Empirically Understanding the Value of Prediction in Allocation

Unai Fischer-Abaigar*[1,2], Emily Aiken[3], Christoph Kern[1,2], and Juan Carlos Perdomo[4, 5]

[1] *LMU Munich*
[2] *Munich Center for Machine Learning*
[3] *University of California San Diego*
[4] *New York University*
[5] *Massachusetts Institute of Technology*

Abstract

Institutions increasingly use prediction to allocate scarce resources. From a design perspective, better predictions compete with other investments, such as expanding capacity or improving treatment quality. Here, the big question is not *how* to solve a specific allocation problem, but rather *which* problem to solve. In this work, we develop an empirical toolkit to help planners form principled answers to this question and quantify the bottom-line welfare impact of investments in prediction versus other policy levers such as expanding capacity and improving treatment quality. Applying our framework in two real-world case studies on German employment services and poverty targeting in Ethiopia, we illustrate how decision-makers can reliably derive context-specific conclusions about the relative value of prediction in their allocation problem. We make our software toolkit, `rvp`, and parts of our data available in order to enable future empirical work in this area.

1 Introduction

Public institutions worldwide increasingly use predictive systems to guide the allocation of scarce social goods, in applications ranging from identifying students at risk of dropout for education-focused interventions, to allocating public housing units to unhoused families, to prioritizing vulnerable communities for early access to new vaccines. Prediction in these settings is not an end in itself, but a means to improve the targeting of limited resources and, ultimately, advance institutional goals and downstream social welfare.

Historically, allocation problems have often been framed as optimizing a specified objective subject to given constraints, e.g. maximizing expected welfare subject to budget limits. A substantial literature addresses how to solve such problems once formulated: learning optimal treatment assignment rules through empirical welfare maximization [Athey and Wager, 2021, Kitagawa and Tetenov, 2018], training constrained predictors that satisfy operational or fairness requirements [Cotter et al., 2019], or establishing calibration conditions under which simple post-processing leads to optimal downstream decisions [Hu et al., 2023].

*Corresponding author: `Unai.FischerAbaigar@stat.uni-muenchen.de`

However, this previous work takes the problem definition—which features are observed, what institutional capacity exists, and the quality of allocated resources—as exogenously given. We argue a the central question is not *how do we solve a specific allocation problem*, but *which allocation problem should we be solving?* These meta-design choices (such as capacity constraints, data quality, feature selection, intervention intensity, and more) are themselves optimization variables—and they are just as, if not more consequential to overall success than the downstream algorithmic choices about how to solve the resulting optimization problem.

Once these design choices are recognized as optimization variables themselves, the design space is vast: a planner might invest in better prediction for a demographic subgroup, expand program capacity, or improve service quality. How should practitioners navigate this space?

Recent work has begun to study this design space from a theoretical perspective [Perdomo, 2024, Fischer-Abaigar et al., 2025, Shirali et al., 2024]. These mathematical frameworks provide valuable insights, characterizing, for instance, when prediction improvements dominate capacity expansion or vice versa. However, they typically focus on specific comparisons—for example, aggregate prediction improvements versus aggregate capacity expansion—under stylized distributional assumptions.

Many practically relevant questions remain unexplored: When are prediction improvements for specific subgroups worthwhile? How do conclusions change under different welfare objectives? What about policy levers beyond prediction and capacity? Furthermore, practitioners may rightly question whether existing theoretical results apply to their particular domains, where the relevant comparisons, cost structures, distributions, and welfare objectives may differ substantially from those analyzed in theory.

Here, we take an empirical approach to understanding the value of prediction in allocation problems. We conceptually map out the space of questions practitioners face when comparing policy levers, and develop a software toolkit `rvp` that enables them to answer these questions using data from their own programs. Rather than relying on theoretical results derived under stylized assumptions, planners can simulate concrete comparisons that apply to their specific domains. We demonstrate this toolkit on real-world data from German employment services and cash transfer distribution in Ethiopia, illustrating how to generate actionable insights on when and where prediction investments pay off.

