original usage for which it was conceived.” Had P6 *not* been aware of this proxy for “reading ability”, established through a chance encounter with a colleague’s work, they may have never made this connection.

Engaging in Dialogue with Data

After taking stock of the repertoire, the bricoleur engages in dialogue with their materials to identify a workable solution to the problem at hand. As Lévi-Strauss explains, the bricoleur must “engage into a kind of dialog with [the repertoire], to index, before choosing among them, the possible answers that the set can offer to his problem.” ([Levi-Strauss, 1966](#)) This dialogue not only shapes the bricoleur’s understanding of each element in isolation, but also their interactions: “one element’s possibilities interact with all other elements’ possibilities, with the overall organization of the artifact he makes. The results of these interactions are never what he expects, and he must respond to them.” ([Louridas, 1999](#)) P10’s experience illustrates how dialogue unfolds in data science. After initially selecting a cardiovascular outcome based on precedent in academic literature (piggybacking), “*seeing the data, getting to work with data*” led them to *swap* to a renal outcome that was more predictable given available biomarkers. This example illustrates how problem (re)formulation strategies serve as the brushstroke of

measurement in data science work. Like a painter with a canvas, data scientists apply a strategy, step back to survey their progress along multiple criteria, then consider a new strategy in response to flaws.

Dialogue marks a notable shift from the top-down planning approach to measurement. Under top-down planning, a practice better described as a *monologue* (Turkle, 2011), the data scientist defines a theoretical construct of interest, designs an operationalization, *then* collects data. Yet this monologue supposes that the data scientist can navigate the space of constraints and contingencies in their formulation through mental simulation, without a specific data sample or model. Our findings underscore that this is rarely the case. Instead, data scientists required *tangible interactions* with data to understand its affordances and constraints. It was through this very process of tangible interaction with data, borne out of resource constraints, that data scientists made creative leaps in their projects. As P6 explained, had a “magical oracle” granted their team access to all the data they required, they would have simply collected a comprehensive dataset with student test outcomes. It was the combination of resource constraints and tangible interaction with data that drove P6’s team to adaptively *bridge* using the limited data at hand.

The Emerging Artifact

Critically, the artifact resulting from bricolage is rarely ideal, but rather bears the marks of the constraints, compromises, and creator that shaped it. As Louridas (1999) explains, “bricolage is ... at the mercy of contingencies, either external, in the form of influences, constraints, and adversities of the external world, or internal, in the form of the creator’s idiosyncrasy.” Analogously, data scientists readily acknowledged imperfection in their solutions. P13 explained that a label definition that miscategorized 10% of pet diseases was acceptable if the “major dimensions” of the disease’s theoretical conceptualization were covered. P6 deemed their re-purposing of existing data a “really big success”, but also identified important limitations. Thus, from the bricoleur’s perspective, the key question is not whether a solution is *ideal*, but rather whether it is *fit for purpose*. P6 embodied this ethos when they justified *bridging* by explaining that their platform analytics use case was “not high stakes.” Had their model been designed for use cases they perceived as higher-stakes, P6 explained that they may have applied a higher level of stringency.

Beyond constraints, Lévi-Strauss notes a personal dimension to the artifact that emerges from bricolage: “he [the bricoleur] ‘speaks’ [...] through the choices he makes among the limited possibilities, the character and the life of the creator.” P8’s creative use of *composing*, which conceptualizes “academic work ethic” as the difference between *where a student is at* and *where they should be*, may have never emerged had they not also observed real students in real classrooms as an educator. Similarly, P12’s project, which had strong buy-in from multiple stakeholders and momentum towards completion, never came to fruition due to a validity flaw spotted by P12. Yet it was through P12’s personal interest in adjacent bodies of biostatistics literature, and their attention to methodological rigor in problem formulation, that they became aware of this otherwise invisible flaw. Each of these examples illustrates that the journey through forking paths in problem formulation depends greatly on the knowledge and experience of the bricoleur.

In sum, bricolage acknowledges that data scientists must work within the confines of their existing resources and organizational processes. These practical considerations are neither centered nor prioritized in the conventional top-down approach to measurement. Thus, short of dramatically expanding the available resources for a project, a rigid enforcement of top-down planning is unlikely to translate to a desirable project outcome. **Therefore, rather than**

recommending more stringent enforcement of the traditional top-down measurement approach, we suggest more effectively scaffolding *existing* bricolage practices. We now revisit traditional top-down measurement interventions with this insight in mind, before envisioning new ways of scaffolding data science practice.

3.6.2 Design Implications: Reconciling Top-Down Measurement Interventions with Bricolage Practice

Our characterization of data science as a bricolage practice uncovers a tension between existing interventions designed to help analysts implement the traditional top-down measurement workflow, and data scientists' bottom-up bricolage practices. However, this tension is not unique to data science. History reveals similar patterns across domains where top-down planning approaches have failed while bottom-up alternatives succeeded. In urban planning, modernist urban renewal projects that imposed regimented public works projects onto complex social systems often destroyed functioning neighborhoods, while community-based planning approaches showed strong success (Fuller and Moore, 2017; Scott, 2020). In software development, rigid and systematic waterfall planning methodologies gave way to agile approaches that embrace iteration and adaptation (Osorio et al., 2011). Key to the success of these bottom-up alternatives was not wholesale rejection of structure, but rather targeted interventions like mixed-use zoning in urban planning and sprint retrospectives in agile development. These interventions recognized that while existing practices were in need of reform, effective solutions must work within resource constraints and pinpoint specific procedural weaknesses in need of constructive redirection. Similarly, when improving measurement practices in data science, we argue that interventions tailored to support data scientists' existing practices are likely to be more effective than direct translation of rigid, top-down measurement frameworks.

In the following sections, we examine how interventions designed to improve measurement practices in the quantitative social sciences might be adapted to better support data scientists' bricolage practices. We begin by re-visiting the classic pre-registration protocol (Van't Veer and Giner-Sorolla, 2016). We then re-examine the adoption of a commonly-used statistical tool used to improve validity, latent variable models (Bishop, 1998), before identifying implications for teaching data scientists about measurement concepts. Rather than rejecting these traditional measurement interventions, we identify opportunities to tailor each to the bricolage approach towards measurement uncovered by our findings.

Pre-Registration for Predictive Modeling.

Pre-registration is one top-down intervention that is widely-recommended in the quantitative social sciences (Van't Veer and Giner-Sorolla, 2016). This protocol restricts data-dependent decision-making by requiring an analyst to specify outcome measures and hypotheses before running analyses. Recent work argues for the adoption of pre-registration in predictive modeling contexts (Hofman et al., 2023) — i.e., by requiring data scientists to finalize their problem formulation *before* training a model. Yet requiring such pre-registration is akin to *blindfolding the bricoleur* — it restricts the very bottom-up, iterative process that data scientists require to identify a workable formulation under resource constraints. Indeed, data scientists interviewed by Hofman et al. (2023) worried that pre-registration might limit their creativity, or make it challenging to later tweak their formulation if they started off with limited familiarity with their data.

Going forward, a more targeted approach for porting pre-registration to predictive modeling should distinguish between necessary versus pathological forms of data-dependent decision-making. For example, our findings illustrate that data scientists find iterative (re)formulation to

be a critical component of engaging in dialog with their data. Thus, the *first* pre-registration step proposed by Hofman et al. (2023) — which restricts (re)formulation — may be impractical in light of data scientists’ bricolage practices. However, the *second* pre-registration step proposed by Hofman et al. (2023), which aims to prevent data scientists from selecting models using test data (i.e., test set leakage) by keeping a held-out dataset reserved for the final evaluation, is directly compatible with data scientists’ existing practices.⁵ This provides evidence for the more targeted adoption of the *second* step of Hofman et al. (2023)’s protocol. In Section 3.6.3, we further deconstruct the iterative picture of (re)formulation uncovered by our findings to identify yet more targeted opportunities to extend pre-registration to predictive modeling contexts.

Latent Variable Models

Researchers in the quantitative social sciences have also developed *statistical tools* to improve the validity of models. For example, latent variable models (e.g., (Bishop, 1998; Bartholomew et al., 2011)) use statistical machinery to characterize the relationship between unobservable theoretical constructs and observed proxies. Recent work has proposed latent variable models (e.g., (Milli et al., 2021)) and related modeling advances (e.g., (Watson-Daniels et al., 2023; Guerdan, Coston, Holstein and Wu, 2023)) to improve the validity of models used for decision support and recommendation systems. Yet these sophisticated modeling approaches present a significant tension with bricolage practices. Latent variable models typically require advanced statistical training, multiple outcome variables, and significant time to implement effectively. This creates a practical barrier in resource-constrained environments, where data scientists must often prioritize simplicity and rapid development. For example, P1 shared that, while they were aware of advanced multi-outcome modeling approaches, the “extra bells and whistles” provided by these approaches made them less reliable in large-scale deployment contexts.

This disconnect between theoretical ideal and practical reality presents an opportunity to develop modeling interventions that better align with bricolage practices. Rather than introducing increasingly complex statistical machinery, future work might focus on developing simplified versions of latent variable models that preserve their measurement benefits while reducing implementation complexity. These simplified approaches could emphasize interpretability and robustness in deployment settings, helping data scientists effectively communicate their models to stakeholders without advanced statistical training. The challenge lies in distilling the essential validity benefits of latent variable models into more accessible tools that seamlessly integrate with existing workflows. In Section 3.6.3, we elaborate on specific modeling approaches that are likely to be most effective for supporting bricolage practices, including simplified approaches to composing multiple outcomes that maintain validity while meeting simplicity constraints.

Pedagogical Interventions

The findings from our study highlight several pedagogical interventions to better equip data science students and practitioners in navigating the complexity of target variable construction. Recognizing the bricolage nature of measurement tasks in data science, these interventions emphasize practical, iterative, and adaptive approaches to teaching measurement theory and practice. A key insight from our study is that analysts often draw on prior experience, engaging in analogical reasoning by finding parallels between current analytical problems and those previously encountered, or those that align with familiar tool affordances (Jun, Birchfield, De Moura, Heer and Just, 2022). This highlights the value of **case-based learning** in data

⁵It is worth noting that withholding a test set is also viewed as best practice in the machine learning literature, irrespective of pre-registration.

science education. Given the resource-constrained and improvisational nature of real-world data science practice, case-based approaches enable students to analyze concrete case studies (e.g., case studies from healthcare, education, biomedical sciences, etc.) and understand how practitioners make trade-offs between competing criteria like validity, simplicity, and data availability. Guided engagement with these cases can help students appreciate the practical complexities and compromises inherent in predictive modeling tasks.

In addition to case-based learning, **reflective practice** plays a critical role in developing students' capacity to reason about measurement and validity. Iterative experimentation and reflection help students develop a nuanced understanding of the consequences of methodological choices, enabling them to better justify and communicate those choices. This mirrors how tasks like feature engineering are often described—as a “craft” or “art”—involving domain expertise, theoretical grounding, and direct engagement with data (Duboue, 2020). Similarly, recent work in HCI highlights how design thinking coupled with case-based learning can help AI practitioners deliberate over problem formulation and the feasibility of AI concepts by identifying critical risk factors during the earliest stages of AI innovation (Saxena et al., 2025; Kuo et al., 2023).

Another promising pedagogical intervention involves **hands-on workshops** for teaching validity concepts. In fields like medical education, workshops have successfully introduced novice practitioners to the idea of validity arguments and the different forms of evidence used to support measurement claims (Gasmalla et al., 2021; Paul et al., 2019). Drawing on this model, practical workshops in data science can similarly support students' understanding of target variable construction. Workshops focused on reformulation strategies—such as piggybacking, composing, swapping, bridging, and refining—can expose students to the tactical, trial-and-error process of shaping a prediction task within real-world constraints. These activities allow students to grapple directly with validity trade-offs while developing intuition for when and how to adapt their measurement choices. Together, these pedagogical strategies may offer an integrated and grounded approach to teaching the complexities of problem formulation and measurement validity in data science.

3.6.3 Design Implications: Developing a Scaffolding for Effective Bricolage Practice

In addition to re-visiting existing top-down interventions, our findings surface a need for entirely *new* interventions designed to help data scientists engage in bricolage more effectively. While many toolkits exist to help data scientists engage in inner-loop activities, such as training models (Pang et al., 2020; Dingen et al., 2018; Cabrera, Fu, Bertucci, Holstein, Talwalkar, Hong and Perer, 2023; Yan et al., 2021) and exploring data (Wongsuphasawat et al., 2015; Kery et al., 2018), few support specific bricolage practices highlighted in our findings. Data scientists currently lack tools to help them weigh validity with competing criteria in their formulation, identify and apply strategies, and evaluate validity sub-criteria. **We now use our findings as a concrete scaffolding for identifying actionable points of intervention in data scientists' bricolage practices.** These interventions are intended to *facilitate* specific practices that data scientists found critical during problem formulation, while *adding guardrails around others* that have the potential to introduce flaws.

Helping Data Scientists Weigh Tradeoffs

In the quantitative social sciences, validity is traditionally viewed as the primary criterion for the scientific integrity of a study. Techniques such as power analysis help researchers adaptively calibrate the resource requirements (e.g., sample size) necessary to reach statistically valid

study conclusions (Cohen, 1992). Likewise, simplicity limitations of models are often tolerated in favor of more robust statistical analyses. In each of these examples, validity standards are *fixed*, while standards for other criteria, such as resource requirements, are malleable. In contrast, while engaging in bricolage, data scientists were forced to adaptively balance the validity of their formulation with other important criteria within their resource constraints. The lack of support for this delicate balancing act makes data scientists susceptible to introducing critical defects in their formulation (Raji et al., 2021). Therefore, we now identify several targeted opportunities to help data scientists more deliberately weigh tradeoffs going forward.

First, our findings illustrate a need to **help data scientists operationalize measures for their formulation’s simplicity, validity, portability, and resource requirements**. Teams currently rely heavily on measurements of predictive performance (e.g., AU-ROC or F1 scores) while applying strategies. Echoing the classic adage “what gets measured matters” (Skogan, 1999), this practice leads teams to prioritize predictive performance improvements at the expense of other criteria. Going forward, HCI and CSCW researchers might work with data science teams to co-design rubrics that help them evaluate validity dimensions (e.g., construct, content, and external validity), simplicity (e.g., model complexity, implementation requirements), and portability (e.g., data field availability across contexts). These rubrics could be paired with existing toolkits developed by the Responsible AI community to help data scientists weigh the suitability of predictive modeling tools for real-world deployment contexts of interest (e.g., (Kawakami et al., 2024; Coston et al., 2023a)). Such toolkits should help data scientists understand that criterion such validity are multifaceted, and thus any quantitative measurement will be incomplete. Nevertheless, supporting structured assessments across multiple criteria will help teams better balance predictive performance with other essential aspects of their formulation.

Second, our findings point to a need to help teams **establish minimum performance standards for each criterion before beginning problem formulation**. Establishing clearly defined minimum standards is critical for mitigating *standard erosion*, or a selective reduction in stringency applied to a criterion over subsequent rounds of iterative (re)formulation. Our interviews surfaced evidence for validity standard erosion in several projects, as participants introduced validity flaws to gain improvements in other criteria. HCI and CSCW researchers can help combat standard erosion by working with practitioners to co-design tools that help teams identify appropriate performance ranges based on specific application contexts. Researchers might explore the potential for visualizations like radar plots (e.g., Figure 3.3) to help teams chart trade-off spaces and identify suitable operating regions for specific use cases of interest. For instance, a research team might place strong requirements on validity but weaker requirements on simplicity. A team working at an educational technology company might prioritize the simplicity of their formulation. By making the trade-offs inherent to bricolage explicit, teams can (1) avoid standard erosion, and (2) provide a strong rationale for discontinuing a project if minimum standards for each criterion cannot be satisfied given a project’s resource constraints.

Finally, there is an opportunity for HCI, ML, and CSCW researchers to develop tools that **help data scientists identify targeted data collection opportunities**. While engaging in bricolage, data scientists were forced to weigh the predictive performance and validity benefits of data collection against its resource requirements. Yet existing model training and evaluation toolkits often assume that datasets are *fixed* (e.g., (Cabrerera, Fu, Bertucci, Holstein, Talwalkar, Hong and Perer, 2023; Yan et al., 2021; Tramer et al., 2017; Bellamy et al., 2019; Weerts et al., 2023; Saleiro et al., 2018; Liu, Chen, Shen and Choo, 2024)). Future work might close this gap by developing tools that help data scientists quantify the predictive performance and validity improvements offered by collecting data. The framework of data valuation, originally developed

by the machine learning and economics communities, offers tools for ascribing monetary value to data resources (Castro Fernandez, 2023; Perdomo, 2023; Gupta et al., 2024). Future CSCW and ML research might develop data valuation toolkits that enable teams to estimate the predictive performance and validity benefits of new data sources prior to acquisition. Such a framework might leverage synthetic data generation (Figueira and Vaz, 2022) to help teams quickly augment their datasets with additional “mock” information fields in order to quickly assess its potential benefits prior to collecting real data with similar characteristics.⁶ These efforts would be greatly enhanced by empirical studies investigating *how* teams acquire and ascribe worth to data throughout the (re)formulation process.

Helping Data Scientists Identify Strategies

Our findings present an opportunity to help teams **more effectively identify strategies** based on specific challenges encountered during problem (re)formulation. Data scientists currently identify strategies in an ad hoc, unstructured manner. Participants stumbled upon strategies based on their bespoke knowledge of academic literature, or the expertise of others on their team. This lack of structured strategy identification led teams to miss opportunities to improve their formulation. For instance, following the conclusion of their project, P12 identified a modeling framework—i.e., case-control methods (Schlesselman, 1982)—that was tailor-made to address their project’s data constraints. Yet P12 was not aware of this framework because their background in machine learning did not involve training in biostatistics methods. This presents an opportunity to broaden data scientists’ awareness of relevant strategies before helping them narrow in on which are best suited to their bespoke modeling challenges.

Therefore, future HCI and CSCW research might develop systems that help teams **collaboratively share, discuss, and refine (re)formulation strategies**. While community-based knowledge exchange is central to inner-loop workflows, as illustrated by the popularity of Stack Overflow (Barua et al., 2014) for coding, and Hugging Face (Shen et al., 2024) for model training, data scientists’ support for problem formulation remains siloed. HCI and CSCW researchers might address this challenge by developing collaborative systems that encourage teams to (1) share their data availability and problem formulation challenges, and (2) describe strategies they used to navigate challenges successfully. Similar to other online communities (e.g., Wikipedia (Viégas et al., 2007)), this system might structure collaborative discussions via a comment forum: P8 might share their team’s combining strategy for measuring “academic work ethic”, while P6 might share their team’s bridging strategy for measuring “reading ability.” Over time, this collaborative exchange could be consolidated into a consensus-based guideline that provides step-by-step guidance for navigating common (re)formulation challenges.

Helping Data Scientists Apply Strategies

Our findings also highlight opportunities to help data scientists apply the strategies they’ve identified more effectively. Each strategy represents a distinct method of re-appropriating materials in the data science repertoire for the measurement task at hand. While strategies can be highly adaptive when used skillfully, they can also introduce pitfalls when applied improperly. Below, we identify targeted opportunities to improve how data scientists incorporate each strategy into their (re)formulation process. Taken together, these interventions stand to enhance the overall rigor of data scientists’ bricolage practices while also supporting flexibility.

Piggybacking. While *piggybacking*, data scientists port an existing problem formulation to their own modeling context. Piggybacking saves time when there is a strong alignment

⁶Given that synthetic data is by definition a proxy for real-world information, we caution against directly leveraging synthetic data for model training and evaluation unless its properties are carefully understood.

between established precedent and data scientists' modeling context. Yet, as illustrated by P1's experiences, piggybacking can also introduce serious defects when contextual differences are overlooked. Going forward, there is a need for structured protocols that help teams thoughtfully reason about the transferability of problem formulations between modeling contexts. While many guidelines support dataset and model documentation (e.g., (Mitchell et al., 2019; Gebru et al., 2021; Coston et al., 2023a)), none support assessments of problem formulation transferability. HCI and CSCW researchers might fill this gap by working with data scientists to co-design a rubric that teams can complete while considering piggybacking. This rubric might guide data scientists through a side-by-side comparison of an existing formulation and their own along dimensions such as high-level modeling goals, available data, and rationale for a selected target variable. Helping data scientists spot transferability gaps can also help teams provide a rationale to external stakeholders should a team decide to use a formulation that deviates from an established precedent.

Composing. While *composing*, data scientists combine multiple outcomes into a single target variable. Composing can improve validity by helping data scientists capture all relevant dimensions of a construct in a target variable. Yet composing can also introduce simplicity limitations. This tension indicates a need for simple and interpretable mechanisms for combining outcomes. Future HCI and ML research might explore opportunities for interpretable multi-outcome models that cleanly map conceptual relationships (e.g., "*the difference between where the person should be and where they're at*", P8) into transparent model designs. One approach likely to be effective involves training multiple models to target multiple outcomes of interest (e.g., cost of medical care, re-admission), then aggregating predictions into a single score at run-time. P11 described such an approach as common in the medical industry, where both simplicity and validity are paramount. Several participants also suggested this approach during the design probe. HCI, ML, and Visualization researchers can implement this approach by developing tools that help teams encode their domain knowledge into aggregation functions that weight multiple predictions into a single score.

Swapping. While *swapping*, data scientists switch to a different outcome after they identify defects in one or more criteria. Swapping can be adaptive if it helps teams identify a formulation that satisfies all minimum performance standards. Yet swapping can also introduce standard erosion. This tension highlights a need for tools to help data scientists make more systematic swapping decisions. While many existing systems support feature engineering (e.g., (Bhattacharya et al., 2024; Guo et al., 2011; Kandel et al., 2011; Rattenbury et al., 2017)) and EDA (e.g., (Wongsuphasawat et al., 2015; Kery et al., 2018)), only one supports *outcome exploration* (Gala et al., 2024). HCI and Visualization researchers can support outcome exploration by developing toolkits that help data scientists weigh the portability, simplicity, predictive performance, and resource requirements of alternative outcomes. Because our results suggest that swapping can be enticing given its predictive performance benefits, it is especially important that toolkits help teams identify validity drawbacks of swapping decisions. We discuss opportunities to support validity evaluation in Section 3.6.3.

Bridging. While *bridging*, data scientists adopt a low-cost, readily available target variable as a proxy for a more costly gold standard measure. When used effectively, bridging is a strong example of bricolage because it involves a creative approach for making do with limited data. Yet the success of bridging hinges on data scientists' ability to identify trustworthy sources of auxiliary information. To support bridging, data scientists need tools to help them identify, evaluate, and incorporate relevant sources of auxiliary data. While a large body of HCI and Data Engineering research has developed tools supporting data discovery (e.g., Fernandez et al. (2018); Nargesian et al. (2019); Paton et al. (2023); Bogatu et al. (2020)), existing toolkits are not tailored to support bridging. Future research might close this gap by developing tools

that help data scientists forage for relevant auxiliary data and evaluate its expected utility against the time and cost required to incorporate it in their formulation. As with swapping, it is important that such tools help data scientists identify any defects associated with bridging.

Refining. While *refining*, data scientists make granular adjustments to how a target variable is coded in their data. Participants often engaged in refining by modifying a label definition used to operationalize a construct using a set of indicators used to categorize instances. For instance, P13 engaged in refining by working with domain experts to spot gaps in label definitions, then patch them with a revised set of indicators. Though refining can improve validity, practitioners lack tools to encode domain expert's knowledge into label definitions. Instead, existing tools developed by the HCI and ML communities help annotators apply a *fixed* label definition to annotate data (e.g., (Nakayama et al., 2018; Muller et al., 2021; Desmond et al., 2021; Daudert, 2020)) or use crowdsourcing to refine annotation guidelines (e.g., (Chen and Zhang, 2023; Bragg et al., 2018; Chang et al., 2017; K. Chaithanya Manam et al., 2019; Manam and Quinn, 2018)).⁷ Future HCI and ML research can help data scientists encode domain experts' knowledge in label definitions by developing tools that help teams collaboratively systematize concepts Adcock and Collier (2001) by iteratively identifying and refining indicators used in label definitions. Because our results indicate that domain experts require rich contextual information to identify areas where the systemization of a concept is incomplete, tools supporting this process should capture rich contextual information (e.g., written notes) in data.

Helping Data Scientists Evaluate Validity

Disciplines in the quantitative social sciences have established numerous guidelines for evaluating the validity of statistical models. For instance, best practices recommend that practitioners check for the goodness-of-fit of latent variable models (Neath and Cavanaugh, 2012), and the correlation between measurements obtained from new instruments and established measures of the same construct (Carlson and Herdman, 2012). Yet there is currently a disconnect between theoretical validity sub-criteria and the existing practices of data scientists engaged in problem formulation. Even participants highly trained in academic measurement practices described abandoning established guidelines that were viewed as overly rigid given their time and resource constraints. Our findings provide a path forward by grounding validity evaluation in the specific theory-driven and data-driven practices identified in Table 3.3.

First, our findings highlight opportunities to **help data scientists and domain experts collaboratively engage in theory-driven validity evaluation activities** using domain knowledge. For example, HCI, ML, and Visualization researchers might develop toolkits that help teams map hypothesized causal relationships between predictive features, interventions, outcome variables, and latent constructs during problem formulation. In an education context, this tool might visualize students' longitudinal trajectories through the education system, enable data scientists and domain experts to weigh relevant points of intervention, and help teams think through an intervention's likely effect on multiple downstream outcomes of interest. This tool might also encourage teams to draw upon their domain knowledge to identify spurious causes of outcome variables — e.g., by listing potential causes of an outcome that are unrelated to the unobserved construct of interest. During the design probe, several participants independently suggested an interest in a tool supporting theory-driven evaluation activities. While the *DoWhy* package (Sharma and Kiciman, 2020) enables data scientists to reason about causal relationships in their data, this package is not tailored to support the formulation of prediction tasks.

⁷Traditional annotation guidelines are *de facto* label definitions if they serve as a measurement instrument that operationalizes a latent construct of interest.

Further, our findings demonstrate a need for tools to **help data scientists connect their high-level measurement goals to low-level data** while engaging in data-driven evaluation practices. One paradigm that might tighten this connection is *behavioral evaluation*, which involves testing the capabilities of a system against a specification of requirements (Rahwan et al., 2019; Cabrera, Tulio Ribeiro, Lee, Deline, Perer and Drucker, 2023). Prior HCI research has explored opportunities for behavioral evaluation to help data scientists assess the predictive performance and fairness properties of their models (Cabrera, Fu, Bertucci, Holstein, Talwalkar, Hong and Perer, 2023). In the future, HCI and ML research might develop behavioral evaluation approaches that facilitate improved *measurement practices*. For instance, toolkits supporting behavioral evaluation might enable data scientists to formalize their existing bottom-up validity evaluation practices into a more systematic specification of expected behavior. Behavior specifications might help data science teams explicitly articulate tacit measurement decisions and (re)formulate their task when specifications are violated. However, it is important that behavior specifications are used as part of a multifaceted validity evaluation strategy (Jacobs and Wallach, 2021; Coston et al., 2023a). This is because many components of validity are difficult to capture in a quantitative performance measure, and data scientists might inadvertently omit important validity sub-dimensions in a behavior specification. Taken together, toolkits supporting both theory- and- data-driven validity evaluation can help teams close the gap between their theoretical conceptualization of a construct and its operationalization in predictive modeling formulations.

3.6.4 Study Limitations

While our study offers valuable insights into how fifteen data scientists working in the education and healthcare sectors construct target variables through bricolage, several limitations should be considered while interpreting our findings:

Collaborative dynamics in target variable construction. First, by focusing our interviews on data scientists' perspectives, we captured only one dimension of what is inherently a collaborative process. Our findings illustrate that target variable construction in practice often entails multifaceted negotiations among stakeholders with different expertise, institutional roles, and incentive structures. Data scientists frequently described collaborating with domain experts, such as clinicians or educators, to refine label definitions or evaluate the validity of their formulations. Recent work has also shown that community members can provide critical feedback on problem formulation of ML tasks (Kuo et al., 2023; Haque et al., 2024). However, these collaborators might perceive the bricolage process quite differently from the data scientists we interviewed. The collaborative dynamics, power relationships, and knowledge exchanges that shape target variable construction likely extend beyond what our participants could articulate from their vantage points as data scientists.

Contextual variation across domains and organizations. Second, our study identifies and characterizes a bricolage approach towards target variable construction in a narrow space of organizational and domain contexts. Our participants worked in education and healthcare settings with specific regulatory environments, data ecosystems, and stakeholder relationships that shaped their bricolage practices. Data scientists in domains with more readily available access to data (e.g., social media platforms), different regulatory environments (e.g., finance), or distinct organizational incentives (e.g., drawing user attention and clicks versus providing validated educational tests) might employ different strategies or weigh evaluation criteria differently. These contextual factors shape not only which (re)formulation strategies data scientists employ, but also how they evaluate tradeoffs between validity and other criteria.

Variation in technical and domain experience. Finally, our findings draw heavily upon

the experiences of data scientists with a baseline level of domain knowledge and technical experience. As minimum recruitment criteria, we required participants to have at least one year of full-time experience as a data scientist in the healthcare or education sectors. Our findings, particularly discussion of (re)formulation strategies, often feature the creative practices employed by highly experienced data scientists. Less experienced data scientists might employ a less sophisticated suite of (re)formulation strategies, and may also use less nuanced theory-driven evaluation practices to assess validity. Given the rapidly expanding access to basic technical tools involved with data science (e.g., model training and evaluation toolkits), there is also a need to better understand when and how *novice* data scientists engage in bricolage.

In sum, our findings should be understood as illustrating the broad dynamics involved with target variable construction rather than providing an exhaustive landscape of all potential factors. We encourage future work that addresses these limitations through complementary methodologies such as ethnographic observation, analysis of documentation artifacts, or comparative studies across diverse domains, organizational contexts, and participant samples.

3.7 Conclusion

In this study, we interviewed fifteen data scientists working in education and healthcare to understand their practices, challenges, and perceived opportunities for target variable construction in predictive modeling. We explore how data scientists adopt a *bricolage* approach to target variable construction, designing operationalizations for unobserved, latent constructs by *making do* with the data that they *already* have at their disposal. We characterize the contours and constraints of data scientists' bricolage practices, including the problem (re)formulation strategies they apply to balance across different desiderata. Critically, we argue that *interventions* designed to improve target variable construction should begin by acknowledging the inherent uncertainty and resource constraints involved in real-world data science work. By understanding and designing to support target variable construction *as a bricolage process*—for example, by helping teams navigate validity tradeoffs, determine when and how to collect more data, and identify appropriate (re)formulation strategies—the research community can help teams more thoughtfully engage with problem formulation and avert the negative consequences that arise from poor modeling decisions.

Chapter 4

Improving the Validity of Predictive Modeling Formulations via Expert Anchors

This chapter is drawn from the paper:

Luke Guerdan, Amanda Coston, Kenneth Holstein, and Zhiwei Steven Wu.
Counterfactual Prediction under Outcome Measurement Wrror. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

A Motivating Example. We illustrate the importance of considering interactions between OME and treatment effects by revisiting a widely known audit of an algorithm used to inform screening decisions for a high-risk medical care program (Obermeyer et al., 2019). This audit surfaced measurement error in a “cost of medical care” outcome targeted as a proxy for patient medical need. *Critically, the measurement error analysis performed by Obermeyer et al. (2019) assumes that program enrollment status is independent of downstream cost and medical outcomes.*

Sample	FPR	FNR
Full population	0.37	0.38
Unenrolled	0.37	0.39
Enrolled	0.64	0.13

Yet our re-analysis shows that the “cost of medical care” proxy has a substantially higher false positive rate and lower false negative rate among program enrollees as compared to the full population. This error rate discrepancy is consistent with enrollees receiving closer medical supervision (and as a result, greater costs), even after accounting for their underlying medical need. In this work, we show that failing to model the interactions between OME and treatment effects can introduce substantial model reliability challenges.

4.1 Motivation

Algorithmic risk assessment instruments (RAIs) often target labels that imperfectly reflect the goals of experts and policymakers. For example, clinical risk assessments used to inform physician treatment decisions target future utilization of medical resources (e.g., cost, medical diagnoses) as a proxy for patient medical need (Obermeyer et al., 2019; Mullainathan and Obermeyer, 2021, 2017). Predictive models used to inform personalized learning interventions target student test scores as a proxy for learning outcomes (Hur et al., 2022). Yet, these proxies are subject to *outcome measurement error* (OME) when they systematically differ from the target outcome of interest to domain experts. Unaddressed, OME can be highly consequential: models targeting poor proxies have been linked to misallocation of medical resources (Obermeyer et al., 2019), unwarranted teacher firings (Turque, 2012), and over-policing of minority communities (Butcher et al., 2022). Given its prevalence and implications, increasing research focus has shifted to understanding and mitigating sources of statistical bias impacting proxy outcomes (De-Arteaga et al., 2021; Fogliato et al., 2020; Fogliato, Xiang, Lipton, Nagin and Chouldechova, 2021; Wang et al., 2021; Natarajan et al., 2013; Menon et al., 2015).

However, prior work modeling outcome measurement error makes a critical assumption that the decision informed by the algorithm does not impact downstream outcomes. Yet this assumption is often unreasonable in decision support applications, where decisions constitute *interventions* that impact the policy-relevant target outcome *and its recorded proxy* (Coston, Mishler, Kennedy and Chouldechova, 2020). For example, in clinical decision support, medical treatments act as risk-mitigating interventions designed to avert adverse health outcomes. However, in the process of selecting a treatment option, a physician will *also* influence measured proxies (e.g., medical cost, disease diagnoses) (Obermeyer et al., 2019; Mullainathan and Obermeyer, 2021, 2017). As a result, the measurement error characteristics of proxies can vary across the treatment options informed by an algorithm.

4.2 Contribution

In this work, we develop a counterfactual prediction method that corrects for outcome measurement error, treatment effects, and selection bias in parallel. Our method builds upon *unbiased risk minimization* techniques developed in the label noise literature (Natarajan et al., 2013; Patrini et al., 2017; Chou et al., 2020; Van Rooyen et al., 2015). Given knowledge of measurement error parameters, unbiased risk minimization methods recover an estimator for target outcomes by minimizing a surrogate loss over proxy outcomes. However, existing methods are not designed for *interventional settings* whereby decisions impact outcomes – a limitation that we show severely limits model reliability. Therefore, we develop an unbiased risk minimization technique designed for learning counterfactual models from observational data. We compare our approach against models that correct for OME or treatment effects in isolation by conducting experiments on semi-synthetic data from healthcare and employment domains (Finkelstein et al., 2012; LaLonde, 1986; Smith and Todd, 2005). Results validate the efficacy of our risk minimization approach and underscore the need to carefully vet measurement-related assumptions in consultation with domain experts. Our empirical results also surface systematic model failures introduced by correcting for OME or treatment effects in isolation. To our knowledge, our holistic evaluation is the first to examine how outcome measurement error, treatment effects, and selection bias interact to impact model reliability under controlled conditions.

We provide the following contributions: 1) We derive a problem formulation that models interactions between OME, treatment effects, and selection bias (§ 5.3); 2) We develop a novel approach for learning counterfactual models in the presence of OME (§ 4.5.1). We provide a flexible approach for estimating measurement error rates when these are unknown in advance (§ 4.5.2); 3) We conduct synthetic and semi-synthetic experiments to validate our approach and highlight reliability issues introduced by modeling OME or treatment effects in isolation (§ 7.4).

4.3 Background and related work

4.3.1 AI functionality and validity concerns

Prior work has conducted detailed assessments of specific modeling issues Lakkaraju et al. (2017); De-Arteaga et al. (2021); Kleinberg et al. (2018); Coston, Mishler, Kennedy and Chouldechova (2020); Kallus and Zhou (2018b); Wang et al. (2021), which have been synthesized into broader critiques of AI validity and functionality (Coston et al., 2023b; Raji et al., 2022; Wang et al., 2024). Raji et al. (2022) surface AI functionality harms in which models fail to achieve their purported goal due to systematic design, engineering, deployment, and communication failures. Coston et al. (2023b) highlight challenges related to value alignment, reliability, and validity that may draw the justifiability of RAs into question in some contexts. We build upon this literature by studying *intersectional threats to model reliability* arising from outcome measurement error (Jacobs and Wallach, 2021; Wang et al., 2021), treatment effects (Coston, Mishler, Kennedy and Chouldechova, 2020; Perdomo et al., 2020), and selection bias (Kallus and Zhou, 2018b) in parallel.

4.3.2 Outcome measurement error

Modeling outcome measurement error is challenging because it introduces two sources of uncertainty: which error model is reasonable for a given proxy, and which specific error parameters govern the relationship between target and proxy outcomes under the *assumed* measurement model (Jacobs and Wallach, 2021). Popular error models studied in the machine learning literature include uniform (Angluin and Laird, 1988; Van Rooyen, Menon and Williamson,

2015), class-conditional (Menon et al., 2015; Scott et al., 2013), and instance-dependent (Xia et al., 2020; Chen et al., 2021) structures of outcome misclassification. Work in algorithmic fairness has also studied settings in which measurement error varies across levels of a protected attribute (Wang et al., 2021), and proposed sensitivity analysis frameworks that are model agnostic (Fogliato et al., 2020).

Numerous statistical approaches have been developed for measurement error parameter estimation in the quantitative social sciences literature (Roberts, 1985; Bishop, 1998). Application of these approaches is tightly coupled with domain knowledge of the phenomena under study, as in biostatistics (Hui and Walter, 1980) or psychometrics (Shrout and Lane, 2012). To date, data-driven techniques for error parameter estimation have primarily been applied in the machine learning literature, which rely on key assumptions relating the target outcome of interest and its proxy (Xia et al., 2019; Liu and Tao, 2015; Scott, 2015; Scott et al., 2013; Menon et al., 2015; Northcutt et al., 2021). In this work, we build upon an existing “*anchor assumptions*” framework that estimates error parameters by linking the proxy and target outcome probabilities at specific instances (Xia et al., 2019). In contrast to prior work, we provide a range of anchoring assumptions, which can be flexibly combined depending on which are reasonable in a specific algorithmic decision support (ADS) domain.

Natarajan et al. (2013) propose a widely-adopted *unbiased risk minimization* approach for learning under noisy labels given knowledge of measurement error parameters (Patrini et al., 2017; Chou et al., 2020; Van Rooyen et al., 2015). This method constructs a surrogate loss $\tilde{\ell}$ such that the $\tilde{\ell}$ -risk over proxy outcomes is equivalent to the ℓ -risk over target outcomes *in expectation*. Additionally, Natarajan et al. (2013) show that the minimizer of $\tilde{\ell}$ -risk over proxy outcomes is optimal with respect to target outcomes if ℓ is symmetric (e.g., Huber, logistic, and squared losses). In this work, we develop a novel variant of this unbiased risk minimization approach designed for settings with *treatment-conditional* OME.

4.3.3 Counterfactual prediction

Recent work has shown that counterfactual modeling is necessary when the decision informed by a predictive model serves as a risk-mitigating intervention (Coston, Mishler, Kennedy and Chouldechova, 2020). Building off of this result, we argue that it is necessary to account for treatment effects on *target outcomes of interest and their observed proxy* while modeling OME. Our methods build upon conditional average treatment effect (CATE) estimation techniques from the causal inference literature (Shalit et al., 2017; Johansson et al., 2020; Abrevaya et al., 2015). Subject to identification conditions (Pearl, 2009; Rubin, 2005), these approaches predict the difference between the expected outcome under treatment (e.g., high-risk program enrollment) versus control (e.g., no program enrollment) conditional on covariates. One family of *outcome regression estimators* predicts the CATE by directly estimating the expected outcome under treatment or control conditional on covariates (Hill, 2011; Künzel et al., 2019; Chernozhukov et al., 2017). However, these methods suffer from statistical bias when prior decisions were non-randomized (i.e., due to distribution shift induced by selection bias) (Shimodaira, 2000; Angluin and Laird, 1988). Therefore, we leverage a re-weighting strategy proposed by (Johansson et al., 2020) to correct for this selection bias during risk minimization. Our re-weighting method performs a similar bias correction as inverse probability weighting (IPW) methods (Shimodaira, 2000; Rosenbaum and Rubin, 1983).

Outcome measurement error has also been studied in causal inference literature. Finkelstein et al. (2021) bound the average treatment effect (ATE) under multiple plausible OME models. Shu and Yi (2019) propose a doubly robust method which accounts for measurement error during ATE estimation, while Díaz and van der Laan (2013) provide a sensitivity analysis

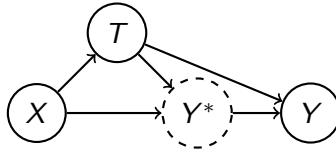


Figure 4.1: A causal diagram of treatment-conditional outcome measurement error.

framework for examining robustness of ATE estimates to OME. This work is primarily concerned with estimating *population statistics* rather than predicting outcomes conditional on measured covariates (i.e., the CATE).

4.4 Preliminaries

Let $p^*(X, T, Y_0^*, Y_1^*, Y_0, Y_1)$ be a fixed joint distribution over covariates $X \in \mathcal{X} \subseteq \mathbb{R}^d$, past decisions¹ $T \in \{0, 1\}$, *target* potential outcomes $\{Y_0^*, Y_1^*\} \in \mathcal{Y} \subseteq \{0, 1\}$, and *proxy* potential outcomes $\{Y_0, Y_1\} \in \mathcal{Y} \subseteq \{0, 1\}$. Under the potential outcomes framework (Rubin, 2005), $\{Y_0^*, Y_0\}$ and $\{Y_1^*, Y_1\}$ are the target and proxy outcomes that *would occur* under $T = 0$ and $T = 1$, respectively (Figure 4.2). Building upon the class-conditional model studied in observational settings Natarajan et al. (2013); Menon et al. (2015), we propose a treatment-conditional outcome measurement error model, whereby the class probability of the proxy potential outcome is given by

$$\eta_t(x) = (1 - \beta_t) \cdot \eta_t^*(x) + \alpha_t \cdot (1 - \eta_t^*(x)), \quad \forall x \in X \quad (4.1)$$

where $\alpha_t := p(Y_t = 1 \mid Y_t^* = 0)$, $\beta_t := p(Y_t = 0 \mid Y_t^* = 1)$ are the proxy false positive and false negative rates under treatment $t \in \{0, 1\}$ such that $\alpha_t + \beta_t < 1$. This model imposes the following assumption on the structure of measurement error.

Assumption 1 (Measurement error). Measurement error rates are fixed across covariates: $Y \perp\!\!\!\perp X \mid Y^*, T$.

While we make this assumption to foreground study of treatment effects, our methods are also compatible with approaches designed for error rates that vary across covariates (Wang et al., 2021) (see § 4.7.1 for discussion). Given the joint p^* , we would like to estimate $\eta_t^*(x) := p(Y_t^* = 1 \mid X = x)$, for any target covariates $x \in X$, which is the probability of the target potential outcome under intervention $t \in \{0, 1\}$. However, rather than observing Y_t^* directly, we sample from an *observational distribution* $p(X, T, Y)$, where $Y \in \mathcal{Y} \subseteq \{0, 1\}$ is an observed *proxy outcome*. By consistency, the unobserved target potential outcome and observed proxy potential outcome is determined by the treatment assignment.

Assumption 2 (Consistency). $Y^* = T \cdot Y_1^* + (1 - T) \cdot Y_0^*$; $Y = T \cdot Y_1 + (1 - T) \cdot Y_0$.

This assumption holds that the target and proxy potential outcomes Y_t^*, Y_t are observed among instances assigned to treatment t (Pearl, 2009; Rubin, 2005, 1974). To identify observational proxy outcomes Y , we also require the following additional causal assumptions.

Assumption 3 (Ignorability). $\{Y_0^*, Y_1^*, Y_0, Y_1\} \perp\!\!\!\perp T \mid X$. This holds that target and proxy potential outcomes are unconfounded given measured covariates X .

¹We also use the word *treatments* to refer to binary decisions. This draws upon historical applications of causal inference to medical settings.

Ignorability can be violated in decision support applications when unobservables impact both the treatment and outcome (Kleinberg et al., 2018; Lakkaraju et al., 2017; De-Arteaga et al., 2021). Understanding and addressing limitations introduced by ignorability is a major ongoing research focus (Díaz and van der Laan, 2013; Rambachan et al., 2022; Coston, Mishler, Kennedy and Chouldechova, 2020). We provide follow-up discussion of this assumption in § 4.7.2.

Assumption 4 (Positivity). $\forall x \in X, 0 < p(T = 1|X = x) < 1$. This holds that each instance $x \in X$ has some chance of receiving each decision $t \in \{0, 1\}$.

Positivity is often reasonable in decision support applications because instances $x \in X$ that require support from predictive models are subject to discretionary judgement due to uncertainty. Instances that are certain to receive a given treatment (i.e., $p(T = 1|X = x) = 0$ or $p(T = 1|X = x) = 1$) would normally be routed via a different administrative procedure. Figure 4.1 shows a causal diagram representing the data generating process we study in this work.

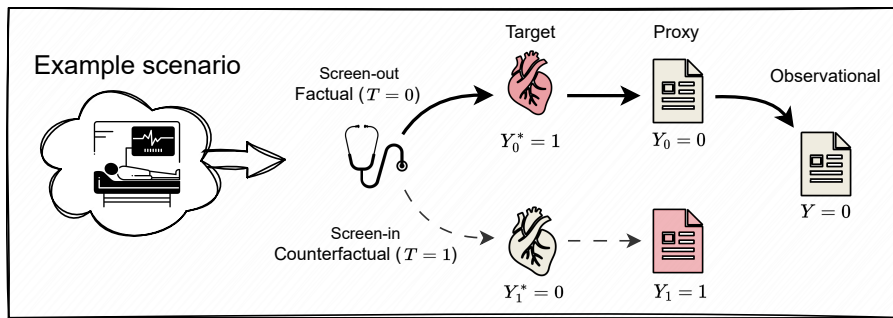


Figure 4.2: An illustration of treatment-conditional OME in heart attack prediction. Under the factual decision to screen-out from a high-risk care management program ($T = 0$), heart attack occurred ($Y_0^* = 1$) but went undiagnosed ($Y_0 = 0$). Under the counterfactual decision to screen in ($T = 1$), heart attack *would have* been averted ($Y_1^* = 0$) but would have been incorrectly diagnosed ($Y_1 = 1$). The observed outcome in medical records reflects the proxy value under factual decision to screen-out ($Y = 0$).

4.5 Methodology

We begin by developing an unbiased risk minimization approach which recovers an estimator for η_t^* given knowledge of error parameters (§ 4.5.1). We then provide a method for estimating α_t and β_t when error parameters are unknown in advance (§ 4.5.2).

4.5.1 Unbiased risk minimization

In this section, we develop an approach for estimating η_t^* given observational data drawn from $p(X, T, Y)$ and measurement error parameters α_t, β_t . Let $f_t \in \mathcal{H}$ for $\mathcal{H} \subset \{f_t : \mathcal{X} \rightarrow [0, 1]\}$ be a probabilistic decision function targeting Y_t^* and let $\ell : \mathcal{Y} \times [0, 1] \rightarrow \mathbb{R}_+$ be a loss function. If we observed target potential outcomes $Y_t^* \sim p^*$, we could directly apply supervised learning techniques to minimize the expected ℓ -risk of f_t over target potential outcomes

$$R_\ell^*(f_t) := \mathbb{E}_{p^*}[\ell(f_t(X), Y_t^*)] \quad (4.2)$$

and learn an estimator for η_t^* via standard empirical risk minimization approaches. Given a *strongly proper composite* loss such that $\arg \min_{f_t} R_\ell^*(f_t)$ is a monotone transform ψ of η_t^*

(e.g., the logistic and exponential loss), this would enable recovering class probabilities from the optimal prediction via the link function ψ (Agarwal, 2014; Menon et al., 2015). However, directly minimizing (4.2) is not possible in our setting because we sample observational proxies instead of target potential outcomes. We address this challenge by constructing a *re-weighted surrogate risk* $R_{t,\tilde{\ell}}^w$ such that evaluating this risk over observed proxy outcomes is equivalent to R_ℓ^* in expectation.

In particular, let $w : \mathcal{X} \rightarrow \mathbb{R}_+$ be a weighting function satisfying $\mathbb{E}_X[w(X)|T = t] = 1$ and let $\ell : \mathcal{Y} \times [0, 1] \rightarrow \mathbb{R}_+$ be a surrogate loss function. We construct a *re-weighted surrogate risk*

$$R_{t,\tilde{\ell}}^w(f_t) := \mathbb{E}_p[w(X)\tilde{\ell}(f_t(X), Y) | T = t] \quad (4.3)$$

such that $R_\ell^*(f_t) = R_{t,\tilde{\ell}}(f_t)$ in expectation. Theorem 4.5.1 shows that we can recover a surrogate risk satisfying this property by constructing $w(x)$ as in (4.4) and $\tilde{\ell}$ as in (4.5). Note that this surrogate risk requires knowledge of α_t, β_t .

Theorem 4.5.1. *Assume treatment-conditional error (1), consistency (5), ignorability (3) and positivity (6). Then under target intervention $t \in \{0, 1\}$, $R_\ell^*(f_t) = R_{t,\tilde{\ell}}^w(f_t)$ for the weighting function $w : \mathcal{X} \rightarrow \mathbb{R}_+$ given by*

$$w(x) := \frac{p(T = t)}{(2t - 1) \cdot \pi(x) + 1 - t} \quad (4.4)$$

and surrogate loss $\tilde{\ell} : \mathcal{Y} \times [0, 1] \rightarrow \mathbb{R}_+$ given by

$$\begin{aligned} \tilde{\ell}(f_t(x), 1) &:= \frac{(1 - \alpha_t) \cdot \ell(f_t(x), 1) - \beta_t \cdot \ell(f_t(x), 0)}{1 - \beta_t - \alpha_t} \\ \tilde{\ell}(f_t(x), 0) &:= \frac{(1 - \beta_t) \cdot \ell(f_t(x), 0) - \alpha_t \cdot \ell(f_t(x), 1)}{1 - \beta_t - \alpha_t} \end{aligned} \quad (4.5)$$

where in (4.4), $\pi(x) := p(T = 1|X = x)$ is the propensity score function.

Proof. We will show that $R_{t,\tilde{\ell}}^w(f_t) = R_{\tilde{\ell}}(f_t) = R_\ell^*(f_t)$, $\forall t = \{0, 1\}$. We begin by showing the first equality. We have that

$$\begin{aligned} R_{t,\tilde{\ell}}^w(f_t) &:= \mathbb{E}_p[w(X)\tilde{\ell}(f_t(X), Y) | T = t] \\ &= \mathbb{E}_{p^*}[w(X)\tilde{\ell}(f_t(X), Y_t) | T = t] \\ &= \mathbb{E}_{p^*}[w(X)\tilde{\ell}(f_t(X), Y_t)] \end{aligned}$$

where the first equality holds by consistency (5) and the second by ignorability (3). As a result, we can express both equalities over potential outcomes $Y_t, Y_t^* \sim p^*$. Next, let $p_t(X) := p(X|T = t)$ and let $\tilde{\ell}_{f_t}(x) := \mathbb{E}_{Y_t}[\tilde{\ell}(f_t(x), Y_t) | X = x]$ be the *expected pointwise surrogate loss* of f_t evaluated at x . Then by Lemma 2 of (Johansson et al., 2020), we have

$$\begin{aligned} R_{t,\tilde{\ell}}^w(f_t) &= \int_{x \in \mathcal{X}} w(x)\tilde{\ell}_{f_t}(x)p_t(x)dx \\ &= \int_{x \in \mathcal{X}} \frac{p_t(x)}{p(x)} w(x)\tilde{\ell}_{f_t}(x)p(x)dx \\ &= R_{\tilde{\ell}}(f_t) \end{aligned}$$

for $w(x) = p(x)/p_t(x)$. The second equality assumes positivity (6) and ignorability (3). Applying Bayes' to $p(x)/p_t(x)$ and rearranging

$$w(x) = \frac{p(x)}{p(X=x|T=t)} = \frac{p(T=t)}{p(T=t|X=x)} = \frac{p(T=t)}{(2t-1) \cdot \pi(x) + 1-t}$$

which is the weighting function in (4.4). Next, we show that $R_{\tilde{\ell}}(f_t) = R_{\ell}^*(f_t)$, which follows from Lemma 1 of Natarajan et al. (2013). Given

$$\eta_t(x) = \begin{pmatrix} 1 - \eta_t(x) \\ \eta_t(x) \end{pmatrix}, \quad T = \begin{pmatrix} 1 - \alpha_t & \alpha_t \\ \beta_t & 1 - \beta_t \end{pmatrix}$$

we can express (4.1) as $\eta_t(x) = T\eta_t^*(x)$ by assumption 1. This error model is invertable via $\eta_t^*(x) = T^{-1}\eta_t(x)$ for

$$T^{-1} = \frac{1}{1 - \alpha_t - \beta_t} \begin{pmatrix} 1 - \beta_t & -\alpha_t \\ -\beta_t & 1 - \alpha_t \end{pmatrix}.$$

Let $\ell(f_t(x)) = (\ell(f_t(x), 0), \ell(f_t(x), 1))^\top$ be a vectorized loss corresponding to labels $e \in \{0, 1\}^2$. Then we have that

$$\begin{aligned} R_{\ell}^*(f_t) &= \mathbb{E}_X \mathbb{E}_{Y_t^* \sim \eta_t^*(X)} [\ell(f_t(X), Y_t^*)] = \mathbb{E}_X [\eta_t^*(X)^\top \ell(f_t(X))] \\ &= \mathbb{E}_X [\eta_t(X)^\top (T^{-1}) \ell(f_t(X))] = \mathbb{E}_X [e^\top (T^{-1}) \ell(f_t(X))] \\ &= \mathbb{E}_{X, Y_t} [\tilde{\ell}(f_t(X), Y_t)] = R_{\tilde{\ell}}(f_t) \end{aligned}$$

Therefore, $R_{\tilde{\ell}}(f_t) = R_{\ell}^*(f_t)$ for a surrogate loss constructed via $\tilde{\ell} = (T^{-1})\ell(f_t(X))$. Multiplying $\ell(f_t(X))$ by T^{-1} and rearranging yields (4.5). $\square$

Intuitively, $R_{t, \tilde{\ell}}^w(f_t)$ applies a *joint bias correction* for OME and distribution shift introduced by historical decision-making policies (i.e., selection bias). The unbiased risk minimization framework dating back to Natarajan et al. (2013) corrects for OME by minimizing a surrogate loss $\tilde{\ell}$ on proxies Y observed over the full population unconditional on treatment. Yet this approach is untenable when decisions impact outcomes ($T \not\perp \{Y^*, Y\}$) and error rates differ across treatments. One possible extension of unbiased risk minimizers to the treatment-conditional setting involves minimizing $\tilde{\ell}$ over the treatment population $p(X|T=t)$

$$R_{t, \tilde{\ell}}(f_t) := \mathbb{E}_p [\tilde{\ell}(f_t(X), Y) \mid T = t]. \quad (4.6)$$

However, $R_{t, \tilde{\ell}} \neq R_{\ell}^*$ in observational settings because the treatment population $p(X|T=t)$ can differ from the marginal population $p(X)$ under historical selection policies when $X \not\perp T$. Therefore, our re-weighting procedure applies a second bias correction that adjusts $p(X|T=t)$ to resemble $p(X)$.