2 Related Work

A growing body of work examines machine learning systems that allocate scarce resources in public institutions. A recurring finding is that prediction, while often the focus of technical development, may not be the most consequential component of these systems. For example, Perdomo et al. [2025] study early warning systems in Wisconsin public schools and find that effective allocation of school resources may not require individual-level risk prediction.

This observation motivates a line of theoretical work on the value of prediction in allocation systems. Casacuberta and Hardt [2026] show that the sample complexity of learning good allocations is far lower than for heterogeneous effect estimation. Shirali et al. [2024] show that individual targeting only outperforms aggregate targeting in high-budget regimes when inequality is low. Mashiat et al. [2025] find that prediction-based targeting of outreach to prevent eviction outperforms neighborhood-based approaches under a spatial cost structure. Perdomo [2024] and Fischer-Abaigar et al. [2025] develop formal frameworks for comparing prediction improvements against other policy interventions such as capacity expansion, introducing the prediction-access ratio to quantify these tradeoffs. Wilder and Welle [2025] examine balancing treating those in need against gathering information on treatment effectiveness; Jain et al. [2024] consider when to introduce randomization into allocation systems; Jain et al. [2025] show that many allocations achieve equal utility while selecting entirely different individuals. These contributions reveal a vast design space, but results are sensitive to problem-specific assumptions, and theoretical analyses necessarily focus on stylized settings.

A related literature focuses on *solving* allocation problems once formulated. A common approach is to learn risk scores that rank individuals for allocation [Liu et al., 2026, Wang et al., 2024], referred to as "prediction policy problems" by Kleinberg et al. [2015]. Recognizing the causal nature of these problems, counterfactual risk prediction models estimate outcomes under no intervention [Coston et al., 2020]. Allocating to those who would benefit most requires estimating heterogeneous treatment effects, drawing on a vast literature on heterogeneous effect estimation [Wager and Athey, 2018, Künzel et al., 2019].

A long history in causal inference focuses on learning optimal treatment rules [Kitagawa and Tetenov, 2018, Athey and Wager, 2021, Manski, 2004, Kallus, 2018, Le et al., 2019]. Complementary work on decision-focused learning optimizes predictors to account for the downstream optimization problem they intend to solve [Elmachtoub and Grigas, 2022, Ren et al., 2024]. Recent work attempts to clarify how these approaches relate, connecting estimation targets to utility under different problem formulations [Sharma and Wilder, 2025, Liu et al., 2026, Fischer-Abaigar et al., 2024, Kern et al., 2025]. Our work is not primarily concerned with estimation or solving a given allocation problem, but rather with the meta-design of the allocation problem itself and the tradeoffs that arise from different design choices.

3 Prediction in Allocation

We focus on allocation problems where a social planner assigns a scarce intervention to individuals in a population:

Definition 3.1 (Allocation Problem). *Let $\mathcal{D}$ be a distribution over pairs $(x,w) \in \mathcal{X} \times \mathcal{W}$ and let $\alpha \in (0,1)$ denote a capacity constraint. Fix a utility function $u : \mathcal{W} \times \{0,1\} \to \mathbb{R}$. We define $\mathsf{Alloc}(\mathcal{D},u,\alpha)$ to be the value of the optimal policy $\pi\colon \mathcal{X} \to \{0,1\}$ with respect to $(\mathcal{D},u,\alpha)$,*

$$\mathsf{Alloc}(\mathcal{D},u,\alpha) := \max_{\pi\colon \mathcal{X}\to\{0,1\}} \mathbb{E}_{\mathcal{D}}\left[u(w,\pi(x))\right] \quad s.t.\ \mathbb{E}_{\mathcal{D}_x}\left[\pi(x)\right] \leqslant \alpha$$

The allocation problem is defined based on three primitives; the utility function u, distribution $\mathcal{D}$, and the capacity constraint α. Prior work largely treats them as exogenous. We instead consider them as design parameters that a policymaker decides on, opening up the optimization space beyond the choice of decision rule π.