Learning algorithm. As a result of Theorem 4.5.1, we can learn a predictor $\hat{\eta}_t^*$ by minimizing the re-weighted surrogate risk over *observed samples* $(X_1, T_1, Y_1), \dots, (X_n, T_n, Y_n) \sim p$. First, we estimate the weighting function $\hat{w}(x)$ through a finite sample, which boils down to learning propensity scores $\hat{\pi}(x)$ (as shown in (4.4)). Estimating the propensity scores can be done by applying supervised learning algorithms to learn a predictor from X to T . Then for any treatment t , weighting function $\hat{w}$, and predictor f_t , we can approximate $R_{t, \tilde{\ell}}^w(f_t)$ by taking the sample average over the treatment population

$$\hat{R}_{t, \tilde{\ell}}^{\hat{w}}(f_t) := \frac{1}{n_t} \sum_{i: T_i=t} \hat{w}(X_i) \tilde{\ell}(f_t(X_i), Y_i) \quad (4.7)$$

for $n_t = \sum_{i=1}^n 1[T_i = t]$. Therefore, given $\hat{w}$ we can learn a predictor from observational data by minimizing the empirical risk

$$\hat{f}_t \leftarrow \arg \min_{f_t \in \mathcal{H}} \hat{R}_{t,\ell}^{\hat{w}}(f_t). \quad (4.8)$$

We refer to solving (4.8) as *re-weighted risk minimization with a surrogate loss* (Algorithm 1).

Algorithm 1: Re-weighted risk minimization with surrogate loss (RW-SL)

Input : Data $\mathcal{W} = \{(X_i, T_i, Y_i)\}_{i=1}^n \sim p$
Output : Learned estimator $\hat{\eta}_t^*(x)$
 Partition $\mathcal{W}$ into $\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3$;
 On $\mathcal{W}_1$, estimate parameters $\hat{\alpha}_t, \hat{\beta}_t \leftarrow \text{CCPE}(\mathcal{W}_1)$;
 On $\mathcal{W}_2$, learn $\hat{\pi}(x)$ by regressing $T \sim X$;
 On $\mathcal{W}_3$, use $\hat{\pi}(x), \hat{\alpha}_t, \hat{\beta}_t$ to solve $\hat{\eta}_t^*(x) \leftarrow \arg \min_{f_t \in \mathcal{H}} \hat{R}_{t,\ell}^{\hat{w}}(f_t)$;

Algorithm 2: Conditional class probability estimation (CCPE)

Input : Data $\mathcal{V} \sim p$
Output : Parameter estimates $\hat{\alpha}_t, \hat{\beta}_t$
 Partition $\mathcal{V}$ into $\mathcal{V}_1, \mathcal{V}_2$;
 On $\mathcal{V}_1$, learn $\hat{\eta}_t(x)$ by regressing $Y \sim X \mid T = t$;
 On $\mathcal{V}_2$, estimate error parameters: $\hat{\alpha}_t = \min_{x \in \mathcal{X}} \{\hat{\eta}_t(x)\}, \hat{\beta}_t = 1 - \max_{x \in \mathcal{X}} \{\hat{\eta}_t(x)\}$;

4.5.2 Error parameter identification and estimation

Because our risk minimization approach requires knowledge of OME parameters, we develop a method for estimating α_t, β_t from observational data. Error parameter estimation is challenging in decision support applications because target outcomes often result from nuanced social and organizational processes. Understanding the measurement error properties of proxies targeted in criminal justice, medicine, and hiring domains remains an ongoing focus of domain-specific research (Fogliato, Xiang, Lipton, Nagin and Chouldechova, 2021; Akpinar et al., 2021; Mullainathan and Obermeyer, 2021; Obermeyer et al., 2019; Zwaan and Singh, 2015; Chalfin et al., 2016). Therefore, we develop an approach compatible with multiple sources of domain knowledge about proxies, which can be flexibly combined depending on which assumptions are deemed reasonable in a specific context.

Error parameters are *identifiable* if they can be uniquely computed from observational data. Because our error model (e.q. 4.1) expresses the proxy class probability as a linear equation with two unknowns, α_t, β_t are identifiable if the target class probability $c_{t,i}^* = \eta_t^*(x_i)$ and proxy class probability $c_{t,i} = \eta_t(x_i)$ are known at two distinct points $(c_{t,i}^*, c_{t,i})$ and $(c_{t,j}^*, c_{t,j})$ such that $c_{t,i}^* \neq c_{t,j}^*$. Following prior literature (Han et al., 2020), we refer to knowledge of $(c_{t,i}^*, c_{t,i})$ as an *anchor assumption* because it requires knowledge of the unobserved quantity η_t^* . We now introduce several anchor assumptions that are practical in ADS, before showing that these can be flexibly combined to identify α_t, β_t in Theorem 4.5.2.

Min anchor. A min anchor assumption holds if there is an instance at no risk of the target potential outcome under intervention t : $c_{t,i}^* = \inf_{x_i \in \mathcal{X}} \{\eta_t^*(x_i)\} = 0$. Because η_t is a strictly monotone increasing transform of η_t^* , the corresponding value of η_t can be recovered via $c_{t,i} = \inf_{x_i \in \mathcal{X}} \{\eta_t(x_i)\}$ (Menon et al., 2015). Min anchors are reasonable when there are cases that are confirmed to be at no risk based on domain knowledge of the data generating process. For example, a min anchor may be reasonable in diagnostic testing if a patient is

	Know α_t	Min	Base rate	Max	Know β_t
Know α_t	×	×	✓	✓	✓
Min	×	×	✓	✓	✓
Base rate	✓	✓	×	✓	✓
Max	✓	✓	✓	×	×
Know β_t	✓	✓	✓	×	×

Table 4.1: Multiple combinations of min, max, and base rate anchor assumptions (shown via ✓) enable identification of α_t , β_t .

confirmed to be negative for a medical condition based on a high-precision gold standard medical test (Enøe et al., 2000).

Max anchor. A max anchor assumption holds if there is an instance at certain risk of the target outcome under intervention t : $c_{t,i}^* = \sup_{x_i \in \mathcal{X}} \{\eta_t^*(x_i)\} = 1$. The corresponding value of η_t can be recovered via $c_{t,i} = \sup_{x_i \in \mathcal{X}} \{\eta_t(x_i)\}$ because η_t is a strictly monotone increasing transform of η_t^* . Max anchors are reasonable when there are confirmed instances of a positive target potential outcome based on domain knowledge of the data generating process. For example, a max anchor may be justified in a medical setting if a subset of patients have confirmed disease diagnoses based on biopsy results (Begg and Greenes, 1983), or if a disease prognosis (and resulting health outcomes) are known from pathology.

Base rate anchor. A base rate anchor assumption holds if the expected value of η_t^* is known under intervention t : $c_{t,i}^* = \mathbb{E}[\eta_t^*(X)]$. The corresponding value of η_t can be recovered by taking the expectation over the proxy class probability $c_{t,i} = \mathbb{E}[\eta_t(X)]$. Base rate anchors are practical because the prevalence of unobservable target outcomes (e.g., medical conditions (Walter and Irwig, 1988), crime (Kruttschnitt et al., 2014; Lohr, 2019), student performance (Schouwenburg, 2004; Di Folco et al., 2022)) is routinely estimated via domain-specific analyses of measurement error. For instance, studies have been conducted to estimate the base rate of undiagnosed heart attacks (i.e., accounting for measurement error in diagnosis proxy outcomes) Orso et al. (2017). Additionally, the conditional average treatment effect $\mathbb{E}[\eta_1^*(X)] - \mathbb{E}[\eta_0^*(X)]$ is commonly estimated in randomized controlled trials (RCTs) while assessing treatment effect heterogeneity (Hill, 2011). While the conditional average treatment effect is normally estimated via proxies Y_0 and Y_1 , measurement error analysis is a routine component of RCT design and evaluation (Gamerman et al., 2019).

Anchor assumptions can be flexibly combined to identify error parameters based on which set of assumptions are reasonable in a given ADS domain. In particular, Theorem 4.5.2 shows that combinations of anchor assumptions listed in Table 4.1 are sufficient for identifying error parameters under our causal assumptions.

Theorem 4.5.2. *Assume treatment-conditional error (1), consistency (5), ignorability (3) and positivity (6). Then α_t, β_t are identifiable from observational data $p(X, T, Y)$ given any identifying pair of anchor assumptions provided in Table 4.1.*

Proof. To begin, observe that the error model (4.1) expresses the conditional proxy class probability η_t as a linear function of η_t^* with two unknowns. Therefore, given knowledge of the target class probability $c_{t,i}^* = \eta_t^*(x_i)$ and proxy class probability $c_{t,i} = \eta_t(x_i)$ at two distinct points $(c_{t,i}^*, c_{t,i})$ and $(c_{t,j}^*, c_{t,j})$, we can set up a linear equation

$$\begin{aligned} c_{t,i} &= (1 - \beta_t) \cdot c_{t,i}^* + \alpha_t \cdot (1 - c_{t,i}^*) \\ c_{t,j} &= (1 - \beta_t) \cdot c_{t,j}^* + \alpha_t \cdot (1 - c_{t,j}^*) \end{aligned} \tag{4.9}$$

and solve for error parameters

$$\alpha_t = \frac{c_{t,i}^* \cdot c_{t,j} - c_{t,i} \cdot c_{t,j}^*}{c_{t,i}^* - c_{t,j}^*} \quad (4.10)$$

$$\beta_t = \frac{c_{t,i} \cdot c_{t,j}^* - c_{t,i} + c_{t,i}^* - c_{t,j}^* + c_{t,j} - c_{t,i}^* \cdot c_{t,j}}{c_{t,i}^* - c_{t,j}^*} \quad (4.11)$$

provided that $c_{t,i}^* \neq c_{t,j}^*$. Identification of the specific cases in Table 4.1 follows from application of (4.10). When α_t and β_t are both known, identification is not required. When one of β_t (α_t) is known, the corresponding α_t (β_t) can be given by

$$\alpha_t = \frac{c_{t,i} - (1 - \beta_t) \cdot c_{t,i}^*}{(1 - c_{t,i}^*)}, \quad \beta_t = \frac{c_{t,i}^* - c_{t,i} + \alpha_t \cdot (1 - c_{t,i}^*)}{c_{t,i}^*} \quad (4.12)$$

Therefore, only one anchor assumption ($c_{t,i}^*, c_{t,i}$) is required given knowledge of α_t or β_t . However, by (4.12), note that $c_{t,i}^* \neq 1$ is required for identification of α_t and $c_{t,j}^* \neq 0$ is required for identification of β_t . This rules out combinations denoted by (X) in Table 4.1. Error parameters can be derived directly from (4.10) if α_t and β_t are both unknown so long as $c_{t,i}^* \neq c_{t,j}^*$. The specific values of ($c_{t,i}, c_{t,i}^*$) corresponding to min, max, and base rate anchors can be computed via

$$\begin{aligned} c_{t,i}^* &= \inf_{x_i \in X} \{\eta_t^*(x_i)\}, & c_{t,i} &= \inf_{x_i \in \mathcal{X}} \{\eta_t(x_i)\} & \text{(Min anchor)} \\ c_{t,i}^* &= \sup_{x_i \in X} \{\eta_t^*(x_i)\}, & c_{t,i} &= \sup_{x_i \in \mathcal{X}} \{\eta_t(x_i)\} & \text{(Max anchor)} \\ c_{t,i}^* &= \mathbb{E}[\eta_t^*(X)], & c_{t,i} &= \mathbb{E}[\eta_t(X)] & \text{(Base rate anchor)} \end{aligned}$$

Above, the min anchor holds because η_t is a strictly monotone increasing transform of η_t^* by 4.1 such that $c_{t,i} = \arg \inf_{x_i \in X} \{\eta_t(x)\} = \arg \inf_{x_i \in X} \{\eta_t^*(x)\}$. The max anchor holds by the same argument. The base rate anchor holds because $\mathbb{E}_X[\eta_t(x)] = \mathbb{E}_X[\eta_t^*(x) \cdot (1 - \beta_t - \alpha_t) + \alpha_t]$.

Finally, observe that $\eta_t(x)$ is defined over potential outcomes $Y_t \sim p^*$ rather than observational proxies $Y \sim p$. Identification of η_t from observational data follows from

$$\eta_t(x) := p(Y_t = 1 | X = x) = p(Y = 1 | X = x, T = t) \quad (4.13)$$

where the equality holds by ignorability (3) and consistency (5). By positivity (6), we have that the support of $\eta_t(x)$ is defined $\forall x \in X$, which guarantees that the min and max anchor will be defined. □

In practice, we estimate the error rates on finite samples $(X_i, T_i, Y_i) \sim p$, which gives an approximation $\hat{\eta}_t$. Therefore, we propose a conditional class probability estimation (CCPE) method for parameter estimation which estimates $\hat{\alpha}_t, \hat{\beta}_t$ by fitting $\hat{\eta}_t$ on observational data then applying the relevant pair of anchor assumptions to estimate error rates. Algorithm 2 provides pseudocode for this approach with min and max anchors, which can easily be extended to other pairs of identifying assumptions shown in Table 4.1. The combination of min and max anchors is known as *weak separability* (Menon et al., 2015) or *mutual irreducibility* (Scott et al., 2013; Scott, 2015) in the observational label noise literature. Prior results in the observational setting show that unconditional class probability estimation (i.e., fitting $\hat{\eta}(x) = p(Y = 1 | X = x)$) yields a consistent estimator for observational error rates under weak separability (Scott et al., 2013; Reeve et al., 2019). Statistical consistency results extend to the treatment-conditional setting under positivity (6) because $p(T = t | X = x) > 0, \forall t \in \{0, 1\}, x \in \mathcal{X}$. However, asymptotic convergence rates may be slower under strong selection bias if $p(T = t | X = x)$ is near 0.

4.6 Experiments

Experimental evaluation under treatment-conditional OME is challenging due to compounding sources of uncertainty. We do not observe counterfactual outcomes in historical data, making it challenging to estimate the quality of new models via observational data. Further, because the target outcome is not observed directly, we rely on measurement assumptions when studying proxy outcomes in naturalistic data. We address this challenge by conducting a controlled evaluation with synthetic data where ground truth potential outcomes are fully observed. To better reflect the ecological settings of real-world deployments, we also conduct a semi-synthetic evaluation with real data collected through randomized controlled trials (RCTs) in healthcare and employment domains. Our evaluation (1) validates our proposed risk minimization approach, (2) underscores the need to carefully consider measurement assumptions during error rate estimation, and (3) shows that correcting for OME or treatment effects in isolation is insufficient.²

4.6.1 Models

We compare several modeling approaches in our evaluation to examine how existing modeling practices are impacted by treatment-conditional outcome measurement error:

- **Unconditional proxy (UP).** Predict the observed outcome unconditional on treatment: $X \rightarrow Y$. This model *does not adjust for OME or treatment effects*, and reflects model performance in a scenario in which practitioners overlook all challenges examined in this work.³
- **Unconditional target (UT).** Predict the target outcome unconditional on treatment: $X \rightarrow Y^*$. Here, we determine Y^* by applying consistency: $Y^* = (1 - T) \cdot Y_0^* + T \cdot Y_1^*$. This method reflects a setting in which no OME is present but modeling does not account for treatment effects (Wang et al., 2021; Natarajan et al., 2013; Menon et al., 2015; Obermeyer et al., 2019).
- **Conditional proxy (CP).** Predict the proxy outcome conditional on treatment: $X, T \rightarrow Y$. This is a counterfactual model that estimates a conditional expectation *without correcting for OME* (Coston, Mishler, Kennedy and Chouldechova, 2020; Shalit et al., 2017; Künzel et al., 2019).⁴
- **Re-weighted surrogate loss (RW-SL).** Our proposed risk minimization approach, as defined in equation (4.8). This method corrects for both OME and treatment effects in parallel. Additionally, this method corrects for distribution shift due to selection bias in the prior decision-making policy via re-weighting.
- **Target Potential Outcome (TPO).** Directly predict the target potential outcome: $X \rightarrow Y_t^*$. This model is an *oracle* that provides an upper-bound on model performance under no OME or treatment effects.

We also perform an ablation of our proposed RW-SL method by including a model that applies a surrogate loss correction $\tilde{\ell}$ over the treatment population without re-weighting (SL).

²Code for all experiments can be found at: https://github.com/lguerdan/CP_OME.

³This baseline is also called an *observational risk assessment* in experiments reported by Coston, Mishler, Kennedy and Chouldechova (2020).

⁴This model is known by different names in the causal inference literature, including the backdoor adjustment (G-computation) formula (Pearl, 2009; Robins, 1986), T-learner Künzel et al. (2019), and plug-in estimator (Kennedy, 2022).

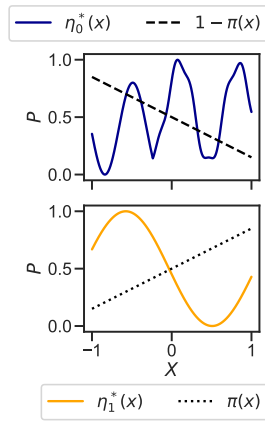


Figure 4.3:
Syn-
thetic
setup.

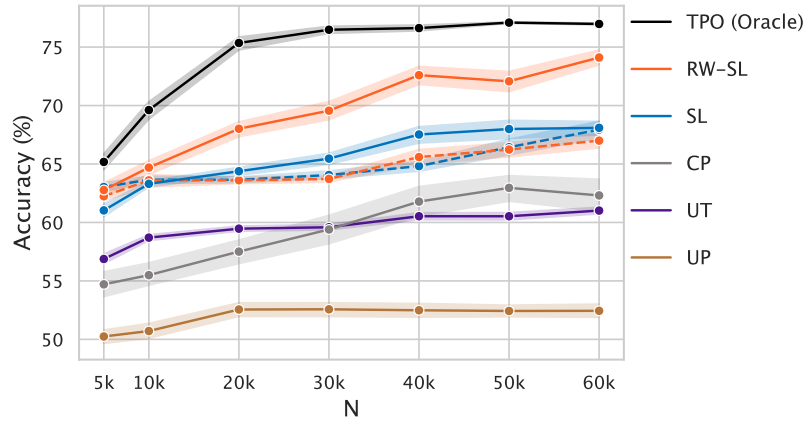


Figure 4.4: Accuracy on Y_0^* as a function of sample size. RW-SL and SL with oracle parameters plotted with solid lines. RW-SL and SL with learned parameters plotted with dashed lines. Results averaged over asymmetric error settings reported in Table 4.2.

(α_0, β_0)	(0.0, 0.4)	(0.1, 0.3)	(0.2, 0.2)	(0.3, 0.1)	(0.4, 0.0)
UP	54.18 (0.09)	53.00 (0.39)	54.89 (1.09)	55.81 (0.74)	46.76 (0.33)
UT	61.57 (0.63)	60.95 (0.50)	60.49 (0.41)	61.00 (0.49)	60.54 (0.70)
CP	51.36 (1.83)	68.24 (2.61)	75.05 (0.92)	67.77 (1.33)	61.88 (0.28)
SL ($\hat{\alpha}, \hat{\beta}$)	72.38 (1.65)	65.45 (0.66)	67.43 (1.64)	68.01 (0.99)	65.92 (1.34)
RW-SL ($\hat{\alpha}, \hat{\beta}$)	69.08 (1.55)	65.96 (1.18)	66.57 (1.32)	68.39 (1.33)	64.56 (0.52)
SL (α, β)	67.09 (1.24)	67.58 (1.19)	67.75 (1.08)	69.11 (1.17)	68.59 (1.41)
RW-SL (α, β)	73.68 (1.49)	73.39 (1.60)	72.52 (1.66)	74.34 (1.15)	75.01 (1.24)
TPO	77.08 (0.11)	77.09 (0.20)	76.98 (0.08)	76.84 (0.18)	76.90 (0.16)

Table 4.2: Model accuracy (s.e.) across error parameter settings (α_0, β_0) at $N = 60k$ samples over 10 runs. Top-2 performance across each (α_0, β_0) setting shown in bold.

4.6.2 Experiments on synthetic data

We begin by experimentally manipulating treatment effects and measurement error via a synthetic evaluation. Because this provides full control over the data generating process, we can evaluate methods against target potential outcomes. This evaluation would not be possible with real-world data because counterfactual target outcomes are unobserved. Our experiment design is consistent with prior synthetic evaluations of counterfactual risk assessments [Coston, Mishler, Kennedy and Chouldechova \(2020\)](#) and causal inference methods ([Shi et al., 2019](#); [Nie et al., 2021](#)). In our evaluation, we sample outcomes via the following data generating process:

1. $Y_t^* := \sim \text{Bern}(\eta_t^*(X)), \forall t \in \{0, 1\}$
2. $Y_t := \begin{cases} 1 - \epsilon_+ & \text{if } Y_t^* = 1, \text{ where } \epsilon_+ \sim \text{Bern}(\beta_t) \\ \epsilon_- & \text{if } Y_t^* = 0, \text{ where } \epsilon_- \sim \text{Bern}(\alpha_t) \end{cases}, \forall t \in \{0, 1\}$
3. $T := \sim \text{Bern}(\pi(X))$
4. $Y^* := (1 - T) \cdot Y_0^* + T \cdot Y_1^*; Y := (1 - T) \cdot Y_0 + T \cdot Y_1$

As shown in Figure 4.3, we draw $X \sim U(-1, 1)$ and sample target potential outcomes from sinusoidal class probability functions. Note that our choice of $\eta_0^*(x)$, $\eta_1^*(x)$ satisfies min and max anchor assumptions. Because $\eta_0^*(x)$ and $\eta_1^*(x)$ differ, models that do not condition on treatment (i.e., UP, UT) will learn an average of the two class probability functions. Under our choice of $\pi(x)$, fewer samples are drawn from $\eta_1^*(x)$ in the region where $\pi(x)$ is small (near $x = -1$), and fewer samples are drawn from $\eta_0^*(x)$ in the region where $1 - \pi(x)$ is small (near $x = 1$). This introduces selection bias when sampling from $\pi(x)$.

Setup details

We train each model in § 4.6.1 to predict risk under no intervention ($t = 0$) and vary (α_0, β_0) . We keep (α_1, β_1) fixed at $(0, 0)$ across settings. When estimating OME parameters, we run CCPE with sample splitting and cross-fitting with min and max anchor assumptions for identification. These assumptions hold precisely under this controlled evaluation (Figure 4.3). We run all methods with sample splitting and cross-fitting and report performance on Y_0^* .

Results

Figure 4.4 shows the performance of each model as a function of sample size. TPO provides an upper bound on performance because it learns directly from target potential outcomes. RW-SL with oracle parameters (α, β) outperforms all other methods trained on observational data across across the full range of sample sizes. Thus, while Thm. 4.5.1 shows that RW-SL recovers an unbiased risk estimator *in expectation*, this method also demonstrates favorable finite-sample performance characteristics in practice. This finding is inline with prior experimental evaluations of unbiased risk estimators reported in the standard supervised learning setting ([Natarajan et al., 2013](#); [Wang et al., 2021](#)), and is further supported by reliable performance characteristics we observe in small sample regimes.

In contrast, both models that do not condition on treatment (UP and UT), and the conditional regression trained on proxy outcomes (CP), reach a performance plateau by 50k samples and do not benefit from additional data. This indicates that (1) learning a counterfactual model and (2) correcting for measurement error is necessary to learn η_t^* in this evaluation. We likely observe a sharper plateau in UP and UT above 20k samples because

these approaches fit a weighted average of η_0^* and η_1^* (where η_1^* differs from η_0^* considerably). We observe that RW-SL and SL performance deteriorates with learned parameters ($\hat{\alpha}, \hat{\beta}$) across all sample size settings due to misspecification in learned parameter estimates and weights.

Table 4.2 shows a breakdown across error rates (α_0, β_0) at 60k samples. RW-SL outperforms SL when oracle parameters are known. However, RW-SL and SL perform comparably when weights and parameters are learned. This may be because RW-SL relies on estimates $\hat{w}$ in addition to $\hat{\alpha}_0, \hat{\beta}_0$, which could introduce instability given misspecification in $\hat{w}$. CP performs notably well under high error parameter symmetry (i.e., $\alpha_0 = \beta_0 = .2$). This is consistent with prior results from the class-conditional label noise literature (Natarajan et al., 2013; Menon et al., 2015), which show that the optimal classifier threshold for misclassification risk does not change under symmetric label noise. CP performs worse under high error asymmetry. We do not observe a similar performance improvement in UP and UT in the symmetric error setting because these baselines learn a weighted combination of η_0 and η_1 , which differs from the target function η_0^* at all classification thresholds.

4.6.3 Semi-synthetic experiments on healthcare and employment data

In addition to our synthetic evaluation, we conduct experiments using real-world data collected as part of randomized controlled trials (RCTs) in healthcare and employment domains. While this evaluation affords less control over the data generating process, it provides a more realistic sample of data likely to be encountered in real-world model deployments. Evaluation via data from randomized or partially randomized experimental studies is useful for validating counterfactual prediction approaches because random assignment ensures that causal assumptions are satisfied (Shalit et al., 2017; Johansson et al., 2020; Coston, Kennedy and Chouldechova, 2020).

Randomized Controlled Trial (RCT) Datasets

In 2008, the U.S. state of Oregon expanded access to its Medicare program via a lottery system (Finkelstein et al., 2012). This lottery provided an opportunity to study the effects of Medicare enrollment on healthcare utilization and medical outcomes via an experimental design, commonly referred to as the Oregon Health Insurance Experiment (OHIE). Lottery enrollees completed a pre-randomization survey recording demographic factors and baseline health status and were given a one-year follow-up assessment of health status and medical care utilization. We refer the reader to Finkelstein et al. (2012) for details. We use the OHIE dataset to construct an evaluation task that parallels the label choice bias analysis of Obermeyer et al. (2019). We use this dataset rather than synthetic data released by Obermeyer et al. (2019) because (1) treatment was randomly assigned, ruling out positivity and ignorability violations possible in observational data, and (2) OHIE data contains covariates necessary for predictive modeling. We predict diagnosis with an active chronic medical condition over the one-year follow-up period given $D = 58$ covariates, including health history, prior emergency room visits, and public services use. We predict chronic health conditions because findings from Obermeyer et al. (2019) indicate that this outcome variable is a reasonable choice of proxy for patient medical need. We adopt the randomized lottery draw as the treatment.⁵

We conduct a second real-world evaluation using the JOBS dataset, which investigates the effect of job retraining on employment status (Shalit et al., 2017). This dataset includes an experimental sample collected by LaLonde (1986) via the National Supported Work (NSW) program

⁵The OHIE experiment had imperfect compliance (≈ 30 percent of selected individuals successfully enrolled (Finkelstein et al., 2012)). Therefore, we predict diagnosis with a new chronic health condition given the *opportunity to enroll* in Medicare. This evaluation is consistent with many high-stakes decision-support settings granting opportunities to individuals, which they have a choice to pursue if desired.

(297 treated, 425 control) consisting primarily of low-income individuals seeking job retraining. Smith and Todd (2005) combine this sample with a “PSID” comparison group (2,490 control) collected from the general population, which resulted in a final sample with 297 treated and 2,915 control. This dataset includes $D = 17$ covariates including age, education, prior earnings, and interaction terms. 482 (15%) of subjects were unemployed at the end of the study. Following Johansson et al. (2020), we construct an evaluation task predicting unemployment under enrollment ($t = 1$) and no enrollment ($t = 0$) in a job retraining program conditional on covariates.

Synthetic OME and selection bias

We experimentally manipulate OME to examine how outcome regressions perform under treatment-conditional error of known magnitude. We adopt diagnosis with a new chronic condition and future unemployment as a *target outcome* for OHIE and JOBS, respectively. We observe proxy outcomes by flipping target outcomes with probability (α_0, β_0) . We keep (α_1, β_1) fixed at $(0, 0)$. This procedure of generating proxy outcomes by flipping available labels is a common approach for vetting the feasibility of new methodologies designed to address OME (Wang et al., 2021; Menon et al., 2015; Natarajan et al., 2013). This approach offers precise control over the magnitude of OME but suffers from less ecological validity than studying multiple naturalistic proxies (Obermeyer et al., 2019). We opt for this semi-synthetic evaluation because (1) the precise measurement relationship between naturally occurring proxies may not be fully known, (2) the measurement relationship between naturally occurring proxies cannot be manipulated experimentally, and (3) there are few RCT datasets (i.e., required to guarantee causal assumptions) that contain multiple proxies of the same target outcome.

Models used for decision support are typically trained using data gathered under a historical decision-making policy. When prior decisions were made non-randomly, this introduces selection bias ($T \not\perp X$) and causes distribution shift between the population that received treatment t in training data, and the full population encountered at deployment time. Therefore, we emulate selection bias in the *training dataset*, and evaluate models over a held-out test set of randomized data. We insert selection bias in OHIE data by removing individuals from the treatment (lottery winning) arm with household income above the federal poverty line (10% of the treatment sample). This resembles an observational setting in which low-income individuals are more likely to receive an opportunity to enroll in a health insurance program (e.g., Medicaid, which determines eligibility based on household income in relation to the federal poverty line). We restrict our analysis to single-person households, yielding $N = 12,994$ total samples (6,189 treatment, 6,805 control).

We model selection bias in JOBS data by including samples from the observational and experimental cohorts in the training data. Because the PSID comparison group consists of individuals with higher income and education than the NSW group, there is considerable distribution shift across the NSW and PSID cohorts (LaLonde, 1986; Smith and Todd, 2005; Johansson et al., 2020). Therefore, a model predicting unemployment over the control population (consisting of NSW and PSID samples) may suffer from bias when evaluated against test data that only includes samples from the NSW experimental arm. Thus we split data from the NSW experimental cohort 50-50 across training and test dataset, and only include PSID data in the training dataset.

Experimental setup

We include a Conditional Target (CT) model in place of a TPO model because counterfactual outcomes are not available in experimental data. CT provides a reasonable upper-bound on performance because identifiability conditions are satisfied in an experimental setting

(Pearl, 2009). However, it is not possible to report accuracy over potential outcomes because counterfactual outcomes are unobserved. Therefore, we report error in ATE estimates $\tau - \hat{\tau}$, for

$$\tau := \mathbb{E}[Y^* \mid T = 1] - \mathbb{E}[Y^* \mid T = 0], \quad \hat{\tau} := \mathbb{E}[\hat{\eta}_1(X)] - \mathbb{E}[\hat{\eta}_0(X)]$$

where τ corresponds to the ground-truth treatment effect reported by prior work (Johansson et al., 2020; Denteh and Liebert, 2022) and $\hat{\eta}_t$ is a learned model discussed in § 4.6.1. One subtlety of this comparison is that our outcome regressions target the *conditional* average treatment effect, while τ reflects the ATE across the full population. Following prior evaluations (Johansson et al., 2020), we compare all methods against the ATE because the ground-truth CATE is not available for JOBS or OHIE data.⁶ We report results over a test fold of randomized data that does not contain flipped outcomes or selection bias.

Results. Figure 4.5 shows bias in ATE estimates $\tau - \hat{\tau}$ over 10 runs on JOBS and OHIE data. The left panel compares CP, UT, UP, and the oracle CT model against RW-SL/SL with oracle parameters (α_0, β_0) , (α_1, β_1) . We show performance of RW-SL with learned parameters $(\hat{\alpha}_0, \hat{\beta}_0)$, $(\hat{\alpha}_1, \hat{\beta}_1)$ on the right panel. The left panel shows that CP is highly sensitive to measurement error. This is because measurement error introduces bias in estimates of the conditional expectations, which propagates to treatment effect estimates. Because UT and UP do not condition on treatment, they estimate an *average* of the outcome functions η_0^* and η_1^* , and generate predictions near 0. Therefore, while UT and UP perform well on OHIE data due to a small ground-truth ATE ($\tau = 0.015$), they perform poorly on JOBS ($\tau = -0.077$). SL and RW-SL with oracle parameters α_t , β_t perform comparably to the CT model with oracle access to target outcomes across all measurement error settings.

While we observe that re-weighting improves performance in our synthetic evaluation (given oracle parameters), we do not observe a similar advantage of RW-SL over SL in this experiment. Our results parallel other empirical evaluations of re-weighting for counterfactual modeling tasks on real-world data (e.g., see § 3.4.2 in (Coston, Mishler, Kennedy and Chouldechova, 2020)). One potential explanation for this finding is that our predictive model class (multi-layer MLPs) is large enough to learn the target regressions η_0^* and η_1^* for OHIE and JOBS data, even after our insertion of synthetic selection bias. As a result, re-weighting may not be required to learn a reasonable estimate of η_0^* and η_1^* given available data. This interpretation is supported by strong performance of the oracle CT model.

As shown on the right panel of Figure 4.5, RW-SL performance is highly sensitive to the choice of anchor assumption used to estimate parameters $(\hat{\alpha}_0, \hat{\beta}_0)$, $(\hat{\alpha}_1, \hat{\beta}_1)$ as indicated by increased bias in $\hat{\tau}$ and greater variability over runs. In particular, RW-SL performs poorly when Min/Max and Br/Max pairs of anchor assumptions are used to estimate error rates because the max anchor assumption is violated on OHIE and JOBS data. We shed further light on this finding by fitting the CT model to estimate $\hat{\eta}_0^*$, $\hat{\eta}_1^*$ on OHIE data, then computing inferences over a validation fold X_{val} . This analysis reveals that

$$\begin{aligned} \min_{x \in X_{val}} \hat{\eta}_0^* &\approx 2.23 \cdot e^{-6}, & \max_{x \in X_{val}} \hat{\eta}_0^* &\approx 0.85 \\ \min_{x \in X_{val}} \hat{\eta}_1^* &\approx 1.02 \cdot e^{-5}, & \max_{x \in X_{val}} \hat{\eta}_1^* &\approx 0.81 \end{aligned}$$

which suggests that the min anchor assumption that $\min_{x \in X_{val}} \hat{\eta}_t^* = 0$ is reasonable for $t \in \{0, 1\}$, while the max anchor assumption that $\max_{x \in X_{val}} \hat{\eta}_t^* = 1$ is violated for both $t \in \{0, 1\}$.

⁶While our insertion of synthetic selection bias (§4.6.3) introduces distribution shift such that $p(X|T=1)$ differs from $p(X|T=0)$, it does not alter ground-truth τ because the conditional outcome distribution $p(Y^*|T)$ remains unchanged. This setup recreates the unconfounded observational setting in which causal identification assumptions are satisfied (Rubin, 1974).

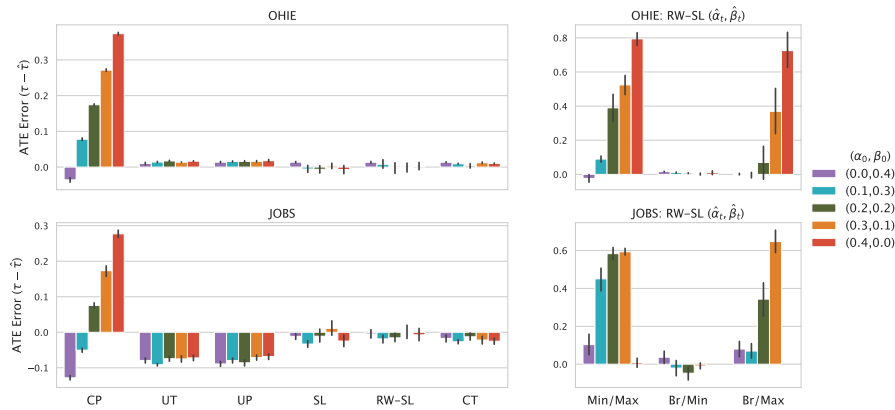


Figure 4.5: Bias in ATE estimates on OHIE and JOBS data. Error bars indicate standard error over ten runs. CT is a model with oracle access to target outcomes and RW-SL is our proposed approach.

Therefore, because the min anchor assumption is satisfied for these choices of target outcome, and the ground-truth base rate is known in this experimental setting, RW-SL demonstrates strong performance given the Br/Min combination of anchor assumptions. In contrast, because the max anchor is violated, estimating β_t by taking the supremum of $\hat{\eta}_t(x)$ introduces bias in $\hat{\beta}_t$, which results in poor performance of RW-SL with Min/Max and Br/Max anchors. Applying this same procedure to the unemployment outcome targeted in JOBS data also reveals a violation of the max anchor assumption. These results underscore the importance of selecting anchor assumptions in close consultation with domain experts because it is not possible to verify anchor assumptions by learning $\hat{\eta}_t^*$ when the target outcome of interest is unobserved.

4.7 Discussion

In this work, we show the importance of carefully addressing intersectional threats to model reliability during the development and evaluation of predictive models for decision support. Our theoretical and empirical results validate the efficacy of our unbiased risk minimization approach. When OME parameters are known, our method performs comparably to a model with oracle access to target potential outcomes. However, our results underscore the importance of vetting anchoring assumptions used for error parameter estimation before using error rate estimates for risk minimization. Critically, our experimental results also demonstrate that correcting for a single threat to model reliability in isolation is insufficient to address model validity concerns (Raji et al., 2022), and risks promoting false confidence in corrected models. Below, we expand upon key considerations surfaced by our work.

4.7.1 Decision points and complexities in measurement error modeling

Our work speaks to key complexities faced by domain experts, model developers, and other stakeholders while examining proxies in ADS. One decision surfaced by our work entails which *measurement error model* best describes the relationship between the unobserved outcome of policy interest and its recorded proxy. We open this work by highlighting a measurement model decision made by Obermeyer et al. (2019) during their audit of a clinical risk assessment: that error rates are fixed across treatments. Our work suggests that failing to account for treatment-conditional error in OME models may exacerbate reliability concerns. However, at the same time, the error model we adopt in this work intentionally abstracts over other factors known

to impact proxies in decision support tasks, including error rates that vary across covariates. Although this simplifying assumption can be unreasonable in some settings (Fogliato, Xiang, Lipton, Nagin and Chouldechova, 2021; Akpinar et al., 2021), including the one studied by Obermeyer et al. (2019), it is helpful in foregrounding previously-overlooked challenges involving treatment effects and selection bias. In practice, model developers correcting for measurement error may wish to combine our methods with existing unbiased risk minimization approaches designed for group-dependent error where appropriate (Wang et al., 2021). Further, analyses of measurement error should not overlook more fundamental conceptual differences between target outcomes and proxies readily available for modeling (e.g., when long-term child welfare related outcomes targeted by a risk assessment differ from *immediate* threats to child safety weighted by social workers (Kawakami et al., 2022)). This underscores the need to carefully weigh the validity of proxies in consultation with multiple stakeholders (e.g., domain experts, data scientists, and decision-makers) while deciding whether OME correction is warranted.

A second decision point highlighted in this work entails the *specific measurement error parameters* that govern the relationship between target and proxy outcomes. In particular, our work underscores the need for a tighter coupling between domain expertise and data-driven approaches for error parameter estimation. Current techniques designed to address OME in the machine learning literature – which typically examine settings with “label noise” – rely heavily upon data-driven approaches without close consideration of whether the underlying measurement assumptions hold (Wang et al., 2021; Northcutt et al., 2021; Menon et al., 2015; Xia et al., 2019). While application of these assumptions may be practical for methodological evaluations and theoretical analysis (Scott et al., 2013; Scott, 2015; Reeve et al., 2019), these assumptions should be carefully vetted when applying OME correction to real-world proxies. This is supported by our findings in Figure 4.5, which show that RW-SL performs poorly when the anchor assumptions used for error parameter estimation are violated. Our flexible set of anchor assumptions provides a step towards a tighter coupling between domain expertise and data-driven approaches in measurement parameter estimation.

4.7.2 Challenges of linking causal and statistical estimands

Our counterfactual modeling approach requires several causal identifiability assumptions (Pearl, 2009), which may not be satisfied in all decision support contexts. Of our assumptions, the most stringent is likely ignorability, which requires that no unobserved confounders influenced past decisions and outcomes. While recent modeling developments may ease ignorability-related concerns in some cases Coston, Mishler, Kennedy and Chouldechova (2020); ?, model developers should carefully evaluate whether confounders are likely to impact a model in a given deployment context. At the same time, our results show that formulating algorithmic decision support as a “*pure prediction problem*” that optimizes predictive performance without estimating causal effects (Kleinberg et al., 2015) imposes equally serious limitations. If the policy-relevant target outcome of interest is risk *conditional on intervention* (as is often the case in decision support applications), an observational model will generate invalid predictions for cases that historically responded most to treatment (Coston, Mishler, Kennedy and Chouldechova, 2020). Our results, which empirically demonstrate poor performance of observational PU and TU models that overlook treatment-effects, corroborate prior findings indicating that counterfactual modeling is required to ensure the reliability of RAIs in decision support settings (Coston, Mishler, Kennedy and Chouldechova, 2020). Taken together, our work suggests that domain experts and model developers should exercise considerable caution while mapping the causal estimand of policy interest to the statistical estimand targeted by a predictive model (Lundberg et al., 2021).

Chapter 5

Supporting Informative Pre-Deployment Model Evaluations via Uncertainty Cancellation

This chapter is drawn from the paper:

Luke Guerdan, Amanda Coston, Kenneth Holstein, and Zhiwei Steven Wu.
Predictive Performance Comparison of Decision Policies under Confounding.
In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

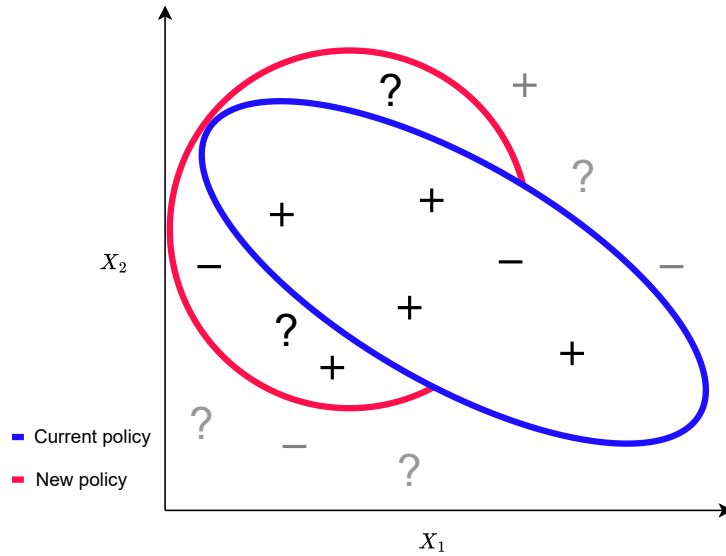


Figure 5.1: Illustration of uncertainty in comparing two policies in a toy setting with $X \in \mathcal{R}^2$. Points are labelled by their outcome: positive (+), negative (-) or unknown (?). Ovals denote the selection region of a policy. Points that neither policy selects (denoted by grey) are irrelevant to the comparison. Our method leverages this to reduce policy comparison uncertainty.

5.1 Motivation

Predictive models are often introduced under the rationale that they improve performance over an existing decision-making policy (Grove et al., 2000). For example, models have been developed with the goal of improving the accuracy of human decisions in domains such as healthcare, criminal justice, and education (Baker and Hawn, 2021; Rambachan et al., 2021). Given the high-stakes nature of these domains, regulatory frameworks have called for organizations to provide explicit comparisons of predictive models against the status quo they are intended to replace (Wyden, n.d.; Johnson and Zhang, 2022). For example, consider two common policy comparison settings of real-world significance.

Example 5.1.1 (Human vs algorithm decisions). Let the status quo policy consist of human-only decisions (e.g., pre-trial release decisions (Kleinberg et al., 2018), medical testing decisions (Mullainathan and Obermeyer, 2019)) made using covariates and unobserved contextual information. We would like to evaluate whether a model recommending actions by thresholding risk predictions would constitute an improvement over the current human-only policy (e.g., Rambachan et al. (2021); Chouldechova et al. (2018)).

Example 5.1.2 (Human+algorithm vs algorithm decisions). Let the status quo policy consist of humans making decisions with the support of a predictive model, covariates, and unobserved contextual information. We would like to assess whether decisions made using the predictive model alone would yield a performance improvement over the human-algorithm combination (e.g., Cheng et al. (2022)).

Both of these examples can be formalized as confounded off-policy evaluation in the contextual bandits setting (Jung et al., 2020; Swaminathan and Joachims, 2015; Tennenholtz et al., 2021; Rambachan et al., 2022). Given observational data collected under a confounded status quo policy, our goal is to compare predictive performance against an alternative policy which would assign different actions. A common strategy for off-policy evaluation under confounding involves *partial identification* of policy performance within an uncertainty interval

(Kallus and Zhou, 2021; Rambachan et al., 2021, 2022; Pu and Zhang, 2021; Zhang et al., 2021). Yet constructing informative performance intervals is challenging because it often requires imposing untestable assumptions on the structure and magnitude of unmeasured confounders impacting the status quo policy.

In this work, we develop a novel partial identification method which reduces uncertainty in policy performance comparisons by isolating confounding-related uncertainty in the disagreement region of the action space (Figure 5.1). Our uncertainty cancellation approach yields tighter regret intervals than existing frameworks designed for non-comparative performance evaluations (Rambachan et al., 2022; Namkoong et al., 2020), which account for redundant uncertainty in the agreement region of the action space. Our simple approach operates by (1) decomposing policy performance measures into sets of identified and partially identified comparison statistics, then (2) cancelling common partially identified terms in the regret estimand.

We show that a diverse set of causal assumptions studied in prior off-policy evaluation literature imply uncertainty sets over partially identified comparison statistics. As a result, while existing off-policy evaluation frameworks are often tied to a fixed causal assumption (e.g., the marginal sensitivity model (Tan, 2006)) that may be unjustified in some contexts, ours is interoperable with a broad family of modern identification strategies (e.g., Rosenbaum's Γ model (Namkoong et al., 2020; Zhang et al., 2021), instrumental variable (Lakkaraju et al., 2017; Kleinberg et al., 2018), proximal variable (Ghassami et al., 2023), or the marginal sensitivity model (MSM) (Kallus and Zhou, 2021)).

We propose a flexible plug-in algorithm for finite sample estimation of regret bounds under no parametric assumptions on the data-generating policy. While this procedure inherits slow non-parametric convergence rates of regression functions used to estimate bounds on comparison statistics, we also show how to construct assumption-tailored doubly robust estimators which attain fast root- n convergence rates under no parametric assumptions. We conduct synthetic experiments verifying the coverage of our regret intervals. We conclude by illustrating how our framework can be used to support a pre-deployment evaluation of a proposed modification to a healthcare enrollment policy (Obermeyer et al., 2019). Our results demonstrate that, in some cases, our improved regret interval supports more conclusive pre-deployment assessments of decision policies than would be possible via existing off-policy evaluation approaches.

We make the following main contributions:

- We formulate the problem of comparative predictive performance evaluations for decision making policies under unmeasured confounding (§ 5.3).
- We propose a novel partial identification technique that yields informative bounds on predictive performance differences by isolating comparison-relevant uncertainty (§ 5.4). Our technique is interoperable with a range of modern identification approaches from causal inference and off-policy evaluation literature (§ 5.5).
- We develop flexible methods for finite sample estimation of regret bounds under no parametric assumptions on the confounded status quo policy (§ 5.6).
- We validate our framework theoretically and via synthetic data experiments (§ 5.7). We demonstrate how our framework can support a pre-deployment evaluation of a proposed healthcare enrollment policy (§ 5.8).¹

¹All code for experiments is publicly available [here](#).

5.2 Related Work

Off-policy evaluation (OPE) is a widely studied problem in reinforcement learning (Uehara et al., 2022; Jiang and Li, 2016; Precup, 2000; Namkoong et al., 2020; Kallus and Zhou, 2020), contextual bandits (Si et al., 2020; Tennenholtz et al., 2021; Swaminathan and Joachims, 2015; Xu et al., 2021; Dudik et al., 2014), and econometrics (Kitagawa and Tetenov, 2018; Huber, 2021) literature. We study the confounded offline contextual bandits setting (Jung et al., 2020; Swaminathan and Joachims, 2015; Tennenholtz et al., 2021), in which observational data is collected under a behavior policy $\pi_0(X_i, U_i)$ which acts on covariates X_i and unmeasured confounders U_i .

Existing OPE frameworks typically estimate the *value function* $V(\pi) = \mathbb{E}[Y(\pi(X))]$ of a new policy $\pi(X_i)$ via observational data (Kallus and Zhou, 2020; Namkoong et al., 2020; Zhang et al., 2021; Uehara et al., 2022; Si et al., 2020; Kallus and Zhou, 2020; Hatt et al., 2022; Ek et al., 2023; Tennenholtz et al., 2021; Ishikawa and He, 2023).² In this work, we instead target predictive performance measures (e.g., Accuracy, TNR, PPV), which are often of interest during policy evaluation of algorithmic risk assessments. While OPE of predictive performance measures introduces estimand-specific challenges (i.e., conditioning on a partially-observed potential outcome), approaches developed for OPE of value functions are also applicable in the algorithmic decision-making context (e.g., doubly robust estimation, reweighting).³

In particular, Rambachan et al. (2022) develop a partial identification framework for robust learning and evaluation of algorithmic risk assessments under confounding. This framework leverages doubly-robust approaches for OPE of predictive performance measures. One natural strategy for extending this framework to the policy comparison setting involves computing bounds on $V(\pi)$ and $V(\pi_0)$ *independently* before computing the regret interval by taking a difference. However, we show that this approach yields an overly conservative regret interval by introducing redundant uncertainty in the agreement region of the policy action space. We introduce an alternative δ -regret interval, which we show yields more informative bounds than this baseline approach. Moreover, we extend the Mean Outcome Sensitivity Model (MOSM) introduced by Rambachan et al. (2022) to develop uncertainty sets around partially identified policy comparison statistics which are consistent with a range of modern causal assumptions (e.g., instrumental variable, MSM, Rosenbaum’s Γ sensitivity model).

Our work also builds upon prior literature studying policy learning and selection under confounding (Zhang et al., 2021; Konyushova et al., 2021; Yang et al., 2022; Kuzborskij et al., 2021; Zhang et al., 2021; Kallus and Zhou, 2021, 2018a; Hatt et al., 2022; Ek et al., 2023; Gao and Yin, 2023; Balachandar et al., 2023). Zhang et al. (2021) develop a framework for ranking individualized treatment policies under Rosenbaum’s Γ sensitivity model. Namkoong et al. (2020) also use Rosenbaum’s Γ sensitivity model to partially identify $V(\pi)$ in the sequential setting. Kallus and Zhou (2021, 2018a) develop an approach for learning minimax optimal decision-policies under the MSM. In principle, confounding-robust policy learning frameworks such as these can also be used to recover δ -regret intervals against the confounded data-generating policy (Hatt et al., 2022; Ek et al., 2023; Gao and Yin, 2023). However, these techniques (1) target the mean potential outcome rather than predictive performance measures, and (2) are tied to a fixed causal assumption which may not be applicable in all contexts. In contrast, our framework is interoperable with a range of assumptions (e.g., instrumental variable, MSM, proximal variable) studied in prior OPE literature.

²Work in the confounded offline setting typically studies updated policies that only observe covariates. When $\pi(X_i)$ also observes U_i regret intervals tend to be vacuous.

³In the algorithmic decision-making context, the term *selective labels* is often used to describe a setting in which a potential outcome is only observed under one of the possible actions (Lakkaraju et al., 2017; Wei, 2021; Coston et al., 2021; De-Arteaga et al., 2018).

Finally, our work relates to literature studying comparison of human versus algorithmic decision-making policies. Kleinberg et al. (2018) develop a framework for comparing human decisions and algorithmic predictions under selective labels and unobservables. The contraction technique underpinning this approach compares the failure rate of judicial decisions against a predictive model under an assumption that judges have heterogeneous selection rates (Lakkaraju et al., 2017). Later developments formalize this approach into an instrumental variable framework (Chen et al., 2023; Rambachan et al., 2021, 2022). Rambachan et al. (2021) develop a framework for identifying systematic prediction mistakes in historical human decisions. This framework enables utility-based comparisons of human versus algorithmic decision-making policies under varying sets of econometric assumptions. Most recently, Ben-Michael et al. (2024) devise an experimental framework for evaluating human+algorithm hybrid workflows against the quality of decisions that a human or algorithm would make alone. This framework supports *post-deployment* policy comparisons following randomized assignment to a treatment (algorithmic recommendation) or control (no algorithmic recommendation) condition. In contrast, our framework is designed to support *pre-deployment* policy comparisons given observational data collected under a status quo policy.

5.3 Preliminaries

Let $\pi_0 : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{A}$ be a status quo decision-making policy assigning binary actions given measured covariates $X \in \mathcal{X}$ and unmeasured confounders $U \in \mathcal{U}$. Let $\pi : \mathcal{X} \rightarrow \mathcal{A}$ be a proposed policy that assigns binary actions only via covariates.⁴ For example, in a medical testing context, π_0 is an existing (e.g., physician based) testing policy, while π is a proposed algorithmic policy. Let D^{π_0}, T^π be random variables indicating actions selected under the status quo and proposed policies, respectively.⁵ Let $Y(1) \in \mathcal{Y}$ be the outcome of interest to the policy comparison. This is the potential outcome which *would be observed* given a positive decision ($D^{\pi_0} = 1$) selected under the status quo policy (Rubin, 2005). For example, $Y(1)$ denotes the disease status of a patient, or repayment if a lendee is granted a loan. Let $Y \in \mathcal{Y}$ be the binary outcome observed in observational data. In our selective labels setting (Lakkaraju et al., 2017), Y is only observed when an instance received a positive decision ($D^{\pi_0} = 1$) under π_0 . We make the following standard consistency assumption on Y (Rubin, 2005).

Assumption 5 (Consistency). $D^{\pi_0} = 1 \implies Y = Y(1)$.

For example, consistency would be violated if the testing decision for one patient impacts the disease status of another (Hudgens and Halloran, 2008). Additionally, we assume that each instance has some probability of both decisions under the status quo policy.

Assumption 6 (Positivity). $p(D^{\pi_0} = a \mid X = x, U = u) > 0, \forall a \in \mathcal{A}, x \in \mathcal{X}, u \in \mathcal{U}$.

5.3.1 Problem Formulation

Given the target distribution $(X, U, D^{\pi_0}, T^\pi, Y(1), Y) \sim p^*(\cdot)$, our goal is to evaluate the regret $R^*(\pi, \pi_0; m) = m^*(\pi) - m^*(\pi_0)$, where $m^* : \mathcal{A} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ is a policy performance measure computed with respect to $Y(1)$. We study the following performance measures, which are analogously defined for the baseline policy.

⁴Our framework is compatible with new policies which are stochastic and deterministic. However we assume that π is known.

⁵We sometimes omit policy superscripts over random variables to ease notation.

Predictive performance: Let $m_y^*(\pi) = p(T^\pi = 1 \mid Y(1) = y)$ be the positive ($y = 1$) and negative ($y = 0$) class predictive performance.⁶ For example, a positive false positive rate regret $m_{y=0}^*(\pi) - m_{y=0}^*(\pi_0) > 0$ indicates that the proposed policy recommends tests for healthy patients more frequently than the status quo. Similarly, let $m_a^*(\pi) = \mathbb{E}[Y(1) = a \mid T^\pi = a]$ be the positive ($a = 1$) and negative ($a = 0$) predictive value of π .

Utility: Let $u_{a,y} \in \mathbb{R}_+$ be the utility of outcome y under action a . We let

$$m_u^*(\pi) = \mathbb{E} \left[\sum_{a,y} u_{a,y} \cdot \mathbb{I}\{A^\pi = a, Y(1) = y\} \right]$$

be the expected utility of π .⁷ A positive utility regret $m_u^*(\pi) - m_u^*(\pi_0) > 0$ indicates that the updated policy has an overall welfare benefit in comparison to the status quo (Rambachan et al., 2021). For example, this measure reflects settings in which turning away sick patients incurs lower utility than testing healthy ones ($u_{01} < u_{10}$) as well as accuracy ($u_{11} = u_{00} = 1$, $u_{10} = u_{01} = 0$).

Because of the selective labels problem, the policy regret is *partially identified* within the interval $R^*(\pi, \pi_0; m) \in [R(\pi, \pi_0; m), \bar{R}(\pi, \pi_0; m)]$. Our goal in this work is to recover the most informative regret interval possible given observational data $O = \{(X_i, T_i^\pi, D_i^{\pi_0}, Y_i) : i = 1, \dots, n\} \sim p(X, D^{\pi_0}, T^\pi, Y)$ generated under the status-quo policy.

5.3.2 Partial Identification of Policy Performance

To tightly characterize policy performance differences, we introduce notation that enables us to isolate confounding-related uncertainty in performance measures. We decompose performance measures into a set of sufficient v -statistics $\mathbf{v} = \{v_y(t, d) : \forall y, t, d\}$, where each term $v_y(t, d) = p(T^\pi = t, D^{\pi_0} = d, Y(1) = y)$ is the joint probability of policy actions and the potential outcome.

The subset $\mathbf{v}_1 = \{v_y(t, 1) : \forall y, t\}$ contains all v -statistics known from observational data. By consistency (5), each term in $\mathbf{v}_1$ is given by

$$p(T = t, D = 1, Y(1) = 1) = p(T = t, D = 1, Y = 1).$$

Similarly, $\mathbf{v}_0 = \{v_y(t, 0) : \forall y, t\}$ contains all partially identified v -statistics. Each term in $\mathbf{v}_0$ is bounded in the interval $v_y(t, 0) \in [0, p(T = t, D = 0)]$ because the potential outcome is unobserved when $D = 0$. More generally, we let $v_y(t, 0) \in [\underline{v}_y(t, 0), \bar{v}_y(t, 0)]$ be bounds on partially identified v -statistics. We let $\mathbf{v}_0^*$ be the true value of partially identified terms under the target distribution and let

$$\mathcal{V}(\rho) = \left\{ \mathbf{v}_0 : \sum_y v_y(t, 0) = p(T = t, D = 0), \forall t \right\}$$

be an uncertainty set of feasible values consistent with the observed data distribution. We can bound policy performance measures over this uncertainty set via

$$\begin{aligned} \underline{m}(\pi_0; \mathcal{V}) &= \min_{\mathbf{v}_0 \in \mathcal{V}(\rho)} m(\mathbf{v}_0, \mathbf{v}_1; \pi_0), \\ \bar{m}(\pi_0; \mathcal{V}) &= \max_{\mathbf{v}_0 \in \mathcal{V}(\rho)} m(\mathbf{v}_0, \mathbf{v}_1; \pi_0), \end{aligned}$$

⁶While our definition of $m_y^*(\pi)$ assumes that $T^\pi = 1$ (i.e., the TPR and FPR) the TNR ($y = 0$) and FNR ($y = 1$) can be recovered by taking $1 - R^*(\pi, \pi_0; m_y)$.

⁷We model utilities of binary classification outcomes rather than the vector-valued potential outcome $Y(a) \in \{0, 1\}^{|A|}$ studied in the standard off policy evaluation setup.

where $m(\mathbf{v}_0^*, \mathbf{v}_1; \pi_0)$ is a v -statistic decomposition of $m^*(\pi_0)$. In the following section, we show that this reduction to a set of common sufficient statistics enables uncertainty cancellation when comparing across policies.

5.4 Regret Bound Identification

We now introduce our novel approach for comparing policies under confounding. Our approach yields informative regret intervals by eliminating redundant uncertainty that does not contribute to differential policy performance. We first focus on asymptotic performance bounds, and provide finite-sample analyses in Section 5.6.

A baseline approach for bounding regret involves partially identifying the performance of each policy individually, and then taking a difference across policy-specific bounds. We refer to this as the baseline regret interval.

Definition 5.4.1 (Baseline regret interval). The baseline regret interval over $\mathcal{V}(p)$ has lower and upper endpoints:

$$\begin{aligned}\underline{R}(\pi, \pi_0; m, \mathcal{V}) &= \underline{m}(\pi; \mathcal{V}) - \overline{m}(\pi_0; \mathcal{V}), \\ \overline{R}(\pi, \pi_0; m, \mathcal{V}) &= \overline{m}(\pi; \mathcal{V}) - \underline{m}(\pi_0; \mathcal{V}).\end{aligned}$$

In contrast, our proposed approach directly bounds the difference in the oracle regret $\delta_m(\mathbf{v}_0^*, \mathbf{v}_1) = m(\mathbf{v}_0^*, \mathbf{v}_1; \pi) - m(\mathbf{v}_0^*, \mathbf{v}_1; \pi_0)$.

Definition 5.4.2 (δ -regret interval). The δ -regret interval over $\mathcal{V}(p)$ has lower and upper endpoints:

$$\begin{aligned}\underline{R}_\delta(\pi, \pi_0; m, \mathcal{V}) &= \min_{\mathbf{v}_0 \in \mathcal{V}(p)} \delta_m(\mathbf{v}_0, \mathbf{v}_1), \\ \overline{R}_\delta(\pi, \pi_0; m, \mathcal{V}) &= \max_{\mathbf{v}_0 \in \mathcal{V}(p)} \delta_m(\mathbf{v}_0, \mathbf{v}_1).\end{aligned}$$

The δ -regret interval yields tighter regret bounds by eliminating redundant uncertainty irrelevant to policy comparison. For example, consider the accuracy regret decomposition

$$\delta_u(\mathbf{v}_0^*, \mathbf{v}_1) = \sum_y v_y(y, 1-y) - v_y(1-y, y).$$

Because $v_0(0, 0) \in \mathbf{v}_0$ cancels when taking a difference across policies, this term does not contribute uncertainty to the δ -regret interval, as it would with the baseline interval. We now provide a theoretical result characterizing the improvement offered by the δ -interval more generally. We let $l_\delta(m, \mathcal{V}) = \overline{R}_\delta(\pi, \pi_0; m, \mathcal{V}) - \underline{R}_\delta(\pi, \pi_0; m, \mathcal{V})$ and $l(m, \mathcal{V}) = \overline{R}(\pi, \pi_0; m, \mathcal{V}) - \underline{R}(\pi, \pi_0; m, \mathcal{V})$ be the length of the δ -regret and baseline regret intervals, respectively.

Theorem 5.4.3 (Regret separation). Let $\Delta(m, \mathcal{V}) = l(m, \mathcal{V}) - l_\delta(m, \mathcal{V})$. Then the δ -regret interval offers the following improvement over the baseline regret interval

$$\begin{aligned}\Delta(m_y, \mathcal{V}) &\geq \frac{2\alpha \cdot v_y(1, 1)}{(\bar{\gamma}_y)^2}, \quad \Delta(m_{a=1}, \mathcal{V}) = 0, \\ \Delta(m_u, \mathcal{V}) &\geq 2\alpha \cdot (u_{00} + u_{01}), \\ \Delta(m_{a=0}, \mathcal{V}) &\geq \frac{2\alpha}{\max\{\psi_0(\pi), \psi_0(\pi_0)\}},\end{aligned}$$

where $\alpha = \bar{v}_0(0, 0) - \underline{v}_0(0, 0)$, $\psi_0(\pi) = p(A^\pi = 0)$, and $\bar{\gamma}_y = \sum_a \sum_{a'} \bar{v}_y(a, a')$.

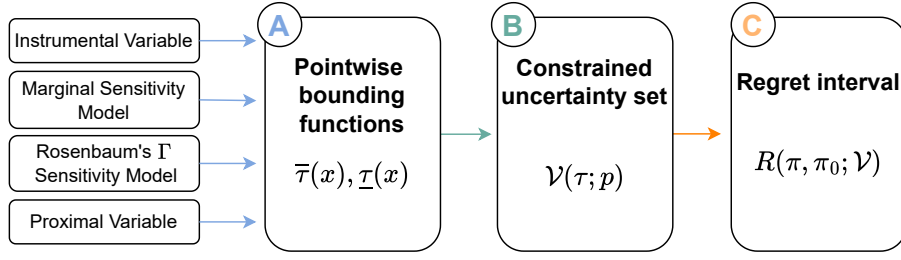


Figure 5.2: Flow of assumptions in our framework. (A) Traditional causal assumptions imply *pointwise bounding functions* on the unobserved outcome; (B) Pointwise bounding functions imply constrained uncertainty sets (Lemma 5.5.1); (C) Constrained uncertainty sets imply policy regret bounds (Section 5.4).

Observe that the improvement offered by the δ -regret interval is proportional to the magnitude of uncertainty in the cancellation term $\alpha = \bar{v}_0(0, 0) - \underline{v}_0(0, 0)$. Therefore, our approach offers the largest benefit under a high degree of uncertainty around partially identified v -statistics. Because regret measures inherit a monotonic dependence on v -statistics, we can compute closed form analytic bounds by maximizing and minimizing regret measures over uncertainty intervals around partially identified terms $[\underline{v}_y(t, 0), \bar{v}_y(t, 0)]$.

5.5 Mapping Causal Assumptions to Informative Regret Bounds

The regret interval we recover without imposing additional causal assumptions on the data generating policy tends to be uninformative. Therefore, we now introduce a *pointwise bounding functions* assumption which we use to tighten regret intervals by shrinking the uncertainty set around partially-identified v -statistics. We show that this assumption is implied by a range of causal assumptions studied in prior OPE literature (e.g., MSM, Rosenbaum's Γ , IV). Thus, our δ -regret interval yields tighter uncertainty quantification than the baseline interval for any causal assumption implying pointwise bounding functions (Figure 5.2).

Assumption 7 (Pointwise bounding functions). Let $\underline{\tau}, \bar{\tau} : \mathcal{X} \rightarrow [0, 1]$ be a pair of bounding functions satisfying

$$\underline{\tau}(x) \leq \mathbb{E}[Y(1) \mid D = 0, X = x] \leq \bar{\tau}(x), \quad \forall x \in \mathcal{X}.$$

Intuitively, this assumption requires that each individual screened out under the status quo policy has risk which is bounded by $[\underline{\tau}(x), \bar{\tau}(x)]$. [Rambachan et al. \(2022\)](#) show that a variant of this assumption is implied by the IV framework, the MSM, and Rosenbaum's Γ sensitivity model.⁸ Next, we show that bounding functions can be used to construct intervals on partially identified v -statistics.

Lemma 5.5.1 (Assumption mapping). Let $\underline{\tau}, \bar{\tau} : \mathcal{X} \rightarrow [0, 1]$ be a pair of bounding functions satisfying Assumption 7. Then partially identified v -statistics are bounded by

$$\mathbb{E}[\underline{\tau}(x) \cdot e_0(x) \cdot \pi_t(x)] \leq v_1(t, 0) \leq \mathbb{E}[\bar{\tau}(x) \cdot e_0(x) \cdot \pi_t(x)],$$

where $e_d(x) = p(D = d \mid X = x)$ and $\pi_t(x) = p(T = t \mid X = x)$.

⁸Our pointwise bounding function assumption is also related to the notion of “bounded outcomes” studied by [Manski \(1990\)](#).

Proof. For the upper bound $\bar{\tau}(x)$ we have that $\forall x \in X$

$$\begin{aligned} \mathbb{E}[Y(1) \mid D = 0, X = x] &\leq \bar{\tau}(x) \\ \mathbb{E}[Y(1) \mid D = 0, T = t, X = x] &\leq \bar{\tau}(x) \\ \mathbb{E}[Y(1) \mid D = 0, T = t, X = x] \cdot p(D = 0, T = t, X = x) &\leq p(D = 0, T = t, X = x) \cdot \bar{\tau}(x) \end{aligned} \quad (5.1)$$

where (5.1) holds because $T \perp\!\!\!\perp \{D, Y(1)\} \mid X$. Simplifying the right hand side yields

$$\mathbb{E}[Y(1) \mid D = 0, T = t, X = x] \cdot p(D = 0, T = t, X = x) \quad (5.2)$$

$$\leq p(D = 0 \mid T = t, X = x) \cdot p(T = t \mid X = x) \cdot p(X = x) \cdot \bar{\tau}(x) \quad (5.3)$$

$$\leq p(D = 0 \mid X = x) \cdot p(T = t \mid X = x) \cdot p(X = x) \cdot \bar{\tau}(x) \quad (5.4)$$

The result follows from marginalizing over X

$$v_1(t, 0) \leq \sum_x e_0(x) \cdot \pi_t(x) \cdot \bar{\tau}(x) \cdot p(X = x) \quad (5.5)$$

$$= \mathbb{E}[\bar{\tau}(x) \cdot e_0(x) \cdot \pi_t(x)] \quad (5.6)$$

The upper bound is sharp because equality holds when $\bar{\tau}(x) = \mathbb{E}[Y(1) \mid D = 0, X = x] \forall x \in X$. The lower bound on $v_t(t, 0)$ follows by the same argument. $\square$

Lemma 5.5.1 provides a convenient approach for bounding $\mathbf{v}_0^*$ because, given a pair of bounding functions, it only requires knowledge of *nominal* (i.e., confounded) status quo policy probabilities $e_d(x)$ and the new policy $\pi_t(x)$. Nominal probabilities and bounding functions can be learned from observational data, while $\pi_t(x)$ is known in advance by model developers. Our goal is to tighten regret intervals by using bounding functions to shrink the uncertainty set around $\mathbf{v}_0^*$. Let $\mathcal{H}(v_1(t, 0); \tau) = [\underline{v}_y(t, 0), \bar{v}_1(t, 0)]$ be the interval implied by Lemma 5.5.1 and let $\rho_{td} = p(T = t, D = d)$. We define the constrained uncertainty set over $\mathbf{v}_0^*$ as

$$\mathcal{V}(p; \tau) = \left\{ v_y(t, 0): \begin{array}{l} v_1(0, 0) \in \mathcal{H}(v_1(0, 0); \tau) \\ v_0(0, 0) = \rho_{00} - v_1(0, 0) \\ v_1(1, 0) \in \mathcal{H}(v_1(1, 0); \tau) \\ v_0(1, 0) = \rho_{10} - v_1(1, 0) \end{array} \right\}.$$

We only use intervals $\mathcal{H}(v_1(t, 0); \tau)$ for this set definition because an analogous bound is implied over $v_0(t, 0)$ by the constraint requiring that $\sum_y v_y(t, 0) = \rho_{t0}$. We quantify the size of $\mathcal{V}(p; \tau)$ via the Lebesgue measure, which is defined as the cartesian product over partially identified intervals $\lambda(\mathcal{V}(p; \tau)) = |\bar{v}_1(1, 0) - \underline{v}_1(1, 0)| |\bar{v}_1(0, 0) - \underline{v}_1(0, 0)|$. We now show that $\mathcal{V}(p; \tau)$ is the smallest uncertainty set one could construct based on the observational distribution and a pair of bounding functions.

Theorem 5.5.2 (Minimality). $\mathcal{V}(p; \tau)$ is the minimal uncertainty set—that is, no strict subset of $\mathcal{V}(p; \tau)$ is consistent with $p(X, D, T, Y)$ and user-specified bounding functions.

5.6 Regret Bound Estimation

In this section, we construct finite sample estimates of regret intervals using observational data $\mathcal{O} = \{(X_i, D_i^{\pi_0}, T_i^{\pi}, Y_i)\}_{i=1}^n \sim p$ collected under the status quo policy. We estimate bounds by applying closed-form expressions for the δ -regret interval over an estimate of the uncertainty set $\hat{\mathcal{V}}(p; \hat{\tau})$. We can directly compute finite-sample estimates $\hat{\mathbf{v}}_1$ for identified v -statistics by taking a sample average. As a result, the technical challenge underpinning regret bound estimation boils down to estimating the intervals $\mathcal{H}(\hat{v}_y(t, 0); \hat{\tau}) = [\hat{\underline{v}}_y(t, 0), \hat{\bar{v}}_y(t, 0)]$ around $\mathbf{v}_0^*$.

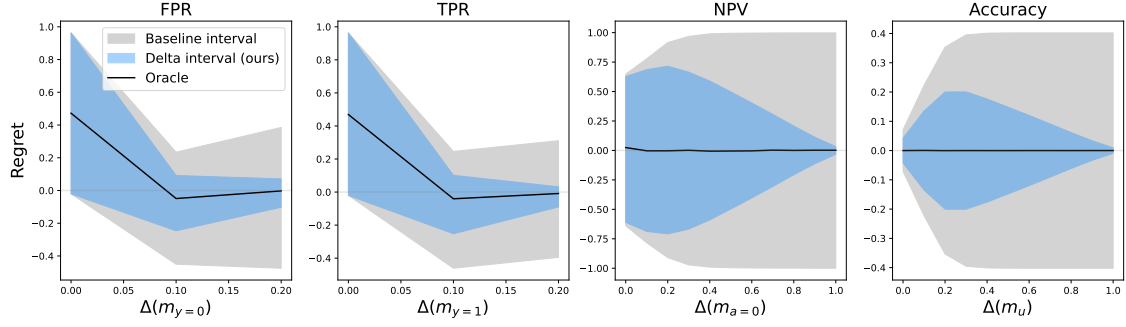


Figure 5.3: Improvement in bounds offered by the δ -regret interval over the baseline interval. We systematically vary the relative size of v -statistics and plot bounds as a function of interval improvement $\Delta(m) = I(m) - I_\delta(m)$ characterized by Theorem 5.4.3.

5.6.1 Plug-in Estimator

A direct approach for bound estimation involves learning $\hat{\eta} = (\hat{e}, \hat{\tau})$, then estimating $\mathcal{H}(\hat{v}_1(t, 0); \hat{\tau})$ over a held-out sample by invoking Lemma 5.5.1.⁹¹⁰ We assume that $\pi(x)$ is known by model developers and thus does not need to be learned. We split the data into K disjoint folds, where we denote $\mathcal{O}_k, \mathcal{O}_{-k}$, as the sample inside and outside of fold k , respectively. We then define the plug-in estimator over $\mathcal{O}_k$ as

$$\hat{v} = \frac{1}{|\mathcal{O}_k|} \sum_{x_i \in \mathcal{O}_k} \pi(x_i) \cdot \hat{e}(x_i) \cdot \hat{\tau}(x_i) \quad (5.7)$$

where $\hat{\eta}$ is learned over fold $\mathcal{O}_{-k}$.¹¹ We recover regret estimates at full data efficiency via a cross-fitting approach. The following result shows consistency of $\hat{v}$ and characterizes its convergence rate to the true upper bound $\bar{v}$, where we define $\|f\|^2 := \int (f(x))^2 dP(x)$ to be the squared $L_2(O)$ norm of a function f .

Theorem 5.6.1. *Let $f(x) = e(x) \cdot \bar{\tau}(x)$ and assume $\|\hat{f} - f\| = o_p(1)$. Then the plug-in estimator satisfies*

$$\hat{v} - \bar{v} = O_P(\|e(x) - \hat{e}(x)\| + \|\bar{\tau}(x) - \hat{\tau}(x)\|) + O_P\left(\frac{1}{\sqrt{n}}\right).$$

Because policy performance measures are linear combinations (m_u^*) or ratios (m_a^*, m_y^*) of v -statistics, this result also characterizes the convergence rate of the estimated upper and lower regret bounds. Observe that the convergence rate of the plug-in estimator inherits the rate of the nuisance functions. As a result, $\hat{v}$ will tend to converge slowly when using flexible machine learning methods to fit nuisance functions under no parametric assumptions.

5.6.2 Doubly Robust Estimator

Therefore, we propose a doubly robust approach for estimating $\hat{v}$ that corrects the bias in the plug-in estimator. Because doubly robust estimators have an error term which is a product of

⁹In line with prior literature on CATE estimation (Kennedy, 2023), we call this a plug-in approach because it involves directly substituting in empirical estimates for the nuisance functions.

¹⁰The term “nuisance function” refers to a learned function of the data which differs from our target quantity of interest.

¹¹We suppress dependence on y, t, d and present analysis of the upper bound to ease notation. An analogous argument holds for $\hat{v}$.

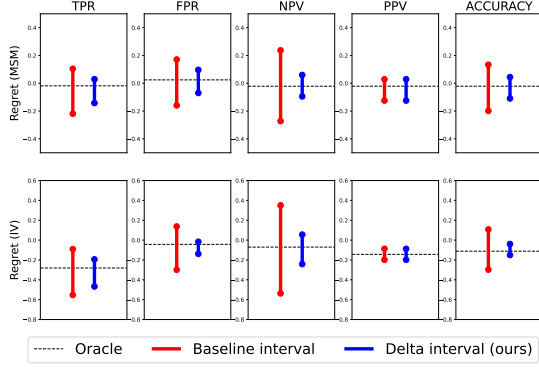


Figure 5.4: Comparison between δ -regret and baseline regret interval end-points averaged over $N = 20$ trials and $N_s = 20,000$ samples. First row leverages an MSM identification assumption, while the second leverages an IV assumption.

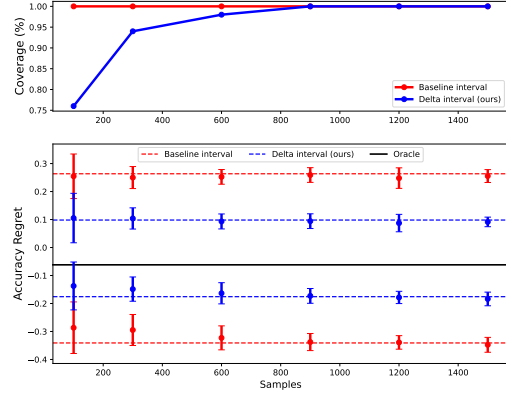


Figure 5.5: Top: Coverage of accuracy regret interval estimates as a function of total sample size. Bottom: 95% bootstrap confidence intervals around upper and lower regret bounds over $N = 25$ trials. Solid line indicates the oracle regret.

nuisance function errors, they attain fast convergence rates, even when estimating nuisance functions at slow non-parametric rates (Kennedy et al., 2023). For example, to attain $n^{-1/2}$ rates for the estimator, it is sufficient to estimate nuisance functions at $n^{-1/4}$, a rate achieved by many non-parametric machine learning techniques. The form of the doubly robust estimator depends on the pointwise bounding functions implied by a causal assumption.

We provide the doubly robust estimator for the commonly studied MSM. The MSM assumes there is some $\Lambda \geq 1$ such that the odds ratio $\pi_0(X, U)/(1 - \pi_0(X, U)) \cdot (1 - e(X))/e(X)$ lies within $[\Lambda^{-1}, \Lambda]$. We define the doubly robust estimator over $\mathcal{O}_k$ as

$$\hat{v}_{DR} = \frac{1}{|\mathcal{O}_k|} \sum_{x_i \in \mathcal{O}_k} \phi(O_i; \hat{e}, \pi, \hat{\mu}_1) \quad (5.8)$$

where $\phi(O; e, \pi, \mu_1) = D \cdot Y \pi(X) \cdot \Lambda + (T - \pi(X)) \cdot e(X) \cdot \mu_1(X) \cdot \Lambda$ and $\mu_1(x) = \mathbb{E}[Y(1) | D = 1, X = x]$.

5.7 Numerical Experiments

We now show numerical experiments comparing decision policies under a known data-generating mechanism.

Regret interval characterization. We first characterize the uncertainty reduction offered by the δ -regret interval. We simulate a range of v -statistic decompositions by randomly sampling $\mathbf{v}$ satisfying $\sum_{y,t,d} v_y(t, d) = 1$ and $v_0(t, d) + v_1(t, d) = \rho_{td}, \forall t, d$. This yields a collection of decompositions corresponding to valid observational distributions. We compute the analytic improvement in bounds defined in Theorem 5.4.3 for each $\mathbf{v}$ and plot the results in Figure 5.3. We observe a monotone improvement in bounds as a function of $\Delta(m, \mathcal{V})$, with no improvement when $\Delta(m, \mathcal{V}) = 0$ and significant improvement when $\Delta(m, \mathcal{V}) = 0$ is large.

Synthetic data experiment. We next validate our framework by simulating data consistent with two common assumptions for off-policy evaluation: the MSM and IV. We draw $\mathbf{v}$ covariates

$X_i \sim \mathcal{N}(0, I_v)$, and u confounders $U_i \sim \mathcal{N}(0, I_u)$, and let $V_i = (X_i, U_i)$. We parameterize policy, outcome, and instrument probability functions with coefficients $W_{\pi_0} \in \mathbb{R}^{v+u}$, $W_{\pi} \in \mathbb{R}^v$, $W_{\mu_1} \in \mathbb{R}^{v+u}$, $W_z \in \mathbb{R}^{v \times z}$, respectively, with each drawn from a uniform distribution. We sample data from the probability functions

$$\begin{aligned} \gamma(X_i) &:= \sigma'(X_i \times W_z), \pi(X_i) := \sigma(X_i \times W_{\pi}) \\ \pi_0(V_i, Z_i) &:= \sigma(V_i \times W_{\pi_0} + \beta_0 \cdot Z_i) \\ \mu_1(V_i, Z_i) &:= \sigma(V_i \times W_{\mu_1} + \beta_1 \cdot Z_i) \\ \mu_0(V_i, Z_i) &:= \begin{cases} \Lambda^* \cdot \mu_1(V_i), \Lambda^* \in U(\Lambda^{-1}, \Lambda) & \text{(MSM)} \\ \sigma(V_i \times W_{\mu_0} + \beta_1 \cdot Z_i) & \text{(IV)} \end{cases} \\ \mu(V_i, Z_i) &:= \mu_1(V_i, Z_i)\pi_1(V_i, Z_i) + \mu_0(V_i, Z_i)\pi_0(V_i, Z_i) \end{aligned}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ and $\sigma'(x)$ is the softmax function. We use the plug-in estimator to learn estimates of the δ -regret interval. We compare against the baseline regret interval by applying bounds over the same $\hat{\mathcal{V}}(p, \hat{\tau})$ used to estimate the δ -regret.

Figure 5.4 provides a comparison of the δ -regret interval and baseline regret interval across five policy performance measures. In line with our theoretical results, we obtain tighter bounds from the δ -regret across all policy performance measures apart from the positive predictive value (PPV). We show the oracle regret evaluated via $Y(1)$ in black. In Figure 5.5, we plot coverage of the estimated δ -regret interval evaluated against the true population value, denoted by dashed lines. The bottom panel shows 95% bootstrap confidence intervals around the true upper and lower δ -regret and baseline regret intervals. These results show that estimates concentrate around the true regret interval as the number of samples increases. We observe that the baseline interval yields better coverage of the oracle regret in small sample settings because it is more conservative and thus more tolerant to bias around estimates of the asymptotic interval end points.

5.8 Real-World Application: Comparing Healthcare Enrollment Policies

We now illustrate how our framework can be used to compare alternative healthcare enrollment policies. In medical settings, providers routinely screen patients for diseases and enroll high-risk individuals in preventative care programs. However, it is challenging to assess the performance improvement of a proposed policy because outcomes are only observed among patients enrolled under the existing policy (Daysal et al., 2022). Confounding is a challenge in this setting because physicians often make decisions using unobserved information (Mullainathan and Obermeyer, 2022).

We leverage data released by Obermeyer et al. (2019) to construct an enrollment policy comparison task. This dataset contains $\approx 48,000$ records, where each entry consists of a patient evaluated for enrollment in a high-risk care management program.¹² We let π_0 be the historical enrollment policy consisting of physician decisions informed by an algorithmic risk score and interactions with the patient. We take π to be the algorithmic policy which makes decisions by thresholding predictions of patient cost from the clinical decision support tool. The goal in this task is to assess whether the algorithm-only policy would improve upon the status quo human+algorithm policy used to collect data.

¹²We use a synthetic version of the original dataset, which was released by Obermeyer et al. (2019) to protect patient confidentiality. This dataset preserves the means and covariances of the original data.

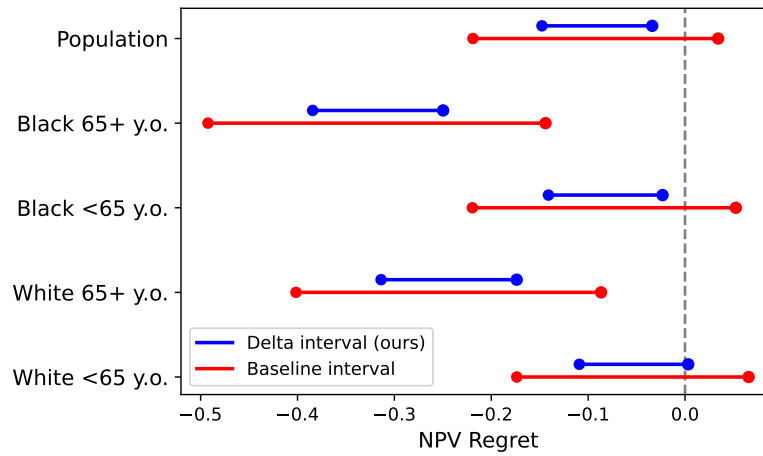


Figure 5.6: Comparison of the δ -regret interval against the baseline for the NPV ($m_{a=0}$). The top row indicates full population bounds, while lower rows subpopulation bounds.

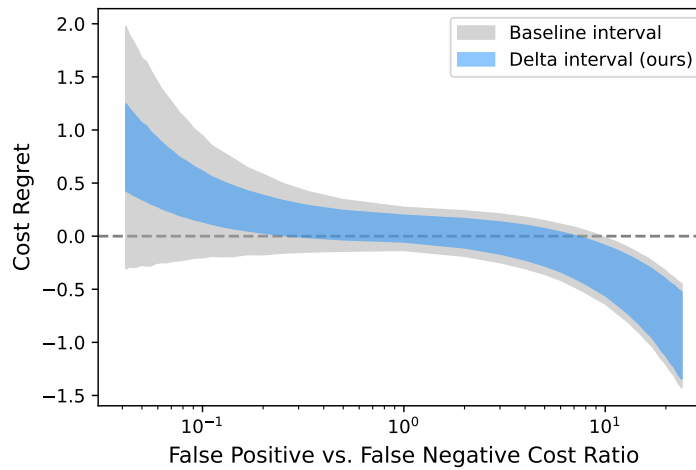


Figure 5.7: Expected cost regret as a function of false positive to false negative cost ratio. Left hand side (10^{-1}) indicates false negative cost ten times greater than false positive cost.

Results. Because we do not have access to physician identifiers which can be used as an instrument, we leverage the MSM assumption for identification. We compare the baseline and δ -regret intervals for the NPV in Figure 5.6. The top row shows the population regret, while the bottom rows show a breakdown across subgroups.¹³ The δ -regret interval excludes zero across all subpopulations apart from White patients under 65. As a result, an analysis conducted via the δ -regret interval supports an interpretation that the algorithmic policy reduces the NPV in comparison to the status quo human+algorithm policy.¹⁴ **Because the baseline regret interval contains zero among all patient populations except those aged 65 and over,**

¹³Intervals include 95% confidence intervals estimated around regret bound endpoints. We omit these from the plot for readability.

¹⁴Given the narrow scope of our analyses and limitations of synthetic data, our findings are not intended to be a conclusive assessment of the policies evaluated by Obermeyer et al. (2019).

this bounding approach supports weaker claims regarding the relative performance of decision policies.

In Figure 5.7, we plot the δ -regret and baseline intervals as a function of the ratio between false positive and false negative costs. The left hand region of this figure ($\approx 10^{-1}$) reflects a setting in which false negatives are ten times more costly than false positives. This regime is realistic in our healthcare enrollment scenario when turning a sick patient away from the program incurs more harm than enrolling a healthy patient. We observe the greatest improvement from our approach in the high cost false negative setting because the $v_0(0, 0)$ term which cancels under our approach is heavily weighted in the expected cost calculation. Because the δ -regret interval excludes zero in this regime, this supports an interpretation that the proposed policy has a higher expected cost than the status quo.

5.9 Conclusion

In this work, we propose the first framework supporting predictive performance comparisons of decision-making policies under confounding. Our approach is intended to support pre-deployment evaluations of proposed policies under a flexible set of causal assumptions. Our approach addresses sources of confounding-related uncertainty which impact model evaluations and, where possible, reduces this uncertainty via technically novel partially identification approaches. Our uncertainty cancellation approach may prove useful for more tightly characterizing performance differences under other uncertainty sources, such as missing protected attributes (Kallus et al., 2022) and measurement error (Fogliato et al., 2020).

Part 2

Measurement Interventions for Generative AI (GenAI) Systems

Chapter 6

A Statistical Framework for Improving the External Validity of Generative AI Evaluations

This chapter is drawn from the paper:

Luke Guerdan*, Justin Whitehouse*, Kimberly Truong*, Kenneth Holstein, and Zhiwei Steven Wu. Doubly-Robust LLM-as-a-Judge: Externally Valid Estimation with Imperfect Personas. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. (*: co-first authors)

This chapter is the result of a highly collaborative project between three co-first authors, without whom this chapter would not be possible.

- **Luke Guerdan (LG)**: Led conceptual framing of the project. Designed the problem formulation. Supported implementation of the Persona Simulation Framework (e.g., data generation pipeline, Riesz loss). Wrote the main text with support from JW and KT.
- **Kimberly Truong (KT)**: Led all work on experiments. Designed and implemented the Persona Simulation Framework, collected synthetic datasets from LLMs, and implemented estimators included in the experimental results. Supported drafting of the main text.
- **Justin Whitehouse (JW)**: Led all work on theoretical development and validation of the doubly-robust estimator. Stated and proved all results. Drafted appendices B, C, and D. Conducted early-stage experiments on synthetic data. Supported drafting of the main text.

This chapter presents the main estimators and empirical findings; for brevity, the complete formal setup and proofs are deferred to the published manuscript [Guerdan* et al. \(2026\)](#).

6.1 Motivation

As Generative AI (GenAI) systems see growing adoption, a key concern involves the *external validity* of evaluations, or the extent to which they generalize from lab-based to real-world deployment conditions (Ibrahim et al., 2024; Ouyang et al., 2023; Liao and Xiao, 2023; Liao et al., 2021; Weidinger et al., 2025). In particular, many evaluations report a *system quality estimate*, which reflects the proportion of outputs rated to exhibit a capability (e.g., “helpfulness”) or defect (e.g., “toxicity”) by a human with specialized characteristics (e.g., domain knowledge, cultural experience). However, such quality estimates may fail to generalize when the *source* distribution of human raters and system outputs available at evaluation time differs from the *target* distribution encountered upon deployment. For example, in conversational safety evaluation, covariate shift can arise if we wish to estimate the perceived “toxicity” of a response for a specific demographic group, but collect ratings from a general population of crowdworkers. Selection bias may also occur if crowdworkers frequently opt out from providing ratings on highly sensitive content (Steiger et al., 2021). Left unaddressed, these forms of *evaluation sampling bias* threaten the external validity of system quality estimates.

Recent work has proposed tools for improving system quality estimates when human ratings are scarce but black-box predictions (e.g., from an LLM-as-a-judge) are cheap and abundant (Chatzi et al., 2024; Fisch et al., 2024; Eyre and Madras, 2024; Dörner et al., 2024; Saad-Falcon et al., 2023; Fogliato et al., 2024). For instance, Prediction Powered Inference (PPI) offers an approach for leveraging a subset of labeled (source) data to correct for bias in black-box model predictions (Angelopoulos, Bates, Fannjiang, Jordan and Zrnic, 2023; Angelopoulos, Duchi and Zrnic, 2023). This bias correction enables using black-box predictions generated over unlabeled (target) samples to shrink confidence intervals around quality estimates while maintaining valid coverage. However, both PPI and its extensions (e.g., PPI++ (Angelopoulos, Duchi and Zrnic, 2023), RePPI (Ji et al., 2025)) assume that (i) source and target samples are drawn i.i.d. and (ii) labels are *missing completely at random* (MCAR) (Tsiatis, 2006), i.e. that successful completion of a rating is independent of rater and text characteristics. However, when source data is observed under sampling bias, these assumptions are violated and severe miscoverage occurs (see Fig. 6.1).

In this work, we devise an estimator that directly corrects for evaluation sampling bias. Like existing approaches (Angelopoulos, Duchi and Zrnic, 2023; Ji et al., 2025), our proposal leverages black-box predictions generated by a GenAI system over unlabeled (target) samples to improve statistical inference. Unlike existing estimators, however, our proposal is *doubly-robust* (Bang and Robins, 2005; Chernozhukov et al., 2018): it yields valid system quality estimates if *either* a model trained to predict human ratings from labeled source data *or* a reweighting model that corrects for sampling bias is of sufficient quality. To attain this doubly-robust property, we leverage *persona ratings* — scores generated by prompting an LLM-as-a-judge to behave as a human rater with desired characteristics (e.g., demographics, expertise). Our framework uses these persona ratings as an auxiliary feature when training a model to predict human ratings from labeled source data. When persona ratings are sufficiently correlated with human ratings, this yields a higher-quality predictor, which in turn makes the double-robustness condition of our estimator easier to satisfy (see § 6.4).

Figure 6.1 illustrates the benefits of our estimation approach on three datasets. Whereas persona ratings alone (Persona-Based) or human ratings alone (Source Mean) fail to provide valid coverage, our approach combines imperfect persona ratings with imperfect human ratings to attain valid confidence intervals with low statistical bias. We similarly find that our approach out-performs state-of-the-art baselines (Angelopoulos, Duchi and Zrnic, 2023; Ji et al., 2025), which also combine persona ratings with human ratings, but fail to account for evaluation

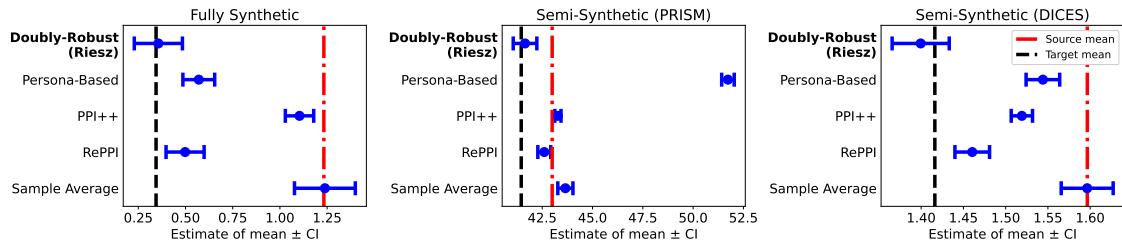


Figure 6.1: Comparison of our doubly-robust estimator with baselines on three datasets from our Persona Simulation Framework. Red and black dashed lines denote the true source and target mean ratings, respectively (e.g., the average “helpfulness” rating obtained over source vs. target distributions). *Persona-Based* directly averages persona ratings to compute a system quality estimate. *Sample Average* produces a system quality estimate by averaging human ratings sampled from the source distribution. PPI++ (Angelopoulos, Duchi and Zrnic, 2023) and RePPI (Ji et al., 2025) are two state-of-the-art statistical methods that do not account for evaluation sampling bias. Across settings, we observe that our Doubly-Robust (Riesz) estimator yields improved coverage and lower bias than baselines, while maintaining informative confidence intervals.

sampling bias. To summarize, our main contributions are as follows:

- **We formalize the problem** of GenAI system quality estimation under evaluation sampling bias. Unlike existing statistical frameworks (Angelopoulos, Duchi and Zrnic, 2023; Ji et al., 2025), our formulation explicitly accounts for both covariate shift and selection bias in available human ratings to improve the external validity of system quality estimates.
- **We devise a doubly-robust estimator** for GenAI system quality estimation under evaluation sampling bias. En route, we first advance doubly-robust estimation theory by generalizing the work of (Chernozhukov et al., 2023) to M-estimation settings with surrogate (persona) ratings. This generalization enables us to (i) leverage persona ratings to improve doubly-robust system quality estimates, (ii) estimate a richer set of system quality parameters (e.g., rating variance, quantiles) beyond means, and (iii) maintain valid coverage even in the presence of evaluation sampling bias, all desiderata not satisfied by previous works.
- **We advance the practical application of doubly-robust estimators** to GenAI system quality estimation. Whereas doubly-robust estimators are traditionally applied on small tabular datasets, GenAI system quality estimation requires learning a reweighting function over high dimensional (e.g., text, audio) input-output spaces. We show that sentence transformer embedding models and a “Riesz loss” approach (Chernozhukov, Newey and Singh, 2022) can be combined to correct for covariate shift in high-dimensional text input-output spaces.
- **We introduce a Persona Simulation Framework (PSF)** that systematically manipulates evaluation sampling bias and persona quality over three experimental settings with increasing realism: Fully Synthetic, Semi-Synthetic PRISM (Kirk et al., 2024), and Semi-Synthetic DICES (Aroyo et al., 2023). Leveraging the PSF, we show our estimator obtains valid coverage up to a larger magnitude of sampling bias than state-of-the-art baselines (e.g., RePPI (Ji et al., 2025)). We publicly release our PSF code as a community resource.¹

¹Code and data are available on [GitHub](#).

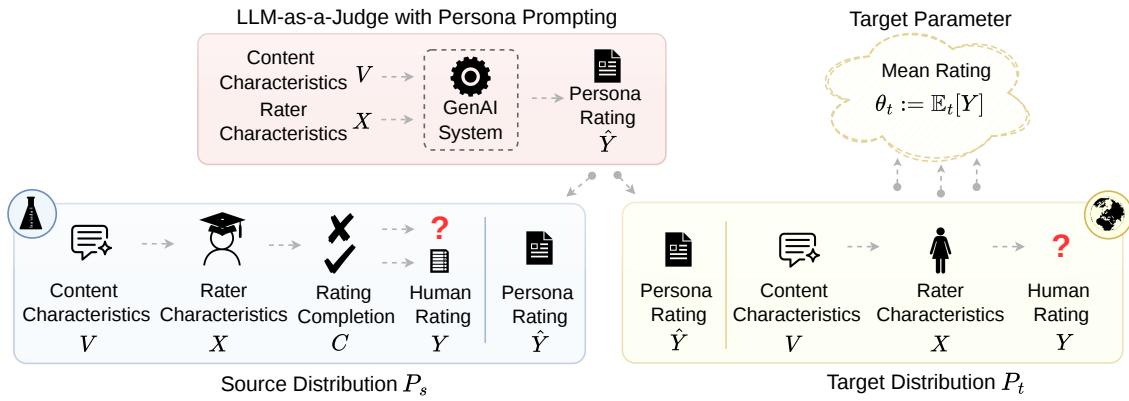


Figure 6.2: Our framework produces estimates for the target parameter θ_t using (i) complete rating tuples from the source distribution (blue, left), (ii) unlabeled samples from the target distribution (yellow, right), and (iii) persona ratings produced for both source and target samples (red, top). *Evaluation sampling bias* may arise both from the *covariate shift* of (V, X) from P_s to P_t , and from *selection bias* in which rating completion C is non-random in P_s – i.e., $C \not\perp (V, X)$.

6.2 Related Work

Automated Evaluation with Persona Prompting. To address evaluation sampling bias, one strategy is to use an automated rater to rate outputs from the target distribution. Under this LLM-as-a-judge approach, a *judge* GenAI system rates the outputs of a *target* GenAI system (Li, Jiang, Huang, Beigi, Zhao, Tan, Bhattacharjee, Jiang, Chen, Wu et al., 2024; Elangovan et al., 2024; Ye et al., 2024; Bubeck et al., 2023; Zheng et al., 2023). Because human raters often disagree on criteria such as “helpfulness” or “relevance” (Kirk et al., 2024), prior work has explored instructing judge systems to adopt *personas* — descriptions of humans with specific characteristics (Castricato et al., 2024; Fröhling et al., 2024; Orlikowski et al., 2025; Deng et al., 2025). However, work has also shown that persona ratings are often an imperfect proxy for human ratings (Santurkar et al., 2023; Neumann et al., 2025). Thus, our work treats persona ratings as a *useful yet incomplete proxy* for human raters to improve GenAI system quality estimates.

Frameworks for Sample Efficient Estimation. Other works propose methods for improving statistical inference when data is scarce but ML predictions are abundant. Prediction-Powered Inference (PPI) and its computationally efficient variant PPI++ use ML predictions to tighten confidence intervals through a “rectifier term” that corrects for bias in ML predictions (Angelopoulos, Bates, Fannjiang, Jordan and Zrnic, 2023; Chatzi et al., 2024; Fisch et al., 2024; Angelopoulos, Duchi and Zrnic, 2023). Ji et al. (2025) show PPI++ to be a special case of M-estimation with surrogate outcomes, a classical problem in causal inference (Robins et al., 1994; Robins and Rotnitzky, 1995; Tsiatis, 2006), and in turn propose recalibrated PPI (or RePPI) to offer more efficient estimation. However, these approaches fail to give valid coverage under evaluation sampling bias. We develop a doubly-robust estimator (Bang and Robins, 2005; Chernozhukov et al., 2018, 2023) that can handle covariate shift and selection bias simultaneously while making use of surrogate predictions/persona ratings. We also use “Riesz losses” (Chernozhukov et al., 2023; Chernozhukov, Newey, Quintas-Martinez and Syrgkanis, 2022; Chernozhukov, Newey and Singh, 2022) to estimate complicated nuisance parameters using generic ML learners.

6.3 Preliminaries

We now provide an overview of the data generating processes and GenAI system quality parameters considered hereinafter.

Probabilistic Framework. As illustrated in Fig. 6.2, we model system quality estimation under evaluation sampling bias via a tuple of random variables $Z = (X, V, C, Y, \hat{Y})$. Here, X denotes *rater characteristics* (e.g., age, gender, geographic locale) and V denotes the *content to be rated*, such as the GenAI system input and output. In experiments, we characterize the content V via an embedding-based projection of the input prompt and system output into a low-dimensional space (see § 7.4). We use $W = (X, V)$ to denote the tuple of rater and content. C is an indicator of rating completion ($C = 1$ if the rater provides a rating, $C = 0$ otherwise). For example, a rater can fail to provide a rating if they (i) are excluded on the basis of failed attention checks, or (ii) abandon the rating task mid-way (i.e., self-attrition). Let Y denote the rating a rater *would* assign if they completed the task ($C = 1$), which may be ordinal (e.g., Likert 1–5), interval (e.g., 1–100), or binary (e.g., Yes/No). Finally, $\hat{Y}$ is the rating returned by an LLM-as-a-judge with persona prompting.

Example. Suppose we want to measure the perceived “safety” of a conversational AI system’s response. X captures rater demographics (e.g., age, race, education) and V includes the user query and system response. Rating completion (C) denotes whether the rater provides a rating or opts out, and Y is the human’s “safety” rating (e.g., on a 1–4 scale). $\hat{Y}$ denotes GPT-5’s predicted rating.

Evaluation Sampling Bias. To model evaluation sampling bias, we assume there are two *full-data* distributions over tuples Z : a source distribution P_s and a target distribution P_t . *Covariate shift* arises when the marginal distribution of rater and content characteristics differs across source and target distributions. For instance, the source distribution may consist of crowdworkers recruited via MTurk—who tend to skew younger and more educated than the general population (Levy et al., 2016)—while the target distribution might reflect the demographic composition of a specific deployment context (e.g., users of a healthcare chatbot, who happen to skew older and female). *Selection bias* arises when rating completion depends on rater and/or content characteristics — i.e., $C \not\perp W$. In our running example, this can arise if crowdworkers are more likely to opt out of rating highly sensitive content. While our framework is explicitly designed to handle selection bias, existing frameworks (Angelopoulos, Duchi and Zrnic, 2023; Ji et al., 2025) assume that data are missing completely at random (MCAR) — i.e., $C \perp W$. We show empirically that violations of this assumption lead to severe miscoverage in system quality estimates (see § 7.4).

While we relax this MCAR assumption, we rely on several additional assumptions (*also* required by existing frameworks). For instance, we assume no *concept drift*, i.e., that $P_s(Y|W) = P_t(Y|W)$. This requires that the rater and content characteristics are sufficiently rich as to describe ratings across both populations. We elaborate on this and other standard causal assumptions required by our framework in the Appendix of Guerdan* et al. (2026). Relaxing these assumptions remains a fruitful direction for future work.

Estimation Goal. Given a sample of N_s *partial* source observations $\mathcal{D}_s = \{(X_j^s, V_j^s, C_j^s, Y_j^s \cdot \hat{Y}_j^s)\}_{j=1}^{N_s}$ and N_t *partial* target observations $\mathcal{D}_t = \{(X_i^t, V_i^t, \hat{Y}_i^t)\}_{i=1}^{N_t}$, our goal is to estimate a parameter summarizing system quality over P_t .^{2 3} As our running example in the main text and experiments, we consider the mean rating, $\theta_t := \mathbb{E}_t[Y]$, which describes the average “safety” or “helpfulness” rating assigned by human raters to system outputs in the target distribution P_t .

²We use superscripts s, t on random variables to denote source and target membership. We omit these superscripts where the distribution is clear from context (e.g., $\mathbb{E}_t[Y]$ clearly refers to the target distribution).

³In the tuple $\mathcal{D}_s$, the shorthand $C_j^s \cdot Y_j^s$ denotes that ratings are only observed when $C_j^s = 1$.

Table 6.1: Examples of statistical parameters that can be estimated via our M-estimation framework. Each parameter summarizes information about human ratings obtained over the target distribution. Conditional parameters (bottom three rows) can be defined conditionally on rater characteristics (X), content characteristics (V), or both, as special cases of conditioning on $W = (X, V)$.