The outcome w and the utility function u together encode the social planner's objective. In applications, w may take different forms. For example, w may represent an outcome y (e.g., probability of disease occurrence), a baseline outcome under no treatment $w = y(0)$ (e.g., dropout risk without intervention), or a treatment effect $w = y(1) - y(0)$ (e.g., program efficacy for that individual)[1]. When the net utility $\Delta u = u(w,1) - u(w,0)$ is affine in w, individuals can

[1] For instance, if one sets $w = y(0)$ (baseline risk), the utility may encode a constant known effect τ as $u(w,1) = w+\tau$. Alternatively, one could set $w = y(1) - y(0)$ with simpler utility $u(w,1) = w$. The conceptual difference is that w will need to be estimated, whereas we treat u as a known function.

be optimally ranked by $\mathbb{E}[w \mid x]$, and the optimal solution takes a simple form: allocate to the top-ranked individuals, while respecting the budget constraint,

$$\pi^*(x) = \mathbf{1}\{\mathbb{E}[w \mid x] \geqslant t_\alpha\}.$$

Here, t_α is chosen such that $\Pr(\mathbb{E}[w \mid x] \geqslant t_\alpha) = \alpha$, allocating to top-ranked individuals, provided expected net-utility is positive. In practice, $\mathbb{E}[w \mid x]$ is unknown and must be approximated from historical data. We restrict our focus to policies that threshold a scalar predictor $p(x) \approx \mathbb{E}[w \mid x]$. Such policies are the default choice among practitioners and theoretically optimal in some cases (e.g., Perdomo 2024, Fischer-Abaigar et al. 2025).

4 Policy Levers

From a planner's perspective, there are multiple ways to redefine the allocation problem. Following Perdomo [2024] and Fischer-Abaigar et al. [2025], we call these *policy levers*. Although our work does not contain formal theorems, we define these formally for the sake of conceptual clarity.

Definition 4.1 (Policy Lever). *A policy lever is an intervention that modifies one or more components of the allocation problem. Formally, for a given allocation problem, a policy lever induces a mapping $\ell : (\mathcal{D}, u, \alpha) \to (\mathcal{D}', u', \alpha')$, together with associated investment cost[2] $c(\ell) \geqslant 0$. Here, $\mathsf{Alloc}(\mathcal{D}', u', \alpha') - \mathsf{Alloc}(\mathcal{D}, u, \alpha)$ denotes the optimal welfare gain after applying the policy lever.*

This definition is intentionally general. A planner might consider dropping an expensive-to-collect feature set, investing in program capacity at a local office, improving service quality for a subpopulation, or enhancing data collection throughout. All of these can be represented as changes to the distribution, constraint, or utility. Furthermore, policy levers can often be parameterized by a single quantity $\theta \geqslant 0$ that controls the magnitude of intervention with associated costs $\theta \mapsto c(\theta)$. We discuss three instantiations of such parameterized policy levers.

Expanding (Local) Capacity Relaxing the constraint $\alpha \to \alpha + \Delta\alpha$ allows serving a larger fraction of the population. Capacity expansions can also be locally targeted. For example, increasing capacity at specific regional offices or for particular subgroups. This can be easily represented by adding additional constraints to the allocation problem.

Improving Treatment Improving the treatment effectiveness can be represented as a change in utility u. For instance, in a cash transfer program, increasing the cash transfer amount from b to $b + \Delta b$ increases the welfare gain from allocating to a household; a direct change to $u(a, w)$.

Improving Prediction Prediction improvements can be represented as changes to the joint distribution $\mathcal{D}$ that make available features x more informative about outcomes w. In practice, many mechanisms could improve prediction: collecting additional features, expanding data coverage to unobserved populations, or improving data quality.

[2] An important dimension is the time horizon over which costs are incurred. One-time investments (e.g., developing a new data system) should be evaluated differently than recurring costs (e.g., ongoing data collection or expanded capacity). We abstract from this by assuming all comparisons occur over the same time horizon, with recurring costs included in the total investment amount $c(\ell)$ over that period.

4.1 rvp Toolkit: Comparing Policy Levers

Given the broad space of possible policy levers, the central question is: which lever delivers the greatest welfare gain for a given level of investment? We conceptually map out the space of questions practitioners face when making such comparisons and present rvp, a software toolkit that enables practitioners to construct scenarios using their own data to make principled comparisons across policy levers that account for their specific distributions, welfare objectives, and cost structures. Appendix C maps these conceptual components to their implementations in rvp[3].