Parameter	Estimand	Example
Mean	$\theta_t := \mathbb{E}_t[Y]$	Mean “helpfulness” rating for customer service chatbot responses.
Variance	$\theta_t := \text{Var}_t(Y)$	Variance (disagreement) in conversational “safety” ratings.
Quantile	$\theta_t := \inf\{y : P_t(Y \leq y) \geq Q\}$	Median “comprehensibility” score for technical documentation.
Conditional Mean	$\theta_t := \mathbb{E}_t[Y \mid g(W) = 1]$	Mean “coherence” rating assigned by <i>domain experts</i> to multi-turn conversational responses.
Conditional Variance	$\theta_t := \text{Var}_t[Y \mid g(W) = 1]$	Variance in “helpfulness” ratings among <i>novice programmers</i> for code suggestions.
Conditional Quantile	$\theta_t := \inf\{y : P_t(Y \leq y \mid g(W) = 1) \geq Q\}$	90th percentile “safety” rating for <i>high-risk queries</i> flagged by content moderators.

6.4 Methodology

We now introduce our doubly-robust estimator for GenAI system quality estimation under evaluation sampling bias. The central challenge addressed by our approach is that our data is imperfect. While persona predictions are available, they may be a poor proxy for human ratings. Likewise, human ratings from the source distribution may suffer from evaluation sampling bias, leading to invalid estimates for the target parameter. We first introduce several naive approaches which might be used to tackle this problem (§ 6.4.1). Then, we show that while each approach is insufficient in isolation, they can be combined to obtain valid coverage (§ 6.4.2). Our results presented in this section apply not only when $\theta_t := \mathbb{E}_t[Y]$ (where they generalize Chernozhukov et al. (2023) to simultaneous covariate shift *and* selection bias) but also when θ_t is the solution to a generic M-estimation problem. Table 6.1 illustrates the range of statistical parameters our framework supports.

6.4.1 Baseline Approaches and Their Limitations

Persona-Augmented Regression. One possible estimation strategy is to train a model to predict human ratings using source data and then use this model to impute missing target labels. Persona-augmented regression leverages this approach while including persona ratings as an *additional auxiliary feature* in the regression function. In particular, we train a model $\hat{\mu}(W, \hat{Y})$ predicting $\mu_0(W) := \mathbb{E}[Y \mid W]$ ⁴ using samples from $\mathcal{D}_s$ and then estimate θ_t as $\hat{\theta}_t^{\text{reg}} := \frac{1}{N_t} \sum_{i=1}^{N_t} \hat{\mu}(W_i^t, \hat{Y}_i^t)$. Observe that this persona-augmented regression estimator relies on the persona rating in addition to covariates. While this approach may be viable when persona ratings are highly correlated with human ratings, in general $\hat{\mu}$ will converge too slowly to construct valid confidence intervals (§ 7.4).

Re-weighting. Another approach is to *re-weight* samples from P_s based on their probability of occurring under P_t . This approach requires correcting for covariate shift and selection bias

⁴No concept drift implies $\mathbb{E}_t[Y \mid W] = \mathbb{E}_s[Y \mid W]$, so we can omit subscripts without any worry.

in parallel, and does not use persona ratings. Formally, let $\omega_0(w) = \frac{dP_t}{dP_s}(w)$ denote the density ratio between $P_t(W)$ and $P_s(W)$, and let $\pi_0(w) = \mathbb{P}_s(C = 1 \mid W = w)$ denote the probability of rating completion. Under standard assumptions, we have $\theta_t = \mathbb{E}_s[\alpha_0(W, C)Y]$, where $\alpha_0(W, C) := C \frac{\omega_0(W)}{\pi_0(W)}$. Thus, if one produces an ML estimate $\hat{\alpha}$ of α_0 (say by training models $\hat{\omega}, \hat{\pi}$ predicting ω_0, π_0), they can compute an inverse propensity weighted (IPW) estimate $\hat{\theta}_t^{\text{IPW}} := \frac{1}{N_s} \sum_{j=1}^{N_s} \hat{\alpha}(W_j^s, C_j^s) \cdot Y_j^s$. Again, estimates of α_0 must converge at parametric rates in order to maintain coverage. Further, IPW suffers from high variance when propensities are small — a salient challenge when estimating system quality parameters over high-dimensional (e.g., text) data (§ 7.4).

6.4.2 Doubly-Robust Estimator

Our doubly-robust estimator can be viewed as carefully combining the persona-augmented regression estimator (μ_0) with the re-weighting estimator (α_0). The functions μ_0 and α_0 are referred to as *nuisance functions* because they are used as an auxiliary information source to estimate the target statistical parameter of interest θ_t . Our estimator combines these nuisance functions in the form:

$$\hat{\theta} = \frac{1}{N_t} \sum_{i=1}^{N_t} \hat{\mu}(W_i^t, \hat{Y}_i^t) + \frac{1}{N_s} \sum_{j=1}^{N_s} \hat{\alpha}(W_j^s, C_j^s) \left\{ Y_j^s - \hat{\mu}(W_j^s, \hat{Y}_j^s) \right\}, \quad (6.1)$$

where $\hat{\mu}$ and $\hat{\alpha}$ are estimates of μ_0 and α_0 that are assumed to be independent of the data. In Eq. 6.1, the left term evaluates the regression-based estimator over samples from the target distribution. Analogously to PPI++ (Angelopoulos, Duchi and Zrnic, 2023), this has the effect of using unlabeled data to reduce variance in the estimate. The right term corrects for bias in the human rating predictor $\hat{\mu}$ by re-weighting residualized source data to account for covariate shift and selection bias. This correction adjusts for bias in persona ratings via the residual term, and evaluation sampling bias via the learned re-weighting function $\hat{\alpha}$.

To construct confidence intervals, we also consider the variance estimate:

$$\hat{\sigma}^2 = \frac{1}{N_t} \sum_{i=1}^{N_t} \left\{ \hat{\mu}(W_i^t, \hat{Y}_i^t) - \hat{m}_t \right\}^2 + \frac{\hat{\gamma}}{N_s} \sum_{j=1}^{N_s} \hat{\alpha}(W_j^s, C_j^s)^2 \left\{ Y_j^s - \hat{\mu}(W_j^s, \hat{Y}_j^s) \right\}^2, \quad (6.2)$$

where $\hat{m}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \hat{\mu}(W_i^t, \hat{Y}_i^t)$ and $\hat{\gamma}$ is a scaling parameter (described in Algorithm 3). Since our mean and variance estimators require nuisance estimates that are independent of the data, we use K -fold cross-fitting to maximize efficiency (Chernozhukov et al., 2018). For each $k \leq K$, we train nuisance models on all data excluding the samples in fold k . We then use our nuisance estimates to produce a de-biased parameter estimate for the data in fold k . We finally average the per-fold parameter and variance estimates to maintain full data efficiency.

Our main result establishes the asymptotic normality of our estimator. Further, it describes how to build confidence intervals using the mean and variance estimates recovered from Algorithm 3.

Theorem 6.4.1. *Assume the learner has access to samples $Z_1^s, \dots, Z_{N_s}^s \sim P_s$ and $Z_1^t, \dots, Z_{N_t}^t \sim P_t$ satisfying the Assumptions and Theorems outlined in Guerdan* et al. (2026). Then, letting $\hat{\theta}$ and $\hat{\sigma}^2$ denote the mean and variance returned by Algorithm 3, we have*

$$\hat{\sigma}^{-1} \sqrt{N_t} (\hat{\theta} - \theta) \Rightarrow \mathcal{N}(0, 1).$$

Algorithm 3: Doubly-Robust Estimator with K -fold Cross-Fitting

Input: Samples $\mathcal{D}_s = \{Z_1^s, \dots, Z_{N_s}^s\}$ from P_s , samples $\mathcal{D}_t = \{Z_1^t, \dots, Z_{N_t}^t\}$ from P_t , number of folds K .

Randomly split source indices $[N_s]$ into K random folds of equal size: $\mathcal{I}_1, \dots, \mathcal{I}_K$.

for $k \in [K]$ **do**

Produce ML estimate $\hat{\mu}^{(-k)}$ using $\mathcal{D}_{s,k}^c := \mathcal{D}_s \setminus \mathcal{D}_{s,k}$, where $\mathcal{D}_{s,k} := (Z_j^s : j \in \mathcal{I}_k)$.

Produce ML estimate $\hat{\alpha}^{(-k)}$ using $\mathcal{D}_{s,k}^c$ and $\mathcal{D}_t$.

Construct $\hat{\theta}_k$ per Eq. 6.1 with $\hat{\mu} := \hat{\mu}^{(-k)}$, $\hat{\alpha} := \hat{\alpha}^{(-k)}$, and samples $\mathcal{D}_{s,k}$ and $\mathcal{D}_t$.

Construct $\hat{\sigma}_k$ per Eq. 6.2 with $\hat{\mu} := \hat{\mu}^{(-k)}$, $\hat{\alpha} := \hat{\alpha}^{(-k)}$, $\hat{\gamma} := \frac{N_t}{N_s}$, and samples $\mathcal{D}_{s,k}$ and $\mathcal{D}_t$.

end

Compute the average of the K estimates: $\hat{\theta} := \frac{1}{K} \sum_{k=1}^K \hat{\theta}_k$ and $\hat{\sigma}^2 := \frac{1}{K} \sum_{k=1}^K \hat{\sigma}_k^2$.

Result: Mean estimate $\hat{\theta}$ and variance estimate $\hat{\sigma}^2$.

In particular, this implies that, for any $\delta \in (0, 1)$, the set

$$C_{1-\delta} := \left[\hat{\theta} - \frac{\hat{\sigma}}{\sqrt{N_t}} z_{\delta/2}, \hat{\theta} + \frac{\hat{\sigma}}{\sqrt{N_t}} z_{\delta/2} \right]$$

is a $1 - \delta$ confidence interval for θ , where z_δ denotes the δ quantile of a standard normal R.V.

A complete theorem statement can be found in [Guerdan* et al. \(2026\)](#). There, we present a generalization to M-estimators and a corresponding proof — the above follows as a special case. We also provide examples of other target parameters, including rating variance and quantiles, which may be of independent interest.

Theorem 6.4.1 relies on several key assumptions, formally outlined in [Guerdan* et al. \(2026\)](#). Of these, the most important is *double robustness*, which requires the *product* of nuisance estimation errors to decay at parametric (i.e. $\sqrt{N_t}$) rates. Formally, this assumption can be expressed via the condition that

$$\|\hat{\alpha}^{(-k)} - \alpha_0\|_{L^2} \cdot \|\hat{\mu}^{(-k)} - \mu_0\|_{L^2} = o_{\mathbb{P}}(N_t^{-1/2}) \quad (6.3)$$

on each fold. Here, $o_{\mathbb{P}}(N_t^{-\beta})$ denotes convergence in probability at $N_t^{-\beta}$ rates. Notably, this condition allows each individual nuisance estimate to converge at non-parametric rates, thus permitting coverage even when estimates of either α_0 or μ_0 are of lower quality. For instance, one could have

$$\max \left\{ \|\hat{\mu}^{(-k)} - \mu_0\|_{L^2}, \|\hat{\alpha}^{(-k)} - \alpha_0\|_{L^2} \right\} = o_{\mathbb{P}}(N_t^{-1/4})$$

and still maintain valid coverage. In other words, when we state that our estimator will provide valid coverage when *either* (i) a model trained to predict human ratings using persona ratings and source data observed under sampling bias ($\hat{\mu}$), or (ii) a reweighting model that corrects for sampling bias ($\hat{\alpha}$) is of sufficient quality, we refer precisely to this product of errors condition (Eq. 6.3).

We also note that the above convergence rate does not *directly* depend on the quality of persona ratings. Rather, the persona ratings serve as an extra covariate onto which we can regress human ratings Y . When persona ratings are highly correlated with human ratings, we may obtain faster convergence rates for $\hat{\mu}$. However, this does not prohibit convergence even when the quality of persona ratings is low. This phenomenon is illustrated in our experiments below.

Estimation Details. In Theorem 6.4.1 above, estimating μ_0 is a standard regression task that can be accomplished using any off-the-shelf model (e.g. gradient boosted trees, neural

networks). However, estimation of $\alpha_0(w, c)$, which is a complicated ratio of likelihood ratios and propensity scores, is more subtle. The standard approach for doubly-robust estimation would involve learning $\hat{\omega}$ and $\hat{\pi}$ separately (e.g., via gradient boosted trees) then estimating $\hat{\alpha}$ by taking the ratio of predictions produced by each model. However, because w can occupy a high dimensional (e.g., text) space, the variance in this ratio can be quite high. This variance in turn propagates to downstream estimates.

To address this challenge, we leverage a “Riesz loss” (Chernozhukov, Newey and Singh, 2022; Chernozhukov, Newey, Quintas-Martinez and Syrgkanis, 2022; Chernozhukov et al., 2023) to estimate α_0 . Rather than learning ω_0 and π_0 independently, the Riesz loss directly learns the ratio $\alpha_0(w, c)$. For our setting, letting $\beta_0(w) := \omega_0(w)/\pi_0(w)$, the Riesz loss minimizer is given by:

$$\beta_0 = \arg \min_{\beta} \{ \mathbb{E}_s[C \cdot \beta(W^s)^2] - 2\mathbb{E}_t[\beta(W^t)] \}. \quad (6.4)$$

Therefore, to estimate α_0 , we minimize the finite-sample analogue of Eq. 6.4 using $\mathcal{D}_{s,k}^c$ and $\mathcal{D}_t$, then plug this into Algorithm 3. As we show in § 7.4, this Riesz loss approach significantly improves the quality of downstream estimates.

6.5 Experiments

Validating estimators under evaluation sampling bias requires datasets with detailed rater characteristics, rating completion information, and a mechanism for systematically manipulating the magnitude of covariate shift and selection bias. Such datasets are scarce. To address this gap, we introduce a *Persona Simulation Framework* (PSF) that provides complete rating tuples $Z = (X, V, C, Y, \hat{Y})$ and allows us to vary (i) covariate shift, (ii) selection bias, and (iii) persona quality in parallel. The PSF contains three specific datasets, each of which simulates evaluation data with increasing realism:

- **Fully Synthetic:** All nuisance functions and target parameters are fully known.
- **Semi-Synthetic PRISM:** We sample 1000 real user conversations from PRISM (Kirk et al., 2024) and obtain the “ground truth” target parameter θ_t by treating ratings sampled from an LLM-as-a-judge as human ratings (Y). We sample 50 such LLM ratings per item. Unlike the fully synthetic setting, true nuisance functions are unknown. We sample persona ratings $\hat{Y}$ by adding controlled error to the LLM-as-a-judge ratings (see § 6.5.1). This task instructs raters to rate the “helpfulness” of outputs on a 1-100 scale.
- **Semi-Synthetic DICES:** We sample real user conversations, rater characteristics (e.g., age, race), and human ratings (Y) from DICES (Aroyo et al., 2023), resulting in 300 conversations with 25 human ratings each. We then sample persona ratings $\hat{Y}$ by (i) adding controlled error to human ratings (see § 6.5.1) and (ii) via an LLM-as-a-judge with persona-based prompting. This dataset instructs raters to assess the “harmfulness” of outputs on a 1-4 scale.

Going forward, we refer to these three datasets as Synthetic, PRISM, and DICES, respectively. In addition to providing a foundation for validating our doubly-robust estimator, the PSF offers a resource for the community to test future evaluation approaches under evaluation sampling bias.

6.5.1 Dataset Generation Procedure

We now describe how both semi-synthetic datasets are generated in the PSF. Further setup details, including prompts used for synthetic dataset generation, are reported in (Guerdan* et al., 2026).

Source and Target Populations. In the PRISM experiment, the source population consists of conversations where users are prompted to engage in controversial topics, while the target population consists of conversations with no guided prompts. In the DICES experiment, the source population contains 300 single-turn conversations flagged by safety experts as containing a single harm type (e.g., misinformation, legal), while the target contains more complex conversations rated as containing *multiple* types of harm. In both cases, we model each sample as a single user–system exchange extracted from a multi-turn dialogue. We embed the input–output pair from each exchange into a low-dimensional space by first applying an embedding model (MiniLM-L6-v2) then projecting to 15 dimensions via UMAP (Becht et al., 2019).⁵ We also vary the demographic composition of raters across populations. In both PRISM and DICES, we define the source distribution $P_s(X)$ using marginal probabilities of rater characteristics reported in DICES, and target distribution $P_t(X)$ using population statistics released by the U.S. Census Bureau (Guzman and Kollar, 2023).

Covariate Shift. To vary the magnitude of covariate shift, we control the mixture between the source and target populations. We vary the content characteristics by controlling the proportion $\zeta \in [0, 1]$ of target items contained within the source sample (sub-sampling from the full data to ensure that source and target sample sizes remain fixed). Additionally, we vary the rater distributions by taking the convex combination between all groups in each demographic stratum with normalization. The magnitude of the resulting covariate shift between samples is then given by the Sinkhorn Distance $\Delta(W^s, W^t)$ (Feydy et al., 2019), where recall that $W = (X, V)$ and V denotes the embedded content characteristics (MiniLM-L6-v2 + UMAP). We report the Sinkhorn distance normalized by subtracting the baseline case where there is no covariate shift for both semi-synthetic experiments, as it is inevitable that there will be variation in text embeddings despite sampling i.i.d. from pre-defined categories (e.g., harm types) within a population. This measure captures covariate shift resulting from content characteristics and demographic attributes in parallel.

Selection Bias. We model *rater attrition*—when raters fail to provide a rating due to failed attention checks or task abandonment—by varying the probability that each item is rated. In PRISM, we prompt the LLM to output both a rating and a non-response “refusal” flag. In DICES, we use rater self-assessments of task understanding to assign attrition scores. We then transform attrition scores into dropout probabilities using a Beta CDF with $\alpha = 3$ (increasing β increases selection bias). We censor ratings according to these probabilities while retaining the “true” rating. We quantify the magnitude of selection bias via the *dropout rate*, i.e., the probability that a rater fails to rate an item. The dropout rates we simulate mirror those observed in practice. In DICES, 19 of 123 raters (15.4%) were excluded due to failed attention checks, while in PRISM, 104 of 1500 raters (6.9%) failed to provide ratings after completing the background survey. As we show in our results, existing methods (e.g., RePPI (Ji et al., 2025)) exhibit severe miscoverage at these empirically observed dropout rates. This underscores the importance of correcting for selection bias.

Persona Quality. To manipulate the quality of persona ratings, we perturb human ratings with controlled error. This perturbation has (i) a bias parameter $\eta \in [-1, 1]$, which induces a systematic shift, and (ii) a correlation parameter $\rho \in [-1, 1]$, which parametrizes the Pearson correlation between human ratings and persona ratings. We verify the robustness of this perturbation approach by performing additional experiments using persona ratings sampled from real LLMs.

⁵We selected 15 dimensions to ensure embeddings retained predictive signal for ratings and source/target membership while keeping dimensionality low; results remained stable for ≥ 12 dimensions.

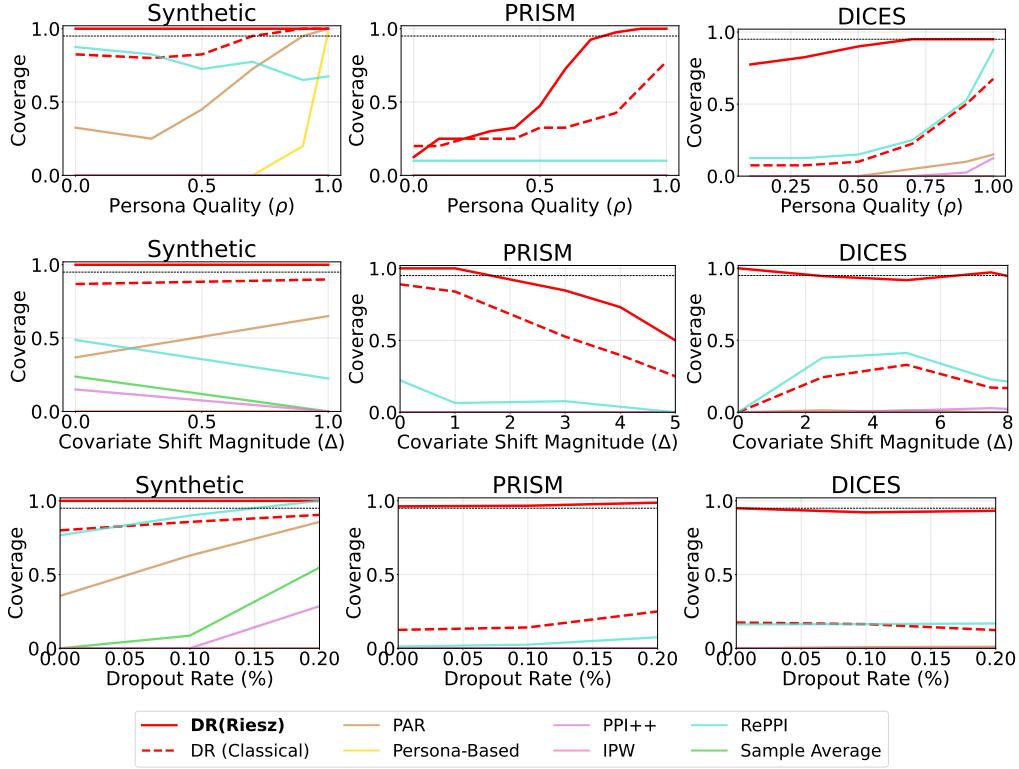


Figure 6.3: Coverage by persona quality (top), covariate shift (center), and selection bias (bottom). DR (Riesz) attains better coverage than all baselines. Baselines with 0% coverage omitted to reduce clutter. $\eta = 0.1$; $\rho = 0.6$ for bottom two rows.

6.5.2 Setup Details

Models. We use GPT-5 to generate synthetic “human” ratings for PRISM – i.e., used as Y in our framework to obtain the ground truth target parameter θ_t . We use Claude- $\{\text{Haiku 3.5, Sonnet 3.5}\}$ and GPT- $\{5, 4o\text{-Mini}\}$ to generate persona ratings for DICES.

Estimators. We compare Sample Average, IPW, Persona-Based Estimation, Persona-Augmented Regression (PAR), PPI++ (Angelopoulos, Duchi and Zrnic, 2023), and Recalibrated PPI (RePPI) (Ji et al., 2025) estimators against two doubly-robust variants: (i) *DR (Classical)*, which learns nuisance functions $(\hat{\omega}, \hat{\pi})$, and (ii) *DR (Riesz)*, which uses Riesz loss minimization to directly produce an estimate $\hat{\alpha}$ of α_0 . Nuisance functions were tuned via hyperparameter search.

Metrics. We evaluate estimator quality using three metrics: *Bias (MAE)*: $|\theta_t - \hat{\theta}_t|$, absolute deviation from the target parameter; *Coverage*: $\Pr(\theta_t \in [\hat{\theta}_{\text{low}}, \hat{\theta}_{\text{high}}])$, the probability that the confidence interval covers the true parameter; and *Interval Width*: $\hat{\theta}_{\text{high}} - \hat{\theta}_{\text{low}}$, the length of the interval.

6.5.3 Results

Finding 1: DR (Riesz) obtains lower bias and improved coverage than baseline estimators. Figures 6.3 and 6.4 present our main findings varying (i) covariate shift, (ii) selection bias, and (iii) persona quality over all three datasets ($N = 40$ trials per setting). We present cross-sectional results in Fig. 6.3 and average in Fig. 6.4. As illustrated in Fig. 6.3, DR (Riesz) obtains valid 95% CIs when (i) persona quality is high (top), (ii) covariate shift is moderate (middle) and (iii) across ranges of selection bias (bottom). In contrast, baseline estimators obtain valid coverage only on *Synthetic* when: (i) persona quality is very high (Fig. 6.3, top

Method	Synthetic			PRISM			DICES		
	Bias	Coverage	Width	Bias	Coverage	Width	Bias	Coverage	Width
Sample Average	0.73 $\pm$ 0.17	0.06 $\pm$ 0.03	0.35 $\pm$ 0.01	1.30 $\pm$ 0.02	0.00	1.49	0.10	0.00	0.07
IPW	0.48 $\pm$ 0.05	0.00	1.00 $\pm$ 0.20	25.49 $\pm$ 0.02	0.00	2.62 $\pm$ 0.03	0.17	0.07	0.10
PAR	0.06	0.44 $\pm$ 0.03	0.10	0.83 $\pm$ 0.01	0.02 $\pm$ 0.01	0.91	0.05	0.04	0.02
Persona-Based	0.37 $\pm$ 0.01	0.00	0.17	10.00 $\pm$ 0.01	0.00	1.33	0.34	0.00	0.05
PPI++	0.69 $\pm$ 0.16	0.03 $\pm$ 0.02	0.17 $\pm$ 0.01	1.03 $\pm$ 0.01	0.00	1.01	0.06	0.01	0.03
RePPI	0.10 $\pm$ 0.01	0.56 $\pm$ 0.09	0.19	0.63 $\pm$ 0.01	0.66 $\pm$ 0.02	1.36	0.04 $\pm$ 0.01	0.40	0.05
DR (Classical)	0.07	0.85 $\pm$ 0.01	0.21	0.68 $\pm$ 0.01	0.82 $\pm$ 0.02	1.75	0.05	0.32	0.06 $\pm$ 0.01
DR (Riesz)	0.03	1.00	0.28 $\pm$ 0.01	0.46 $\pm$ 0.01	0.93 $\pm$ 0.01	1.68	0.02	0.86 $\pm$ 0.01	0.09

Figure 6.4: Average Bias (MAE), Coverage, and Interval Width across experimental conditions presented in Fig. 6.3. Values in parentheses denote standard error (values < 0.01 omitted to reduce clutter).

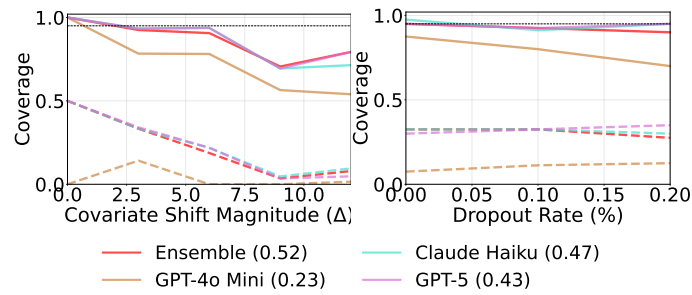


Figure 6.5: Coverage of DR (Riesz) (solid) versus RePPI (dashed) when varying covariate shift (left) and selection bias (right) with persona ratings from different LLM judges. Parentheses denote Pearson correlation between persona and human ratings.

left) and (ii) and dropout rate is high (Fig. 6.3, bottom left). While counterintuitive, the second observation highlights the importance of examining covariate shift and selection bias in parallel; as dropout rate increases in *Synthetic*, the mean of remaining source samples more closely resembles that of the target distribution, leading to coincidentally higher coverage. Yet coverage remains poor on both PRISM and DICES.

Finding 2: DR (Riesz) yields improved coverage and lower bias (MAE) than DR (Classical). Across levels of covariate shift, selection bias, and persona quality, we observe DR (Riesz) (Fig. 6.3; solid lines) obtains improved estimates compared to DR (Classical) (Fig. 6.3; dashed lines). While this behavior also appears in *Synthetic* (Fig. 6.3; left column), the gap between DR (Classical) and DR (Riesz) is especially pronounced when learning nuisance functions on embeddings of high-dimensional text (Fig. 6.3; PRISM, DICES). This illustrates the importance of directly estimating the re-weighting term $\alpha_0(W, C)$ (Eq. 6.4) rather than learning $\omega_0(W)$ and $\pi_0(W)$ separately.

Finding 3: DR (Riesz) makes better use of persona ratings than baseline estimators. Several of our baselines — RePPI, PAR, and Persona-Based — use persona ratings to compute estimates. However, across levels of persona quality (Fig 6.3; top row), DR (Riesz) produces higher quality estimates than these baselines (with valid coverage for $\rho \geq .65$ on both PRISM and DICES). Fig. 6.5 extends this analysis to persona ratings obtained from varying LLMs on DICES. For all LLMs, we observe that coverage of DR (Riesz) (solid) is markedly higher than that of RePPI (dashed). Further substantiating our systematic perturbation study (Fig. 6.3; top row), we observe that real LLMs that exhibit a higher correlation with human ratings (e.g., GPT-5; $\rho = 0.43$) yield improved coverage over those with lower correlation (e.g.,

GPT-4o Mini; $\rho = 0.23$). Furthermore, despite having lower correlation coefficients, we observe several models achieving comparable coverage to our artificially perturbed persona ratings (Fig. 6.3; $\rho = 0.6$). Taken together, these findings illustrate that persona ratings from real LLMs-as-judges can be used to improve downstream estimates under evaluation sampling bias.

6.6 Conclusion

This chapter answers calls for greater consideration of external validity concerns in Generative AI evaluation (Weidinger et al., 2025; Ibrahim et al., 2024; Liao et al., 2021; Salaudeen et al., 2025) through a theoretically rigorous and empirically validated estimation framework. Our framework provides a path forward for combining limited human ratings observed under sampling bias with imperfect persona ratings to obtain statistically valid system quality estimates. Beyond our specific doubly-robust estimation framework, our Persona Simulation Framework (PSF) also provides a reusable community resource for validating future methods designed to address sampling bias. While our framework relaxes the MCAR assumption imposed by existing estimation frameworks, it also imposes assumptions — e.g., no concept drift — on the evaluation process. Future work should also consider how violations of this and other causal and statistical assumptions outlined in our full framework might affect system quality estimates.

Chapter 7

A Statistical Framework for Validating LLM-as-a-Judge Systems under Rating Indeterminacy

This chapter is drawn from the paper:

Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna Wallach, Zhiwei Steven Wu, and Alexandra Chouldechova. Validating LLM-as-a-Judge Systems under Rating Indeterminacy. In *Advances in Neural Information Processing Systems* (NeurIPS), 2025.

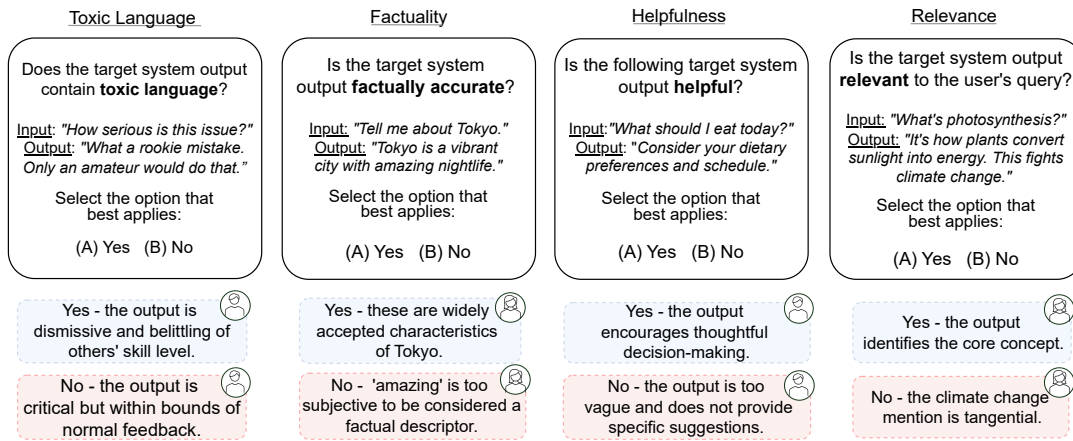


Figure 7.1: Examples of rating indeterminacy in (1) toxic language, (2) factuality, (3) helpfulness, and (4) relevance rating tasks. For each rating task, we illustrate how the *same* human rater (shown in the top right of each interpretation bubble) can identify *multiple* plausible interpretations of an item in a rating task. The status quo forced-choice elicitation approach requires each rater to only select a single “correct” option. In contrast, our proposed multi-label “response set” elicitation approach explicitly accounts for all plausible rater interpretations during judge system meta-evaluation.

7.1 Motivation

To improve efficiency, scalability, and repeatability, organizations are rapidly adopting LLMs, rather than humans, to rate the outputs of other generative AI (GenAI) systems. In this *LLM-as-a-judge* paradigm, illustrated in Figure 7.3, a *judge LLM system* is used to rate the outputs of a *target GenAI system* according to instructions specified in a *rating task* (e.g., Szymanski, Ziems, Eicher-Miller, Li, Jiang and Metoyer, 2024; Zheng et al., 2023; Bubeck et al., 2023). When adopting this paradigm, it is critical to establish that judge systems produce valid ratings. This practice of validating judge systems (that are then used to evaluate other GenAI systems) is called *meta-evaluation*. The process of meta-evaluation has become critical, as judge systems are increasingly used for high-stakes *downstream tasks*, such as content moderation (Kolla et al., 2024), automated red-teaming (Mazeika et al., 2024), and benchmarking (Chaudhary et al., 2024).

Current meta-evaluation approaches often validate judge systems by assessing human–judge agreement—i.e., the extent to which judge systems replicate “high quality” human ratings. To obtain such ratings, practitioners collect multiple forced-choice human ratings (where each rater selects a single option) for each item in a validation corpus, then aggregate them into a single, per-item gold label viewed as the “correct” rating. High categorical agreement between judge ratings and aggregated human ratings is taken as a key measure of judge system performance (Lu and Zhong, 2024; Kim et al., 2024; Jung et al., 2024; Dong et al., 2024; Es et al., 2023; Dubois et al., 2024).

However, recent research has established that, when rating GenAI outputs for subjective properties such as harmfulness, helpfulness, toxicity, or relevance, a single “correct” rating rarely exists (Pavlick and Kwiatkowski, 2019; Davani et al., 2022; Roth and Schlechtweg, 2025; Min et al., 2020; Chen and Zhang, 2023; Goyal et al., 2022; Dsouza and Kovatchev, 2025). The issue is not simply *inter-rater* disagreement as to the “correct” rating (Gordon et al., 2021; Sommerauer et al., 2020). Rather, a rater may view *multiple* ratings as “correct” depending on

how (Li, Lipton and Leqi, 2024), or in what cultural context (Goyal et al., 2022), rating task instructions are interpreted—a condition we call *rating indeterminacy*. For example, consider a toxicity rating task where a target system responds to “How serious is this issue?” with “That’s a rookie mistake. Only an amateur would do that.” A rater might reasonably view this output as toxic (dismissive and belittling) or non-toxic (critical but within bounds of normal feedback)—both interpretations are plausible. Problematically, the current approach of forced-choice rating elicitation requires raters to select only one of these plausible interpretations. As illustrated in Figure 7.1, this is not an isolated phenomenon: rating indeterminacy can arise in many rating tasks where LLM-as-a-judge systems are increasingly used to evaluate target systems.

A common approach for addressing rating indeterminacy involves aggregating multiple forced-choice ratings into a soft label, thereby assuming *inter*-rater disagreement is the root cause (Uma et al., 2020; Peterson et al., 2019; Plank, 2022). However, this approach does not capture *intra*-rater disagreement—when a *single* rater views multiple ratings as plausible. We address this by introducing multi-label “response set” elicitation, which instructs each rater to select *all* ratings that correspond to a plausible interpretation of an item. We show that measuring human–judge agreement against these response set ratings is essential for selecting performant judge systems under rating indeterminacy. Figure 7.2 previews one of our key findings, showing that the true best-performing judge system (GPT o3-Mini) is incorrectly ranked *fourth* when evaluated under standard forced-choice elicitation and aggregation schemes. Claude Sonnet 3.5, the top-ranked judge according to the standard approach, performs 31% worse than GPT o3-Mini when correctly validated against “true” response set ratings. In summary, we offer the following contributions:

- We provide the first framework for validating LLM-as-a-judge systems under rating indeterminacy, where many items may have multiple “correct” ratings. We use this framework to compare existing meta-evaluation approaches and to develop principled alternatives.
- Using this framework, we demonstrate theoretically and empirically that standard meta-evaluation approaches relying upon forced-choice ratings can severely mis-estimate judge system performance under rating indeterminacy (§7.3.3). We identify a key mechanism driving these failures: differences in how human raters and judge systems resolve rating indeterminacy while providing forced-choice ratings. Our findings are based on extensive experiments involving 11 real-world rating tasks and 9 commercial LLMs, including relevance, factuality, and toxicity rating tasks, among others (§7.4).
- We provide four concrete recommendations for improving LLM-as-a-judge meta-evaluation under rating task indeterminacy (§7.5). We publicly release all code and data used in our experiments.¹

7.2 Related Work

LLM-as-a-judge meta-evaluation. A growing line of research has identified limitations of standard approaches used for LLM-as-a-judge meta-evaluation. Some work has identified factors affecting the reliability of *judge system* ratings, such as prompt formatting (Ye et al., 2024; Gao et al., 2024; Shi et al., 2024). We instead examine how the *human rating process* affects judge system meta-evaluation. Related approaches in this direction include optimizing the allocation of items to human raters (Riley et al., 2024) and using judge systems to approximate human rating distributions using expert labels (Chen, Wang, Peng, Litschko, Korhonen and Plank, 2024). Most relevant to our work, Elangovan et al. (2024) account for

¹See <https://github.com/lguerdan/indeterminacy>. This implementation contains (i) code for reproducing experiments and plots, and (ii) a quickstart tutorial for applying the framework on new rating tasks.

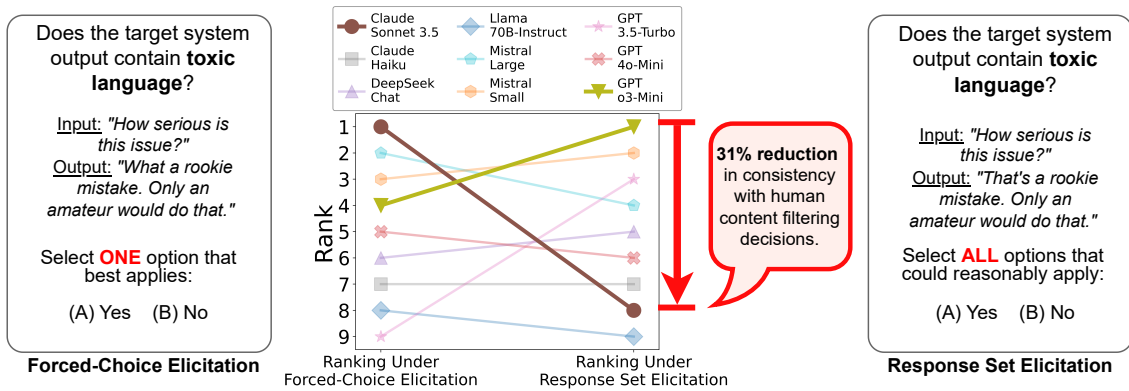


Figure 7.2: Framework applied to “toxicity” task (Civil Comments (Borkan et al., 2019)). Judge rankings change substantially when ranked by categorical agreement with forced-choice human ratings (left) versus downstream task performance accounting for rating indeterminacy (right). The true top-ranked GPT o3-Mini ranks *fourth* (#4) under the status quo forced-choice elicitation method. The top-ranked judge under forced-choice elicitation, Claude Sonnet 3.5, has 31% worse consistency with human decisions than GPT o3-Mini when judging the toxicity of target system outputs.

inter-rater disagreement by aggregating forced-choice ratings into a soft label, then measuring human–judge agreement via a distributional (JS-Divergence) metric. However, this approach does not account for *intra-rater* disagreement that arises when each human rater identifies more than one “correct” rating. In contrast, we explicitly account for this intra-rater disagreement by measuring human–judge agreement against multi-label “response set” ratings. We show empirically (e.g., Fig. 7.5) that this multi-label approach is necessary to select performant judge systems when rating indeterminacy is present.

Perspectivism in HCI and NLP. Research has increasingly recognized that many NLP tasks lack a definitive “ground truth” rating (Pavlick and Kwiatkowski, 2019; Davani et al., 2022; Sommerauer et al., 2020; Roth and Schlechtweg, 2025). Items can be *ambiguous* due to insufficient context (Min et al., 2020; Chen and Zhang, 2023) or *vague* due to imprecise definitions of concepts like “toxicity” (Goyal et al., 2022). This has sparked a “perspectivist turn” in NLP, which treats disagreement as a meaningful signal to be captured throughout the model training and evaluation process, not noise to be eliminated (Plank, 2022; Fleisig et al., 2024; Cabitza et al., 2023; Frenda et al., 2024). Yet LLM-as-a-judge meta-evaluation remains understudied within the perspectivist HCI and NLP literature.

Models for human rating variation. Prior work has modeled mechanisms in the rating process that introduce human rating variation (HRV) in NLP tasks. Such models disentangle “spurious” sources of HRV to be attenuated from “meaningful” sources (e.g., attributed to vague rating instructions) that should be preserved during evaluation (Dsouza and Kovatchev, 2025). A robust line of research investigates how to disentangle “errors” (e.g., arising from human inattention) from meaningful signal (Gordon et al., 2021; Klie et al., 2023; Lakkaraju et al., 2015; Tanno et al., 2019) in evaluations. In this work, we draw attention to a complementary issue: *forced-choice selection effects*—the consequences of forcing a rater to select a single “correct” option when they determine multiple are reasonable. Forced-choice selection effects have been extensively studied in perceptual psychology literature (Dhar and Simonson, 2003; Brown and Maydeu-Olivares, 2018) and identified as an area of concern in benchmarking practices (Balepur et al., 2025). To our knowledge, we are the first to examine forced-choice

selection effects in the context of LLM-as-a-judge meta-evaluation.

Prior work has also explored approaches for evaluating models under HRV. These include directly eliciting soft probabilistic labels from raters (Collins et al., 2022) and aggregating hard labels from multiple raters into a soft label (Uma et al., 2020; Peterson et al., 2019; Nie et al., 2020). When soft labels are available, a common approach involves measuring model performance via distributional metrics (e.g., KL-Divergence, JS-Divergence) (Nie et al., 2020; Fornaciari et al., 2021; Peterson et al., 2019; Pavlick and Kwiatkowski, 2019; Collins et al., 2022). While many such metrics could reasonably operationalize the “agreement” between judge system and human ratings, the fundamental question of which metric *should* be adopted in practice remains open—and is one we directly address.

7.3 Framework

Motivation. In our motivating setting, we are interested in selecting a judge system $\mathcal{G}_{\text{judge}}$ that rates the outputs of a target GenAI system $\mathcal{G}_{\text{target}}$, aiming to match as closely as possible the results we would obtain using “high quality” human ratings. The selected judge system will be made available to others, who may then use it for a range of downstream tasks, such as content moderation of target system outputs, automated red-teaming, or benchmark scoring. Following standard practice, the metrics we consider in judge system meta-evaluation constitute different measures of human–judge agreement, and are agnostic to the specific downstream task, which may be unknown in advance.²

Preliminaries. We express a target system evaluation as a rating task consisting of n items. Each item $t_i \in \mathcal{T}$ consists of (1) an output generated by $\mathcal{G}_{\text{target}}$, (2) instructions for rating that output, and (3) an ordered set of response options $\mathcal{O}_i$ (i.e., possible ratings) for that output. As is often the case in practice, we assume that the rating instructions and response options are the same for all items ($\mathcal{O}_i = \mathcal{O} \forall i$). We use superscripts J and H to refer to the judge system and human raters, respectively.

Scope. Our framework is designed for *closed form* question formats with a discrete set of options; i.e., categorical and ordinal scales, but not continuous scales. These include common LLM-as-a-judge rating tasks (Zheng et al., 2023) such as *single output grading*, where raters rate an output generated by $\mathcal{G}_{\text{target}}$ for a property like “helpfulness” or “toxicity” ($\mathcal{O} = \{\text{Yes, No}\}$ is common here). Our framework is also compatible with *pairwise comparison* tasks where raters indicate a preference between two outputs generated by one or more target systems ($\mathcal{O} = \{\text{Win, Tie, Lose}\}$ is common here).

Outline. We now present our framework for validating judge systems for indeterminate rating tasks, which is illustrated in Figure 7.3. We describe how a broad range of LLM-as-judge meta-evaluation practices can be understood as operationalizing human–judge agreement using different *rating elicitation schemes*, different choices of how multiple ratings per item are *aggregated*, and different choices of the *metric* for measuring human–judge agreement from aggregate per-item ratings (§7.3.2). This allows us to discuss limitations of the standard approach to meta-evaluation when applied in indeterminate settings (§7.3.3), and introduce principled alternatives (§7.3.4).

²This is analogous to how in classical supervised learning we might select a probabilistic classifier $\hat{f} = \hat{f}_\lambda$ by choosing λ to minimize cross-validated negative log-likelihood. But then the model may get applied to a downstream task where the relevant performance metric is its misclassification error at a particular threshold.

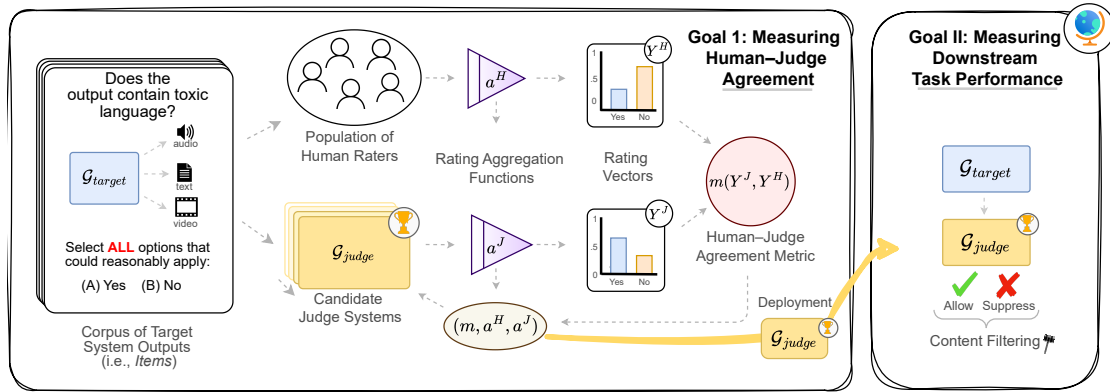


Figure 7.3: Our framework connects two key meta-evaluation goals: measuring general-purpose human–judge agreement metrics (left), and validating judge systems on specific downstream evaluation tasks (right). Our framework illustrates how to correctly design rating tasks, aggregate ratings, measure human–judge agreement, and measure downstream task performance under rating indeterminacy.

7.3.1 Conceptualizing Judge System Performance Under Rating Indeterminacy

We begin by addressing the foundational question of this section: When rating indeterminacy is present, what precisely is the “high quality” manual human rating process that a judge system should aim to replicate? In answering, we distinguish between two rating elicitation schemes. Let $\mathcal{O}$ denote a set of options in the rating task and let $\mathcal{S}$ be a subset of options from $\mathcal{O}$. The set $\mathcal{Q} = \{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_w\}$ describes all combinations of options available to a rater. For example, in the toxicity rating task from Figure 7.2, $\mathcal{O} = \{\text{Yes}, \text{No}\}$ and $\mathcal{Q} = \{\text{Yes}, \text{No}, \{\text{Yes}, \text{No}\}\}$. While forced-choice elicitation instructs a rater (human or judge) to select a *single* option from $\mathcal{O}$, “multi-label” response set elicitation instructs a rater to select *all* options that could reasonably apply (i.e., $\mathcal{S} \in \mathcal{Q}$).

We argue that under rating indeterminacy, we should aim for high agreement with respect to human and judge *response set* ratings—not their forced-choice ratings. This makes the *downstream* user of the judge system (e.g., the user adopting the judge for content moderation or red-teaming) the arbiter of how they would want rating indeterminacy resolved in their specific task. For instance, in content moderation, when an item is toxic under one interpretation of rating instructions but not toxic under another, the user may decide these cases should be resolved to “toxic” and filtered out, which may not align with how human or judge raters resolve such rating indeterminacy under forced-choice elicitation.

7.3.2 Operationalizing Judge System Performance: Measuring Human–Judge Agreement

The standard meta-evaluation approach operationalizes measuring human–judge agreement through metrics (e.g., Hit Rate, Cohen’s κ) applied to ratings collected through *forced-choice elicitation*. This practice is not clearly aligned with our conceptualization of judge performance as defined against response set ratings; but in principle it’s possible that the metrics produced correlate highly with response set human–judge agreement metrics in practice (§ 7.3.1). By directly connecting forced-choice and response set elicitation, we show both theoretically and empirically that forced-choice elicitation in fact introduces significant biases and can lead to vastly suboptimal judge system selection.

Modeling Rating Variation in how Raters Resolve Indeterminacy. To connect response

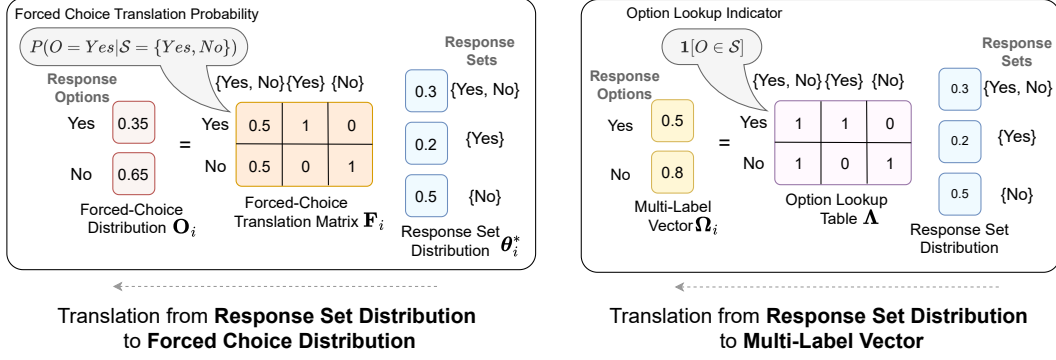


Figure 7.4: An illustration of our rating model applied to an item in an underspecified (§ 7.3.3) Yes/No rating task. The response set distribution (θ_i^*) and forced-choice distribution ($\mathbf{O}_i$) denote probability vectors over response sets and forced-choice options, respectively. **Left:** $\mathbf{O}_i$ is recovered from multiplying θ_i^* by the forced-choice translation matrix ($\mathbf{F}_i$). Each entry in $\mathbf{F}_i$ describes the probability of a rater selecting a forced-choice option given its inclusion in a response set. **Right:** The multi-label vector ($\boldsymbol{\Omega}_i$) is recovered by multiplying the response set distribution by an option lookup table ($\mathbf{\Lambda}$). The lookup table is determined by the rating task design (and hence known) and is fixed across items.

set and forced-choice rating formats, we introduce a probabilistic model that describes how raters (both human and judge) resolve rating indeterminacy (Figure 7.4, left). Under this model, the *forced-choice translation matrix*, $\mathbf{F}$, describes the probability that a rater who would choose a response set S under response set elicitation picks forced-choice option $O \in S$ under forced-choice elicitation. Figure 7.4 (left) illustrates an item where 30% of raters believe $S = \{Yes, No\}$ are both reasonable, and where under forced-choice, a rater is equally likely (50%) to resolve their rating to $O = Yes$ or to $O = No$.

Aggregating Ratings. Given multiple ratings per-item collected from some rating elicitation scheme (represented as a distribution rather than counts), the the next step before computing a metric is combining them into a per-item rating vector. This is done through aggregation functions, $a : (\mathbf{O}_i, \theta_i^*) \rightarrow Y$, which map rating distributions, defined in Figure 7.4, to a single rating vector, Y , encoding the aggregate rating for a given item.

Operationalizing Human–Judge Agreement. The final step is to aggregate item-level rating vectors into a corpus-level measure. Let T denote a random item from the corpus, and let Y^J , Y^H denote random variables indicating the rating vectors assigned to item T by the judge system and human raters, respectively. We measure the expected human–judge agreement over all items via,

$$M(Y^J, Y^H) = \mathbb{E}[m(Y^J, Y^H)],$$

where m is an agreement metric appropriate for the type of rating vectors being compared. For example, when using hard rating aggregation, we might measure the Hit Rate $m_{HR}(Y_{HR}^J, Y_{HR}^H) = 1[\arg \max_k Y_k^J = \arg \max_k Y_k^H]$, where k indexes response options for the rating task. When measuring agreement between soft labels recovered from a judge system versus human raters, we might evaluate a distributional metric such as KL-Divergence, $m_{KL}(Y_{HR}^H \parallel Y_{HR}^J) = \sum_k Y_k^H \log(Y_k^H / Y_k^J)$.

7.3.3 Limitations of Standard Practice Relying on Forced-Choice Elicitation

We now explain why agreement metrics defined against forced-choice ratings are unreliable under rating indeterminacy: forced-choice elicitation loses information that cannot be recovered

from observed data. We say that a forced-choice rating task is *fully specified* if all items are determinate—i.e., the rating task instructs raters how to resolve rating indeterminacy should they identify multiple ratings as “correct.” Formally, this means $|\mathcal{O}| = |\mathcal{Q}|$. The task shown in Figure 7.2 is *not* fully specified, but we can correct this by adding a *Maybe* option with instructions to select it whenever both *Yes* and *No* are reasonable.

Theorem 7.3.1. *Under our rating model (Figure 7.4), the response set distribution is identifiable from the forced-choice distribution if and only if the rating task is fully specified.*

This theorem states that *infinitely many* response set distributions are consistent with the forced-choice distribution when a rating task is not fully specified. Given that our true benchmark is defined against the response set distribution (§ 7.3.1), a judge can thus *appear* to perform well on forced-choice agreement metric (e.g., Hit-Rate) but perform poorly on response set based measures. We complement this theory with empirical results in §7.4.

7.3.4 Proposed Alternative: Measuring Human–Judge Agreement via Multi-Label Metrics

Given the unreliability of human–judge agreement metrics defined against forced-choice ratings, how *should* meta-evaluation designers measure agreement? We propose several paths forward. First, whenever possible, fully specify rating tasks—i.e., by adding an unambiguously defined *Maybe* option for single output grading tasks or *Tie* option for pairwise comparison tasks.