In practice, such comparisons are often difficult because costs or benefits may be underspecified. A planner may not know what investment yields a given prediction improvement, or how capacity expansion costs scale. However, even with incomplete information, we show that useful comparisons remain possible. We partition the scenarios rvp supports based on what cost information is available to the planner.

(Q1): How should a fixed budget be allocated across levers? When costs are known for all levers, the planner can compare levers directly. Given a total budget B, the problem becomes allocating resources across levers to maximize total welfare gain. This joint optimization produces budget-allocation curves showing how resources should be distributed as a function of total budget. *See, e.g., Section 6.1 and Section 6.2.*

(Q2): What improvement would one lever need to deliver to match an investment in another? When costs are known for one lever but uncertain for another, we can use the known lever as a benchmark. Suppose a planner knows the cost of expanding capacity by a fixed amount. What prediction improvement would generate equivalent welfare gains? This identifies a break-even point, allowing the planner to assess whether such gains are plausibly achievable at comparable cost. *See, e.g., Section 5.1.*

(Q3): How much more should we be willing to pay for one lever over the other? When cost information is unavailable for both levers, we can still compare their relative impact. Following Perdomo [2024] and Fischer-Abaigar et al. [2025], we consider the ratio of welfare gains from each lever. This ratio provides a threshold. For example, suppose a given prediction improvement generates twice the welfare gain of a given capacity expansion. Then the prediction improvement is the better investment as long as it costs less than twice as much. *See, e.g., Section 5.2.*

5 Case Study: German Employment Offices

Having introduced our empirical toolkit, we now demonstrate how practitioners can use rvp to derive actionable insights around the design of prediction-based allocation systems. All analyses in this section were conducted using rvp. Our first case study focuses on identifying long-term unemployment, using real administrative data from the German Federal Employment Agency. Prediction-based profiling of jobseekers is a topic of substantial policy relevance. To our knowledge, such systems have not been analyzed empirically using the kind of framework we develop here.

[3]Full documentation and worked examples are available at github.com/unai-fa/relative-value-of-prediction.

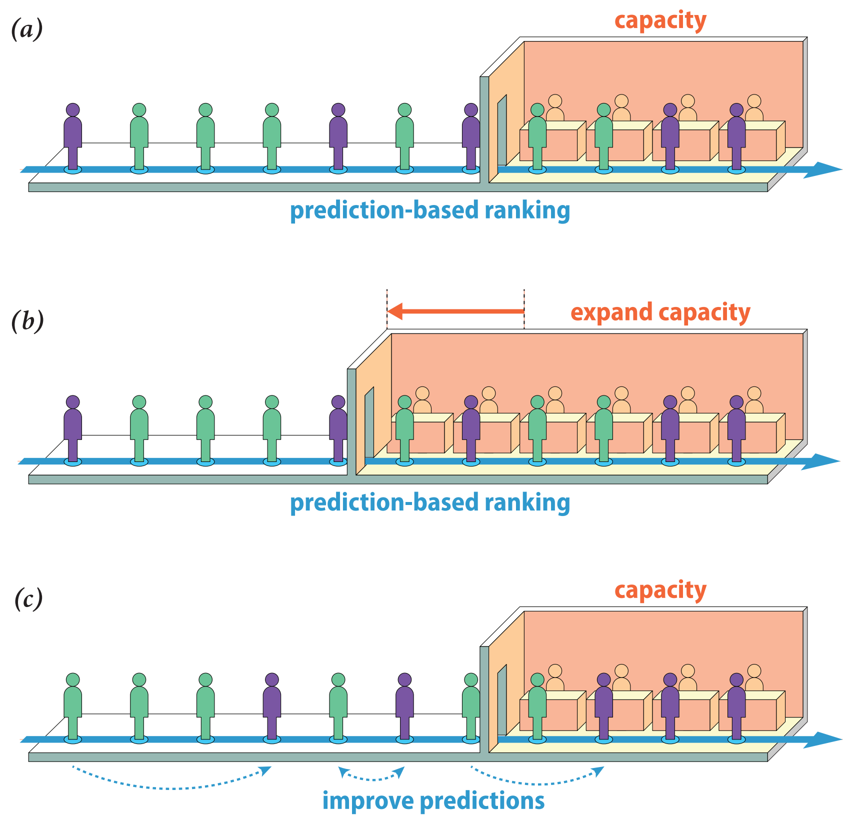


Figure 1: **(a)** Illustrates the task of identifying jobseekers at risk of long-term unemployment (purple) under a fixed capacity constraint. The employment office observes only imperfect predictions of risk, so the ranking of individuals is uncertain. **(b)** Increasing capacity expands the set of individuals who can be served. **(c)** Improving prediction sharpens the ranking, allowing limited resources to be targeted more effectively toward those truly at risk.

Background. Public Employment Services (PES) across countries deploy predictive models to support one of their most critical task: the allocation of support programs to job seekers at risk of long-term unemployment (LTU) [Körtner and Bonoli, 2023, Desiere et al., 2019]. As LTU incurs considerable costs for public welfare, the support programs allocated to at-risk job seekers account for large shares of PES spending. The need for improved efficiency and effectiveness is high [Bach et al., 2023, Kern et al., 2024]. We use a 2% random sample of all German governmental labor market records comprising over 60 million entries from 1970–2021. More information on the data and experimental setup can be found in Appendix A.

Problem Setup. Agencies use pre-unemployment covariates $x \in \mathcal{X}$, such as employment and benefits history, to predict unemployment duration w (measured in days). A common utility is $u(w,a) = a \cdot \mathbf{1}\{w \geqslant t_\beta\}$ which rewards correctly targeting jobseekers in the top-β fraction of unemployment duration (see Figure 1). This assumes no harm from misallocating, i.e. $u(w,0) = 0$. The planner can prioritize at most an α-fraction of the population and sets $\pi(x) = \mathbf{1}\{p(x) \geqslant t_\alpha\}$, where t_α is the appropriate quantile.

What is the value of prediction for identifying jobseekers at risk? Before comparing policy levers, we first ask: what is the value of predictive targeting at all? Given the utility defined above, we can bound the value of prediction. Random targeting achieves welfare $\alpha\beta$, while perfect targeting achieves $\min(\alpha, \beta)$. For example, with $\beta = 0.15$ and $\alpha = 0.1$, perfect targeting correctly identifies 100 at-risk individuals per 1,000 jobseekers, compared to only 15 under random assignment. In practice, prediction quality lies between these extremes. Using a gradient boosting model trained on administrative records, predictive targeting achieves 42 correct identifications per 1,000 jobseekers. Figure 9(a) in the Appendix shows how this value declines with rising capacity; when resources are abundant, accurate targeting matters less.

To explore the value of prediction along this continious range, we interpolate between current performance and perfect prediction, $\tilde{p} = p + \eta(w - p)$, where $0 \leqslant \eta \leqslant 1$ controls the magnitude of improvement, implying $\mathrm{RMSE}(\tilde{p}) = (1-\eta) \times \mathrm{RMSE}(p)$. In settings like employment services, where agencies already operate on large administrative datasets, further predictive gains require collecting new features and data sources, investments whose costs and effects are difficult to anticipate. With this approach, planners can evaluate whether a given improvement in predictive accuracy would be worthwhile relative to other levers. By allowing η to vary across the population, we can simulate targeted improvements for subgroups of the population.

- *Covariate Subgroup.* Setting $\eta(x) = \eta \cdot \mathbf{1}\{g(x) = 1\}$ concentrates improvement in a particular demographic group, region, or administrative office.
- *Prediction Subgroup.* And, $\eta(p) = \eta \cdot \mathbf{1}\{|p(x) - \tau| < \varepsilon\}$ targets marginal cases near the decision boundary.