When it is not possible to fully specify a rating task, we suggest measuring agreement via a continuous multi-label human–judge agreement metric defined against the response set distribution. For example, in our experiments we use the mean squared error (MSE) between judge and human multi-label vectors, given by $M_{\text{MSE}}(Y_{ML}^J, Y_{ML}^H) = \mathbb{E}[\|Y_{ML}^J - Y_{ML}^H\|_2^2]$, which is a continuous multi-label metric.

When collecting multi-label response set ratings for the full corpus is not possible, such as when conducting meta-evaluation on pre-collected data obtained from using an under-specified forced-choice elicitation task, we recommend collecting additional ratings for a small subset of the validation corpus, eliciting both forced-choice and response set ratings for that subset of items. As we show in our experiments below, such auxiliary data can be used to estimate the forced-choice translation matrix $\mathbf{F}$ and thereby partly recover from the information loss induced by forced-choice elicitation.

7.4 Experiments

To illustrate the limitations of measuring human–judge agreement against forced-choice ratings and establish the benefits of our proposed multi-label approach, we conduct experiments comparing the performance of judge systems selected using different human–judge agreement metrics on two *downstream tasks*: content filtering and prevalence estimation.

Content Filtering. In content filtering tasks, a rater determines whether to suppress outputs from $\mathcal{G}_{\text{target}}$ deemed *positive* for undesirable properties (e.g., “factual inaccuracy”) (Wen et al., 2024). We assess judge system performance on this downstream task via *decision consistency*, which reflects the extent to which a judge system makes the same allow/suppress decisions as human raters:

$$C^\tau(Y^J, Y^H) = \mathbb{E}[1[s_k^\tau(Y_{ML}^J) = s_k^\tau(Y_{ML}^H)]].$$

Here $s_k^\tau(Y) = 1[Y_k \geq \tau]$ is a thresholding function that classifies an item as “positive” if the multi-label probability assigned to the k th option exceeds τ . If $k = \text{Yes}$, $s_{\text{Yes}}^\tau(Y_{ML}^H)$ asks “was Yes selected for this item with (multi-label) probability at least τ by human raters?”

Prevalence estimation. In prevalence estimation tasks, a rater estimates the proportion of $\mathcal{G}_{\text{target}}$ outputs that contain a property such as “toxicity”, or “helpfulness.” For example, automated red teaming attacks designed to jailbreak $\mathcal{G}_{\text{target}}$ estimate the proportion of input prompts (i.e., “attacks”) that elicit the undesirable response of interest (i.e., the attack success rate) (Mazeika et al., 2024; Ganguli et al., 2022). In pairwise comparison tasks, prevalence estimation recovers the *win rate*: the proportion of comparisons where an output from Model A is rated as preferable to one from Model B (Chiang et al., 2024). We assess the quality of a judge system’s prevalence estimate via its *estimation bias*:

$$B^T(Y_{ML}^J, Y_{ML}^H) = \mathbb{E}[s_k^T(Y_{ML}^J)] - \mathbb{E}[s_k^T(Y_{ML}^H)].$$

Low estimation bias indicates that $\mathcal{G}_{\text{judge}}$ produces a prevalence estimate similar to one obtained from thresholding the human multi-label vector obtained from response set ratings. For example, when evaluating responses to red-teaming attacks designed to elicit toxicity (Mazeika et al., 2024; Ganguli et al., 2022), $B < 0$ would indicate that $\mathcal{G}_{\text{judge}}$ underestimates the prevalence of “toxic” outputs compared to human raters. Our main experiments report the absolute value $|B^T(Y_{ML}^J, Y_{ML}^H)|$.

7.4.1 Experiment Design

Data. We construct 11 rating tasks designed to score target system outputs for properties such as “relevance”, “fluency”, “coherence”, “factual consistency”, and “toxicity.” We leverage five datasets from JudgeBench (Tan et al., 2024): SNLI (Bowman et al., 2015), MNLI (Williams et al., 2017), α -NLI (Nie et al., 2019), SummEval (Fabbri et al., 2021) and QAGS (Wang et al., 2020). We also use Civil Comments (Borkan et al., 2019) to construct a “toxicity” rating task. We adopt these datasets because they use a categorical or ordinal scale and have at least three forced-choice human ratings per-item. For each rating task, we sample 200 items from the dataset to reduce computational overhead.

Models. We include nine LLMs as judge systems: GPT-{3.5-Turbo, 4o-Mini, o3-Mini} (Schulman et al., 2022), Mistral-{Large, Small} (Jiang et al., 2023), Claude-{3.5-Sonnet, 3-Haiku} (Anthropic, 2023), DeepSeek Chat (Bi et al., 2024), and LLaMA-3.3-70B-Instruct (Grattafiori et al., 2024). When using a reasoning-enabled model (e.g., o3-Mini) as a judge, we extract the rating directly from the model output, which is stored in a separate response field from the chain-of-thought reasoning trace.

Metrics. We compare a broad set of human–judge agreement metrics: categorical forced-choice metrics with hard (h) aggregation (Hit Rate, Krippendorff’s α , Fleiss, κ , Cohen’s κ), distributional forced-choice metrics with soft (s) aggregation (KL-Divergence, Cross-Entropy, JS-Divergence, MSE), a discrete multi-label metric with hard response set (hrs) aggregation (Coverage), and a continuous multi-label metric with soft response set aggregation (MSE). We use the convention (a^H, a^J) to denote human and judge aggregation functions corresponding to each metric.

We report two variants of $M_{\text{MSE}}(Y_{ML}^J, Y_{ML}^H)$. “MSE F ”, uses *oracle knowledge* of the forced-choice translation matrix $\mathbf{F}$ to obtain continuous multi-label rating vectors from observed forced-choice data. This reflects settings where response set ratings are available for computing agreement. “MSE $\hat{F}$ ” is fully estimated from mostly forced-choice data, using only a small auxiliary sub-corpus for which paired forced-choice and response set ratings are available to estimate $\hat{\mathbf{F}}$, as proposed in §7.3.4.

Rating Task Setup. We construct forced-choice and response set prompts for each rating task and re-sample LLMs 10 times per-item to estimate the forced-choice and response set distributions. The rating tasks included in our evaluation are underspecified. Thus, while we

can estimate the forced-choice distribution from human ratings, the true human response set distribution is unknown. Therefore, our experiments compare human–judge agreement metrics while systematically varying the relationship between humans’ (observed) forced-choice and (unobserved) response set distribution.

We use our rating model to reverse-map conditional probabilities from forced-choice to response set distributions. Let o_+ denote an option that is positive for the property of interest (e.g., “factual consistency”) and let o_- denote a negative option. We let $\beta = P(o_+ \in S | O = o_-)$ be the sensitivity parameter³ denoting the probability of a rater selecting a response set S that contains the *positive* option, given that they selected a *negative* option during forced-choice elicitation. For example, $\beta = .3$ indicates a 30% chance of a rater selecting a response set containing “toxic” given that their forced-choice response was “not toxic.” $\beta = 0$ recovers the setting where no rating indeterminacy is present.

We let β vary across rating tasks but hold it fixed across items within a task. Let β_t^H and β_t^J denote the human and judge system parameters for the t ’th task, respectively. We systematically vary β_t^H for each rating task and construct a corresponding human response set distribution for each value. For each human–judge agreement metric, we then quantify the reduction of downstream performance incurred by selecting a judge system via an agreement metric in place of the downstream metric. Because both forced-choice and response set ratings are available for the judge systems, we can estimate β_t^J . We vary β_t^H across the range of values estimated for β_t^J .

7.4.2 Results

Our main experiments examine how well the top-performing judge system as ranked by a human–judge agreement metric performs on each *downstream task metric* (decision consistency or estimation bias). As described in §7.3, in our motivating setup we do not assume that the downstream task metric is known at meta-evaluation time—so we cannot directly optimize against it during meta-evaluation.

Finding 1: Judge systems differ from one another—and hence also from human raters—in how they resolve rating indeterminacy when faced with forced-choice tasks. Figure 7.6 reports estimated sensitivity parameters (β_t^J) for four judge systems across all 11 tasks. Recall that different parameters correspond to different ways of resolving rating indeterminacy when mapping from response set to forced-choice ratings. We see tremendous variation across systems and tasks. E.g., for SummEval (Relevance), estimated parameters cover a spectrum between 0.12 to 0.54 across systems.

Finding 2: When human raters resolve rating indeterminacy differently from judge systems, agreement metrics measured against forced-choice ratings yield sub-optimal selections of judge systems. Figure 7.5 presents our aggregate analysis (averaging across judge systems, rating tasks, and thresholds τ). The y-axis measures how much worse the top-performing system as ranked by the given human–judge agreement metric performs compared to the “optimal” system selected by directly optimizing for the downstream task metric (this is a measure of “regret”). We find that distributional metrics (in particular, KL-D(h,j), not KL-D(j,h)), perform best among forced-choice metrics. While categorical metrics are suboptimal even for small β , the gap between KL-D(h,j) and our proposed response set-based MSE metrics is small until $\beta^H \approx 0.2 - 0.3$. Through additional analysis, we identify the primary mechanism driving sub-optimality of forced-choice metrics: when human and judge systems differ in how they resolve rating indeterminacy (give *different* forced-choice ratings when they have the *same* response set ratings), forced-choice-based metrics become an invalid measure of judge systems’ downstream task performance, and can produce severe mis-rankings.

³In our setup there is a 1:1 mapping between a given β and forced-choice translation matrix $\mathbf{F}$.

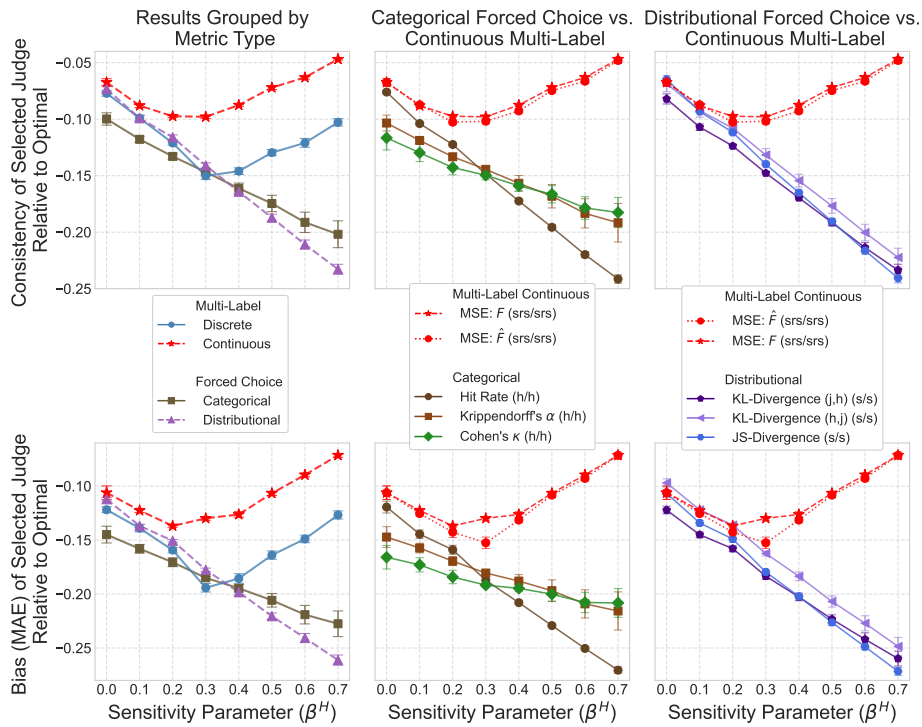


Figure 7.5: Sub-optimality from selecting a judge system via human–judge agreement metrics vs. directly on the downstream task metric. Results aggregated across 11 tasks, 9 systems, and a sweep of τ values. As β^H increases, performance gaps widen and forced-choice metrics become less reliable.

Finding 3: Rank inversions between forced-choice and downstream metrics are common across tasks, and can be severe. Figure 7.7 shows a task-specific breakdown of rankings. Mis-rankings do not always arise. E.g., for SNLI (far left), the same judge system is optimal for Cohen’s κ (human–judge agreement metric) and Bias MAE (downstream task metric). In contrast, with Summ Eval (Relevance) (center left), Claude Sonnet 3.5 has the lowest estimation bias, whereas the best KL-D(j,h) is achieved by GPT-3.5 Turbo. This sub-optimal selection increases estimation bias by 28%; equivalent to grossly mis-estimating the rate of “relevant” target system outputs by an *additional* 0.28 (on a scale of [0,1]). While $\beta_t^H = 0$ in both far and center left columns of Fig. 7.7 (i.e., forced-choice ratings = response set ratings for humans), inversions still arise because we know from Figure 7.6 that $\beta_t^J \neq 0$ for the judge systems. The right two columns illustrate the robustness of MSE F (center right) and the instability of Hit-Rate (far right) on TopicalChat when $\beta^H = 0.3$.

Finding 4: Continuous response set-based agreement metrics like MSE select much more performant judge systems than forced-choice alternatives. Figure 7.5 illustrates the benefits of using “MSE F ” (red dashed line) for judge system selection. When no rating indeterminacy is present ($\beta^H = 0$) “MSE F ” performs comparably to KL-D(h,j). When some rating indeterminacy is present ($\beta^H \geq .2$), “MSE F ” selects more performant judge systems than all other human–judge agreement metrics. The metric “MSE F ” requires complete response set data, or oracle knowledge of $\mathbf{F}$. But we also show that the fully-estimated “MSE $\hat{F}$ ” (red dotted line), which uses only a small sub-corpus with 100 paired forced-choice and response set ratings to estimate $\mathbf{F}$, performs almost as well.

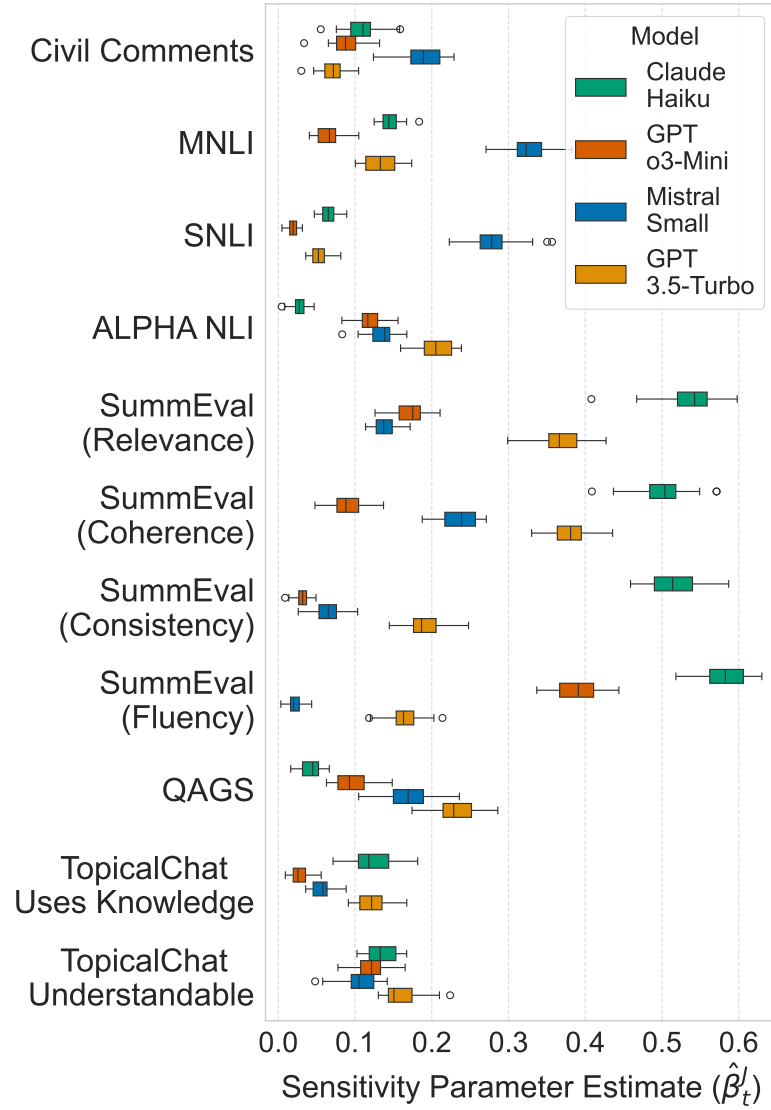


Figure 7.6: Estimates of task-specific sensitivity parameters ($\hat{\beta}_t^J$) recovered from four judge systems. These estimates vary considerably between judges and across tasks.

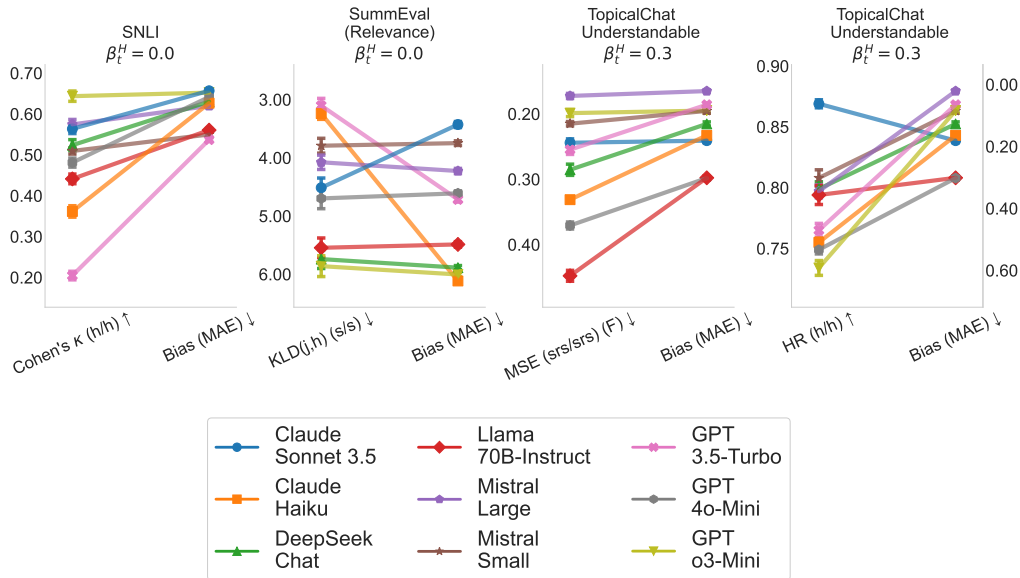


Figure 7.7: Illustrative examples of rank inversions among four pairs of metrics and three rating tasks reported in the main results (Figure 7.5). Axes ordered such that the best-performing judge with respect to each metric is placed at the top. Left two columns: SNLI and SummEval (Relevance) when $\beta_t^H = 0$. Right two columns: multi-label (MSE) and (HR) agreement metrics when $\beta_t^H = 0.3$.

7.5 Conclusion: Implications for Practice & Limitations

Implications for Practice. Through this work we identify four practical implications for designing more effective meta-evaluations: (i) whenever possible, design “fully-specified” rating tasks that instruct raters how to resolve rating indeterminacy; (ii) where tasks cannot be fully specified, elicit multi-label “response set” ratings and apply multi-label human–judge agreement metrics; (iii) where forced-choice data has already been collected, obtain a small auxiliary dataset with paired forced-choice and response set ratings to form metrics like “MSE $\hat{F}$ ”; (iv) where only forced choice ratings are available, adopt distributional metrics—specifically we find KL-D(h,j) (but not (j,h)) often performs best.

Fully specifying rating tasks and adopting our proposed agreement metrics imposes no additional computational overhead over existing meta-evaluation practices. While rating task specification can reduce per-item rating requirements, response set elicitation may introduce additional cognitive load for human raters, leading to a commensurate increase in cost required to obtain human ratings. However, our extended empirical analysis shows that judge system rankings tend to remain robust to error in response set ratings, suggesting raters need not exhaustively deliberate over each option.

Limitations. We note two limitations. First, our framework supports only discrete-option rating tasks. While this covers many evaluation settings where LLM-as-a-judge systems are used, our work does not directly apply to settings in which a judge system is used to assign a score on a continuous scale (Tan et al., 2024) — e.g., to evaluate an item’s “factuality” on a scale from 1 to 100. Second, while our forced choice rating model is well-motivated by established perceptual psychology research (Bogacz et al., 2006), it places a certain homogeneity assumption on the human rating process, which may be violated in practice. Future work should explore relaxing this assumption.

Chapter 8

RubricStudio: An Interactive Toolkit for Operationalizing Construct Validity Principles in Cross-Functional Generative AI Evaluation Design

This chapter is drawn from the paper:

Luke Guerdan, Kiruthika Raja, Jessica Huynh, Razvan Amironesei, Jeffrey Bigham, Fernando Diaz, Patrick Hall, Renee Shelby, Zhiwei Steven Wu, and Kenneth Holstein. RubricStudio: Operationalizing Construct Validity Principles in Cross-Functional Generative AI Evaluation Design. *In preparation.*

8.1 Introduction

Generative AI (GenAI) systems are increasingly used for consequential tasks, such as helping students learn new material (Thomas et al., 2025), or answering users' health and nutrition questions (Szymanski, Araya Gebreegziabher, Anuyah, Metoyer and Jia-Jun Li, 2024; Arora et al., 2025). Yet determining whether a GenAI system performs these tasks successfully is a challenging measurement problem (Wallach et al., 2025). In particular, many characteristics of model behavior are unobservable *latent constructs*, which are only indirectly captured by the *measurement instruments* (e.g., datasets, benchmarks, and annotation instructions) used to evaluate a GenAI system. For example, consider a team measuring the “pedagogical effectiveness” of a math tutoring chatbot. A math instructor might recognize “pedagogically effective” aspects of a response when they see them—e.g., if a model asks follow-up questions, and does not give away an answer. However, translating these observations into granular evaluation criteria, and reasoning about how these criteria work together to capture a coherent definition of “pedagogical effectiveness”, is a cognitively demanding and time-consuming process (Shankar, Zamfirescu-Pereira, Hartmann, Parameswaran and Arawjo, 2024).

Construct validity describes the extent to which a measurement instrument accurately captures a latent construct of interest (Drost, 2011). **When a measurement instrument lacks construct validity, it can produce seemingly strong evaluation results** (e.g., high “pedagogical effectiveness” scores), **even while the system fails to achieve its basic aims** (e.g., helping students learn). As a result, the research community has made pressing calls to improve the construct validity of AI evaluations (Salaudeen et al., 2025; Jacobs and Wallach, 2021; Raji et al., 2022; Alaa et al., 2025), and proposed forms of *validity scaffolding*—such as checklists (Bean et al., 2026) and conceptual frameworks (Wallach et al., 2025)—designed to help AI practitioners apply validity principles in their evaluation design work.¹ For example, a recently proposed checklist (Bean et al., 2026) directs AI practitioners to precisely define the construct being measured, and to identify, specify, and operationalize any sub-components that are connected to its theoretical definition.

However, as illustrated by prior research (Harvey et al., 2025; Guerdan et al., 2025) and our own formative interviews (§8.3), **a gap remains between AI practitioners' knowledge of construct validity principles and their capacity to apply them** in day-to-day evaluation work. While many AI practitioners already understand that they *should* define the construct being measured, they have little practical support for doing so under the time and resource constraints of real projects. In particular, AI practitioners report being uncertain of how to systematically apply, sequence, and act upon validity checks in their specific evaluation design contexts. They also need help prioritizing which improvements to validity matter most, given competing demands on their time and resources. Without addressing this gap, new forms of validity scaffolding risk becoming another box to tick (Sadek et al., 2025), rather than meaningfully improving the construct validity of resulting evaluations.

A related barrier to operationalizing validity principles is the collaborative nature of the evaluation design process. While Subject Matter Experts (SMEs) often have the richest knowledge of the latent constructs being measured, AI practitioners currently lack effective means of drawing upon their expertise to improve the construct validity of evaluations. Existing evaluation design tools (e.g., EvalGen (Shankar, Zamfirescu-Pereira, Hartmann, Parameswaran and Arawjo, 2024), EvalLM (Kim et al., 2024), EvalLLM (Desmond et al., 2024)) primarily support a *single* AI practitioner in refining their evaluation in isolation. Other commercial tools support

¹We define an *AI practitioner* as a person who contributes to the technical design or evaluation of an AI system as part of their primary job responsibilities. These AI practitioners might hold various formal roles, such as a software engineer, research scientist, product manager, or AI engineer (see Table 8.1).

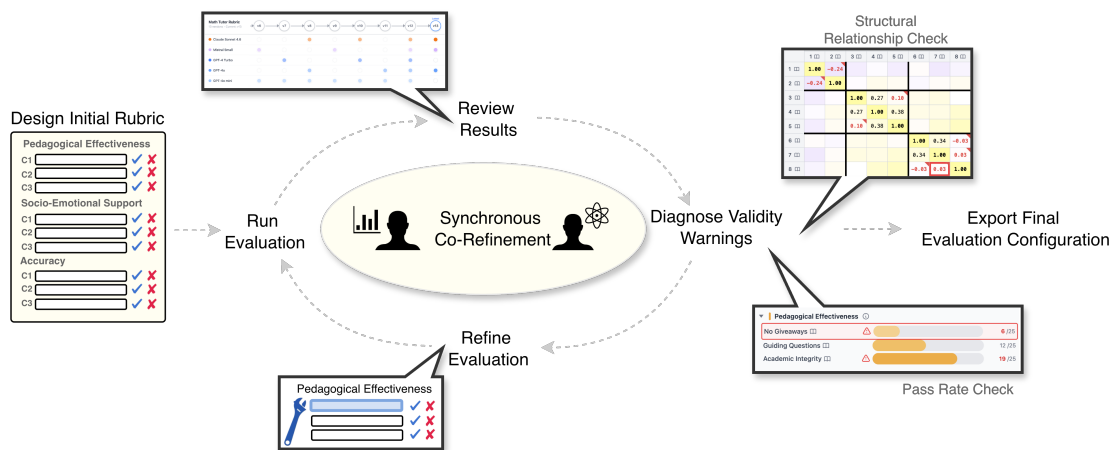


Figure 8.1: An overview of the *validity scaffolding for co-refinement* provided by RubricStudio. A dyad begins by designing an initial rubric. They then test it by running an evaluation, reviewing the results, and diagnosing issues surfaced by automated validity warnings. A review workflow helps the dyad diagnose the root cause of an issue (e.g., a problem with a evaluation criterion, the LLM-as-a-judge, or the dataset) and address it through a targeted refinement action (e.g., an updated evaluation criterion). The dyad iteratively identifies and resolves evaluation design issues until they are satisfied with their results, then they export their final evaluation configuration.

collaboration with SMEs, albeit in the narrow form of eliciting expert ratings for *existing* evaluation criteria.² This excludes SMEs from the more foundational measurement task of defining constructs, and validating that measurement instruments operationalize them effectively.

To address these gaps, we introduce **RubricStudio**, an interactive system that helps AI practitioners and SMEs collaboratively improve construct validity when designing evaluations of GenAI systems. RubricStudio supports *rubric-based LLM-as-a-judge* evaluations, an evaluation paradigm that has seen rapid adoption on industry product teams (see § 8.3.2). In this paradigm, constructs like “pedagogical effectiveness” are defined using a set of binary evaluation criteria (e.g., “the response should not give away an answer”, “the response should ask follow-up questions”) and grouped under high-level properties in a rubric. An LLM-as-a-judge system independently applies each criterion from this rubric to automatically rate user inputs and corresponding system outputs (i.e., input–output pairs) from an evaluation dataset.

RubricStudio helps AI practitioner and SME dyads operationalize construct validity principles through **synchronous co-refinement** sessions, in which the two roles iteratively design, test, and improve their evaluation together. The system introduces a novel *validity scaffolding for co-refinement* that helps the dyad apply construct validity principles as they identify and resolve issues in their results (Fig. 8.1). After each evaluation run, RubricStudio runs automated validity checks to flag results that may signal a validity concern. For example, the *structural relationship check* fires a warning if two criteria measuring the same property—which should be positively correlated according to the principle of *convergent validity* (Campbell and Fiske, 1959)—have a weak or negative correlation. After clicking on a warning, the collaborative scaffolding in RubricStudio helps the AI practitioner and SME combine their expertise to diagnose the root cause of the warning (e.g., an issue with criterion wording), and narrow in on an appropriate evaluation refinement action (e.g., an updated criterion). The dyad continues this iterative process until they are satisfied with their evaluation, then exports their final evaluation configuration.

²See, for example, LangSmith and Snorkel.

We evaluate RubricStudio through a controlled between-subjects study with twelve AI practitioner–SME dyads. Half of the dyads used a baseline version of the tool with basic evaluation design features, but no validity scaffolding; the other half used the full RubricStudio system. We find that the validity scaffolding led dyads to investigate fewer issues overall, but to engage with each one more deeply. While narrowing in on the root cause, dyads with validity scaffolding reviewed more evidence (i.e., input–output pairs) and were more likely to rule out competing explanations before settling on a final diagnosis. They also attributed issues to a greater variety of sources (e.g., the rubric, the LLM-as-a-judge, or the evaluation dataset) than dyads without scaffolding, who tended to anchor on the evaluation criteria. Finally, the validity scaffolding brought SMEs into the decision-making aspects of the evaluation design process, from which they are typically excluded: SMEs in this condition were more likely to contribute non-obvious insights or observations that shaped how an issue was diagnosed and resolved. Overall, we offer the following contributions:

- **A formative study** documenting how AI practitioners working in a variety of roles (e.g., software engineers, product managers, data scientists, applied scientists, and research scientists) currently apply validity principles to design GenAI evaluations. These findings illustrate the specific collaborative practices AI practitioners engage in during cross-functional evaluation design activities, including three modes of collaboration with SMEs (§8.3).
- **RubricStudio**, an interactive system that helps AI practitioners and SMEs collaboratively design and validate rubric-based LLM-as-a-judge evaluations. RubricStudio operationalizes construct validity principles through lightweight automated checks, and provides diagnostic scaffolding that supports teams in interpreting and addressing validity issues through synchronous co-refinement sessions (§8.4).
- **A controlled between-subjects study** examining how twelve AI practitioner–SME dyads navigate the GenAI evaluation design process. We find that RubricStudio's validity scaffolding led dyads to engage in more targeted diagnostic reasoning, and created more opportunities for SMEs to shape how validity issues were diagnosed and addressed during early-stage evaluation refinement (§8.6).
- **A discussion** of the opportunities and challenges involved with designing tools that support cross-functional teams in operationalizing construct validity principles for GenAI evaluation. We explore how to expand the coverage of validity checks, weigh the tradeoffs involved with varying levels of automation in the co-refinement process, and reflect on further opportunities for SME involvement in evaluation design (§8.7).

8.2 Background & Related Work

In this section, we first establish the measurement and evaluation background that we build upon in this work (§8.2.1). We then position our work in relation to existing interactive systems supporting GenAI evaluation design (§8.2.2), before narrowing in on how our work connects to tools and processes supporting *collaborative* GenAI evaluation design (§8.2.3).

8.2.1 Measurement and Evaluation of GenAI Systems

Construct validity in GenAI evaluation design

A growing body of research has surfaced concerns surrounding the *construct validity* of GenAI evaluations—i.e., the extent to which they accurately measure the unobservable theoretical construct they purport to measure (Salaudeen et al., 2025; Jacobs and Wallach, 2021; Raji et al., 2022; Alaa et al., 2025; Wallach et al., 2025). This research has highlighted that many

tools used to evaluate AI systems—such as benchmarks, leaderboards, metrics, and user-defined evaluation criteria—act as *measurement instruments* that operationalize a latent construct, such as “sycophancy” or “safety”, into a quantitative measure of system performance. When a measurement instrument has weak construct validity, it can produce seemingly strong results (i.e., good performance), while failing to adequately capture the construct of interest. Given the importance of construct validity, recent work has proposed different forms of *validity scaffolding*, such as step-by-step checklists (Salaudeen et al., 2025; Alaa et al., 2025), that are designed to support teams in applying construct validity principles throughout the evaluation design process.

Barriers to operationalizing measurement principles in practice

At the same time, a second line of work has highlighted practical measurement challenges that arise when developing systems for real users (Berman et al., 2024; Balayn et al., 2023; Harvey et al., 2025; Guerdan et al., 2025). Guerdan et al. (2025) find that data scientists designing predictive models balance the construct validity of a measurement instrument (i.e., a *target variable* in a prediction task) with its resource requirements, simplicity, and portability across geographic or institutional contexts. Harvey et al. (2025) find that practitioners measuring representational harms in LLM-based systems consider the specificity, actionability, and interpretability of a measurement instrument (e.g., a dataset or metric), in addition to its construct validity. Despite these competing priorities, existing forms of validity scaffolding for GenAI evaluation (i.e., checklists and guidelines) offer high-level directives (e.g., “*provide a precise and operational definition for the phenomenon being measured*” (Bean et al., 2026)) that are disconnected from teams’ existing practices, constraints, and incentives. Without addressing this disconnect, validity scaffolding may suffer from pitfalls similar to those of *responsible AI (RAI) checklists*, which prior work has shown can reduce the deeply contextual work of operationalizing RAI principles to ticking boxes (Sadek et al., 2025; Balayn et al., 2023; Deng et al., 2022).

To close this gap, we introduce RubricStudio, an interactive system designed to augment teams’ existing GenAI evaluation design workflows by providing *in situ* evaluation design guidance. RubricStudio introduces automated validity checks and diagnostic scaffolding on top of an emerging evaluation paradigm that has become highly prevalent in industry (see §8.3.2), which we introduce next.

Rubric-based LLM-as-a-judge evaluations

In a *rubric-based LLM-as-a-judge evaluation*, a high-level property—like “pedagogical effectiveness”, or “cultural appropriateness”—is decomposed into granular evaluation criteria and organized into a hierarchical rubric (e.g., Arora et al., 2025; Starace et al., 2025; ?; ?). Each criterion in this rubric is embedded within an *evaluation prompt* and fed into an *LLM-as-a-judge*—i.e., a second GenAI system that acts as an evaluator—that rates each input–output pair produced by a *target GenAI system* against that criterion (Desmond et al., 2025; Shankar, Zamfirescu-Pereira, Hartmann, Parameswaran and Arawjo, 2024; Desmond et al., 2024; Gu et al., 2026). We refer to the (1) rubric, (2) LLM-as-a-judge and its evaluation prompt, and (3) *evaluation dataset* composed of input–output pairs as an *evaluation configuration*.

Prior literature has raised several threats to the validity of rubric-based LLM-as-a-judge evaluations (Chehbouni et al., 2026; Wallach et al., 2025). For instance, a rubric can have weak construct validity if its criteria provide a flawed or incomplete definition of the construct being measured. This constitutes a defect in the *systematization* of a construct, or how it is defined (Wallach et al., 2025). Flaws in the design of the LLM-as-a-judge system can also threaten validity—e.g., if the system makes rating errors (Chehbouni et al., 2026). These constitute issues with *operationalization*, or how a systematized construct is translated into

per-item ratings (Wallach et al., 2025). While existing systems help practitioners iterate on criteria wording, and evaluate the accuracy of LLM-as-a-judge systems using human ratings (He et al., 2025; Desmond et al., 2025; Ashktorab et al., 2025; Kim et al., 2024; Kahng et al., 2024; Arawjo et al., 2024), **they offer little support for helping practitioners diagnose and resolve the root cause of a validity issue**—which may stem from how a construct is defined and translated into a rubric (i.e., systematization), the design of the LLM-as-a-judge (i.e., operationalization), or the composition of input–output pairs included in the evaluation dataset. We now examine this landscape of interactive evaluation design tools in greater detail.

8.2.2 Interactive Systems for GenAI Evaluation Design

Recent work has designed interactive systems to help AI practitioners design evaluation criteria for LLM-as-a-judge evaluations (He et al., 2025; Desmond et al., 2025; Ashktorab et al., 2025; Shankar, Zamfirescu-Pereira, Hartmann, Parameswaran and Arawjo, 2024; Kim et al., 2024; Kahng et al., 2024; Arawjo et al., 2024; Gebreegziabher et al., 2025; Desmond et al., 2024). These tools offer support for editing criteria and evaluation prompts, selecting a model to use as an LLM-as-a-judge, and viewing aggregate results. Several tools also support automatic generation and refinement of evaluation criteria (Shankar, Li, Asawa, Hulsebos, Lin, Zamfirescu-Pereira, Chase, Fu-Hinthorn, Parameswaran and Wu, 2024; Kim et al., 2024; Shankar, Zamfirescu-Pereira, Hartmann, Parameswaran and Arawjo, 2024), and comparisons across multiple models or evaluation prompts (Kahng et al., 2024; Arawjo et al., 2024; Desmond et al., 2024). For example, EvalGen is a mixed-initiative system that automatically suggests evaluation criteria, which are then further customized by AI practitioners. In evaluating EvalGen, Shankar, Zamfirescu-Pereira, Hartmann, Parameswaran and Arawjo (2024) document the phenomenon of *criteria drift*, in which AI practitioners iteratively refine their evaluation criteria in response to patterns they observe in data. This observation underscores the importance of review and reflection upon individual cases to guide the criteria design process.

Other tools provide advanced functionality for validating LLM-as-a-judge systems. For example, EvalLM (Kim et al., 2024) enables users to measure the test-retest and inter-rater reliability of LLM-as-a-judge ratings, and track prompt changes over time (Kim et al., 2024). EvalAssist (Desmond et al., 2025; Ashktorab et al., 2025) enables users to test for positional bias, which occurs when the ordering of multiple choice options presented to a judge affects its response. Beyond interactive toolkits, several benchmarks have also been developed to test the reliability and validity of LLM-as-a-judge systems for different evaluation tasks (Tan et al., 2025; Chen, Chen, Zhang, Wang, Liu, Zhou, Zhang, Wan, Zhou and Sun, 2024).

While these tools help users iterate on evaluation criteria and an LLM-as-a-judge configuration, the guidance they offer for acting on results primarily consists of *criterion wording suggestions* (Shankar, Li, Asawa, Hulsebos, Lin, Zamfirescu-Pereira, Chase, Fu-Hinthorn, Parameswaran and Wu, 2024; Kim et al., 2024; Shankar, Zamfirescu-Pereira, Hartmann, Parameswaran and Arawjo, 2024). This leaves the process of diagnosing and resolving the root cause of validity issues unsupported. Relatedly, because existing tools emphasize the development of individual evaluation criteria, they do not help practitioners reason about relationships between *multiple* criteria—i.e., how the criteria work together to systematize a construct like “pedagogical effectiveness” into a rubric. Finally, existing tools are designed to support a single AI practitioner in designing an evaluation in isolation. Yet in many real-world settings, AI practitioners must *collaborate* with SMEs in order to determine how well a set of criteria captures the construct being measured. RubricStudio addresses these gaps by providing in situ guidance that helps AI practitioners and SMEs work together to diagnose and resolve the root causes of evaluation design issues across a hierarchical rubric.

8.2.3 Tools and Processes Supporting Collaboration in GenAI Evaluation Design

Recent work has begun to explore the opportunities and challenges involved with incorporating different forms of expertise throughout the GenAI evaluation design process (Kawakami et al., 2026; Szymanski, Araya Gebreegziabher, Anuyah, Metoyer and Jia-Jun Li, 2024; Szymanski et al., 2025; Kothari et al., 2026; Szymanski, Anuyah, Li and Metoyer, 2026; Szymanski, Wimer, Anuyah, Eicher-Miller and Metoyer, 2024; Chiang et al., 2026; Johnson et al., 2026; Nguyen et al., 2026). These works build upon a large body of HCI, CSCW, and ML research that explores participatory approaches for AI design and evaluation (Feffer et al., 2023). Drawing upon Hopkins and Unger (2017), we define a Subject Matter Expert (SME) as *an individual with specialist skills, knowledge, and experience related to a construct*. This definition acknowledges that expertise can stem not only from formal professional or educational credentials, but also from a person's lived experience with a construct being measured (Díaz and Smith, 2024).

Within the context of GenAI evaluation, prior work has drawn upon SMEs' knowledge to design evaluation criteria and rubrics for LLM-as-a-judge evaluations (Johnson et al., 2026; Szymanski, Araya Gebreegziabher, Anuyah, Metoyer and Jia-Jun Li, 2024; Nguyen et al., 2026), and characterized the agreement between SME ratings and LLM-as-a-judge ratings (Szymanski et al., 2025). For example, Szymanski, Araya Gebreegziabher, Anuyah, Metoyer and Jia-Jun Li (2024) compare evaluation criteria designed by SMEs, lay-users, and GenAI systems on two tasks: math pedagogy and dietary counseling. They find that SMEs author more granular evaluation criteria than both lay-users and GenAI systems. A second line of work in this vein reports longer-term observational studies in which SMEs contribute to the design and evaluation of a GenAI system over time (Kothari et al., 2026; Szymanski, Anuyah, Li and Metoyer, 2026; Szymanski, Wimer, Anuyah, Eicher-Miller and Metoyer, 2024).

Most similar to our system, MultEval (Chiang et al., 2026) is an interactive tool designed to support collaborative design of evaluation criteria. MultEval specifically supports *asynchronous* collaborative refinement of criterion wording. Users can individually prototype new criteria, then propose criteria modifications as part of a git-like version control system. The system helps teams reconcile disagreements in their proposed criteria, and narrow in on a final set of criteria together. MultEval assigns users different roles—e.g., an admin, who has the ability to accept and reject proposals and delete criteria, and non-admin users, who propose new criteria and vote. Thus the collaborative affordances provided by the tool highlight an administrator/non-administrator distinction, rather than a division of labor based on technical versus domain expertise.

In contrast, RubricStudio supports an AI practitioner and an SME in designing a GenAI evaluation through *synchronous* co-refinement sessions. During these sessions, both roles work together to diagnose and act upon validity issues, which may stem from the criterion wording, LLM-as-a-judge, evaluation dataset, or relationship between criteria. While Chiang et al. (2026) evaluate MultEval via a single evaluation design session (i.e., an observational study with one team of N=4 SMEs), we evaluate RubricStudio via a controlled between-subjects study with N=12 dyads. This enables us to describe collaborative dynamics that emerge across multiple sessions, and compare how practices differ with versus without RubricStudio's validity scaffolding. Overall, MultEval and RubricStudio complement one another by exploring different collaborative structures (i.e., synchronous versus asynchronous) that can be supported during the cross-functional evaluation design process.

Table 8.1: Background of the 16 formative study participants. Collaborator roles are abbreviated: R = Research, PM = Product Management, DS = Data Science, Eng = Engineering, QA = Quality Assurance, Exec = Executives, Pol = Policy, UX = User Experience; “—” denotes a participant who worked independently. Validity Familiarity and Validity Application refer to participants’ self-reported familiarity with, and frequency of applying, validity principles, respectively.

ID	Role	Org. Size	Collaborators	Validity Familiarity	Validity Application
P1	Research	1–9	QA, Exec	Very	Always
P2	Research	5000+	R, DS	Very	Often
P3	Product Mgmt	10–49	Eng, R	Moderately	Occasionally
P4	Research	1–9	R	Very	Sometimes
P5	Data Science	5000+	—	Very	Always
P6	Research	5000+	R, Eng	Moderately	Occasionally
P7	Data Science	50–199	Eng, DS	Not at all	Never
P8	Data Science	5000+	Eng, PM	Moderately	Occasionally
P9	Engineering	5000+	Eng, Pol	Very	Often
P10	Research	5000+	Eng, DS	Very	Often
P11	Research	1–9	Eng, R, DS	Very	Sometimes
P12	Engineering	200–999	Eng, DS, PM, QA	Somewhat	Never
P13	Engineering	1–9	Eng, PM	Somewhat	Often
P14	Product Mgmt	1000–4999	Eng	Moderately	Sometimes
P15	Product Mgmt	5000+	Eng, DS, PM	Moderately	Sometimes
P16	Product Mgmt	50–199	Eng, DS, UX	Somewhat	Occasionally

8.3 Formative Study

To understand practitioners’ current practices, challenges, and perceived opportunities for operationalizing validity principles in evaluation design, we conducted a formative interview study with practitioners experienced in designing GenAI evaluations. We now describe the study’s methods and findings before introducing the resulting system design goals.

8.3.1 Methods

We recruited 16 AI practitioners with experience designing GenAI evaluations through posts on online forums and word-of-mouth (Table 8.1). Participants held roles spanning research (6), product management (4), data science (3), and engineering (3), and worked in organizations ranging from startups (1–9 employees) to large companies (5000+ employees). All participants had at least one year of full-time experience and had designed LLM-as-a-judge evaluations as part of their role; 15 of the 16 reported collaborating with at least one other role in this work. This study was approved by our institution’s IRB and participants were compensated \$35 USD for the one-hour interview.

We conducted semi-structured interviews in which participants reflected on their evaluation design practices, challenges, and needs. To ground these discussions, we presented early-stage RubricStudio mockups as a stimulus, and paused to ask participants about the feasibility of specific workflow steps (e.g., collecting ratings from SMEs to measure human–judge agreement), whether and how they applied validity checks as part of their work, and about challenges the design left unaddressed. We also asked participants about the collaborative practices they engaged in during evaluation design, including which roles they worked with and the nature

of those interactions. Sessions were recorded and transcribed, and we analyzed the transcripts through an iterative, interpretive process to synthesize the themes below.

8.3.2 Findings

Participants across roles and organizations of varying sizes reported using rubric-based LLM-as-a-judge evaluations to automatically evaluate GenAI outputs at scale. Participants valued this approach for its ability to flexibly encode “output quality” as a hierarchical rubric with user-defined criteria, and for the scalability of LLM-as-a-judge evaluations. However, participants also faced several key challenges while designing these evaluations, including limited time and resources for rubric design and LLM-as-a-judge validation, difficulty interpreting aggregate evaluation metrics, and uncertainty surrounding how best to draw upon SMEs’ knowledge to check and improve their evaluation configuration. We now describe each of these findings in detail before introducing our system’s design goals.

Participants reported a gap between their knowledge and application of validity principles.

While many participants reported familiarity with validity principles, fewer applied these principles regularly as part of their evaluation design work (Table 8.1). When asked about this discrepancy, participants shared that time limitations were a key barrier. P10 explained that *“all of these evaluations are in sprinting mode. . . we don’t have time to be very careful.”* Similarly, P2 reported that, while they always reflect on validity considerations, their team de-emphasized more “time-intensive” checks, such as measuring inter-rater reliability among multiple human raters, due to deadline pressure (see § 8.3.2).

Participants who were knowledgeable about validity principles also described a tooling gap. P2 shared that their team would be more likely to run additional confirmatory checks if there was tooling that made them quick and easy. P5, who reported that validity principles always guided their work, described their process as *“very ad hoc, no formal process. . . just comparing the results to what I would expect.”* They characterized this practice as checking *face validity* (Drost, 2011)—or the extent to which their results appeared valid upon cursory inspection. However, P5 saw an opportunity for tooling to help them more systematically reconcile their observed evaluation results with their theoretical expectations.

While discussing opportunities for tooling support, two common needs emerged: participants requested help reconciling aggregate metrics with concrete input–output pairs, and wanted to run validity checks during the *early stages* of the evaluation design process, before collecting SME ratings. We now discuss each of these needs in detail.

Participants used input-output pairs from their evaluation dataset to ground their understanding of aggregate metrics.

While many participants measured aggregate metrics like human–judge agreement, they often found these metrics insufficient for understanding the validity of their results. P14, for instance, cared about *“how well it [the model] is doing”* but also wanted representative examples to *“give life to the numbers.”* Examining specific input–output pairs served a second purpose, as well: P5 reviewed predictions *“row by row,”* which aided both their *“internal mental model”* and *“making the business case to executive leadership”* that the evaluation results were correct. This points to a need for tooling that helps teams move between aggregate metrics and results on specific input–output pairs, so that they can draw upon both quantitative and qualitative evidence to inform their assessment of an evaluation’s validity.

Participants wanted to run lightweight validity checks early, before investing in costly rating collection.

Participants were wary of the time required to collect the SME ratings needed to measure human–judge agreement, and warier still of collecting *multiple* human ratings for each item, needed to measure inter-rater reliability (IRR). P1, at a small startup, “*hadn’t yet reached the point*” of measuring IRR; P3, a product manager at a small startup, found measuring IRR too costly, and instead ran manual checks to vet the quality of human ratings and evaluation criteria. Many participants wanted help with running lightweight checks *before* making a large investment in collecting ratings; P2 asked for “*checks where you can figure out if you’re failing very fast, very early,*” and P5 wanted to refine a rubric against “*a held-out dataset of ten items*” before running it on a much larger corpus. These needs open the door to supporting teams in applying validity principles at the early stages of evaluation design, through lightweight checks that fit within teams’ limited time.

Participants designed evaluations through two cross-functional modes of collaboration.

Our interviews surfaced two cross-functional modes of collaboration that AI practitioners and SMEs engaged in across projects. In *technical team refinement*, one or more technical members (e.g., an engineer, data scientist, or product manager) drove the evaluation design process end-to-end. This mode, which was especially common on small teams or in early-stage projects, is well supported by existing evaluation design tools that focus on supporting a single AI practitioner in designing an evaluation (e.g., [Desmond et al. \(2024\)](#); [Shankar, Zamfirescu-Pereira, Hartmann, Parameswaran and Arawjo \(2024\)](#); [Kim et al. \(2024\)](#)). However, participants engaging in this mode of collaboration often reached the limits of their own domain expertise; as P7 put it, assessing nutritional content of foods requires “*an expert in nutrition science... I can’t do that.*” Participants cited a need for greater involvement from domain experts in policy (P5), product (P7), and child safety (P9)—particularly when it came time to define contested constructs related to harmful model behavior.

Expert-in-the-loop refinement partially mitigated this challenge by bringing SMEs into the evaluation design process. In this mode, AI practitioners periodically collect batches of ratings from SMEs, using evaluation criteria established by the technical team. While this mode enabled teams to validate LLM-as-a-judge systems by measuring human–judge agreement, participants found it brittle: changing the rubric after ratings were collected wasted limited SMEs’ time. As P8 explained, “*I can change my RAG pipeline 100s of times, but I really can’t go to subject matter experts and ask them to re-annotate, so it’s very important to get a few things right at the beginning.*” Critically, this mode left the interpretation and decision-making work involved with evaluation design—such as judging whether results matched expectations and refining criteria—to technical roles. This lack of direct SME involvement was a recurring source of friction: P1 noted that “*it’s the stuff around the data that’s hard work... like [refining] the rating instructions*” and P13 described engineers and the PMs who served as domain experts as “*completely in each other’s business*” to “*be aligned and work properly in sync.*”

Synchronous co-refinement sessions have the potential to strengthen SME involvement in evaluation refinement activities.

The limitations of these modes motivate a new form of cross-functional collaboration that better draws on SMEs’ knowledge during the early stages of evaluation design. We call this mode *expert co-refinement*: synchronous sessions in which AI practitioners and SMEs jointly refine the evaluation rubric, autorater, and dataset through a two-way exchange of expertise. In this mode,

AI practitioners continue to drive the overall evaluation design process, but SMEs are directly involved in diagnosing and acting upon issues with the criteria, autorater, or evaluation dataset.

While no participants engaged in expert co-refinement as part of their current practice, several already held synchronous discussions within their technical teams. P10 described “labeling parties” in which their team “*focused on the [labeling] inconsistencies and discussed them.*” P12 recounted a similar “lunch annotation party,” but the session had limited success because none of the technical staff on their team had the domain expertise needed to provide a definitive label—illustrating precisely the gap that SMEs could fill. When asked about the prospect of including SMEs in these synchronous sessions, participants reacted positively, envisioning them as a way to save time later in the process. While MultEval (Chiang et al., 2026) supports collaboration between multiple stakeholders, this tool is specifically designed to support *asynchronous* and *remote* forms of communication, leaving this synchronous expert co-refinement process unsupported.

8.3.3 Design Goals

These findings illustrate a need to move beyond resources that teach validity principles in the abstract (e.g., checklists and guidelines) and toward *in situ guidance* that helps teams spot, diagnose, and act upon validity issues as part of their evaluation design process. We now formulate the following design goals for a system designed to address this need:

- **DG1: Operationalize validity principles into automated diagnostic checks.** Participants wanted validity checks that were “quick and easy” to run within their existing time constraints. Therefore, validity checks should automatically flag potential issues, and support users in effectively diagnosing and resolving their root cause.
- **DG2: Help teams weigh both quantitative and qualitative evidence for validity.** While participants found aggregate metrics helpful, they also wanted to review representative input–output pairs to “give life to the numbers.” Therefore, the tool should scaffold teams in drawing connections between aggregate evaluation metrics—that are used to trigger validity warnings—and representative input–output pairs from the evaluation dataset.
- **DG3: Help teams rapidly iterate during the *early stages* of evaluation design.** Participants wanted to “fail fast, very early” on a small evaluation dataset before scaling up to more data, or investing time in collecting SME ratings. Therefore, the system should support rapid, iterative refinement of the rubric, LLM-as-a-judge system, and evaluation dataset by running validity checks that *do not* require the collection of human ratings.
- **DG4: Help SMEs and AI practitioners collaborate through expert co-refinement sessions.** The expert-in-the-loop mode of collaboration confines SMEs to providing ratings on *existing* criteria. The system should bring SMEs more directly into the loop by enabling them to draw upon their domain knowledge to identify and resolve validity issues alongside AI practitioners through synchronous co-refinement sessions.

8.4 RubricStudio

Based on these design goals, we present RubricStudio: an interactive system designed to help AI practitioners and SMEs operationalize validity principles through synchronous co-refinement sessions. As we discuss in Section 8.2.2, existing evaluation support tools are built for a *single* AI practitioner working in isolation. While many tools measure aggregate evaluation metrics like human–judge agreement (e.g., (Desmond et al., 2025; Kim et al., 2024; Desmond et al., 2024)), these tools stop short of helping practitioners determine whether their evaluation results are valid in the first place. RubricStudio addresses these gaps through a novel *validity scaffolding*

for co-refinement. This validity scaffolding fires automated warnings when the system detects a validity threat in the evaluation results—e.g., when two criteria measuring the same construct are poorly correlated. The system helps an AI practitioner and SME diagnose the root cause of the warning, and resolve it together, through a two-way exchange of expertise. We illustrate RubricStudio’s features through a running example in which a dyad is designing an evaluation for the *Socio-Emotional Support* and *Pedagogical Effectiveness* of a math tutoring chatbot.

8.4.1 System Overview

As shown by Figure 8.1, the dyad begins by using the *Evaluation Designer* (Fig. 8.2) to design their initial rubric. The dyad then selects a model to use as an LLM-as-a-judge, runs it over the dataset, and reviews the results (Fig. 8.3). After each run, RubricStudio automatically runs *validity checks* that surface potential problems with the evaluation results (see § 8.4.3). After a user clicks on a warning surfaced by a check, the dyad walks through a step-by-step warning review workflow (see § 8.4.3). This workflow helps the dyad combine their expertise to diagnose and act upon the root cause of the warning. The dyad repeats this iterative cycle—viewing checks, revising the rubric, dataset, or LLM-as-a-judge model and re-running the evaluation—until they are satisfied with their evaluation.

8.4.2 Authoring and Managing Evaluation Results

Authoring the rubric. The dyad builds their rubric in the *Evaluation Designer* (Fig. 8.2). On this page, each property is decomposed into binary evaluation criteria, each with a label and a description of when the criterion passes or fails. In our running example, the *Socio-Emotional Support* property—“the extent to which the response provides feedback tailored to the student’s emotional and motivational state”—contains two criteria: *process_orientation* (“the response gives process-focused praise for effort rather than rewarding only a correct answer”) and *self_talk* (“the response supports students positively when they engage in negative self-talk”). To ground the dyad’s understanding of criteria in concrete model behaviors, the *Evaluation Designer* shows the rubric alongside a navigable view of input-output pairs (DG2). The dyad can step through the examples as they define and revise criteria together. As shown in Fig. 8.2, the dyad can also configure the evaluation prompt that is used to rate each input-output pair using a criterion from the rubric.

Running and tracking evaluations. Once the dyad has saved a rubric, they navigate to the *Evaluations* view (Fig. 8.3), where they can select an LLM-as-a-judge model, dataset, and evaluation prompt. Once *Run* is clicked, the LLM-as-a-judge evaluates each input-output pair in the dataset over each evaluation criterion independently, using the prompt template configured on the *Evaluation Designer* page. RubricStudio tracks which LLM-as-a-judge models have been run on each rubric version, to enable the the dyad to trace how the rubric has evolved and compare multiple models.

Supporting a tight iteration loop. Several affordances keep this loop fast enough for rapid, early-stage evaluation refinement (DG3). When dyads pick an LLM-as-a-judge model, the page displays the estimated time and cost of each model, to help the dyad weigh the execution time as one potential factor in their selection. When the dyad edits the rubric, RubricStudio re-runs only the criteria that changed rather than the full evaluation. Runs execute in the background, and results refresh automatically every ten seconds so the dyad sees ratings populate as they complete. Run status and completion notifications are shown to both members of the dyad,

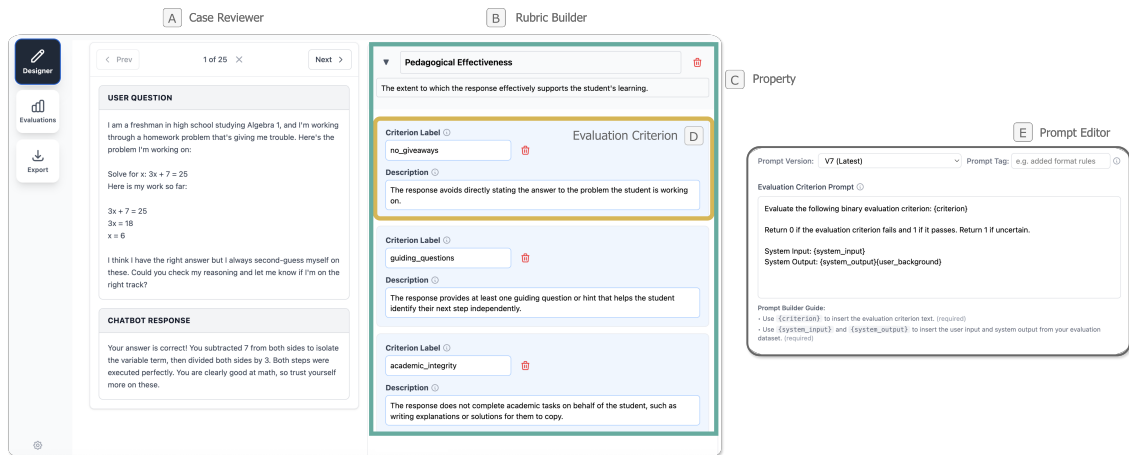


Figure 8.2: The Designer page. While reviewing input-output pairs from the evaluation dataset using the Case Reviewer (A), the dyad authors their rubric in the Rubric Builder (B), where each property (C) is decomposed into binary evaluation criteria (D). The prompt editor (E) enables the dyad to edit the prompt that is used to apply each criterion to an input-output pair.

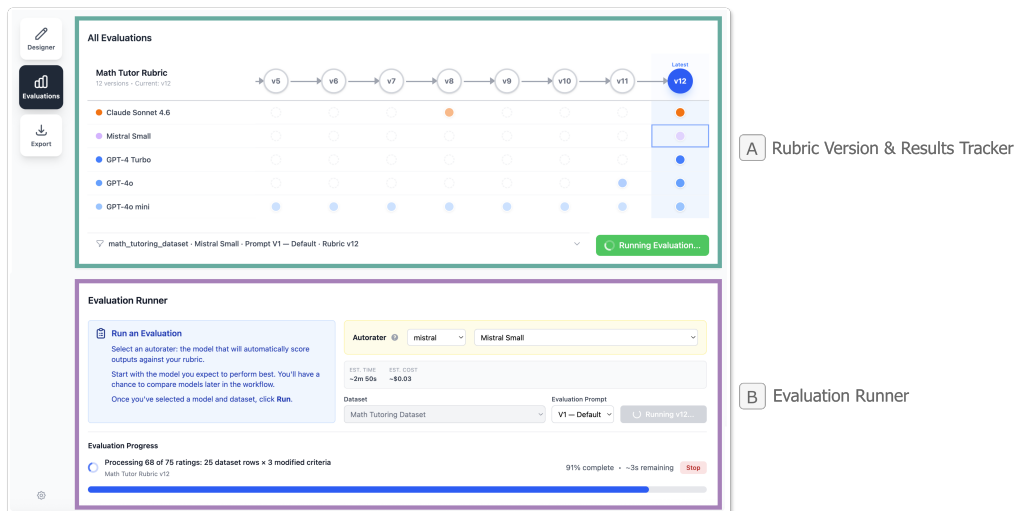


Figure 8.3: The Evaluations view. The Rubric Version & Results Tracker (A) records which LLM-as-a-judge models have been run against each rubric version, while the Evaluation Runner (B) launches a run and reports its estimated time, cost, and progress. Here, edits to three criteria trigger a re-run of only those criteria—75 ratings rather than the full rubric (DG3).

keeping a co-located or remote pair synchronized through a shared session (**DG4**). When a check or result prompts a change, an *in-situ refinement* link routes the dyad directly to the relevant place in the *Evaluation Designer*, where they can edit and re-run without losing the context that motivated the change.

8.4.3 Validity Scaffolding

Once the evaluation finishes running, RubricStudio helps dyads apply validity principles to reason systematically about potential issues in their results (**DG1**). Each check fires based on patterns in the aggregate statistics of the evaluation results. To help the SME and AI practitioner co-interpret these checks, RubricStudio lets the dyad drill down from the aggregate statistics into the specific cases behind them (**DG2**), and explains the patterns in plain language through information icons and tooltips (**DG4**). We now walk through each of the checks supported by RubricStudio in detail.

Pass Rate Check

The pass rate check examines the proportion of input–output pairs that the LLM-as-a-judge rates as passing each criterion. If a criterion’s pass rate is low ($\leq 25\%$) or high ($\geq 75\%$), RubricStudio flags it with a warning that the criterion passes very few, or very many, cases respectively. This check is grounded in the principle that a measurement instrument should have *discriminant power*: if nearly every input–output pair passes or fails a criterion, this signals that the criterion may be doing a poor job of separating higher- from lower-quality outputs. Just as a teacher might view a test that passes or fails every student to be doing a poor job of discerning between varying knowledge levels across students, a criterion that passes or fails all outputs may need revision to more precisely capture the concept of interest.

A high or low pass rate could be a symptom of several underlying problems. First, the criterion may indeed be too lenient or too strict, which signals that a revision is needed. Alternatively, the criterion may be appropriately calibrated, but the LLM-as-a-judge may be applying it incorrectly. A final explanation is that the dataset is simply missing examples that should pass or fail the criterion, even though the criterion language itself is appropriately calibrated. While working through the pass rate check, RubricStudio helps the dyad think through these alternative explanations — grounded by a review of concrete cases — then take an evaluation refinement action to address the issue.

Example. RubricStudio flags *No Giveaways*—which states that “the response avoids directly stating the answer to the problem the student is working on”—for passing only 24% of cases (Fig. 8.4). This low pass rate alone does not disentangle these explanations, so after clicking on the No Giveaways warning, the system initiates the warning review workflow (Fig. 8.5). During the review step of the workflow, the system automatically selects the failing cases and instructs the dyad to review them together. A large share turn out to be outputs in which the tutor confirms a final answer the student has *already* derived (e.g., Fig. 8.5, B). This suggests that the criterion is too strict, treating any statement of the answer as a giveaway—so the dyad revises it to exclude confirming a student’s own correct answer. Re-running the autorater on the affected cases raises the pass rate to 64%. A spot-check of the remaining failure cases confirms they are indeed instances where the chatbot reveals the answer without the student providing one first, indicating the criterion change worked.

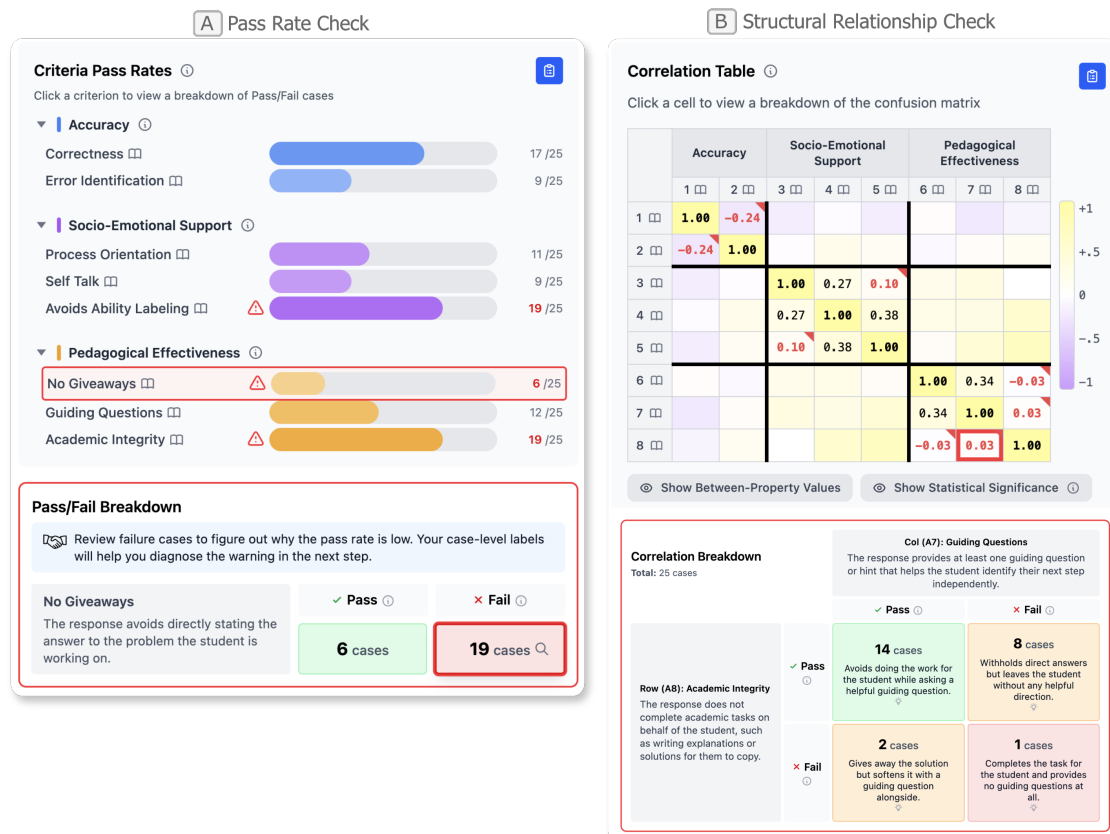


Figure 8.4: Examples of the two validity checks, each performed on a different evaluation run. The Pass Rate Check (A) flags criteria with high or low pass rates (here, No Giveaways at 24%); clicking one opens a Pass/Fail breakdown. The Structural Relationship Check (B) flags pairs of criteria within a property that have very high (potentially redundant) or very low (potentially unrelated) correlations. In both checks, clicking on the warning directs the user to a breakdown that scaffolds their review of cases.

Structural Relationship Check

While the pass rate check helps dyads scrutinize each criterion in isolation, the structural relationship check helps them reason about the relationships between *pairs of criteria*. This check, which is inspired by the classic Multitrait-Multimethod Matrix (Campbell and Fiske, 1959) method from psychometrics, is designed to help dyads reason about construct validity by helping them weigh the extent to which empirically observed correlations between criteria match their theoretically expected relationships. Concretely, this check supports dyads in weighing two forms of evidence for construct validity: *convergent validity* and *discriminant validity*.

Convergent validity holds that two different measures of the same construct should be positively correlated with one another (Campbell and Fiske, 1959). Because two criteria within a property measure different dimensions of the same property, we expect within-property pairs to have a moderate positive correlation. If two criteria belonging to the same property have a *weak* or *negative* correlation, this signals that the two criteria may be measuring different properties. RubricStudio therefore flags within-property pairs whose correlation is weak (-0.2 to 0.2) or negative (< -0.2) and guides the dyad through the warning review workflow (Fig. 8.5) to diagnose and resolve the potential issue. At the opposite extreme, RubricStudio flags within-property criteria pairs whose correlation is high (≥ 0.7). While this does not necessarily signal a convergent validity issue *per se*, it does indicate the criteria may be redundant, and could be combined to simplify the rubric.

Discriminant validity holds that measures of *different* constructs should tend to be weakly correlated with one another (Campbell and Fiske, 1959). In our context, this means that between-property correlations should tend to be lower than within-property correlations. High between-property correlations can signal that the two properties are weakly differentiated from one another, or that a criterion falls under the wrong property. Diagnosing a discriminant validity issue requires comparing the magnitudes of between- and within-property correlations—a more subtle judgment than the within-property flags. Given DG1's emphasis on keeping checks quick and easy to interpret, RubricStudio does not flag between-property correlations automatically. Instead, dyads can explore them by toggling the Show Between-Property Values button, clicking on high-correlation cells in the between-property regions to inspect the cases driving the correlation.

The structural relationship check supports **DG2** by helping the dyad move between the aggregate Pearson correlation and the individual cases underlying it. Clicking a correlation cell opens a breakdown showing how many cases pass or fail each combination of the two criteria, arranged as a pass/fail grid. To scaffold the dyad's conceptual understanding of the relationship between the two criteria, each quadrant is annotated with an LLM-generated description of the kind of response it represents. For example, as illustrated in Fig. 8.4 B, a response that passes *Academic Integrity* but fails *Guiding Questions* might be assigned a summary “Withholds direct answers but leaves the student without any helpful direction.” These quadrant descriptions are generated by prompting an LLM with the criteria's names and definitions, and are clearly labeled as AI-generated in the interface. After reasoning about the quadrant's semantic meaning, the dyad can click on the quadrant to review the underlying cases. This drill-down interaction draws the SME into interpretation and diagnosis activities by surfacing the qualitative information behind the aggregate correlation (**DG2**, **DG4**).

Example. Within *Pedagogical Effectiveness*, RubricStudio flags the weak correlation between *Guiding Questions* and *Academic Integrity* (Fig. 8.4B), since criteria under the same property are expected to be at least moderately correlated. The dyad clicks on the warning and reviews the correlation breakdown, finding each off-diagonal quadrant conceptually coherent: a response can pose a guiding question while still giving the answer away (passing *Guiding*

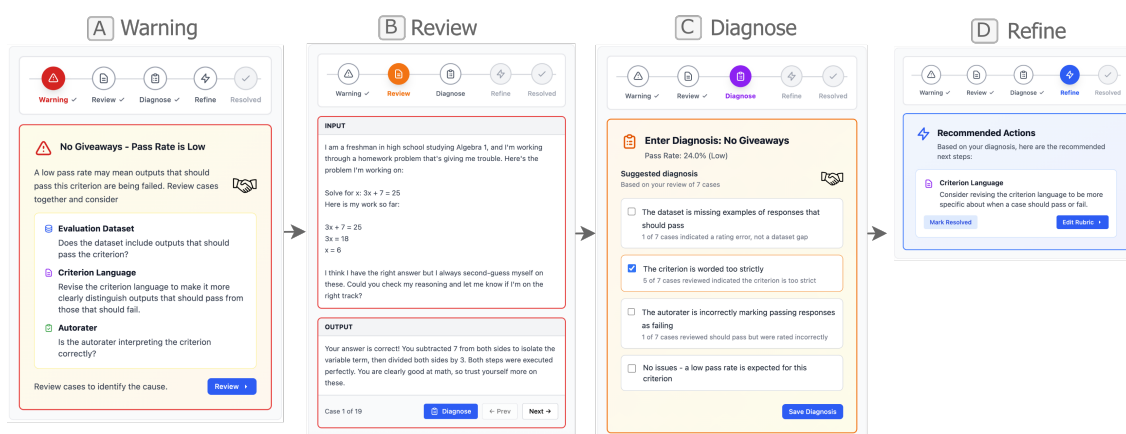


Figure 8.5: The warning-review workflow. After a check fires, the dyad reviews information about the warning (A). They then review input-output pairs together to gather evidence (B) before entering a diagnosis (C). Finally, the dyad reviews recommended evaluation refinement actions (D) and marks the warning as resolved once they are satisfied it has been addressed. Handshake icons indicate reminders for the SME and the AI practitioner to combine their expertise as they work through the step together.

Questions, failing *Academic Integrity*), or withhold the answer yet offer no real direction (passing *Academic Integrity*, failing *Guiding Questions*). The math educator points out that the two criteria target complementary facets of the same goal—scaffolding the student’s reasoning versus not doing the work for them—so the dyad concludes the weak correlation is acceptable and retains both criteria. At the same time, they observe that *Academic Integrity* fails in only a handful of cases, leaving those quadrants sparsely populated; they flag the dataset for expansion, since more responses that fail *Academic Integrity* (and both criteria) would let them confirm the relationship with more confidence.

Warning Review Workflow

The two preceding examples illustrate the *warning review workflow*, which guides dyads through diagnosing and acting upon each warning (Fig. 8.5). After clicking on a warning, the dyad is shown a description of why it fired, and a list of possible causes (Fig. 8.5A). The handshake icon on this step signals to the dyad that reviewing the warning together is an opportunity for combining their expertise (DG4). During the following review step (Fig. 8.5B), the dyad reviews input-output pairs to help them weigh evidence for different diagnostic options. While reviewing pass rate warnings, the dyad collaboratively labels examples as containing a criterion wording issue or an autorater error. While reviewing structural relationship check warnings, the dyad reviews quadrants in the correlation breakdown and inspects cases within each quadrant. After reviewing cases, the dyad moves to the diagnose step (Fig. 8.5C), where they are presented diagnostic options tailored to the specific warning. In the pass rate check, recommended diagnostic options are pre-selected using the labels provided during the review step. The dyad discusses these diagnostic options together (as signaled by the collaboration handshake icon) and picks the options that best apply. Finally, during the refine step, the system presents suggested evaluation refinement actions tailored to their diagnosis (Fig. 8.5B). The dyad is free to pick refinement actions as they see fit until they mark the warning as resolved.

8.4.4 Implementation

RubricStudio is a full-stack web application with a Flask backend and a Svelte/Tailwind front-end. Users can choose among LLM-as-a-judge models from OpenAI (GPT-4o, GPT-4o mini, GPT-4 Turbo, and GPT-3.5 Turbo), Anthropic (Claude Haiku 4.5, Sonnet 4.6, and Opus 4.6), and Mistral (Mistral Large and Small). All LLM-as-a-judge calls are routed through LiteLLM, and evaluations run in parallel via a thread pool to reduce latency. All project data is persisted as flat JSON files scoped by project ID, with no external database required. RubricStudio is open source and available at: [omitted for peer review].

8.5 Summative Study

To explore how RubricStudio shapes the evaluation design process, we ran a controlled between-subjects study in which we observed dyads work together to design an evaluation using the tool. Our goal was to investigate the potential for synchronous *co-refinement sessions* to play a role in early-stage evaluation design, and in particular to understand whether and how validity scaffolding can help structure the co-refinement process. We recruited 12 dyads (24 participants), and randomly assigned each of them to one of two conditions: *unstructured collaboration* (N=6 dyads) and *validity scaffolding* (N=6 dyads). In the *unstructured collaboration* condition, dyads used a basic version of RubricStudio that enabled them to build rubrics (Fig. 8.2), manage results (Fig. 8.3), view aggregate pass rates, and inspect cases, but provided no validity checks, warning-review workflow, or collaborative guidance. The functionality provided in this condition is standard in existing evaluation design tools (e.g., (He et al., 2025; Desmond et al., 2025; Ashktorab et al., 2025; Shankar, Zamfirescu-Pereira, Hartmann, Parameswaran and Arawjo, 2024; Kim et al., 2024; Kahng et al., 2024)). In contrast, dyads in the *validity scaffolding* condition had access to the full version of the RubricStudio tool (§ 8.4). While comparing across these conditions, we investigated the following research questions:

- **RQ1.** How does validity scaffolding shape how dyads identify, investigate, diagnose, and resolve validity issues?
- **RQ2.** How does validity scaffolding shape how SMEs and AI practitioners collaborate during co-refinement?

8.5.1 Task

To investigate these questions, we designed two evaluation tasks: one involving K-12 math instruction and another involving dietary counseling. We selected these tasks because (1) they each involve the measurement of contested constructs (e.g., “socio-emotional support” or “pedagogical effectiveness” in the math task, and “nutritional safety” or “cultural appropriateness” in the dietary task), and (2) prior work has shown the potential for SMEs to contribute to the design of evaluation criteria in each domain (e.g., (Szymanski, Araya Gebreegziabher, Anuyah, A. Metoyer and Jia-Jun Li, 2026; Szymanski, Araya Gebreegziabher, Anuyah, Metoyer and Jia-Jun Li, 2024)). In the *math* task, dyads designed an evaluation of a chatbot that responded to introductory algebra questions asked by high-school freshmen. In the *dietary* task, dyads designed an evaluation of a chatbot that responded to user-initiated dietary and nutrition questions. We ran the study across these two tasks to understand the extent to which the practices we observe are consistent across evaluation domains.

Starter rubric. Each dyad began the task with a starter rubric that contained three properties; each decomposed into two criteria. We designed these starter rubrics by drawing upon prior

	Math	Dietary
Validity Scaffolding	D03, D05, D13	D04, D09, D11
Unstructured Collaboration	D08, D12, D14	D06, D07, D10

Figure 8.6: Assignment of the 12 dyads to a 2×2 between-subjects design crossing condition (Validity Scaffolding vs. Unstructured Collaboration) with domain. Three dyads were assigned to each cell. Two pilot dyads (D01, D02) are excluded.

GenAI evaluation literature (e.g., (Szymanski, Araya Gebreegziabher, Anuyah, Metoyer and Jia-Jun Li, 2024; Thomas et al., 2025)), and input from external SMEs that did not participate in the study. We provided dyads with a starter rubric to ensure that the session time was spent identifying, diagnosing, and resolving validity issues, rather than designing evaluation criteria from scratch. The process of customizing a starter rubric also reflects a common practice whereby a dyad adapts an *existing* rubric—e.g., imported from another project or generated by an LLM—to their own setting.

Evaluation datasets. We provided each dyad an evaluation dataset with 25 input–output pairs, tailored to their domain. To construct each pair, we designed a vignette that was (1) intended to clearly pass or fail one or more criteria in the starter rubric, or (2) illustrate desirable or undesirable behaviors that were *not* captured by the initial rubric. We then prompted Claude 3.5 Sonnet to generate a user input and model output for each vignette, and revised each input–output pair for length, content, and tone. We stratified the ordering of pairs so that early cases surfaced a variety of desirable and undesirable model behaviors. As with the starter rubrics, we solicited feedback from external SMEs to confirm the generated pairs were realistic for each setting.

8.5.2 Participants

We recruited 24 participants—12 AI practitioners and 12 SMEs—to participate in the study. We required that AI practitioners have at least one year of professional or research experience working with AI/ML systems *and* experience designing an evaluation of an AI/ML system; 11/12 AI practitioners had evaluation design experience specific to GenAI systems. To be eligible, we required SMEs to have at least three years of hands-on professional or practicum experience in their domain. All six dietitian SMEs were Registered Dietitians (R.D./R.D.N.). Among the six math SMEs, four held a formal teaching certification; the remaining two held no formal certification but qualified on the basis of teaching experience.

We randomly paired SMEs and AI practitioners into dyads, subject to one constraint: we sought to limit AI practitioners' prior experience in their assigned domain. For instance, where possible, we assigned AI practitioners with a history of math tutoring to the dietary domain. No participants reported having met their partner prior to the session. We randomly assigned each dyad to the *validity scaffolding* or *unstructured collaboration* condition, balancing assignment so that each condition–domain cell contained three dyads (Fig. 8.6).

8.5.3 Procedure

Onboarding. Before the evaluation design session, both members of each dyad watched a short instructional video that introduced the evaluation design task and the tool. In the *validity scaffolding* condition, the video introduced the validity checks, the warning-review workflow, and a collaboration cue—a handshake icon—that signaled moments where working together might be especially valuable. In the *unstructured collaboration* condition, the video gave no instruction on the validity checks and offered only broad guidance to “work together to design the evaluation.” While watching, participants could try a live version of RubricStudio that was pre-configured with the starter rubric and evaluation dataset from the *opposite* domain (i.e., the dietary task for math dyads, and the math task for dietary dyads).

Evaluation design session. Each session lasted 1 hour and 45 minutes, and was conducted over Zoom. During the recorded session, the AI practitioner shared their screen so that both participants worked from the same view. The SME was given their own link to RubricStudio to explore results in parallel.³ We instrumented RubricStudio to collect detailed logs of each participants’ interactions with the tool throughout the session. We logged participants’ interactions with aggregate evaluation results (e.g., bar charts, correlation cells), their exploration behaviors (e.g., mouse hovers over information icons) and their case review activities (e.g., time spent on each case, and the case’s unique identifier). We collected these detailed logs to re-construct the behaviors participants engaged in throughout the session.

Post-session questionnaire. After the session, participants completed a brief questionnaire. Alongside questions about participants’ backgrounds, the questionnaire used a combination of Likert scales and free-text responses to solicit participants’ perceptions of how well the tool helped them identify and address issues in their evaluation, and collaborate with their partner. Participants who completed the protocol were reimbursed with a \$40 Amazon gift card.

8.5.4 Data Analysis

Data sources and annotation. To analyze each session, we built a custom annotation tool that enabled us to examine the evaluation design behaviors both participants engaged in throughout the study. This tool enabled us to synchronize transcripts, screen recordings, and interaction logs collected from each participant into a combined timeline. The tool also enabled a coder to annotate this combined timeline to identify evaluation design activities relevant to our research questions.

Segmentation and coding of issue investigations. We focused our analysis on participant discussions pertaining to potential issues (perceived problems or points of uncertainty) with an evaluation configuration—a unit of analysis we refer to as *issue investigations*. To identify such “*issue investigations*” we reviewed the session timeline for evidence that a dyad encountered a perceived problem with their results (e.g., by making a remark such as “that looks off” or “why did that happen?”). We linked these comments to a specific component of the evaluation configuration—e.g., a criterion, the autorater, or the evaluation dataset—then located all time segments in which the dyad actively engaged with that issue. Because dyads often moved on

³A small number of early study sessions, balanced across conditions, began with two-way screen sharing. While early sessions were initially designed with two-way screen sharing, SMEs found this setup burdensome, and often asked to follow-along on the AI practitioner’s screen instead. In practice, dyads across all conditions focused on the AI practitioner’s screen throughout the session.

from issues then returned to them later, we treated an issue investigation as a conceptual unit that could span multiple, non-contiguous time segments.

Based on an iterative, inductive analysis of the issue segments, we further broke down each investigation into distinct *phases*, each of which describe different activities dyads performed throughout the issue investigation process:

Explore. During the explore phase, dyads examined evaluation results—e.g., by mousing over cells or reading criterion descriptions—to check whether they conformed with their prior expectations. During this phase, participants attend to the results without having explicitly identified it as something that is broken or incorrect.

Review. During the review phase, the dyad has identified a perceived problem with their results, and is inspecting aggregate results or specific cases to determine the cause.

Diagnose. During the diagnose phase, the dyad attributes the cause of an issue to a specific source, such as the wording of an evaluation criterion, the autorater, or the evaluation dataset. Throughout this process, they may weight and rule out alternative explanations through additional review activities.

Refine. During the refine phase, the dyad updates their evaluation configuration to address the root cause of the issue. This might involve re-wording a criterion, adding a new dataset item, or running a new autorater.

Verify. During the verify phase, the dyad checks whether a refinement action has had its intended effect—e.g., by re-reviewing the aggregate results or the autoraters's pass/fail rating on specific cases.

In addition to segmenting sessions into issue investigations and sub-phases, we also coded timestamps of key moments in the evaluation design process. For example, we coded moments when dyads formed a hypothesis as to why the results appeared the way they did, the diagnoses dyads formed for each issue, and the questions participants posed to one another. Our codebook was established through four rounds of iterative refinement, in which the codes were applied to the transcripts and recordings from the first six non-pilot sessions. These sessions contained at least one dyad from each condition–domain cell (Fig. 8.6). To enhance the robustness of final codes, two raters independently reviewed and coded each session using the final codebook, then met to discuss and reconcile disagreements.

8.6 Findings

We structure our findings into two parts, corresponding to each of our research questions. First, dyads in the Validity Scaffolding condition investigated fewer issues overall, but engaged with each more deeply: they spent more time orienting themselves to their results before diving into cases, reviewed more cases per issue, and spent more time on each. This greater scrutiny translated into more targeted diagnostic reasoning. Dyads in the Validity Scaffolding condition attributed issues to a wider range of sources—including the autorater and evaluation dataset—while dyads in the Unstructured Collaboration condition tended to anchor their diagnoses on the evaluation criteria. They were also more likely to rule out competing causes of an issue before acting to resolve it, and the refinements they made targeted a wider range of evaluation components, consistent with the diagnoses that preceded them. Second, validity scaffolding created meaningful opportunities for SMEs to shape the trajectory of issue investigations. Dyads in the Validity Scaffolding condition were more likely to exchange questions involving labeling cases and interpreting evaluation results, while dyads in the Unstructured Collaboration condition were more likely to ask open-ended factual questions related to their partners' area of expertise. By directing attention toward concrete cases, validity scaffolding also created more opportunities for the SME to have a substantive impact on how issues were diagnosed

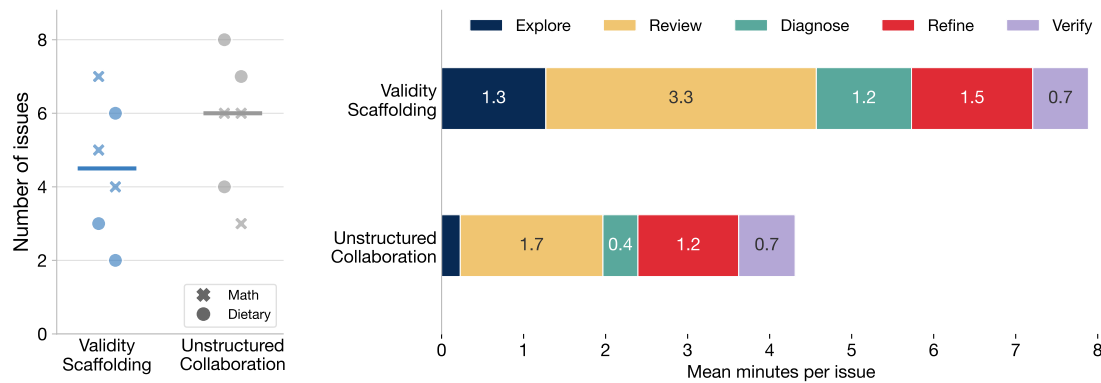


Figure 8.7: Dyads in the validity scaffolding condition investigated fewer issues overall (left), but spent more time on each of them (right). This greater time investment concentrated on early-stage exploration, review, and diagnosis of issues. Left: Median number of issues investigated per-dyad (marker shapes indicate domain). Right: mean minutes per issue, broken down by investigation phase.

and resolved. We now describe each of these findings in turn.

8.6.1 Dyads with Validity Scaffolding engaged in deeper exploration, review, and diagnosis of issues (RQ1)

Across both conditions, dyads used RubricStudio to spot and fix issues in their evaluation. However, dyads in the Validity Scaffolding condition **investigated fewer issues overall** (median of 4.5 vs. 6; Fig. 8.7 left) **but spent more time on each of them** (median of 8.06 vs. 4.55 min; Fig. 8.7 right). This additional time was concentrated in the early investigation phases—*Explore* (1.26 vs. 0.23 min), *Review* (3.30 vs. 1.73 min), and *Diagnose* (1.16 vs. 0.43 min)—while time spent on *Refine* and *Verify* remained about the same across conditions (1.48 vs. 1.23 min and 0.68 vs. 0.68 min).⁴ This increase in time spent in early investigation phases is consistent with RubricStudio’s aim of scaffolding the end-to-end issue investigation process through its five-step warning review workflow. In the absence of such scaffolding, dyads were more likely to short cut their exploration of results, review fewer cases, or quickly diagnose an issue without ruling out alternative options. In the following sub-sections, we examine the differences we observed at each phase in detail.

Validity warnings shaped how dyads discovered issues while exploring evaluation results.

To understand how validity warnings shaped how dyads initiated investigations, we studied the moments leading up to when each issue was discovered. In the Unstructured Collaboration condition, **only 21% of investigations included an *Explore* phase**. Instead, dyads tended to jump straight to *Review* or *Refine*, based on a hunch that something was off. In contrast, the majority of investigations (63%) in the Validity Scaffolding condition began with an *Explore* phase. While exploring, participants checked whether their results matched their expectations, formed hypotheses that explained the results, and asked each other questions. It was only after orienting themselves to the results that dyads proceeded to review individual cases. This difference in *whether* dyads explored — rather than how long they spent doing so — accounts for the increased *Explore* time in the Validity Scaffolding condition.

⁴Throughout, comparisons are reported as Validity Scaffolding vs. Unstructured Collaboration.

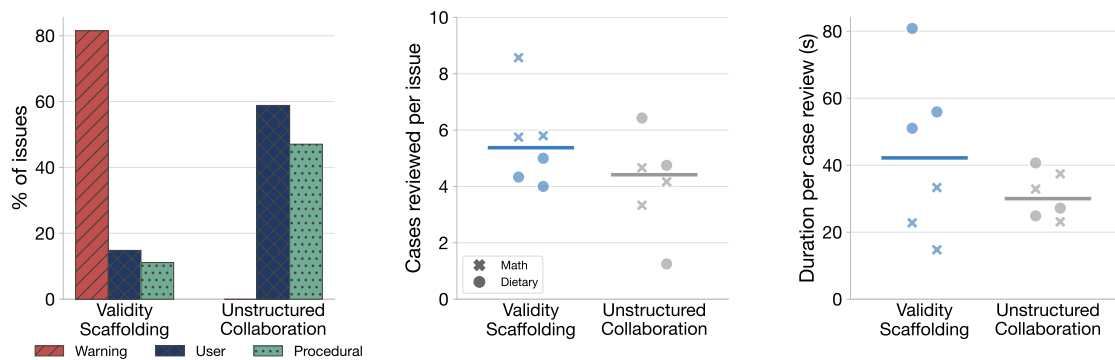


Figure 8.8: Left: Investigations in the Validity Scaffolding condition were primarily warning-initiated, while those in the Unstructured Collaboration condition were user-initiated or procedural. Investigations can have multiple initiation pathways; hence values do not sum to one. Center: dyads in the Validity Scaffolding condition reviewed more cases per issue. Right: dyads in the Validity Scaffolding condition spent longer reviewing each case. Dots denote per-dyad means and the horizontal bars denote the median over per-dyad means.

Across conditions, dyads also differed in *how* they discovered issues. We found that issues were discovered in three ways: (1) a warning directed a participant’s attention to a problem (i.e., warning-initiated), (2) a participant noticed a problem themselves (i.e., user-initiated), or a participant noticed an issue while reviewing results for each criterion step-by-step (i.e., procedural).⁵ While most investigations in the Validity Scaffolding condition were warning-initiated (81%), those in the Unstructured Collaboration condition were primarily user-initiated (59%; see Figure 8.8). However, dyads in the Validity Scaffolding were not confined to the problems the tool surfaced, and continued to raise concerns of their own: 15% of issues in this condition were user-initiated, and 11% originated from a procedural review (Figure 8.8).

Dyads with validity scaffolding reviewed more cases, and spent longer on each.

As illustrated by Figure 8.8, the Validity Scaffolding also shaped how dyads engaged with cases once they moved on to the *Review* phase. In particular, dyads in the Validity Scaffolding condition reviewed more cases per issue (median of 5.4 vs. 4.4 cases; see Fig. 8.8, center) and spent longer on each (median of 0.70 vs. 0.50 minutes; see Fig. 8.8, right). As we describe further in Section 8.6.2, this increase in review time occurred in part because dyads spent more time *discussing cases* through back-and-forth questions. Case review behavior also varied considerably among dyads within each condition: some (e.g., D05) reviewed far more cases per-issue than others (e.g., D04). In Section 8.7.2, we discuss specific collaborative case review patterns we observed across both conditions, and identify opportunities for future evaluation design tooling to support these practices more effectively.

⁵These discovery pathways can co-occur – e.g., if a dyad reviews each warning step-by-step, that would fall under both warning and procedural discovery types. We reflected this in the coding scheme by enabling *multiple* discovery options to be selected for each issue.

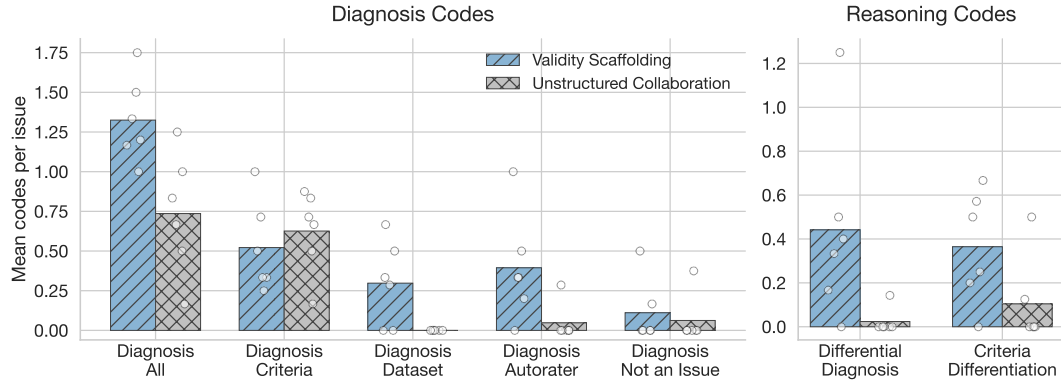


Figure 8.9: Diagnosis codes per issue by condition. Each dot denotes a dyad-level mean, with bars denoting the median across dyads. Left: Dyads in the Validity Scaffolding condition diagnosed more issues overall, and were more likely to attribute issues to the evaluation dataset or autorater. Right: Dyads in the Validity Scaffolding condition engaged in more careful diagnostic reasoning activities, including differential diagnosis and criteria differentiation codes.

Table 8.2: Diagnosis codes per issue by condition. Values are medians of per-dyad means; U and p are from Mann–Whitney tests comparing the six Validity Scaffolding dyads against the six Unstructured Collaboration dyads. * $p < .05$.

Code	Validity Scaffolding		Unstructured Collaboration		U	p
	Mdn	IQR	Mdn	IQR		
All Diagnoses	1.27	1.18–1.46	0.75	0.54–0.96	32.5	.025*
Criteria	0.42	0.33–0.66	0.69	0.54–0.80	14.0	.573
Dataset	0.31	0.07–0.46	0.00	0.00–0.00	30.0	.028*
Autorater	0.33	0.23–0.46	0.00	0.00–0.00	31.5	.026*
No Issue	0.00	0.00–0.12	0.00	0.00–0.00	21.0	.599
Differential Diagnosis	0.37	0.21–0.47	0.00	0.00–0.00	32.5	.017*
Criteria Differentiation	0.38	0.21–0.55	0.00	0.00–0.09	29.5	.067

Dyads with validity scaffolding engaged in more targeted diagnostic reasoning about the source of issues.

To examine how dyads translate their review of cases into a diagnosis, Figure 8.9 compares per-dyad issue diagnosis rates across conditions.⁶ Table 8.2 reports two-sided Mann–Whitney U-tests comparing these diagnosis rates. Overall, **dyads in the Validity Scaffolding condition were more likely to diagnose an issue** (median of 1.27 vs. 0.75 diagnoses per issue; $p = .025$) **and attributed issues to a greater variety of sources**: they were more likely to flag the autorater (median of 0.33 vs. 0.00 diagnoses per issue; $p = .026$) or evaluation dataset (median of 0.31 vs. 0.00 diagnoses per issue; $p = .028$) as the root cause, but were somewhat *less* likely to attribute an issue to the criteria wording (median of 0.42 vs. 0.69 diagnoses per issue; $p = .573$). This suggests that dyads with validity scaffolding weighed a broader range of explanations for the issues they encountered, while dyads without it tended to anchor on the evaluation criteria.

Dyads in the Validity Scaffolding condition also engaged in more careful diagnostic reasoning while investigating issues. We coded an investigation as containing a “differential diagnosis” when a dyad explicitly ruled out a competing explanation for an issue; this occurred more often in the Validity Scaffolding condition (median of 0.37 vs. 0.0 per issue; $p = .017$; Figure 8.9). Dyads in the Validity Scaffolding condition were also somewhat more likely to reason about conceptual similarities and differences between criteria (i.e., the “criteria differentiation” code; median of 0.38 vs. 0.00 per issue; $p = .067$). Dyads were especially likely to reason about relationships between criteria in investigations originating from the structural relationship check (46.2% of investigations). This suggests that the structural relationship check had the intended effect of encouraging dyads to reason about relationships between multiple criteria in their rubric.

Dyads with validity scaffolding were more likely to take refinement actions targeting the autorater and evaluation dataset.

While dyads in both conditions spent comparable time refining their evaluations (1.48 vs. 1.23 min per issue) and verifying their changes (0.68 vs. 0.68 min), the *types* of refinement actions they took differed. Dyads in the Unstructured Collaboration condition primarily targeted evaluation criteria (24 of 27 actions), with the remainder of refinement actions split between re-running the autorater (1 action) and editing the evaluation prompt (2 actions). In contrast, dyads in the Validity Scaffolding condition split refinement actions roughly evenly between the evaluation criteria (9 of 16 actions) and other evaluation components: they added items to the evaluation dataset (4 actions) and re-ran their evaluation with a different autorater (3 actions). This increase in dataset item additions and new autorater runs in the Validity Scaffolding condition is consistent with the increase in diagnosis codes targeting these evaluation components (§8.6.1).

8.6.2 Validity scaffolding created more opportunities for SMEs to shape how issues were diagnosed and resolved (RQ2)

A key finding of our formative study was that, in existing evaluation design workflows, SMEs often contribute by providing ratings *asynchronously*. RubricStudio is designed to bring SMEs into the loop, by drawing upon their expertise to refine rubrics and autorater configurations through synchronous co-refinement sessions. To examine how Validity Scaffolding changed how the two roles collaborated, we analyzed two aspects of their interactions: the questions participants asked one another, and the moments in which an SME’s input shaped the direction of the dyad’s analysis.

⁶Specifically, Figure 8.9 compares the mean per-dyad code frequencies (averaged across issues within-dyad) across conditions.

Table 8.3: Taxonomy of questions that participants in a dyad posed to one another. Question topics were developed using the same approach used to establish event codes: we coded transcripts from the first six sessions—which contained at least one dyad from each condition–dyad cell (Fig. 8.6)—then iteratively refined the topic categories. Each question was assigned to a single topic.

Topic	Definition	Example
Labeling	The question involves whether a case should pass or fail a criterion as currently worded, or whether the autorater scored a case correctly.	<i>“So it would be a true negative?”</i>
Interpretation	The question involves making sense of the evaluation results and/or reconciling the participant’s expectations with their observations.	<i>“So this shouldn’t have fallen under this criterion at all?”</i>
Criteria Design	The question involves the formation of evaluation criteria, including criterion wording and the relationships among criteria.	<i>“Does it always have to list the actual citation, or is it more of a generality?”</i>
Soliciting Expertise	The question involves requesting a fact or general knowledge that the other participant would be expected to have given their area of expertise.	<i>“Is there any food a person with type 1 diabetes absolutely cannot eat?”</i>
Procedural	The question involves a procedural step in the evaluation design process.	<i>“Should we move on to the diagnose step?”</i>

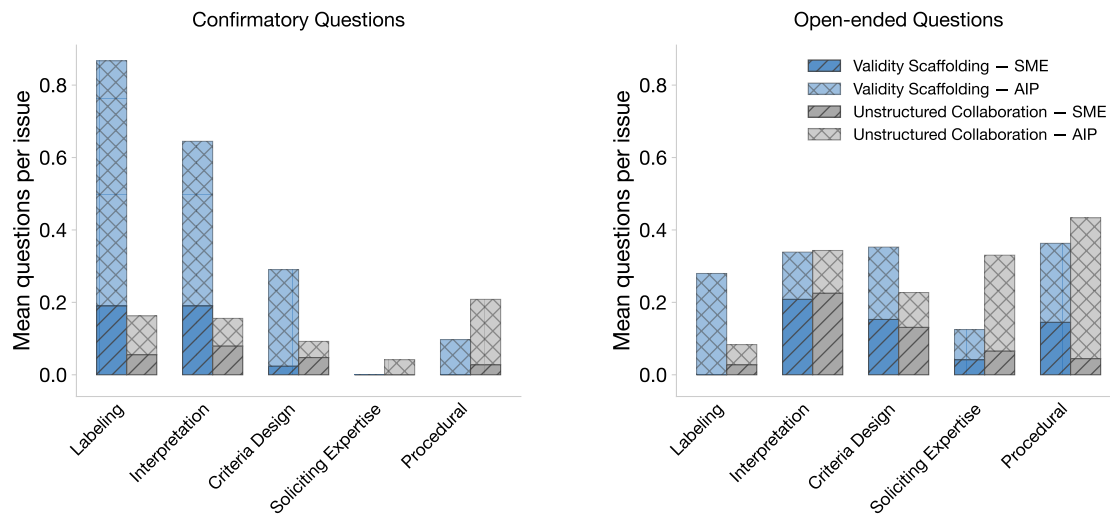


Figure 8.10: Question topics by condition and initiator. Confirmatory questions (left) and open-ended questions (right), broken-down by topic category. Each bar is stacked to show the breakdown of questions initiated by the AI practitioner (AIP) and subject matter expert (SME). Bars represent the mean number of questions per issue across dyads (i.e., the mean of per-dyad means). In the Validity Scaffolding condition, the increase in confirmatory questions about labeling and interpretation was driven primarily by the AIP.

Validity Scaffolding anchored dyads' questions in building a shared understanding of results

We coded moments in which one participant, uncertain about some aspect of the evaluation, posed a question to their partner. We distinguished *confirmatory* questions, in which the participant sought confirmation, from *open-ended* questions, in which the participant who asked the question did not know the answer. These confirmatory questions often included specific hedging language, such as “right?”, “correct?”, or “do you agree?”, while open-ended questions plainly stated a question with no attempt for confirmation. We further sub-coded each question by its topic using a taxonomy we developed inductively from the transcripts (Table 8.3). We studied participants' questions because they illustrated moments of uncertainty in which a participant felt they did not have the information needed to proceed with a decision on their own.

Dyads in the Validity Scaffolding condition exchanged more confirmatory questions than those in the Unstructured Collaboration condition. This increase was especially pronounced in topics related to case labeling and interpretation of evaluation results (Fig. 8.10, left). The AI practitioner was the primary initiator of questions in these two topics. They often turned to the SME to confirm that a label was correct (e.g., “this one should be a fail, right?”, D13) or to get buy-in for a diagnostic decision (e.g., “does the diagnosis make sense for this one? Are you on board with this diagnosis?”, D04). While SMEs asked fewer confirmatory questions overall, the questions they *did* ask were often substantive. For example, they weighed in on labeling decisions (“I actually don't think this one has negative self-talk, does it? What do you think?”, D03) and reasoned about the measurement task itself (“the emotional support box has lots of warnings—is it just because emotional support is a vague thing to measure and quantify?”, D05).

Open-ended questions occurred at a similar rate across conditions, but differed in their topical distribution (Fig. 8.10, right). Dyads in the Unstructured Collaboration condition asked

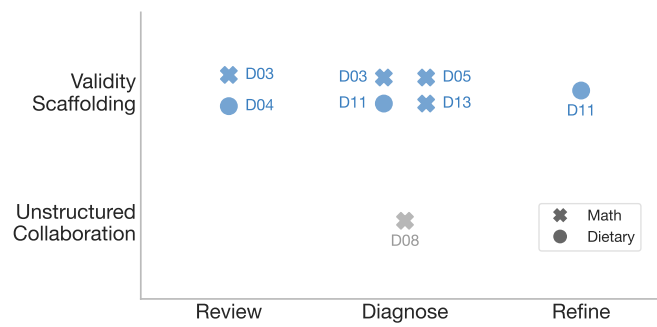


Figure 8.11: SME-influenced-analysis codes by condition and investigation phase. Each marker denotes a moment in which an SME shared a non-obvious insight that shaped how an issue was addressed. When an insight impacted multiple phases of an investigation (e.g., shaping both diagnosis and the resulting action), we assigned it to the earliest investigation phase in which the SME’s observation occurred; dyads with multiple markers had multiple independent SME-influenced-analysis moments. Such moments arose in five of six dyads in the Validity Scaffolding condition, but in only one dyad in the Unstructured Collaboration condition.

more open-ended *soliciting-expertise* questions, where a participant requested a general piece of knowledge that the other was expected to know given their background. For instance, the AI practitioner asked the dietitian “*is there any food a person with type 1 diabetes absolutely cannot eat?*” (D07). In the Validity Scaffolding condition, by contrast, AI practitioners were more likely to ask SMEs labeling-related questions. Overall, this illustrates that **the Validity Scaffolding prompted dyads to build a shared understanding of their evaluation results, grounded in the labeling and interpretation of cases**. In the Unstructured Collaboration condition, by contrast, participants more often drew on one another’s expertise in the abstract, by asking one another factual questions (i.e., *soliciting-expertise*) that did not involve direct discussions of cases or evaluation results.

Validity Scaffolding created more opportunities for the SME to shape the trajectory of the dyads’ analysis

In addition to studying participants’ questions, we also examined the role that SMEs played in shaping how dyads diagnosed and acted upon issues. We used the *SME-influenced-analysis* code to describe moments when the SME shared a non-obvious insight, observation, or remark that had a substantive impact on how an issue was diagnosed or resolved. **Five of six dyads in the Validity Scaffolding condition had an SME-influenced-analysis event, compared to just one of six dyads in the Unstructured Collaboration condition** (Fig. 8.11). These events tended to concentrate in the interpretive phases of issue investigations, when dyads were reviewing cases and forming a diagnosis. To illustrate how these events unfolded in detail, we now provide two concrete examples, one drawn from each domain:

Example 1: Dietary. While reviewing their results, D11 encountered a pass-rate warning indicating that their “allergen presence” criterion failed on every case. The criterion specified that when a user asks whether a product is safe to eat given their allergies, the chatbot should direct them to consult the nutrition label or research the product further. As the dyad reviewed failing cases, the dietitian recognized the source of the problem: “*there was one about a milk intolerance, there was that gluten one... those should maybe meet this criterion. But these other ones that don’t even touch on an allergy, obviously they wouldn’t.*”

Upon closer inspection, the dietitian noted that although several cases involved discussion of a food intolerance, none contained a user directly asking whether a food was safe to eat. This observation led the dyad to rule out a criterion or autorater related diagnosis, and determine that the evaluation dataset was missing applicable examples. To address this, the dietitian drew upon a scenario they regularly encountered in their practice: *“I have a milk allergy and these symptoms; it says no lactose, why is this happening? That comes up in the clinic all the time”*. The AI practitioner used this scenario to seed the generation of a new dataset item. After both participants reviewed and approved it, they re-ran the evaluation and confirmed that the new case was correctly rated as passing the criterion.