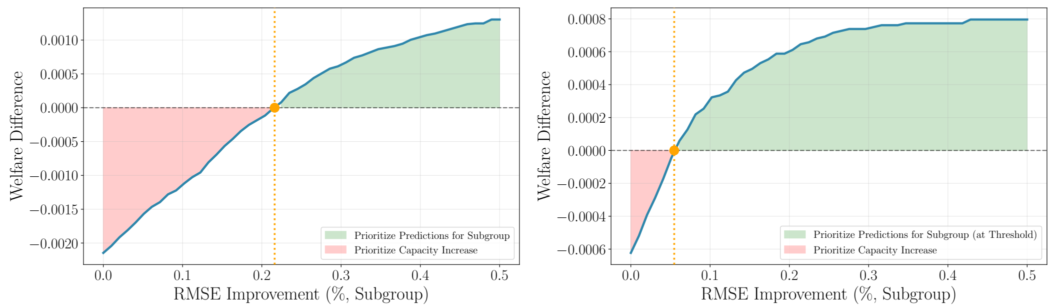
Figure 9(b) in the Appendix shows the welfare gains induced by uniform prediction improvements across the population.

5.1 Is collecting additional information for jobseekers with limited employment history worth it? (Q2)

Next, we consider a scenario in which we aim to simulate the value of improving the predictor for a specific subgroup. We focus on jobseekers over 35 for whom information on last job is missing or unavailable ($\approx$ 3% of cases), either because records are incomplete or because they have limited formal employment experience. These individuals are harder to assess using standard administrative data. Specifically, we set $\eta(x) = \eta \cdot \mathbf{1}\{\text{age} > 35 \wedge \text{missing job information}\}$. In our setting, we observe that RMSE would need to decrease by 17% to match the RMSE of the overall population.

Specifically, we compare two investment options: (1) allocating caseworker hours to conduct intake interviews to collect additional information for this subgroup, thereby improving p, or (2) using those same caseworker hours to expand α, prioritizing a larger fraction of the overall population with existing prediction quality. Suppose each additional screening slot requires 4 hours of caseworker time, while collecting missing employment information takes approximately 1 hour per jobseeker with unknown job history. The cost of data collection is known, but it is unclear what prediction improvement it will generate.

What prediction improvement would justify investing in data collection rather than capacity expansion? We use `rvp` to simulate this tradeoff. Figure 2(a) shows the welfare difference between investing in prediction improvements versus capacity expansion, as a function of the RMSE reduction achieved in the subgroup. The yellow line marks the break-even point. RMSE would need to decrease by at least 20% to make prediction investment welfare-equivalent to



(a) (b)

Figure 2: Comparison of prediction improvements versus capacity expansion for targeting long-term unemployment risk ($\alpha = \beta = 0.15$). We simulate improving predictions for jobseekers over 35 years with missing employment history versus expanding screening capacity (assuming 4 hours per slot). **(a)** Welfare difference as a function of RMSE reduction in the subgroup assuming that collecting missing employment information costs 1 hour per jobseeker. The yellow line marks break-even. **(b)** Welfare difference when prediction improvements are concentrated on subgroup members near the decision threshold (10% bandwidth).

capacity expansion, while a 17% reduction is needed to match population-average prediction quality. Since break-even requires exceeding this threshold, bringing the subgroup to better-than-average prediction quality, the planner must assess whether such gains are realistic.

Figure 3 shows for each level of RMSE improvement the equivalent capacity expansion (in caseworker hours) that would generate the same welfare gain. This provides intuition about the "opportunity cost" of prediction investments. For instance, a 10% RMSE reduction in this subgroup generates welfare equivalent to expanding capacity by approximately $1,000$ hours, or 250 additional screening slots.

However, more targeted improvements in prediction may still be worthwhile. Social planners could focus data collection on individuals near the decision threshold, where better information is most likely to change allocation decisions. We model this as $\eta(x,p) = \eta \cdot \mathbf{1}\{\text{age} > 35 \wedge \text{missing job information}\} \cdot \mathbf{1}\{|p(x) - \tau| < \varepsilon\}$, where ε defines a 10% bandwidth. Figure 2(b) (bottom) shows that this targeted approach requires a much smaller RMSE reduction to break even – less than 5% for those near the threshold. For this employment office, the analysis suggests that intake interviews for all jobseekers with missing history are less likely to pay off, but targeted data collection for those near the threshold could be worthwhile.