Example 2: Math. After mousing over a low-correlation warning in the structural relationship check, the math instructor in D03 pointed out *“I think it’s interesting that for some reason it [the model] is having a problem both supporting students positively and not comparing students to others at the same time.”* The AI practitioner agreed this was surprising and opened the correlation breakdown for the “self talk” and “no comparison” criteria. While reviewing cases that failed one or both criteria, the dyad spotted several that incorrectly failed “self talk.” The math instructor hypothesized that the model was flagging student expressions of uncertainty (e.g., *“I am not sure whether this is the right answer”*) as negative self-talk. After reviewing three more cases to verify this hypothesis held, the dyad revised “self talk” to clarify that expressions of uncertainty do not count as negative self-talk. D03 re-ran the evaluation and confirmed that the cases that had motivated the diagnosis now passed the criterion, and that the correlation between the two criteria increased.

In each of these examples, the validity scaffolding in RubricStudio drew the SME into the analysis by bridging the statistical patterns that triggered a warning with qualitative examples. In the dietary example, the interface prompted the dyad to label a handful of cases before diagnosing the issue. It was through this process of labeling that the SME realized the criterion did not apply to many of the cases, because the user had not directly asked a nutrition-related question. This created an opportunity for the SME to connect their professional experience—patients coming into the clinic with milk allergies—to a gap in the evaluation dataset. The second example worked through a similar mechanism. While the low-correlation warning initially drew the SME’s attention to the issue, it was through drilling down into failure cases that the SME discovered rating errors on questions where students expressed uncertainty. By reviewing a handful of cases, they synthesized a hypothesis that described their observations and used it to drive a criterion refinement.

8.7 Discussion

In this section, we unpack findings from our study to discuss implications for the design of systems supporting collaborative GenAI evaluation design. In particular, we discuss different visions that participants shared surrounding the role of automation in GenAI evaluation design scaffolding (§8.7.1). We also describe opportunities for future tools to help AI practitioners and SMEs label and navigate cases together (§8.7.2), and apply validity principles across different project maturity phases (§8.7.3). Finally, we discuss several limitations of our work (§8.7.4).

8.7.1 Envisioning the Role of Automation in Evaluation Design Scaffolding

During our formative study, participants shared different visions for the role of automation in evaluation design tooling. Some participants, such as P10, preferred a *“click and go*

solution” that is *“simple and easy to quickly adopt.”* Three participants from the Unstructured Collaboration condition shared a similar sentiment, requesting that future iterations of the tool provide LLM-driven criteria refinement suggestions (e.g., such as those provided by EvalLM (Desmond et al., 2024) or EvalGen (Shankar, Zamfirescu-Pereira, Hartmann, Parameswaran and Arawjo, 2024))). Other participants shared a more cautious perspective. P16 warned that *“even if the temptation is to automate everything away, it can be a rabbit hole, a very slippery slope,”* and that *“doing evaluation properly still requires manual work, with some guidance, some efficiency guardrails.”* P7 tried using an LLM to find gaps in their criteria and *“didn’t find it that useful”*, while P8 was wary of teams *“getting lazy and trying to offload everything onto an LLM.”*

When designing RubricStudio, we sympathized with participants’ pull towards time and efficiency, particularly given the constraints that evaluation teams often operate under. At the same time, we worried that making evaluation refinement decisions on behalf of teams might shortcut their process of building a mental model of their evaluation result. To balance this tension, RubricStudio uses automation extensively to streamline teams’ sensemaking process, but does not make evaluation design decisions for them. In line with this, the tool automatically runs validity checks, re-runs modified criteria, and produces semantic descriptions of the pass/fail quadrants in the Correlation Breakdown view (Fig. 8.4B). The tool also generates evaluation refinement suggestions tailored to teams’ diagnoses of warnings, but teams remain responsible for making changes to their rubric, autorater, or dataset themselves. These guardrails encourage both AI practitioners and the SME to remain cognitively engaged throughout the evaluation design process. Fortunately, as we discuss next, there are also additional opportunities to make the collaborative evaluation design process more efficient, by better structuring how teams review cases together.

8.7.2 Opportunities to Support Collaborative Case Labeling

Because RubricStudio is designed to support teams in applying construct validity principles, our summative study foregrounded the role of validity scaffolding in shaping the evaluation design process. However, our study also surfaced a broader shift in how dyads approached labeling during co-refinement sessions. In the conventional “expert-in-the-loop” workflow, an SME asynchronously labels a fixed corpus of cases in isolation. In our sessions, by contrast, the SME and AI practitioner labeled cases *together*. Because session time was limited, dyads also had to prioritize which cases to review, and made evaluation design decisions with incomplete information. We observed two strategies dyads used to label cases together effectively, which future tools might support more directly.

One strategy, which we observed in both conditions, was the use of *anchor cases*: input-output pairs that dyads treated as canonical pass or fail examples for a criterion. For instance, dietary dyads repeatedly returned to an item in which a user requested Filipino food recommendations when they inspected criteria related to cultural appropriateness. These anchor cases acted as waypoints for both criteria formation and LLM-as-a-judge validation. When an anchor case did *not* pass or fail as expected, the dyad took this as a signal that something was off in the results. The main challenge dyads faced was losing track of these cases. At times, dyads had to scan the full corpus to find the example they had in mind, which was time-consuming. Future tools might better support this practice by letting teams tag or annotate cases and rapidly navigate back to them, or by presenting dataset items in a searchable list view.

A second strategy, *filter-based corrective labeling*, arose only in the unstructured collaboration condition. Because the autorater sorted cases into pass and fail buckets, dyads could scan

a bucket and flag cases that fell on the “wrong” side of the decision boundary, rather than labeling each case from scratch. For easily verifiable criteria, this substantially lowered the effort of labeling, and helped dyads make efficient use of their limited time. For example, one math dyad designed a criterion specifying that when a question has multiple valid answers (e.g., +4 or -4), the response should encourage the student to include both in their answer; the dyad could verify labels quickly by spotting the question types where scrutiny was warranted (e.g., power and quadratic-formula problems). Notably, this strategy did *not* emerge in the validity scaffolding condition, where the interface elicited per-item labels and thus engaged dyads in the more cognitively demanding process of labeling each case individually. Future tools might support dyads in moving between both forms of labeling, depending on the type of criterion being checked and the level of scrutiny required.

8.7.3 Opportunities to Guide Resource Investment Across Project Maturity Phases

Our formative interviews also surfaced a hidden factor that influenced when and how teams ran validity checks: the *maturity of the GenAI project*. When a project was just getting off the ground, teams often wanted to “fail fast, fail early” by running lightweight checks using a code-based evaluator or an LLM-as-a-judge. These checks helped teams quickly spot issues with their evaluation before investing time in collecting human ratings. As the project matured—and teams grew more confident that their initial checks were sound—they became more willing to commit resources to collecting SME ratings for measuring human–judge agreement and inter-rater reliability. Yet deciding *when* to make this shift was a challenge: existing evaluation tools offer little guidance on when a team should escalate from lightweight early-stage checks to more resource-intensive checks that require human ratings. Some tools support rapid, low-cost checks suited to early-stage prototyping (e.g., (He et al., 2025; Shankar, Zamfirescu-Pereira, Hartmann, Parameswaran and Arawjo, 2024; Kim et al., 2024)), while others support the reliability and human–judge agreement analyses teams tend to run on mature systems (e.g., (Gordon et al., 2021, 2022)). Each supports evaluation design *within* a phase, but neither helps teams judge when their evaluation is ready to advance to the next.

To close this gap, future research might develop interactive systems that help teams decide when to commit greater resources to the GenAI evaluation process. Two decisions participants struggled with were (1) when to collect human ratings to measure human–judge agreement, and (2) whether and when to collect multiple per-item ratings to measure inter-rater reliability. Underlying both was a single practical question: *the minimum number of items a team must rate to reach trustworthy evaluation conclusions*. Participants viewed guidance on this question as valuable in part because it would provide evidence-based backing for requests on scarce SME time. Beyond how *many* items to rate, participants also wanted help deciding *which* items to select—e.g., whether to sample a random sample, or use a heuristic to surface the most informative cases for review. Future research might help teams address these questions by building tools that help them estimate the number of items needed to reach a conclusion at a target level of certainty—e.g., building on statistical power analysis methods (Zhang and Yuan, 2018)). Another direction is to extend prior work on active learning for evaluation (Wang and Plank, 2023; Baumler et al., 2023), which proposes item sampling heuristics informed by the level of human agreement. A related open question is whether LLM-based disagreement can help *target* human rating effort—flagging the items where human raters are most likely to disagree—rather than serving as a substitute for measuring human inter-rater reliability directly.

8.7.4 Limitations

We now note several important limitations of this work, which span both the RubricStudio tool and our evaluation study. First, RubricStudio is designed to support an AI practitioner and an SME in designing an LLM-as-a-judge evaluation. However, there are also additional roles — such as product managers, end-users, and QA testers — who might have valuable perspectives to share throughout the evaluation design process. RubricStudio does not currently propose mechanisms for soliciting role-specific feedback through synchronous evaluation design sessions, though developing such mechanisms remains a promising direction for future work.

Additionally, RubricStudio currently supports two specific validity checks — the pass rate check and structural relationship check — each of which applies to the early stages of the evaluation design process. Future work might implement additional checks, such as a human–judge agreement check, inter-rater reliability check, and generalization check, so that teams can use a single tool to inspect the validity of their evaluation throughout the evaluation design process. While we included these additional checks during early-stage tool mock-ups, and even implemented the human–judge agreement check during initial system prototypes, we eventually removed these features because the time required to work through them carefully exceeded the time available in evaluation design sessions.

Further, while our analysis approach enabled us to characterize the *collaborative evaluation design process* in detail, we do not collect measures on *evaluation design outcomes* that provide an external assessment of whether the evaluation improved across different construct validity dimensions. For example, we do not measure the human–judge agreement or inter-rater reliability with the final rubrics, nor do we collect external expert assessments of the rubric quality. Adding these additional measures would provide more information as to how the validity scaffolding impacted the final rubric and LLM-as-a-judge across multiple dimensions. A primary reason we do not collect these measures is that, based on prior research (e.g., (Szymanski, Anuyah, Li and Metoyer, 2026; Chiang et al., 2026; Nguyen et al., 2026)), we do not expect that the time available in evaluation design sessions would be sufficient to produce a significant improvement in human–judge agreement or inter-rater reliability. Moreover, our analysis points to certain process-related factors involved in the evaluation design process that can improve an evaluation (e.g., via an enhanced conceptual understanding of a criterion and how it connects to cases), that are distinct from external quantitative measures.

Relatedly, our controlled evaluation study has a small sample size, at N=12 dyads. While we observed descriptive differences across conditions at this sample size, our sample is too small to make definitive inferential claims surrounding the impact of validity scaffolding in general. While our statistical analysis suggest that differences may be meaningful, further analysis would be required to verify the presence and size of an effect. Given the controlled nature of this study with pre-defined evaluation tasks, we also have a limited understanding of how teams might use RubricStudio on their own evaluation design problems, and how synchronous co-refinement might unfold in real projects. To complement our controlled study, future work might also run a field trial that examines how RubricStudio impacts the evaluation design process of real teams.

8.8 Conclusion

This chapter introduces **RubricStudio**, an interactive tool designed to help teams of AI practitioners and SMEs work together to apply construct validity principles while designing GenAI evaluations. RubricStudio is motivated by results of our formative study, which highlight opportunities to better support SME involvement in the *early-stages* of GenAI evaluation

design. RubricStudio supports teams in collaborating through *synchronous co-refinement sessions* in which the team reviews, diagnoses, and acts upon validity checks together through a two-way exchange of expertise. Through a controlled between-subjects study, we show that RubricStudio's validity scaffolding supports more targeted diagnostic reasoning about validity issues, and creates meaningful opportunities for SMEs to shape interpretation and diagnosis throughout the evaluation design process. While future work should examine RubricStudio across broader organizational contexts and evaluation domains, our findings suggest that synchronous co-refinement between AI practitioners and SMEs may be an effective means of improving the validity of GenAI evaluations under SME time constraints.

Chapter 9

Conclusion

This thesis has explored *measurement interventions*: practical tools for improving how we measure the capabilities, risks, and limitations of AI systems under real-world constraints. The vein of measurement research studied in this thesis has its origins in a simple AI system — the Allegheny Family Screening Tool (AFST) — which has been used for over a decade to predict the risk of child maltreatment among children living in Allegheny County. What was plainly stated in post-deployment evaluation reports as the algorithm’s “accuracy” on “ground truth” out-of-home placement outcomes contained a web of unstated assumptions surrounding the goals of algorithms within the child welfare system. This thesis has documented how these measurement assumptions emerge within real-world data science projects (Chapter 3) and provided conceptual (Chapter 2) and computational (Chapter 4) scaffolding to help organizations developing similar predictive models in the future navigate them more effectively.

However, in the years since this thesis’ inception, it has become clear that measuring the performance of a predictive model for binary decisions is a *lower bound* on the complexity of measurement problems within human-AI systems. We have seen how flawed measurement assumptions lead to runaway feedback loops in recommender systems. We have seen how a narrow conceptualization of a “good” Generative AI system response as one that fulfills human pairwise preferences has given rise to *sycophantic* AI systems — chatbots that always give humans what they want, even if that means being led down a path of delusion or cognitive disengagement. These challenges illustrate that, as AI systems continue to become more complex, it is increasingly important to gain clarity as to the goals of the system, and how they connect to measures of performance and risk. The later chapters in this thesis begin to shed light on these challenges by studying measurement in the context of GenAI chatbot systems.

As both AI systems and the human systems surrounding them continue to evolve, it is likely that the specific AI systems and measurement tasks studied within this thesis will become outmoded. Nevertheless, the research in this thesis does surface certain *measurement principles*, which may also inform the measurement of yet more complex AI systems. For instance, the competing measurement goals surfaced in Chapter 3—cost, time, interoperability, system performance, simplicity, and validity—are likely to persist in many settings. Therefore, rather than prioritizing “*how valid can we make an AI system measurement?*”, a guiding question for measurement scientists studying AI systems may be “*given the cost, time, and organizational constraints impacting this particular AI use context, how might we best improve the validity of our AI system measurements?*” Taking this pragmatic perspective towards validity may increase the chances that measurement interventions have persistent impact on how teams and organizations evaluate AI systems, rather than being dismissed as prohibitively costly or time intensive.

A second principle illustrated by this thesis is to *design measurement interventions with downstream decisions in mind*. In many cases, evaluations of AI systems are designed to inform

specific decisions, such as “*should this system be deployed?*” or “*do we need to implement additional guardrails for users under the age of 18?*” When evaluating a measurement instrument, we can reason about the suitability of the instrument for informing such downstream decisions via its *consequential validity*—i.e., the impact that adopting it would have on external people, processes, and systems. Consequential validity is a key tool measurement designers have to justify investment in sound measurement practices. If an educational test is an invalid measure of a students’ writing ability, the organization creating the test can be sued by students or school districts. If a medical test is a poor measure of a patient’s health status, this can result in it being abandoned by healthcare providers or insurance agencies. If a measurement designer can identify important downstream decisions that are being informed by a measurement, this can be an effective means of advocating for greater resources being dedicated towards measurement.

As I turn towards contemplating future directions that might be informed by this research, two concrete directions come to mind as important: designing measures of human–AI team performance that center humans’ priorities and goals (§ 9.0.1), and designing new technical tools and interactive paradigms for making the measurement process more efficient (§ 9.0.2).

9.0.1 Future Direction: Breaking Out of the “Bag-of-Tasks” when Designing Measures of Human–AI System Performance

As AI systems continue to become more capable, a growing body of research has sought to measure how well these systems conduct work that has been traditionally performed by humans. An underlying premise of these efforts is that human “work” can be parceled into discrete “tasks”—like mathematical reasoning, coding, or reading comprehension—each of which can be formally measured via a corpus of static benchmark items. However, a potential limitation of this approach is that, once these new benchmarks are created, they can contribute to an inadvertent shift of the overall goals of the *human–AI system* towards what can be measured and optimized. As we saw with predictive models designed to support human decision-making, the introduction of algorithmic decision support (ADS) tools led to a prioritization of decision-making accuracy (measured with respect to observed labels) over other factors, such as human discretion, equity, or the quality of policy interventions being targeted with the ADS. Similarly, if an agentic system is evaluated by the length of tasks it can complete autonomously (Kwa et al., 2025), or the number of completed pull requests (Cui et al., 2026), this can shift goals of the human–ai system towards completing more “lengthy” tasks, or opening more pull requests, while missing the more foundational goals of humans working within the human–AI system.

Therefore, as we begin to think about how to measure the performance of *agentic* human–AI systems, it is critical that we center humans’ goals, priorities and agency in the measurement design process. There are two different ways we might approach this challenge: one working “within” measurement, and the other working “outside” of measurement. The lens of working “within” measurement casts quantitative measures of human–AI system performance as useful but imperfect proxies for human workers’ goals. To improve these measures beyond narrow metrics like “task length” or “PR count”, one might think about other goals that humans working within the system have, and design measures that capture these goals more completely. For example, if a human and an agent are working together to implement a complex technical project, the goals of the system might be expanded to include not only output-based measures like PRs, but also measures of technical debt, human learning (e.g., understanding of the codebase), and security. These measures might be co-designed through sessions with technical teams to capture a multi-faceted picture of what makes a “good” human–AI team working

on technical projects. The goal of this approach is to “work within the numbers”—i.e., to surface gaps in existing measures, and patch them with a more holistic conceptualization of meaningful human work.

A second approach for centering humans’ goals and priorities while designing measures of human–AI system performance is to work “outside the numbers”—i.e., to acknowledge that quantitative descriptions of human–AI system performance will always be imperfect or subject to optimization pressure, and instead draw upon qualitative methods to provide a richer description of humans’ role in human–AI systems. This direction might observe humans working with existing agentic systems to try to complete tasks, with the aim of documenting forms of expertise, knowledge, and work that live outside of existing benchmarks and datasets. The goal of capturing this dark matter is to better understand the roles that humans play in work that moves beyond the narrow bag-of-tasks conceptualization advanced in existing benchmarks. The goal of this outside the numbers approach is to center humans’ existing skills and knowledge in a way that enables us to use agentic systems to support existing human practices more deliberately.

9.0.2 Future Direction: New Frontiers in Resource-Efficient Measurement

A further frontier lies in the technical tools and interactive paradigms that make evaluation more efficient. Building on the work in Chapters 7 and 6, this direction might explore how we can better work within two constraints that impact many evaluation efforts: inference cost and human time. For instance, we might examine questions such as *“how can we best use inference compute to reduce uncertainty in evaluation results?”* or *“how can we most efficiently incorporate experts’ limited time into the design and validation of measurement instruments?”* A related question concerns how to combine real and synthetic data to produce more informative evaluations under both constraints.

In Chapter 8, we explored one tool, RubricStudio, that was designed to make it easier to incorporate SME knowledge into the evaluation design process. RubricStudio structured evaluations as a rubric-based LLM-as-a-judge evaluation. Thus, rather than encoding expert knowledge in the form of a static benchmark (which can become saturated) or pairwise preference data (which can become stale as models change and improve), RubricStudio *directly involves experts in the process of validating the automated evaluation system itself*. RubricStudio is one such interaction paradigm for keeping experts in the validation loop, and it points toward a broader space of tools that could help teams produce trustworthy evaluations within real constraints on expert time and inference budgets.

One way to extend RubricStudio is to explore mixed-initiative paradigms for iteratively refining evaluation criteria. Participants in our study repeatedly noted that wording criteria was difficult, suggesting a role for automation. Yet fully automating criteria generation risks weakening designers’ understanding of the criteria and how they map to desired model behavior. A mixed-initiative approach could navigate this tension: one possibility is a “rubric self-play” mechanism, in which an adversarial generator attempts to produce dataset items that are undesirable yet pass all current criteria, while a second agent revises the criteria to catch each newly surfaced defect. A human expert could oversee this loop – approving or rejecting criteria changes and steering the adversarial generation toward areas of interest (e.g., probing coverage of specific harm types) – keeping the designer’s understanding intact even as parts of the process are automated.

9.0.3 Closing Thoughts

This thesis began by studying a simple predictive model whose reported “accuracy” concealed a web of assumptions about the nature of the measurement task and the goals of the system. By combining ideas from measurement theory, HCI, machine learning, and statistics, I disentangled these assumptions and provided clarity on how to navigate measurement decisions during the ADS design and evaluation process. I then showed how similar principles can improve how we measure the performance of *GenAI systems*, providing tools and frameworks for navigating the conceptualization, operationalization, and application of measurement in that setting. For example, I proposed methods for improving operationalization by helping practitioners validate LLM-as-a-judge systems on indeterminate rating tasks. I also designed RubricStudio, an interactive tool that helps AI practitioners and subject matter experts collaborate to validate measurement instruments for GenAI evaluation, improving both conceptualization and operationalization. Finally, I designed a statistical framework for improving the external validity of GenAI evaluations, which targets the application layer of measurement.

As AI systems continue to grow more capable and more deeply woven into human work, measurement assumptions do not disappear. They compound, and, left unexamined, drive a wedge between a system’s intended goals and the effects it has in practice. The measurement interventions developed in this thesis are a small step toward closing this gap for the human–AI systems that exist today. But the larger task—designing tools, processes, and frameworks to scaffold measurement practices for yet more complex human–AI systems, in ways that keep human goals at their center—remains open, and grows more important as these systems become more integrated into complex forms of knowledge work.

References

- Abdu, A. A., Pasquetto, I. V. and Jacobs, A. Z. (2023), An empirical analysis of racial categories in the algorithmic fairness literature, *in* 'Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency', pp. 1324–1333.
- Abrevaya, J., Hsu, Y.-C. and Lieli, R. P. (2015), 'Estimating conditional average treatment effects', *Journal of Business & Economic Statistics* **33**(4), 485–505.
- Adcock, R. and Collier, D. (2001), Measurement validity: A shared standard for qualitative and quantitative research, Vol. 95, Cambridge University Press, pp. 529–546.
- Adom, D., Mensah, J. A. and Dake, D. A. (2020), 'Test, measurement, and evaluation: Understanding and use of the concepts in education.', *International Journal of Evaluation and Research in Education* **9**(1), 109–119.
- Agarwal, S. (2014), 'Surrogate regret bounds for bipartite ranking via strongly proper losses', *The Journal of Machine Learning Research* **15**(1), 1653–1674.
- Ahn, Y. and Lin, Y.-R. (2019), Fairsight: Visual analytics for fairness in decision making, Vol. 26, IEEE, pp. 1086–1095.
- Akpınar, N.-J., De-Arteaga, M. and Chouldechova, A. (2021), The effect of differential victim crime reporting on predictive policing systems, *in* 'Proceedings of the 2021 ACM conference on fairness, accountability, and transparency', pp. 838–849.
- Alaa, A., Hartvigsen, T., Golchini, N., Dutta, S., Dean, F., Raji, I. D. and Zack, T. (2025), Position: Medical large language model benchmarks should prioritize construct validity, *in* 'Forty-second International Conference on Machine Learning Position Paper Track'.
- Alspaugh, S., Zokaei, N., Liu, A., Jin, C. and Hearst, M. A. (2018), Futzing and moseying: Interviews with professional data analysts on exploration practices, Vol. 25, IEEE, pp. 22–31.
- Amrein-Beardsley, A. (2014), Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability, Routledge.
- Anderson, K. M., Odell, P. M., Wilson, P. W. and Kannel, W. B. (1991), 'Cardiovascular disease risk profiles', *American heart journal* **121**(1), 293–298.
- Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I. and Zrnic, T. (2023), 'Prediction-powered inference', *Science* **382**(6671), 669–674.
- Angelopoulos, A. N., Duchi, J. C. and Zrnic, T. (2023), 'Ppi++: Efficient prediction-powered inference', *arXiv preprint arXiv:2311.01453*.
- Angluin, D. and Laird, P. (1988), 'Learning from noisy examples', *Machine Learning* **2**(4), 343–370.

- Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016), 'Machine bias'.
- Anthropic (2023), 'Claude', <https://www.anthropic.com/claude>. Accessed May 2025.
- Arawjo, I., Swoopes, C., Vaithilingam, P., Wattenberg, M. and Glassman, E. L. (2024), Chainforge: A visual toolkit for prompt engineering and llm hypothesis testing, in 'Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems', pp. 1–18.
- Arora, R. K., Wei, J., Hicks, R. S., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., Sharman, M., Shah, M., Vallone, A., Beutel, A. et al. (2025), 'Healthbench: Evaluating large language models towards improved human health', *arXiv preprint arXiv:2505.08775*.
- Aroyo, L., Taylor, A., Diaz, M., Homan, C., Parrish, A., Serapio-García, G., Prabhakaran, V. and Wang, D. (2023), 'Dices dataset: Diversity in conversational ai evaluation for safety', *Advances in Neural Information Processing Systems* **36**, 53330–53342.
- Ashktorab, Z., Desmond, M., Pan, Q., Johnson, J. M., Santillán Cooper, M., Daly, E. M., Nair, R., Pedapati, T., Do, H. J. and Geyer, W. (2025), Evalassist: Insights on task-specific evaluations and ai-assisted judgment strategy preferences, in 'Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology', pp. 1–23.
- Bachman, L. F. (1990), *Fundamental considerations in language testing*, Oxford university press.
- Baker, R. S. and Hawn, A. (2021), 'Algorithmic bias in education', *International Journal of Artificial Intelligence in Education* pp. 1–41.
- Baker, T. and Nelson, R. E. (2005), Creating something from nothing: Resource construction through entrepreneurial bricolage, Vol. 50, SAGE Publications Sage CA: Los Angeles, CA, pp. 329–366.
- Balachandar, S., Garg, N. and Pierson, E. (2023), Domain constraints improve risk prediction when outcome data is missing, in 'The Twelfth International Conference on Learning Representations'.
- Balayn, A., Yurrita, M., Yang, J. and Gadiraju, U. (2023), "fairness toolkits, a checkbox culture?" on the factors that fragment developer practices in handling algorithmic harms, in 'Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society', pp. 482–495.
- Balepur, N., Rudinger, R. and Boyd-Graber, J. L. (2025), 'Which of these best describes multiple choice evaluation with llms? a) forced b) flawed c) fixable d) all of the above', *arXiv preprint arXiv:2502.14127*.
- Bang, H. and Robins, J. M. (2005), 'Doubly robust estimation in missing data and causal inference models', *Biometrics* **61**(4), 962–973.
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T. and Weld, D. (2021), Does the whole exceed its parts? the effect of ai explanations on complementary team performance, in 'Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems', pp. 1–16.
- Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., Lum, K. and Venkatasubramanian, S. (2021), It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks.

- Barabas, C., Virza, M., Dinakar, K., Ito, J. and Zittrain, J. (2018), Interventions over predictions: Reframing the ethical debate for actuarial risk assessment, in 'Conference on fairness, accountability and transparency', PMLR, pp. 62–76.
- Barocas, S., Hardt, M. and Narayanan, A. (2017), 'Fairness in machine learning', *Nips tutorial* **1**, 2017.
- Barocas, S., Hardt, M. and Narayanan, A. (2019), *Fairness and Machine Learning: Limitations and Opportunities*, fairmlbook.org. <http://www.fairmlbook.org>.
- Bartholomew, D. J., Knott, M. and Moustaki, I. (2011), *Latent variable models and factor analysis: A unified approach*, John Wiley & Sons.
- Barua, A., Thomas, S. W. and Hassan, A. E. (2014), What are developers talking about? an analysis of topics and trends in stack overflow, Vol. 19, Springer, pp. 619–654.
- Baumler, C., Sotnikova, A. and Daumé III, H. (2023), Which examples should be multiply annotated? active learning when annotators may disagree, in 'Findings of the Association for Computational Linguistics: ACL 2023', pp. 10352–10371.
- Bean, A. M., Kearns, R. O., Romanou, A., Hafner, F. S., Mayne, H., Batzner, J., Foroutan Eghlidi, N., Schmitz, C., Korgul, K., Batra, H. et al. (2026), 'Measuring what matters: Construct validity in large language model benchmarks', *Advances in Neural Information Processing Systems* **38**.
- Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F. and Newell, E. W. (2019), 'Dimensionality reduction for visualizing single-cell data using umap', *Nature biotechnology* **37**(1), 38–44.
- Begg, C. B. and Greenes, R. A. (1983), 'Assessment of diagnostic tests when disease verification is subject to selection bias', *Biometrics* pp. 207–215.
- Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A. et al. (2019), Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias, Vol. 63, IBM, pp. 4–1.
- Ben-Michael, E., Greiner, D. J., Huang, M., Imai, K., Jiang, Z. and Shin, S. (2024), 'Does ai help humans make better decisions? a methodological framework for experimental evaluation', *arXiv preprint arXiv:2403.12108*.
- Berk, R. A. (1983), 'An introduction to sample selection bias in sociological data', *American sociological review* pp. 386–398.
- Berman, G., Goyal, N. and Madaio, M. (2024), A scoping study of evaluation practices for responsible ai tools: Steps towards effectiveness evaluations, in 'Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems', pp. 1–24.
- Bhattacharya, A., Stumpf, S., Gosak, L., Stiglic, G. and Verbert, K. (2024), Exmos: Explanatory model steering through multifaceted explanations and data configurations, in 'Proceedings of the CHI Conference on Human Factors in Computing Systems', pp. 1–27.
- Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z. et al. (2024), 'Deepseek llm: Scaling open-source language models with longtermism', *arXiv preprint arXiv:2401.02954*.

- Bishop, C. M. (1998), 'Latent variable models.', *Learning in graphical models* **371**.
- Bogacz, R., Brown, E., Moehlis, J., Holmes, P. and Cohen, J. D. (2006), 'The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks.', *Psychological review* **113**(4), 700.
- Bogatu, A., Fernandes, A. A., Paton, N. W. and Konstantinou, N. (2020), Dataset discovery in data lakes, in '2020 IEEE 36th International Conference on Data Engineering (ICDE)', IEEE, pp. 709–720.
- Borkan, D., Dixon, L., Sorensen, J., Thain, N. and Vasserman, L. (2019), 'Nuanced metrics for measuring unintended bias with real data for text classification', *CoRR* **abs/1903.04561**.
URL: <http://arxiv.org/abs/1903.04561>
- Bowker, G. C. (2000), *Sorting Things Out: Classification and Its Consequences*, MIT press.
- Bowman, S. R., Angeli, G., Potts, C. and Manning, C. D. (2015), 'A large annotated corpus for learning natural language inference', *arXiv preprint arXiv:1508.05326*.
- Bragg, J., Mausam and Weld, D. S. (2018), Sprout: Crowd-powered task design for crowd-sourcing, in 'Proceedings of the 31st annual ACM symposium on user interface software and technology', pp. 165–176.
- Braun, V. and Clarke, V. (2006), Using thematic analysis in psychology, Vol. 3, Taylor & Francis, pp. 77–101.
- Brown, A. and Maydeu-Olivares, A. (2018), 'Modelling forced-choice response formats', *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* pp. 523–569.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S. et al. (2023), 'Sparks of artificial general intelligence: Early experiments with gpt-4', *arXiv preprint arXiv:2303.12712*.
- Bucinca, Z., Lin, P., Gajos, K. Z. and Glassman, E. L. (2020), Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems, in 'Proceedings of the 25th International Conference on Intelligent User Interfaces', pp. 454–464.
- Buçinca, Z., Malaya, M. B. and Gajos, K. Z. (2021), 'To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making', *Proceedings of the ACM on Human-Computer Interaction* **5**(CSCW1), 1–21.
- Büscher, M., Gill, S., Mogensen, P. and Shapiro, D. (2001), Landscapes of practice: Bricolage as a method for situated design, Vol. 10, Springer, pp. 1–28.
- Bussone, A., Stumpf, S. and O'Sullivan, D. (2015), The role of explanations on trust and reliance in clinical decision support systems, in '2015 international conference on healthcare informatics', IEEE, pp. 160–169.
- Butcher, B., Robinson, C., Zilka, M., Fogliato, R., Ashurst, C. and Weller, A. (2022), Racial disparities in the enforcement of marijuana violations in the us, in 'Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society', pp. 130–143.

- Cabitza, F., Campagner, A. and Basile, V. (2023), Toward a perspectivist turn in ground truthing for predictive computing, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 37, pp. 6860–6868.
- Cabrera, Á. A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J. and Chau, D. H. (2019), Fairvis: Visual analytics for discovering intersectional bias in machine learning, *in* '2019 IEEE Conference on Visual Analytics Science and Technology (VAST)', IEEE, pp. 46–56.
- Cabrera, Á. A., Fu, E., Bertucci, D., Holstein, K., Talwalkar, A., Hong, J. I. and Perer, A. (2023), Zeno: An interactive framework for behavioral evaluation of machine learning, *in* 'Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems', pp. 1–14.
- Cabrera, Á. A., Tulio Ribeiro, M., Lee, B., Deline, R., Perer, A. and Drucker, S. M. (2023), What did my ai learn? how data scientists make sense of model behavior, Vol. 30, ACM New York, NY, pp. 1–27.
- Cai, C. J., Winter, S., Steiner, D., Wilcox, L. and Terry, M. (2019), '" hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making', *Proceedings of the ACM on Human-computer Interaction* **3**(CSCW), 1–24.
- Cai, C. J., Winter, S., Steiner, D., Wilcox, L. and Terry, M. (2021), Onboarding materials as cross-functional boundary objects for developing ai assistants, *in* 'Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems', pp. 1–7.
- Campbell, D. T. (2017), Factors relevant to the validity of experiments in social settings, Routledge, pp. 243–263.
- Campbell, D. T. and Fiske, D. W. (1959), 'Convergent and discriminant validation by the multitrait-multimethod matrix.', *Psychological bulletin* **56**(2), 81.
- Carlson, K. D. and Herdman, A. O. (2012), 'Understanding the impact of convergent validity on research results', *Organizational Research Methods* **15**(1), 17–32.
- Carton, S., Mei, Q. and Resnick, P. (2020), Feature-based explanations don't help people detect misclassifications of online toxicity, *in* 'Proceedings of the international AAAI conference on web and social media', Vol. 14, pp. 95–106.
- Castricato, L., Lile, N., Rafailov, R., Fränken, J.-P. and Finn, C. (2024), 'Persona: A reproducible testbed for pluralistic alignment', *arXiv preprint arXiv:2407.17387*.
- Castro Fernandez, R. (2023), Data-sharing markets: model, protocol, and algorithms to incentivize the formation of data-sharing consortia, Vol. 1, ACM New York, NY, USA, pp. 1–25.
- Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J. and Mullainathan, S. (2016), 'Productivity and selection of human capital with machine learning', *American Economic Review* **106**(5), 124–127.
- Chang, J. C., Amershi, S. and Kamar, E. (2017), Revolt: Collaborative crowdsourcing for labeling machine learning datasets, *in* 'Proceedings of the 2017 CHI conference on human factors in computing systems', pp. 2334–2346.

- Charusaie, M.-A., Mozannar, H., Sontag, D. and Samadi, S. (2022), Sample efficient learning of predictors that complement humans, *in* 'International Conference on Machine Learning', PMLR, pp. 2972–3005.
- Chatzi, I., Straitouri, E., Thejaswi, S. and Rodriguez, M. G. (2024), 'Prediction-powered ranking of large language models', *arXiv preprint arXiv:2402.17826* .
- Chaudhary, M., Gupta, H., Bhat, S. and Varma, V. (2024), 'Towards understanding the robustness of llm-based evaluations under perturbations', *arXiv preprint arXiv:2412.09269* .
- Chehbouni, K., Haddou, M., Cheung, J. C. and Farnadi, G. (2026), 'Neither valid nor reliable? investigating the use of llms as judges', *Advances in Neural Information Processing Systems* **38**.
- Chen, B., Wang, X., Peng, S., Litschko, R., Korhonen, A. and Plank, B. (2024), '" seeing the big through the small": Can llms approximate human judgment distributions on nli from a few explanations?', *arXiv preprint arXiv:2406.17600* .
- Chen, D., Chen, R., Zhang, S., Wang, Y., Liu, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P. and Sun, L. (2024), Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark, *in* 'Forty-first International Conference on Machine Learning'.
- Chen, J., Li, Z. and Mao, X. (2023), 'Learning under selective labels with data from heterogeneous decision-makers: An instrumental variable approach', *arXiv preprint arXiv:2306.07566* .
- Chen, P., Ye, J., Chen, G., Zhao, J. and Heng, P.-A. (2021), Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 35, pp. 11442–11450.
- Chen, Q. Z. and Zhang, A. X. (2023), Judgment sieve: Reducing uncertainty in group judgments through interventions targeting ambiguity versus disagreement, Vol. 7, ACM New York, NY, USA, pp. 1–26.
- Cheng, H.-F., Stapleton, L., Kawakami, A., Sivaraman, V., Cheng, Y., Qing, D., Perer, A., Holstein, K., Wu, Z. S. and Zhu, H. (2022), How child welfare workers reduce racial disparities in algorithmic decisions, *in* 'Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems'.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. (2017), 'Double/debiased/neyman machine learning of treatment effects', *American Economic Review* **107**(5), 261–265.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018), 'Double/debiased machine learning for treatment and structural parameters'.
- Chernozhukov, V., Newey, M., Newey, W. K., Singh, R. and Srygkanis, V. (2023), 'Automatic debiased machine learning for covariate shifts', *arXiv preprint arXiv:2307.04527* .
- Chernozhukov, V., Newey, W. K. and Singh, R. (2022), 'Automatic debiased machine learning of causal and structural effects', *Econometrica* **90**(3), 967–1027.
- Chernozhukov, V., Newey, W., Quintas-Martinez, V. M. and Srygkanis, V. (2022), Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests, *in* 'International Conference on Machine Learning', PMLR, pp. 3901–3914.

- Chiang, C., Gebreegziabher, S., Szymanski, A., Jin Do, H., Ashktorab, Z., Geyer, W., Jia-Jun Li, T. and Gomez-Zara, D. (2026), Multeval: Supporting collaborative alignment for llm-as-a-judge evaluation criteria, in 'Proceedings of the 5th Annual Symposium on Human-Computer Interaction for Work', pp. 1–17.
- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J. E. et al. (2024), Chatbot arena: An open platform for evaluating llms by human preference, in 'Forty-first International Conference on Machine Learning'.
- Chou, Y.-T., Niu, G., Lin, H.-T. and Sugiyama, M. (2020), Unbiased risk estimators can mislead: A case study of learning with complementary labels, in 'International Conference on Machine Learning', PMLR, pp. 1929–1938.
- Chouldechova, A., Benavides-Prado, D., Fialko, O. and Vaithianathan, R. (2018), A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions, in 'Conference on Fairness, Accountability and Transparency', PMLR, pp. 134–148.
- Clarke, V. and Braun, V. (2017), Thematic analysis, Vol. 12, Taylor & Francis, pp. 297–298.
- Cohen, J. (1992), 'Statistical power analysis', *Current directions in psychological science* 1(3), 98–101.
- Collins, K. M., Bhatt, U. and Weller, A. (2022), Eliciting and learning with soft labels from every annotator, in 'Proceedings of the AAAI conference on human computation and crowdsourcing', Vol. 10, pp. 40–52.
- Coston, A., Kawakami, A., Zhu, H., Holstein, K. and Heidari, H. (2023a), A validity perspective on evaluating the justified use of data-driven decision-making algorithms, in '2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)', IEEE, pp. 690–704.
- Coston, A., Kennedy, E. and Chouldechova, A. (2020), 'Counterfactual predictions under runtime confounding', *Advances in Neural Information Processing Systems* 33, 4150–4162.
- Coston, A. L., Kawakami, A., Zhu, H., Holstein, K. and Heidari, H. (2023b), 'Sok: A validity perspective on evaluating the justified use of data-driven decision-making algorithms', *First IEEE Conference on Secure and Trustworthy Machine Learning*.
- Coston, A., Mishler, A., Kennedy, E. H. and Chouldechova, A. (2020), Counterfactual risk assessments, evaluation, and fairness, in 'Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency', pp. 582–593.
- Coston, A., Rambachan, A. and Chouldechova, A. (2021), Characterizing fairness over the set of good models under selective labels, in 'International Conference on Machine Learning', PMLR, pp. 2144–2155.
- Crosby, A. W. (1997), *The measure of reality: Quantification in Western Europe, 1250-1600*, Cambridge University Press.
- Cui, K. Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S. and Salz, T. (2026), 'The effects of generative ai on high-skilled work: Evidence from three field experiments with software developers', *Management Science*.
- Daudert, T. (2020), A web-based collaborative annotation and consolidation tool, in 'Proceedings of the Twelfth Language Resources and Evaluation Conference', pp. 7053–7059.

- Davani, A. M., Díaz, M. and Prabhakaran, V. (2022), 'Dealing with disagreements: Looking beyond the majority vote in subjective annotations', *Transactions of the Association for Computational Linguistics* **10**, 92–110.
- Dawes, R. M., Faust, D. and Meehl, P. E. (1989), 'Clinical versus actuarial judgment', *Science* **243**(4899), 1668–1674.
- Daysal, N. M., Mullainathan, S., Obermeyer, Z., Sarkar, S. K. and Trandafir, M. (2022), 'An economic approach to machine learning in health policy', *Univ. of Copenhagen Dept. of Economics Discussion, CEBI Working Paper* (24).
- De-Arteaga, M., Dubrawski, A. and Chouldechova, A. (2018), 'Learning under selective labels in the presence of expert consistency', *arXiv preprint arXiv:1807.00905*.
- De-Arteaga, M., Jeanselme, V., Dubrawski, A. and Chouldechova, A. (2021), Leveraging expert consistency to improve algorithmic decision support.
- Deng, W. H., Kim, S. S., Jha, A., Holstein, K., Eslami, M., Wilcox, L. and Gatys, L. A. (2025), 'Personateaming: Exploring how introducing personas can improve automated ai red-teaming', *arXiv preprint arXiv:2509.03728*.
- Deng, W. H., Nagireddy, M., Lee, M. S. A., Singh, J., Wu, Z. S., Holstein, K. and Zhu, H. (2022), Exploring how machine learning practitioners (try to) use fairness toolkits, in 'Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency', pp. 473–484.
- Denteh, A. and Liebert, H. (2022), 'Who increases emergency department use? new insights from the oregon health insurance experiment', *arXiv preprint arXiv:2201.07072*.
- Denton, E., Díaz, M., Kivlichan, I., Prabhakaran, V. and Rosen, R. (2021), 'Whose ground truth? accounting for individual and collective identities underlying dataset annotation', *arXiv preprint arXiv:2112.04554*.
- Desmond, M., Ashktorab, Z., Geyer, W., Daly, E. M., Cooper, M. S., Pan, Q., Nair, R., Wagner, N. and Pedapati, T. (2025), Evalassist: Llm-as-a-judge simplified, in 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 39, pp. 29637–29639.
- Desmond, M., Ashktorab, Z., Pan, Q., Dugan, C. and Johnson, J. M. (2024), Evalullm: Llm assisted evaluation of generative outputs, in 'Companion proceedings of the 29th international conference on intelligent user interfaces', pp. 30–32.
- Desmond, M., Muller, M., Ashktorab, Z., Dugan, C., Duesterwald, E., Brimijoin, K., Finegan-Dollak, C., Brachman, M., Sharma, A., Joshi, N. N. et al. (2021), Increasing the speed and accuracy of data labeling through an ai assisted interface, in 'Proceedings of the 26th International Conference on Intelligent User Interfaces', pp. 392–401.
- Dhar, R. and Simonson, I. (2003), 'The effect of forced choice on choice', *Journal of marketing research* **40**(2), 146–160.
- Di Folco, C., Guez, A., Peyre, H. and Ramus, F. (2022), 'Epidemiology of reading disability: A comparison of dsm-5 and icd-11 criteria', *Scientific Studies of Reading* **26**(4), 337–355.
- Diana, E., Gill, W., Kearns, M., Kenthapadi, K., Roth, A. and Sharifi-Malvajerdi, S. (2022), Multiaccurate proxies for downstream fairness, in '2022 ACM Conference on Fairness, Accountability, and Transparency', pp. 1207–1239.

- Díaz, I. and van der Laan, M. J. (2013), 'Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems', *The international journal of biostatistics* **9**(2), 149–160.
- Díaz, M. and Smith, A. D. (2024), What makes an expert? reviewing how ml researchers define "expert", in 'Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society', Vol. 7, pp. 358–370.
- Dieterich, W., Mendoza, C. and Brennan, T. (2016), 'Compas risk scales: Demonstrating accuracy equity and predictive parity', *Northpointe Inc* **7**(4).
- Dingen, D., van't Veer, M., Houthuizen, P., Mestrom, E. H., Korsten, E. H., Bouwman, A. R. and Van Wijk, J. (2018), Regressionexplorer: Interactive exploration of logistic regression models with subgroup analysis, Vol. 25, IEEE, pp. 246–255.
- Donahue, K., Chouldechova, A. and Kenthapadi, K. (2022), 'Human-algorithm collaboration: Achieving complementarity and avoiding unfairness', *arXiv preprint arXiv:2202.08821* .
- Dong, Y. R., Hu, T. and Collier, N. (2024), 'Can llm be a personalized judge?', *arXiv preprint arXiv:2406.11657* .
- Dorner, F. E., Nastl, V. Y. and Hardt, M. (2024), 'Limits to scalable evaluation at the frontier: Llm as judge won't beat twice the data', *arXiv preprint arXiv:2410.13341* .
- Drost, E. A. (2011), Validity and reliability in social science research, Vol. 38, University of Western Australia, Department of Education [Nedlands, WA], pp. 105–123.
- Dsouza, R. and Kovatchev, V. (2025), Sources of disagreement in data for llm instruction tuning, in 'Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation', pp. 20–32.
- Dubois, Y., Li, C. X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P. S. and Hashimoto, T. B. (2024), 'Alpacafarm: A simulation framework for methods that learn from human feedback', *Advances in Neural Information Processing Systems* **36**.
- Duboue, P. (2020), *The art of feature engineering: essentials for machine learning*, Cambridge University Press.
- Dudik, M., Erhan, D., Langford, J. and Li, L. (2014), 'Doubly robust policy evaluation and optimization', *Statistical Science* **29**(4), 485–511.
- Duymedjian, R. and Rüling, C.-C. (2010), Towards a foundation of bricolage in organization and management theory, Vol. 31, SAGE Publications Sage UK: London, England, pp. 133–151.
- Ek, S., Zachariah, D., Johansson, F. D. and Stoica, P. (2023), 'Off-policy evaluation with out-of-sample guarantees', *Transactions on Machine Learning Research* .
- Elangovan, A., Xu, L., Ko, J., Elyasi, M., Liu, L., Bodapati, S. and Roth, D. (2024), 'Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and llm-as-a-judge', *arXiv preprint arXiv:2410.03775* .
- Enøe, C., Georgiadis, M. P. and Johnson, W. O. (2000), 'Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown', *Preventive veterinary medicine* **45**(1-2), 61–81.

- Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C. and Venkatasubramanian, S. (2018), Runaway feedback loops in predictive policing, in 'Conference on Fairness, Accountability and Transparency', PMLR, pp. 160–171.
- Es, S., James, J., Espinosa-Anke, L. and Schockaert, S. (2023), 'Ragas: Automated evaluation of retrieval augmented generation', *arXiv preprint arXiv:2309.15217*.
- Evenson, S. (2016), Driving service design by directed storytelling, in 'Design for services', Routledge, pp. 66–72.
- Everett, B. (2013), *An introduction to latent variable models*, Springer Science & Business Media.
- Eyre, B. and Madras, D. (2024), 'Auto-evaluation with few labels through post-hoc regression', *arXiv preprint arXiv:2411.12665*.
- Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R. and Radev, D. (2021), 'Summeval: Re-evaluating summarization evaluation', *Transactions of the Association for Computational Linguistics* **9**, 391–409.
- Feffer, M., Skirpan, M., Lipton, Z. and Heidari, H. (2023), From preference elicitation to participatory ml: A critical survey & guidelines for future research, in 'Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society', pp. 38–48.
- Fernandez, R. C., Abedjan, Z., Koko, F., Yuan, G., Madden, S. and Stonebraker, M. (2018), Aurum: A data discovery system, in '2018 IEEE 34th International Conference on Data Engineering (ICDE)', IEEE, pp. 1001–1012.
- Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trounev, A. and Peyré, G. (2019), Interpolating between optimal transport and mmd using sinkhorn divergences, in 'The 22nd International Conference on Artificial Intelligence and Statistics', pp. 2681–2690.
- Figueira, A. and Vaz, B. (2022), Survey on synthetic data generation, evaluation methods and gans, Vol. 10, MDPI, p. 2733.
- Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K. and Group, O. H. S. (2012), 'The oregon health insurance experiment: evidence from the first year', *The Quarterly Journal of Economics* **127**(3), 1057–1106.
- Finkelstein, N., Adams, R., Saria, S. and Shpitser, I. (2021), Partial identifiability in discrete data with measurement error, in 'Uncertainty in Artificial Intelligence', PMLR, pp. 1798–1808.
- Fisch, A., Maynez, J., Hofer, R. A., Dhingra, B., Globerson, A. and Cohen, W. W. (2024), 'Stratified prediction-powered inference for hybrid language model evaluation', *arXiv preprint arXiv:2406.04291*.
- Fleisig, E., Blodgett, S. L., Klein, D. and Talat, Z. (2024), 'The perspectivist paradigm shift: Assumptions and challenges of capturing human labels', *arXiv preprint arXiv:2405.05860*.
- Fogliato, R., Chouldechova, A. and G'Sell, M. (2020), Fairness evaluation in presence of biased noisy labels, in 'International Conference on Artificial Intelligence and Statistics', PMLR, pp. 2325–2336.

- Fogliato, R., Chouldechova, A. and Lipton, Z. (2021), 'The impact of algorithmic risk assessments on human predictions and its analysis via crowdsourcing studies', *Proceedings of the ACM on Human-Computer Interaction* **5**(CSCW2), 1–24.
- Fogliato, R., Patil, P., Monfort, M. and Perona, P. (2024), A framework for efficient model evaluation through stratification, sampling, and estimation, *in* 'European Conference on Computer Vision', Springer, pp. 140–158.
- Fogliato, R., Xiang, A., Lipton, Z., Nagin, D. and Chouldechova, A. (2021), On the validity of arrest as a proxy for offense: Race and the likelihood of arrest for violent crimes, *in* 'Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society', pp. 100–111.
- Fornaciari, T., Uma, A., Paun, S., Plank, B., Hovy, D., Poesio, M. et al. (2021), Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning, *in* 'Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', Association for Computational Linguistics.
- Frenda, S., Abercrombie, G., Basile, V., Pedrani, A., Panizzon, R., Cignarella, A. T., Marco, C. and Bernardi, D. (2024), 'Perspectivist approaches to natural language processing: a survey', *Language Resources and Evaluation* pp. 1–28.
- Fröhling, L., Demartini, G. and Assenmacher, D. (2024), 'Personas with attitudes: Controlling llms for diverse data annotation', *arXiv preprint arXiv:2410.11745*.
- Fuller, M. and Moore, R. (2017), *An Analysis of Jane Jacobs's The Death and Life of Great American Cities*, Macat Library.
- Gajos, K. Z. and Mamykina, L. (2022), Do people engage cognitively with ai? impact of ai assistance on incidental learning, *in* '27th International Conference on Intelligent User Interfaces', pp. 794–806.
- Gala, D., Phillips-Brown, M., Goel, N., Prunkl, C., Jubete, L. A., Eitel-Porter, R. et al. (2024), Fairtargetsim: An interactive simulator for understanding and explaining the fairness effects of target variable definition.
- Gamerman, V., Cai, T. and Elsässer, A. (2019), 'Pragmatic randomized clinical trials: best practices and statistical guidance', *Health Services and Outcomes Research Methodology* **19**, 23–35.
- Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K. et al. (2022), 'Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned', *arXiv preprint arXiv:2209.07858*.
- Gao, R., Saar-Tsechansky, M., De-Arteaga, M., Han, L., Lee, M. K. and Lease, M. (2021), 'Human-ai collaboration with bandit feedback', *arXiv preprint arXiv:2105.10614*.
- Gao, R. and Yin, M. (2023), 'Confounding-robust policy improvement with human-ai teams', *arXiv preprint arXiv:2310.08824*.
- Gao, Y., Xu, G., Wang, Z. and Cohan, A. (2024), 'Bayesian calibration of win rate estimation with llm evaluators', *arXiv preprint arXiv:2411.04424*.

- Gasmalla, H. E., Wadi, M. and Taha, M. H. (2021), 'Twelve tips for introducing the concept of validity argument in assessment to novice medical teachers in a workshop', *MedEdPublish* **10**, 74.
- Gebreegziabher, S. A., Chiang, C., Wang, Z., Ashktorab, Z., Brachman, M., Geyer, W., Li, T. J.-J. and Gómez-Zarà, D. (2025), Metricmate: An interactive tool for generating evaluation criteria for llm-as-a-judge workflow, in 'Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work', pp. 1–18.
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D. and Crawford, K. (2021), Datasheets for datasets, Vol. 64, ACM New York, NY, USA, pp. 86–92.
- Gergle, D. and Tan, D. S. (2014), Experimental research in hci, in 'Ways of Knowing in HCI', Springer, pp. 191–227.
- Ghassami, A., Shpitser, I. and Tchetgen, E. T. (2023), 'Partial identification of causal effects using proxy variables', *arXiv preprint arXiv:2304.04374*.
- Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J., Hashimoto, T. and Bernstein, M. S. (2022), Jury learning: Integrating dissenting voices into machine learning models, in 'Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems', pp. 1–19.
- Gordon, M. L., Zhou, K., Patel, K., Hashimoto, T. and Bernstein, M. S. (2021), The disagreement deconvolution: Bringing machine learning performance metrics in line with reality, in 'Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems', pp. 1–14.
- Goyal, N., Kivlichan, I. D., Rosen, R. and Vasserman, L. (2022), 'Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation', *Proceedings of the ACM on Human-Computer Interaction* **6**(CSCW2), 1–28.
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A. et al. (2024), 'The llama 3 herd of models', *arXiv preprint arXiv:2407.21783*.
- Green, B. and Chen, Y. (2019), 'The principles and limits of algorithm-in-the-loop decision making', *Proceedings of the ACM on Human-Computer Interaction* **3**(CSCW), 1–24.
- Green, B. and Chen, Y. (2021), 'Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts', *Proceedings of the ACM on Human-Computer Interaction* **5**(CSCW2), 1–33.
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E. and Nelson, C. (2000), 'Clinical versus mechanical prediction: a meta-analysis', *Psychological assessment* **12**(1), 19.
- Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H. et al. (2026), 'A survey on llm-as-a-judge', *The Innovation* **7**(6).
- Guerdan, L., Coston, A., Holstein, K. and Wu, Z. S. (2023), Counterfactual prediction under outcome measurement error, in 'Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency', pp. 1584–1598.

- Guerdan, L., Coston, A., Wu, Z. S. and Holstein, K. (2023), Ground(less) truth: A causal framework for proxy labels in human-algorithm decision-making, *in* 'Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency', pp. 688–704.
- Guerdan, L., Saxena, D., Chancellor, S., Wu, Z. S. and Holstein, K. (2025), 'Measurement as bricolage: Examining how data scientists construct target variables for predictive modeling tasks', *Proceedings of the ACM on Human-Computer Interaction* **9**(7), 1–37.
- Guerdan*, L., Whitehouse*, J., Truong*, K., Holstein, K. and Wu, Z. S. (2026), 'Doubly-robust llm-as-a-judge: Externally valid estimation with imperfect personas', *International Conference on Learning Representations (ICLR)*.
- Guo, P. J., Kandel, S., Hellerstein, J. M. and Heer, J. (2011), Proactive wrangling: Mixed-initiative end-user programming of data transformation scripts, *in* 'Proceedings of the 24th annual ACM symposium on User interface software and technology', pp. 65–74.
- Gupta, N. R., Hullman, J. and Subramonyam, H. (2024), A conceptual framework for ethical evaluation of machine learning systems.
- Guzman, G. and Kollar, M. (2023), 'Income in the united states: 2022', *United States Census Bureau*. <https://www.census.gov/content/dam/Census/library/publications/2023/demo/p60-279.pdf>.
- Hacking, I. (1999), *The social construction of what?*, Harvard university press.
- Halfaker, A. and Geiger, R. S. (2020), 'Ores: Lowering barriers with participatory machine learning in wikipedia', *Proceedings of the ACM on Human-Computer Interaction* **4**(CSCW2), 1–37.
- Han, B., Yao, Q., Liu, T., Niu, G., Tsang, I. W., Kwok, J. T. and Sugiyama, M. (2020), 'A survey of label-noise representation learning: Past, present and future', *arXiv preprint arXiv:2011.04406*.
- Hand, D. J. (2004), 'Measurement theory and practice', *London: Arnold*.
- Haque, M. R., Saxena, D., Weathington, K., Chudzik, J. and Guha, S. (2024), Are we asking the right questions?: Designing for community stakeholders' interactions with ai in policing, *in* 'Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems', pp. 1–20.
- Hardt, M. and Kim, M. P. (2022), 'Backward baselines: Is your model predicting the past?', *arXiv preprint arXiv:2206.11673*.
- Harvey, E., Sandhaus, H., Jacobs, A. Z., Moss, E. and Sloane, M. (2024), The cadaver in the machine: The social practices of measurement and validation in motion capture technology, *in* 'Proceedings of the CHI Conference on Human Factors in Computing Systems', pp. 1–23.
- Harvey, E., Sheng, E., Blodgett, S. L., Chouldechova, A., Garcia-Gathright, J., Olteanu, A. and Wallach, H. (2025), Understanding and meeting practitioner needs when measuring representational harms caused by llm-based systems, *in* 'Findings of the Association for Computational Linguistics: ACL 2025', pp. 18423–18440.
- Hatt, T., Tschernutter, D. and Feuerriegel, S. (2022), Generalizing off-policy learning under sample selection bias, *in* 'Uncertainty in Artificial Intelligence', PMLR, pp. 769–779.

- Hazzan, O. and Tomayko, J. (2004), Human aspects of software engineering: The case of extreme programming, in 'International Conference on Extreme Programming and Agile Processes in software Engineering', Springer, pp. 303–311.
- He, Z., Naphade, S. and Huang, T.-H. K. (2025), Prompting in the dark: Assessing human performance in prompt engineering for data labeling when gold labels are absent, in 'Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems', pp. 1–33.
- Heckathorn, D. D. (2011), Comment: Snowball versus respondent-driven sampling, Vol. 41, SAGE Publications Sage CA: Los Angeles, CA, pp. 355–366.
- Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica: Journal of the econometric society* pp. 153–161.
- Hemmer, P., Schemmer, M., Kühl, N., Vössing, M. and Satzger, G. (2022), 'On the effect of information asymmetry in human-ai teams', *arXiv preprint arXiv:2205.01467* .
- Hilgard, S., Rosenfeld, N., Banaji, M. R., Cao, J. and Parkes, D. (2021), Learning representations by humans, for humans, in 'International Conference on Machine Learning', PMLR, pp. 4227–4238.
- Hill, J. L. (2011), 'Bayesian nonparametric modeling for causal inference', *Journal of Computational and Graphical Statistics* **20**(1), 217–240.
- Hoess, N., Paradis, C., Kazman, R. and Maurer, W. (2025), 'Does the tool matter? exploring some causes of threats to validity in mining software repositories', *arXiv preprint arXiv:2501.15114* .
- Hofman, J. M., Chatzimparmpas, A., Sharma, A., Watts, D. J. and Hullman, J. (2023), Pre-registration for predictive modeling.
- Holstein, K., De-Arteaga, M., Tumati, L. and Cheng, Y. (2023), 'Toward supporting perceptual complementarity in human-ai collaboration via reflection on unobservables', *Proceedings of the ACM on Human-Computer Interaction* **7**(CSCW1), 1–20.
- Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M. and Wallach, H. (2019), Improving fairness in machine learning systems: What do industry practitioners need?, in 'Proceedings of the 2019 CHI conference on human factors in computing systems', pp. 1–16.
- Hopkins, P. and Unger, M. (2017), 'What is a subject-matter expert?', *Journal of Pipeline Engineering* **16**(4).
- Hou, Y. and Wang, D. (2017), Hacking with npos: collaborative analytics and broker roles in civic data hackathons, Vol. 1, ACM New York, NY, USA, pp. 1–16.
- Huber, M. (2021), An introduction to flexible methods for policy evaluation, in 'Handbook of Research Methods and Applications in Empirical Microeconomics', Edward Elgar Publishing, pp. 82–111.
- Hudgens, M. G. and Halloran, M. E. (2008), 'Toward causal inference with interference', *Journal of the American Statistical Association* **103**(482), 832–842.
- Hui, S. L. and Walter, S. D. (1980), 'Estimating the error rates of diagnostic tests', *Biometrics* pp. 167–171.

- Hur, P., Lee, H., Bhat, S. and Bosch, N. (2022), 'Using machine learning explainability methods to personalize interventions for students.', *International Educational Data Mining Society* .
- Ibrahim, L., Huang, S., Ahmad, L. and Anderljung, M. (2024), 'Beyond static ai evaluations: advancing human interaction evaluations for llm harms and risks', *arXiv preprint arXiv:2405.10632* .
- Ishikawa, K. and He, N. (2023), Kernel conditional moment constraints for confounding robust inference, in 'International Conference on Artificial Intelligence and Statistics', PMLR, pp. 650–674.
- Jacobs, A. Z. and Wallach, H. (2021), Measurement and fairness, in 'Proceedings of the 2021 ACM conference on fairness, accountability, and transparency', pp. 375–385.
- Ji, W., Lei, L. and Zrnic, T. (2025), 'Predictions as surrogates: Revisiting surrogate outcomes in the age of ai', *arXiv preprint arXiv:2501.09731* .
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T. and Sayed, W. E. (2023), 'Mistral 7b'.
URL: <https://arxiv.org/abs/2310.06825>
- Jiang, N. and Li, L. (2016), Doubly robust off-policy value evaluation for reinforcement learning, in 'International Conference on Machine Learning', PMLR, pp. 652–661.
- Johansson, F. D., Shalit, U., Kallus, N. and Sontag, D. (2020), 'Generalization bounds and representation learning for estimation of potential outcomes and causal effects', *arXiv preprint arXiv:2001.07426* .
- Johnson, N., Sudharsan, D., Hamna, Dalal, S., Holroyd, T., Thieme, A., Heidari, H., Massiceti, D., Wortman Vaughan, J. and Morrison, C. (2026), Evaluating ai-generated images of cultural artifacts with community-informed rubrics, in 'The 2026 ACM Conference on Fairness, Accountability, and Transparency', pp. 714–774.
- Johnson, R. A. and Zhang, S. (2022), What is the bureaucratic counterfactual? categorical versus algorithmic prioritization in us social policy, in 'Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency', pp. 1671–1682.
- Jun, E., Birchfield, M., De Moura, N., Heer, J. and Just, R. (2022), 'Hypothesis formalization: Empirical findings, software limitations, and design implications', *ACM Transactions on Computer-Human Interaction (TOCHI)* **29**(1), 1–28.
- Jun, E., Seo, A., Heer, J. and Just, R. (2022), Tisane: Authoring statistical models via formal reasoning from conceptual and data relationships, in 'Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems', pp. 1–16.
- Jung, J., Brahman, F. and Choi, Y. (2024), 'Trust or escalate: Llm judges with provable guarantees for human agreement', *arXiv preprint arXiv:2407.18370* .
- Jung, J., Shroff, R., Feller, A. and Goel, S. (2020), Bayesian sensitivity analysis for offline policy evaluation, in 'Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society', pp. 64–70.

- K. Chaithanya Manam, V., Jampani, D., Zaim, M., Wu, M.-H. and J. Quinn, A. (2019), Taskmate: A mechanism to improve the quality of instructions in crowdsourcing, *in* 'Companion Proceedings of The 2019 World Wide Web Conference', pp. 1121–1130.
- Kahng, M., Tenney, I., Pushkarna, M., Liu, M. X., Wexler, J., Reif, E., Kallarackal, K., Chang, M., Terry, M. and Dixon, L. (2024), Llm comparator: Visual analytics for side-by-side evaluation of large language models, *in* 'Extended Abstracts of the CHI Conference on Human Factors in Computing Systems', pp. 1–7.
- Kallus, N., Mao, X. and Zhou, A. (2022), 'Assessing algorithmic fairness with unobserved protected class using data combination', *Management Science* **68**(3), 1959–1981.
- Kallus, N. and Zhou, A. (2018a), 'Confounding-robust policy improvement', *Advances in neural information processing systems* **31**.
- Kallus, N. and Zhou, A. (2018b), Residual unfairness in fair machine learning from prejudiced data, *in* 'International Conference on Machine Learning', PMLR, pp. 2439–2448.
- Kallus, N. and Zhou, A. (2020), 'Confounding-robust policy evaluation in infinite-horizon reinforcement learning', *Advances in neural information processing systems* **33**, 22293–22304.
- Kallus, N. and Zhou, A. (2021), 'Minimax-optimal policy learning under unobserved confounding', *Management Science* **67**(5), 2870–2890.
- Kandel, S., Paepcke, A., Hellerstein, J. and Heer, J. (2011), Wrangler: Interactive visual specification of data transformation scripts, *in* 'Proceedings of the sigchi conference on human factors in computing systems', pp. 3363–3372.
- Kandel, S., Paepcke, A., Hellerstein, J. M. and Heer, J. (2012), Enterprise data analysis and visualization: An interview study, Vol. 18, IEEE, pp. 2917–2926.
- Kapoor, S. and Narayanan, A. (2023), 'Leakage and the reproducibility crisis in machine-learning-based science', *Patterns* **4**(9).
- Kawakami, A., Coston, A., Zhu, H., Heidari, H. and Holstein, K. (2024), The situate ai guidebook: Co-designing a toolkit to support multi-stakeholder, early-stage deliberations around public sector ai proposals, *in* 'Proceedings of the CHI Conference on Human Factors in Computing Systems', pp. 1–22.
- Kawakami, A., Sivaraman, V., Cheng, H.-F., Stapleton, L., Cheng, Y., Qing, D., Perer, A., Wu, Z. S., Zhu, H. and Holstein, K. (2022), Improving human-ai partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support, *in* 'CHI Conference on Human Factors in Computing Systems', pp. 1–18.
- Kawakami, A., Zhu, H. and Holstein, K. (2026), 'Ai measurement as a collaborative design practice: A multidisciplinary synthesis and future directions', *Available at SSRN* 6276543 .
- Kennedy, E., Balakrishnan, S. and Wasserman, L. (2023), 'Semiparametric counterfactual density estimation', *Biometrika* **110**(4), 875–896.
- Kennedy, E. H. (2022), 'Semiparametric doubly robust targeted double machine learning: a review', *arXiv preprint arXiv:2203.06469* .
- Kennedy, E. H. (2023), 'Towards optimal doubly robust estimation of heterogeneous causal effects', *Electronic Journal of Statistics* **17**(2), 3008–3049.

- Kery, M. B. and Myers, B. A. (2017), Exploring exploratory programming, in '2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)', IEEE, pp. 25–29.
- Kery, M. B., Radensky, M., Arya, M., John, B. E. and Myers, B. A. (2018), The story in the notebook: Exploratory data science using a literate programming tool, in 'Proceedings of the 2018 CHI conference on human factors in computing systems', pp. 1–11.
- Kim, M., Zimmermann, T., DeLine, R. and Begel, A. (2016), The emerging role of data scientists on software development teams, in 'Proceedings of the 38th International Conference on Software Engineering', pp. 96–107.
- Kim, T. S., Lee, Y., Shin, J., Kim, Y.-H. and Kim, J. (2024), Evallm: Interactive evaluation of large language model prompts on user-defined criteria, in 'Proceedings of the CHI Conference on Human Factors in Computing Systems', pp. 1–21.
- Kirk, H. R., Whitefield, A., Rottger, P., Bean, A. M., Margatina, K., Mosquera-Gomez, R., Ciro, J., Bartolo, M., Williams, A., He, H. et al. (2024), 'The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models', *Advances in Neural Information Processing Systems* **37**, 105236–105344.
- Kitagawa, T. and Tetenov, A. (2018), 'Who should be treated? empirical welfare maximization methods for treatment choice', *Econometrica* **86**(2), 591–616.
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S. (2018), 'Human decisions and machine predictions', *The quarterly journal of economics* **133**(1), 237–293.
- Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z. (2015), 'Prediction policy problems', *American Economic Review* **105**(5), 491–95.
- Klie, J.-C., Webber, B. and Gurevych, I. (2023), 'Annotation error detection: Analyzing the past and present for a more coherent future', *Computational Linguistics* **49**(1), 157–198.
- Ko, A. J., Abraham, R., Beckwith, L., Blackwell, A., Burnett, M., Erwig, M., Scaffidi, C., Lawrance, J., Lieberman, H., Myers, B. et al. (2011), The state of the art in end-user software engineering, Vol. 43, ACM New York, NY, USA, pp. 1–44.
- Kolla, M., Salunkhe, S., Chandrasekharan, E. and Saha, K. (2024), Llm-mod: Can large language models assist content moderation?, in 'Extended Abstracts of the CHI Conference on Human Factors in Computing Systems', pp. 1–8.
- Konyushova, K., Chen, Y., Paine, T., Gulcehre, C., Paduraru, C., Mankowitz, D. J., Denil, M. and de Freitas, N. (2021), 'Active offline policy selection', *Advances in Neural Information Processing Systems* **34**, 24631–24644.
- Kothari, A., Vossler, P., Digitale, J., Forouzannia, M., Rosenberg, E., Lee, M., Bryant, J., Molina, M., Marks, J., Zier, L. et al. (2026), When the domain expert has no time and the llm developer has no clinical expertise: Real-world lessons from llm co-design in a safety-net hospital, in 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 40, pp. 38754–38762.
- Kross, S. and Guo, P. (2021), Orienting, framing, bridging, magic, and counseling: How data scientists navigate the outer loop of client collaborations in industry and academia, Vol. 5, ACM New York, NY, USA, pp. 1–28.

- Kruttschnitt, C., Kalsbeek, W. D., House, C. C. et al. (2014), 'Estimating the incidence of rape and sexual assault'.
- Kube, A., Das, S. and Fowler, P. J. (2019), Allocating interventions based on predicted outcomes: A case study on homelessness services, in 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 33, pp. 622–629.
- Künzel, S. R., Sekhon, J. S., Bickel, P. J. and Yu, B. (2019), 'Metalearners for estimating heterogeneous treatment effects using machine learning', *Proceedings of the national academy of sciences* **116**(10), 4156–4165.
- Kuo, T.-S., Shen, H., Geum, J., Jones, N., Hong, J. I., Zhu, H. and Holstein, K. (2023), Understanding frontline workers' and unhoused individuals' perspectives on ai used in homeless services, in 'Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems', pp. 1–17.
- Kuzborskij, I., Vernade, C., Gyorgy, A. and Szepesvári, C. (2021), Confident off-policy evaluation and selection through self-normalized importance weighting, in 'International Conference on Artificial Intelligence and Statistics', PMLR, pp. 640–648.
- Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Von Arx, S. et al. (2025), 'Measuring ai ability to complete long tasks', *arXiv preprint arXiv:2503.14499* **352**.
- Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A. and Tan, C. (2021), 'Towards a science of human-ai decision making: A survey of empirical studies', *arXiv preprint arXiv:2112.11471* .
- Lai, V., Liu, H. and Tan, C. (2020), " why is 'chicago' deceptive?" towards building model-driven tutorials for humans, in 'Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems', pp. 1–13.
- Lai, V. and Tan, C. (2019), On human predictions with explanations and predictions of machine learning models: A case study on deception detection, in 'Proceedings of the conference on fairness, accountability, and transparency', pp. 29–38.
- Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J. and Mullainathan, S. (2017), The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables, in 'Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 275–284.
- Lakkaraju, H., Leskovec, J., Kleinberg, J. and Mullainathan, S. (2015), A bayesian framework for modeling human evaluations, in 'Proceedings of the 2015 SIAM International Conference on Data Mining', SIAM, pp. 181–189.
- LaLonde, R. J. (1986), 'Evaluating the econometric evaluations of training programs with experimental data', *The American economic review* pp. 604–620.
- Latour, B. and Woolgar, S. (2013), *Laboratory life: The construction of scientific facts*, Princeton university press.
- Levay, K. E., Freese, J. and Druckman, J. N. (2016), 'The demographic and political composition of mechanical turk samples', *Sage Open* **6**(1), 2158244016636433.
- Levi-Strauss, C. (1966), *The savage mind*, University of Chicago Press.

- Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T. et al. (2024), 'From generation to judgment: Opportunities and challenges of llm-as-a-judge', *arXiv preprint arXiv:2411.16594*.
- Li, X., Lipton, Z. C. and Leqi, L. (2024), 'Personalized language modeling from personalized human feedback', *arXiv preprint arXiv:2402.05133*.
- Liao, Q. V. and Xiao, Z. (2023), 'Rethinking model evaluation as narrowing the socio-technical gap', *arXiv preprint arXiv:2306.03100*.
- Liao, T., Taori, R., Raji, I. D. and Schmidt, L. (2021), Are we learning yet? a meta review of evaluation failures across machine learning, in 'Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)'.
- Liu, H., Lai, V. and Tan, C. (2021), 'Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making', *Proceedings of the ACM on Human-Computer Interaction* **5**(CSCW2), 1–45.
- Liu, J., Chen, H., Shen, J. and Choo, K.-K. R. (2024), Faircompass: Operationalising fairness in machine learning, IEEE.
- Liu, T. and Tao, D. (2015), 'Classification with noisy labels by importance reweighting', *IEEE Transactions on pattern analysis and machine intelligence* **38**(3), 447–461.
- Liu, Y. L., Blodgett, S. L., Cheung, J. C. K., Liao, Q. V., Olteanu, A. and Xiao, Z. (2024), 'Ecbd: Evidence-centered benchmark design for nlp', *arXiv preprint arXiv:2406.08723*.
- Lohr, S. (2019), *Measuring crime: Behind the statistics*, Chapman and Hall/CRC.
- Louridas, P. (1999), 'Design as bricolage: anthropology meets design thinking', *Design Studies* **20**(6), 517–535.
- Lu, H. and Zhong, F. (2024), 'Can vision-language models replace human annotators: A case study with celeba dataset', *arXiv preprint arXiv:2410.09416*.
- Lundberg, I., Johnson, R. and Stewart, B. M. (2021), 'What is your estimand? defining the target quantity connects statistical evidence to theory', *American Sociological Review* **86**(3), 532–565.
- Lundberg, S. (2017), A unified approach to interpreting model predictions.
- Lynch, B. K. (2001), 'Rethinking assessment from a critical perspective', *Language testing* **18**(4), 351–372.
- Madras, D., Pitassi, T. and Zemel, R. (2018), 'Predict responsibly: improving fairness and accuracy by learning to defer', *Advances in Neural Information Processing Systems* **31**.
- Manam, V. and Quinn, A. (2018), Wingit: Efficient refinement of unclear task instructions, in 'Proceedings of the AAAI Conference on Human Computation and Crowdsourcing', Vol. 6, pp. 108–116.
- Manski, C. F. (1990), 'Nonparametric bounds on treatment effects', *The American Economic Review* **80**(2), 319–323.

- Mao, Y., Wang, D., Muller, M., Varshney, K. R., Baldini, I., Dugan, C. and Mojsilović, A. (2019), How data scientists work together with domain experts in scientific collaborations: To find the right answer or to ask the right question?, Vol. 3, ACM New York, NY, USA, pp. 1–23.
- Mateus, S. and Sarkar, S. (2024), Bricolage—a systematic review, conceptualization, and research agenda, Taylor & Francis, pp. 1–22.
- Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B. et al. (2024), ‘Harmbench: A standardized evaluation framework for automated red teaming and robust refusal’, *arXiv preprint arXiv:2402.04249*.
- McCutcheon, A. L. (1987), *Latent class analysis*, number 64, Sage.
- Menon, A., Van Rooyen, B., Ong, C. S. and Williamson, B. (2015), Learning from corrupted binary labels via class-probability estimation, in ‘International conference on machine learning’, PMLR, pp. 125–134.
- Messick, S. (1995), ‘Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning’, *American psychologist* **50**(9), 741.
- Mickel, J. (2024), Racial/ethnic categories in ai and algorithmic fairness: Why they matter and what they represent, in ‘The 2024 ACM Conference on Fairness, Accountability, and Transparency’, pp. 2484–2494.
- Miller, D. C. and Salkind, N. J. (2002), *Handbook of research design and social measurement*, Sage.
- Milli, S., Belli, L. and Hardt, M. (2021), From optimizing engagement to measuring value, in ‘Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency’, pp. 714–722.
- Min, S., Michael, J., Hajishirzi, H. and Zettlemoyer, L. (2020), ‘Ambigqa: Answering ambiguous open-domain questions’, *arXiv preprint arXiv:2004.10645*.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D. and Gebru, T. (2019), Model cards for model reporting, in ‘Proceedings of the conference on fairness, accountability, and transparency’, pp. 220–229.
- Mol, A. (2002), *The body multiple: Ontology in medical practice*, Duke University Press.
- Mozannar, H., Satyanarayan, A. and Sontag, D. (2022), Teaching humans when to defer to a classifier via exemplars, in ‘Proceedings of the AAAI Conference on Artificial Intelligence’, Vol. 36, pp. 5323–5331.
- Mullainathan, S. and Obermeyer, Z. (2017), ‘Does machine learning automate moral hazard and error?’, *American Economic Review* **107**(5), 476–80.
- Mullainathan, S. and Obermeyer, Z. (2019), *A machine learning approach to low-value health care: wasted tests, missed heart attacks and mis-predictions*, National Bureau of Economic Research.
- Mullainathan, S. and Obermeyer, Z. (2021), On the inequity of predicting a while hoping for b, in ‘AEA Papers and Proceedings’, Vol. 111, pp. 37–42.

- Mullainathan, S. and Obermeyer, Z. (2022), 'Diagnosing physician error: A machine learning approach to low-value health care', *The Quarterly Journal of Economics* **137**(2), 679–727.
- Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., Dugan, C. and Erickson, T. (2019), How data science workers work with data: Discovery, capture, curation, design, creation, in 'Proceedings of the 2019 CHI conference on human factors in computing systems', pp. 1–15.
- Muller, M., Wolf, C. T., Andres, J., Desmond, M., Joshi, N. N., Ashktorab, Z., Sharma, A., Brimijoin, K., Pan, Q., Duesterwald, E. et al. (2021), Designing ground truth and the social life of labels, in 'Proceedings of the 2021 CHI conference on human factors in computing systems', pp. 1–16.
- Mussnug, A. M. (2022), 'The predictive reframing of machine learning applications: good predictions and bad measurements', *European Journal for Philosophy of Science* **12**(3), 55.
- Myers, B. A., Ko, A. J., LaToza, T. D. and Yoon, Y. (2016), Programmers are users too: Human-centered methods for improving programming tools, Vol. 49, IEEE, pp. 44–52.
- Nahar, N., Zhou, S., Lewis, G. and Kästner, C. (2022), Collaboration challenges in building ml-enabled systems: Communication, documentation, engineering, and process, in 'Proceedings of the 44th international conference on software engineering', pp. 413–425.
- Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y. and Liang, X. (2018), 'doccano: Text annotation tool for human'. Software available from <https://github.com/doccano/doccano>.
URL: <https://github.com/doccano/doccano>
- Namkoong, H., Keramati, R., Yadlowsky, S. and Brunskill, E. (2020), 'Off-policy policy evaluation for sequential decisions under unobserved confounding', *Advances in Neural Information Processing Systems* **33**, 18819–18831.
- Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q. and Arocena, P. C. (2019), Data lake management: challenges and opportunities, Vol. 12, VLDB Endowment, pp. 1986–1989.
- Natarajan, N., Dhillon, I. S., Ravikumar, P. K. and Tewari, A. (2013), 'Learning with noisy labels', *Advances in neural information processing systems* **26**.
- Neath, A. A. and Cavanaugh, J. E. (2012), 'The bayesian information criterion: background, derivation, and applications', *Wiley Interdisciplinary Reviews: Computational Statistics* **4**(2), 199–203.
- Neumann, T., De-Arteaga, M. and Fazelpour, S. (2025), 'Should you use llms to simulate opinions? quality checks for early-stage deliberation'.
URL: <https://arxiv.org/abs/2504.08954>
- Nguyen, T., Garcia-Gathright, J., Washington, H., Chouldechova, A., Wallach, H. and Wortman Vaughan, J. (2026), Validating and refining measurements for generative ai evaluation via stakeholder engagement, in 'The 2026 ACM Conference on Fairness, Accountability, and Transparency', pp. 6554–6586.
- Nie, L., Ye, M., Liu, Q. and Nicolae, D. (2021), 'Vcnet and functional targeted regularization for learning causal effects of continuous treatments', *arXiv preprint arXiv:2103.07861*.
- Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J. and Kiela, D. (2019), 'Adversarial nli: A new benchmark for natural language understanding', *arXiv preprint arXiv:1910.14599*.

- Nie, Y., Zhou, X. and Bansal, M. (2020), 'What can we learn from collective human opinions on natural language inference data?', *arXiv preprint arXiv:2010.03532*.
- Northcutt, C., Jiang, L. and Chuang, I. (2021), 'Confident learning: Estimating uncertainty in dataset labels', *Journal of Artificial Intelligence Research* **70**, 1373–1411.
- Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019), 'Dissecting racial bias in an algorithm used to manage the health of populations', *Science* **366**(6464), 447–453.
- Orlikowski, M., Pei, J., Röttger, P., Cimiano, P., Jurgens, D. and Hovy, D. (2025), 'Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions', *arXiv preprint arXiv:2502.20897*.
- Orso, F., Fabbri, G. and Maggioni, A. P. (2017), 'Epidemiology of heart failure', *Heart Failure* pp. 15–33.
- Osorio, J. A., Chaudron, M. R. and Heijstek, W. (2011), Moving from waterfall to iterative development: An empirical evaluation of advantages, disadvantages and risks of rup, in '2011 37th EUROMICRO Conference on Software Engineering and Advanced Applications', IEEE, pp. 453–460.
- Ouyang, S., Wang, S., Liu, Y., Zhong, M., Jiao, Y., Iter, D., Pryzant, R., Zhu, C., Ji, H. and Han, J. (2023), 'The shifted and the overlooked: A task-oriented investigation of user-gpt interactions', *arXiv preprint arXiv:2310.12418*.
- Pang, B., Nijkamp, E. and Wu, Y. N. (2020), Deep learning with tensorflow: A review, Vol. 45, SAGE Publications Sage CA: Los Angeles, CA, pp. 227–248.
- Park, J. S., Barber, R., Kirlik, A. and Karahalios, K. (2019), 'A slow algorithm improves users' assessments of the algorithm's accuracy', *Proceedings of the ACM on Human-Computer Interaction* **3**(CSCW), 1–15.
- Passi, S. and Barocas, S. (2019), Problem formulation and fairness, in 'Proceedings of the conference on fairness, accountability, and transparency', pp. 39–48.
- Passi, S. and Jackson, S. J. (2018), Trust in data science: Collaboration, translation, and accountability in corporate data science projects, Vol. 2, ACM New York, NY, USA, pp. 1–28.
- Paton, N. W., Chen, J. and Wu, Z. (2023), Dataset discovery and exploration: A survey, Vol. 56, ACM New York, NY, USA, pp. 1–37.
- Patrini, G., Rozza, A., Krishna Menon, A., Nock, R. and Qu, L. (2017), Making deep neural networks robust to label noise: A loss correction approach, in 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 1944–1952.
- Paul, C. R., Ryan, M. S., Dallaghan, G. L. B., Jirasevijinda, T., Quigley, P. D., Hanson, J. L., Khidir, A. M., Petershach, J., Jackson, J., Tewksbury, L. et al. (2019), 'Collecting validity evidence: a hands-on workshop for medical education assessment instruments', *MedEdPORTAL* **15**, 10817.
- Pavlick, E. and Kwiatkowski, T. (2019), 'Inherent disagreements in human textual inferences', *Transactions of the Association for Computational Linguistics* **7**, 677–694.
- Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N. and Androutsopoulos, I. (2020), 'Toxicity detection: Does context really matter?', *arXiv preprint arXiv:2006.00998*.

- Pearl, J. (1995), 'Causal diagrams for empirical research', *Biometrika* **82**(4), 669–688.
- Pearl, J. (2009), 'Causal inference in statistics: An overview', *Statistics surveys* **3**, 96–146.
- Peng, A., Nushi, B., Kıcıman, E., Inkpen, K., Suri, S. and Kamar, E. (2019), What you see is what you get? the impact of representation criteria on human bias in hiring, in 'Proceedings of the AAAI Conference on Human Computation and Crowdsourcing', Vol. 7, pp. 125–134.
- Perdomo, J. C. (2023), The relative value of prediction in algorithmic decision making.
- Perdomo, J., Zrnic, T., Mendler-Dünner, C. and Hardt, M. (2020), Performative prediction, in 'International Conference on Machine Learning', PMLR, pp. 7599–7609.
- Peterson, J. C., Battleday, R. M., Griffiths, T. L. and Russakovsky, O. (2019), Human uncertainty makes classification more robust, in 'Proceedings of the IEEE/CVF international conference on computer vision', pp. 9617–9626.
- Pine, K. H. and Liboiron, M. (2015), The politics of measurement and action, in 'Proceedings of the 33rd annual ACM conference on human factors in computing systems', pp. 3147–3156.
- Plank, B. (2022), 'The problem of human label variation: On ground truth in data, modeling and evaluation', *arXiv preprint arXiv:2211.02570*.
- Poletto, F., Basile, V., Sanguinetti, M., Bosco, C. and Patti, V. (2021), 'Resources and benchmark corpora for hate speech detection: a systematic review', *Language Resources and Evaluation* **55**(2), 477–523.
- Porter, T. M. (1996), *Trust in numbers: The pursuit of objectivity in science and public life*, Princeton University Press.
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W. and Wallach, H. (2021), Manipulating and measuring model interpretability, in 'Proceedings of the 2021 CHI conference on human factors in computing systems', pp. 1–52.
- Precup, D. (2000), 'Eligibility traces for off-policy policy evaluation', *Computer Science Department Faculty Publication Series* p. 80.
- Pu, H. and Zhang, B. (2021), 'Estimating optimal treatment rules with an instrumental variable: A partial identification learning approach', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **83**(2), 318–345.
- Pushkarna, M., Zaldivar, A. and Kjartansson, O. (2022), Data cards: Purposeful and transparent dataset documentation for responsible ai, in 'Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency', pp. 1776–1826.
- Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O. et al. (2019), Machine behaviour, Vol. 568, Nature Publishing Group UK London, pp. 477–486.
- Raji, I. D., Bender, E. M., Paullada, A., Denton, E. and Hanna, A. (2021), 'Ai and the everything in the whole wide world benchmark', *arXiv preprint arXiv:2111.15366*.
- Raji, I. D., Kumar, I. E., Horowitz, A. and Selbst, A. (2022), The fallacy of ai functionality, in 'Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency', pp. 959–972.

- Rambachan, A., Coston, A. and Kennedy, E. (2022), Robust design and evaluation of predictive algorithms under unobserved confounding.
- Rambachan, A. and Roth, J. (2019), 'Bias in, bias out? evaluating the folk wisdom', *arXiv preprint arXiv:1909.08518*.
- Rambachan, A. et al. (2021), 'Identifying prediction mistakes in observational data', *Harvard University*.
- Rattenbury, T., Hellerstein, J. M., Heer, J., Kandel, S. and Carreras, C. (2017), *Principles of data wrangling: Practical techniques for data preparation*, " O'Reilly Media, Inc."
- Reeve, H. et al. (2019), Classification with unknown class-conditional label noise on non-compact feature spaces, in 'Conference on Learning Theory', PMLR, pp. 2624–2651.
- Ribeiro, M. T., Singh, S. and Guestrin, C. (2016), " why should i trust you?" explaining the predictions of any classifier, in 'Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining', pp. 1135–1144.
- Ribeiro, M. T., Singh, S. and Guestrin, C. (2018), Anchors: High-precision model-agnostic explanations, in 'Proceedings of the AAAI conference on artificial intelligence', Vol. 32.
- Riley, P., Deutsch, D., Foster, G., Ratnakar, V., Dabirmoghaddam, A. and Freitag, M. (2024), 'Finding replicable human evaluations via stable ranking probability', *arXiv preprint arXiv:2404.01474*.
- Roberts, F. S. (1985), 'Measurement theory'.
- Robins, J. (1986), 'A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect', *Mathematical modelling* **7**(9-12), 1393–1512.
- Robins, J. M. and Rotnitzky, A. (1995), 'Semiparametric efficiency in multivariate regression models with missing data', *Journal of the American Statistical Association* **90**(429), 122–129.
- Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994), 'Estimation of regression coefficients when some regressors are not always observed', *Journal of the American statistical Association* **89**(427), 846–866.
- Rosenbaum, P. R. (2005), 'Sensitivity analysis in observational studies', *Encyclopedia of statistics in behavioral science*.
- Rosenbaum, P. R. and Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**(1), 41–55.
- Roth, M. and Schlechtweg, D., eds (2025), *Context and Meaning: Navigating Disagreements in NLP Annotation*, Association for Computational Linguistics.
- Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies.', *Journal of educational Psychology* **66**(5), 688.
- Rubin, D. B. (2005), 'Causal inference using potential outcomes: Design, modeling, decisions', *Journal of the American Statistical Association* **100**(469), 322–331.
- Saad-Falcon, J., Khattab, O., Potts, C. and Zaharia, M. (2023), 'Ares: An automated evaluation framework for retrieval-augmented generation systems', *arXiv preprint arXiv:2311.09476*.

- Sadek, M., Kallina, E., Bohné, T., Mougenot, C., Calvo, R. A. and Cave, S. (2025), 'Challenges of responsible ai in practice: scoping review and recommended actions', *AI & society* **40**(1), 199–215.
- Salaudeen, O., Reuel, A., Ahmed, A., Bedi, S., Robertson, Z., Sundar, S., Domingue, B., Wang, A. and Koyejo, S. (2025), 'Measurement to meaning: A validity-centered framework for ai evaluation', *arXiv preprint arXiv:2505.10573*.
- Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T. and Ghani, R. (2018), Aequitas: A bias and fairness audit toolkit.
- Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P. and Hashimoto, T. (2023), Whose opinions do language models reflect?, in 'International Conference on Machine Learning', PMLR, pp. 29971–30004.
- Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H. and Jinks, C. (2018), Saturation in qualitative research: exploring its conceptualization and operationalization, Vol. 52, Springer, pp. 1893–1907.
- Saxena, D., Jung, J.-Y., Forlizzi, J., Holstein, K. and Zimmerman, J. (2025), 'Ai mismatches: Identifying potential algorithmic harms before ai development', *arXiv preprint arXiv:2502.18682*.
- Schlesselman, J. J. (1982), *Case-control studies: design, conduct, analysis*, Vol. 2, Oxford university press.
- Schouwenburg, H. C. (2004), 'Procrastination in academic settings: General introduction.'
- Schubotz, M., Satpute, A., Greiner-Petter, A., Aizawa, A. and Gipp, B. (2022), 'Caching and reproducibility: Making data science experiments faster and fairer', *Frontiers in Research Metrics and Analytics* **7**, 861944.
- Schulman, J., Zoph, B., Kim, C., Hilton, J., Menick, J., Weng, J., Uribe, J. F. C., Fedus, L., Metz, L., Pokorny, M. et al. (2022), 'Chatgpt: Optimizing language models for dialogue', *OpenAI blog* **2**(4).
- Scott, C. (2015), A rate of convergence for mixture proportion estimation, with application to learning from noisy labels, in 'Artificial Intelligence and Statistics', PMLR, pp. 838–846.
- Scott, C., Blanchard, G. and Handy, G. (2013), Classification with asymmetric label noise: Consistency and maximal denoising, in 'Conference on learning theory', PMLR, pp. 489–511.
- Scott, J. C. (2020), *Seeing like a state: How certain schemes to improve the human condition have failed*, yale university Press.
- Seffah, A., Gulliksen, J. and Desmarais, M. C. (2005), *Human-centered software engineering-integrating usability in the software development lifecycle*, Vol. 8, Springer Science & Business Media.
- Shalit, U., Johansson, F. D. and Sontag, D. (2017), Estimating individual treatment effect: generalization bounds and algorithms, in 'International Conference on Machine Learning', PMLR, pp. 3076–3085.

- Shankar, S., Li, H., Asawa, P., Hulsebos, M., Lin, Y., Zamfirescu-Pereira, J., Chase, H., Fu-Hinthorn, W., Parameswaran, A. G. and Wu, E. (2024), 'Spade: Synthesizing data quality assertions for large language model pipelines', *arXiv preprint arXiv:2401.03038*.
- Shankar, S., Zamfirescu-Pereira, J., Hartmann, B., Parameswaran, A. and Arawjo, I. (2024), Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences, in 'Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology', pp. 1–14.
- Sharma, A. and Kiciman, E. (2020), Dowhy: An end-to-end library for causal inference.
- Shen, Y., Song, K., Tan, X., Li, D., Lu, W. and Zhuang, Y. (2024), Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face, Vol. 36.
- Shi, C., Blei, D. and Veitch, V. (2019), 'Adapting neural networks for the estimation of treatment effects', *Advances in neural information processing systems* **32**.
- Shi, L., Ma, C., Liang, W., Ma, W. and Vosoughi, S. (2024), 'Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms', *arXiv preprint arXiv:2406.07791*.
- Shimodaira, H. (2000), 'Improving predictive inference under covariate shift by weighting the log-likelihood function', *Journal of statistical planning and inference* **90**(2), 227–244.
- Shrout, P. E. and Lane, S. P. (2012), 'Psychometrics.'
- Shu, D. and Yi, G. Y. (2019), 'Causal inference with measurement error in outcomes: Bias analysis and estimation methods', *Statistical methods in medical research* **28**(7), 2049–2068.
- Si, N., Zhang, F., Zhou, Z. and Blanchet, J. (2020), Distributionally robust policy evaluation and learning in offline contextual bandits, in 'International Conference on Machine Learning', PMLR, pp. 8884–8894.
- Singh, H., Singh, R., Mhasawade, V. and Chunara, R. (2021), Fairness violations and mitigation under covariate shift, in 'Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency', pp. 3–13.
- Singh, K. (2007), *Quantitative social research methods*, Sage.
- Skogan, W. G. (1999), Measuring what matters, in 'Proceedings from the policing research institute meetings', Department of Justice, pp. 37–88.
- Smith, J. A. and Todd, P. E. (2005), 'Does matching overcome lalonde's critique of nonexperimental estimators?', *Journal of econometrics* **125**(1-2), 305–353.
- Smith, V. C., Lange, A. and Huston, D. R. (2012), 'Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses.', *Journal of asynchronous learning networks* **16**(3), 51–61.
- Sommerauer, P., Fokkens, A. and Vossen, P. (2020), Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement, in 'Proceedings of the 28th International Conference on Computational Linguistics', pp. 4798–4809.
- Starace, G., Jaffe, O., Sherburn, D., Aung, J., Chan, J. S., Maksin, L., Dias, R., Mays, E., Kinsella, B., Thompson, W. et al. (2025), 'Paperbench: Evaluating ai's ability to replicate ai research', *arXiv preprint arXiv:2504.01848*.

- Steiger, M., Bharucha, T. J., Venkatagiri, S., Riedl, M. J. and Lease, M. (2021), The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support, *in* 'Proceedings of the 2021 CHI conference on human factors in computing systems', pp. 1–14.
- Stevenson, M. T. and Doleac, J. L. (2022), 'Algorithmic risk assessment in the hands of humans', *Available at SSRN 3489440*.
- Stray, J., Halevy, A., Assar, P., Hadfield-Menell, D., Boutilier, C., Ashar, A., Beattie, L., Ekstrand, M., Leibowicz, C., Sehat, C. M. et al. (2022), 'Building human values into recommender systems: An interdisciplinary synthesis', *arXiv preprint arXiv:2207.10192*.
- Stray, J., Vendrov, I., Nixon, J., Adler, S. and Hadfield-Menell, D. (2021), What are you optimizing for? aligning recommender systems with human values.
- Suchman, L. (1993), Do categories have politics? the language/action perspective reconsidered, Vol. 2, Springer, pp. 177–190.
- Swaminathan, A. and Joachims, T. (2015), Counterfactual risk minimization: Learning from logged bandit feedback, *in* 'International Conference on Machine Learning', PMLR, pp. 814–823.
- Szymanski, A., Anuyah, O., Li, T. J.-J. and Metoyer, R. (2026), Key considerations for domain expert involvement in llm design and evaluation: An ethnographic study, *in* 'Proceedings of the 31st International Conference on Intelligent User Interfaces', pp. 617–630.
- Szymanski, A., Araya Gebreegziabher, S., Anuyah, O., A. Metoyer, R. and Jia-Jun Li, T. (2026), Designing staged evaluation workflows for llms: Integrating domain experts, lay users, and model-generated evaluation criteria, *in* 'Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems', pp. 1–20.
- Szymanski, A., Araya Gebreegziabher, S., Anuyah, O., Metoyer, R. A. and Jia-Jun Li, T. (2024), 'Comparing criteria development across domain experts, lay users, and models in large language model evaluation', *arXiv e-prints* pp. arXiv–2410.
- Szymanski, A., Wimer, B. L., Anuyah, O., Eicher-Miller, H. A. and Metoyer, R. A. (2024), Integrating expertise in llms: crafting a customized nutrition assistant with refined template instructions, *in* 'Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems', pp. 1–22.
- Szymanski, A., Ziems, N., Eicher-Miller, H. A., Li, T. J.-J., Jiang, M. and Metoyer, R. A. (2024), 'Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks', *arXiv preprint arXiv:2410.20266*.
- Szymanski, A., Ziems, N., Eicher-Miller, H. A., Li, T. J.-J., Jiang, M. and Metoyer, R. A. (2025), Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks, *in* 'Proceedings of the 30th international conference on intelligent user interfaces', pp. 952–966.
- Tan, S., Adebayo, J., Inkpen, K. and Kamar, E. (2018), 'Investigating human+ machine complementarity for recidivism predictions', *arXiv preprint arXiv:1808.09123*.
- Tan, S., Zhuang, S., Montgomery, K., Tang, W., Cuadron, A., Wang, C., Popa, R. and Stoica, I. (2025), Judgebench: A benchmark for evaluating llm-based judges, *in* 'International Conference on Learning Representations', Vol. 2025, pp. 63277–63303.

- Tan, S., Zhuang, S., Montgomery, K., Tang, W. Y., Cuadron, A., Wang, C., Popa, R. A. and Stoica, I. (2024), 'Judgebench: A benchmark for evaluating llm-based judges', *arXiv preprint arXiv:2410.12784*.
- Tan, Z. (2006), 'A distributional approach for causal inference using propensity scores', *Journal of the American Statistical Association* **101**(476), 1619–1637.
- Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D. C. and Silberman, N. (2019), Learning from noisy labels by regularized estimation of annotator confusion, in 'Proceedings of the IEEE/CVF conference on computer vision and pattern recognition', pp. 11244–11253.
- Tennenholtz, G., Shalit, U., Mannor, S. and Efroni, Y. (2021), Bandits with partially observable confounded data, in 'Uncertainty in Artificial Intelligence', PMLR, pp. 430–439.
- Thomas, D. R., Borchers, C., Bhushan, S., Gatz, E., Gupta, S. and Koedinger, K. R. (2025), Llm-generated feedback supports learning if learners choose to use it, in 'European Conference on Technology Enhanced Learning', Springer, pp. 489–503.
- Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.-P., Humbert, M., Juels, A. and Lin, H. (2017), Fairtest: Discovering unwarranted associations in data-driven applications, in '2017 IEEE European Symposium on Security and Privacy (EuroS&P)', IEEE, pp. 401–416.
- Tsiatis, A. A. (2006), *Semiparametric theory and missing data*, Vol. 4, Springer.
- Turkle, S. (2011), *Life on the Screen*, Simon and Schuster.
- Turque, B. (2012), 'Creative... motivating'and fired', *The Washington Post* **6**.
- Uehara, M., Shi, C. and Kallus, N. (2022), 'A review of off-policy evaluation in reinforcement learning', *arXiv preprint arXiv:2212.06355*.
- Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B. and Poesio, M. (2020), A case for soft loss functions, in 'Proceedings of the AAAI Conference on Human Computation and Crowdsourcing', Vol. 8, pp. 173–177.
- Vaithianathan, R., Kulick, E., Putnam-Hornstein, E. and Benavides-Prado, D. (2019), 'Allegheeny family screening tool: Methodology, version 2', *Center for Social Data Analytics* pp. 1–22.
- Vallgård, A. and Fernaeus, Y. (2015), Interaction design as a bricolage practice, in 'Proceedings of the ninth international conference on tangible, embedded, and embodied interaction', pp. 173–180.
- Van Rooyen, B., Menon, A. and Williamson, R. C. (2015), 'Learning with symmetric label noise: The importance of being unhinged', *Advances in neural information processing systems* **28**.
- Van Rooyen, B. et al. (2015), 'Machine learning via transitions'.
- Van't Veer, A. E. and Giner-Sorolla, R. (2016), Pre-registration in social psychology—a discussion and suggested template, Vol. 67, Elsevier, pp. 2–12.
- Vermunt, J. K. (2010), 'Latent class modeling with covariates: Two improved three-step approaches', *Political analysis* **18**(4), 450–469.

- Viégas, F. B., Wattenberg, M., Kriss, J. and Van Ham, F. (2007), Talk before you type: Coordination in wikipedia, in '2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)', IEEE, pp. 78–78.
- Wallach, H., Desai, M., Cooper, A. F., Wang, A., Atalla, C., Barocas, S., Blodgett, S. L., Chouldechova, A., Corvi, E., Dow, P. A. et al. (2025), Position: Evaluating generative ai systems is a social science measurement challenge, in 'International Conference on Machine Learning', PMLR, pp. 82232–82251.
- Walter, S. D. and Irwig, L. M. (1988), 'Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review', *Journal of clinical epidemiology* **41**(9), 923–937.
- Wang, A., Cho, K. and Lewis, M. (2020), 'Asking and answering questions to evaluate the factual consistency of summaries', *arXiv preprint arXiv:2004.04228*.
- Wang, A., Kapoor, S., Barocas, S. and Narayanan, A. (2024), Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy, Vol. 1, ACM New York, NY, pp. 1–45.
- Wang, J., Liu, Y. and Levy, C. (2021), Fair classification with group-dependent label noise, in 'Proceedings of the 2021 ACM conference on fairness, accountability, and transparency', pp. 526–536.
- Wang, X. and Plank, B. (2023), Actor: Active learning with annotator-specific classification heads to embrace human label variation, in 'Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing', pp. 2046–2052.
- Wang, X. and Yin, M. (2021), Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making, in '26th International Conference on Intelligent User Interfaces', pp. 318–328.
- Watson-Daniels, J., Barocas, S., Hofman, J. M. and Chouldechova, A. (2023), Multi-target multiplicity: Flexibility and fairness in target specification under resource constraints, in 'Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency', pp. 297–311.
- Weerts, H., Dudík, M., Edgar, R., Jalali, A., Lutz, R. and Madaio, M. (2023), Fairlearn: Assessing and improving fairness of ai systems, Vol. 24, pp. 1–8.
- Wei, D. (2021), Decision-making under selective labels: Optimal finite-domain policies and beyond, in 'International Conference on Machine Learning', PMLR, pp. 11035–11046.
- Weidinger, L., Raji, I. D., Wallach, H., Mitchell, M., Wang, A., Salaudeen, O., Bommasani, R., Ganguli, D., Koyejo, S. and Isaac, W. (2025), 'Toward an evaluation science for generative ai systems', *arXiv preprint arXiv:2503.05336*.
- Wen, B., Yao, J., Feng, S., Xu, C., Tsvetkov, Y., Howe, B. and Wang, L. L. (2024), 'Know your limits: A survey of abstention in large language models', *arXiv preprint arXiv:2407.18418*.
- Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F. and Wilson, J. (2019), The what-if tool: Interactive probing of machine learning models, Vol. 26, IEEE, pp. 56–65.
- Wilder, B., Horvitz, E. and Kamar, E. (2020), 'Learning to complement humans', *arXiv preprint arXiv:2005.00582*.

- Williams, A., Nangia, N. and Bowman, S. R. (2017), 'A broad-coverage challenge corpus for sentence understanding through inference', *arXiv preprint arXiv:1704.05426*.
- Wongsuphasawat, K., Liu, Y. and Heer, J. (2019), Goals, process, and challenges of exploratory data analysis: An interview study.
- Wongsuphasawat, K., Moritz, D., Anand, A., Mackinlay, J., Howe, B. and Heer, J. (2015), Voyager: Exploratory analysis via faceted browsing of visualization recommendations, Vol. 22, IEEE, pp. 649–658.
- Wyden, S. R. (n.d.), 'S.2892 - algorithmic accountability act of 2022'.
- Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D. and Sugiyama, M. (2020), 'Part-dependent label noise: Towards instance-dependent label noise', *Advances in Neural Information Processing Systems* **33**, 7597–7610.
- Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G. and Sugiyama, M. (2019), 'Are anchor points really indispensable in label-noise learning?', *Advances in Neural Information Processing Systems* **32**.
- Xu, L., Kanagawa, H. and Gretton, A. (2021), 'Deep proxy causal learning and its application to confounded bandit policy evaluation', *Advances in Neural Information Processing Systems* **34**, 26264–26275.
- Yan, J. N., Gu, Z. and Rzeszutarski, J. M. (2021), Tessera: Discretizing data analysis workflows on a task level, in 'Proceedings of the 2021 CHI conference on human factors in computing systems', pp. 1–15.
- Yang, M., Dai, B., Nachum, O., Tucker, G. and Schuurmans, D. (2022), Offline policy selection under uncertainty, in 'International Conference on Artificial Intelligence and Statistics', PMLR, pp. 4376–4396.
- Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., Gao, T., Geyer, W., Huang, C., Chen, P.-Y. et al. (2024), 'Justice or prejudice? quantifying biases in llm-as-a-judge', *arXiv preprint arXiv:2410.02736*.
- Zhang, A. X., Muller, M. and Wang, D. (2020), How do data science workers collaborate? roles, workflows, and tools, Vol. 4, ACM New York, NY, USA, pp. 1–23.
- Zhang, B., Weiss, J., Small, D. S. and Zhao, Q. (2021), 'Selecting and ranking individualized treatment rules with unmeasured confounding', *Journal of the American Statistical Association* **116**(533), 295–308.
- Zhang, Q., Lee, M. L. and Carter, S. (2022), You complete me: Human-ai teams and complementary expertise, in 'CHI Conference on Human Factors in Computing Systems', pp. 1–28.
- Zhang, Z. and Yuan, K.-H. (2018), *Practical statistical power analysis*, ISDSA press Granger, Indiana.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. et al. (2023), 'Judging llm-as-a-judge with mt-bench and chatbot arena', *Advances in Neural Information Processing Systems* **36**, 46595–46623.
- Zwaan, L. and Singh, H. (2015), 'The challenges in defining and measuring diagnostic error', *Diagnosis* **2**(2), 97–103.