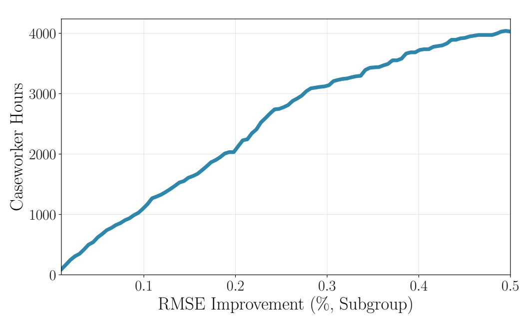
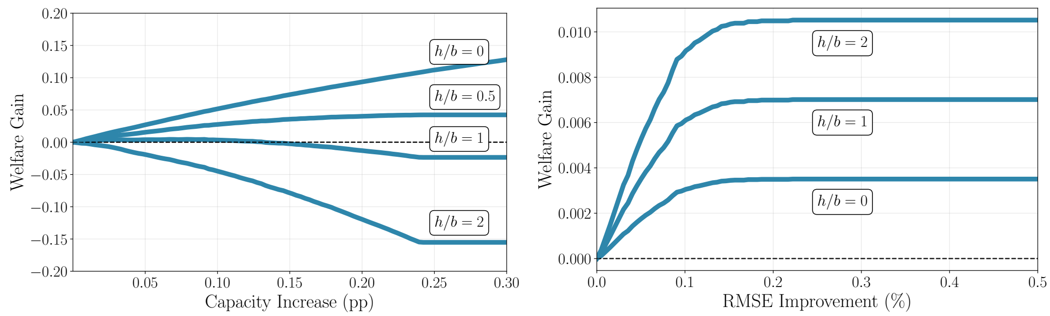


Figure 3: Equivalent capacity expansion (in caseworker hours) that would generate the same welfare gain as each level of RMSE improvement.



(a) (b)

Figure 4: Gain from capacity expansion and uniform prediction improvement under different harm-to-benefit ratios $\frac{h}{b}$. All panels use $\beta = 0.25$, $\alpha = 0.01$, and allocate until capacity is exhausted while excluding individuals with negative expected utility.

5.2 What is the relative value of prediction when allocating both good and harms? (Q3)

Employment agencies often aim to allocate job training programs to individuals expected to benefit most. Empirical evaluations document substantial heterogeneity in training effects across baseline labor market prospects [Osikominu, 2021, 2013, Wunsch, 2016]. Training programs typically require participants to reduce job search activity during enrollment, creating lock-in effects that delay employment. For individuals with strong labor market prospects (those likely to find work quickly) this opportunity cost can outweigh the downstream benefits of training. In contrast, individuals at high risk of long-term unemployment tend to benefit more: they face lower opportunity costs from lock-in and show larger gains in downstream earnings [Osikominu, 2013].

To capture this pattern, we adopt a simple utility $u(w,a) = a \cdot u(w)$ based on predicted unemployment duration y where $u(w) = -h$ if $w < t_\beta$, and $u(w) = b$ if $w \geq t_\beta$ and b and h represent the net benefit and harm respectively (e.g., in income prospects or employment duration). Employment offices have used similar simple targeting logics to guide intake decisions in the past [Allhutter et al., 2020].

When does prediction become more valuable than capacity? Previous work has shown that in low-capacity regimes, the marginal value of serving more individuals greatly exceeds the marginal value of better targeting [Perdomo, 2024, Fischer-Abaigar et al., 2025]. In other words, the relative value of prediction is low. But this conclusion critically assumes that misallocation is merely wasteful, not harmful.

We use rvp to trace out welfare gains across a range of harm-to-benefit ratios $\frac{h}{b}$ (see Figure 4(a). When $\frac{h}{b} = 0$, capacity expansion is beneficial and dominates prediction improvements. As $\frac{h}{b}$ increases, returns diminish and turn negative; expanding capacity *reduces* welfare because the harm from misallocating to low-risk individuals outweighs the benefits of reaching additional high-risk individuals.

Figure 4(b) shows the opposite pattern. The value of better prediction *increases* with $\frac{h}{b}$. When misallocation is costly, better prediction allows the system to reach those at risk without causing harm. The relative value of prediction rises accordingly. When allocation can cause harm, the conclusion that capacity dominates prediction may reverse, even in low-capacity regimes where

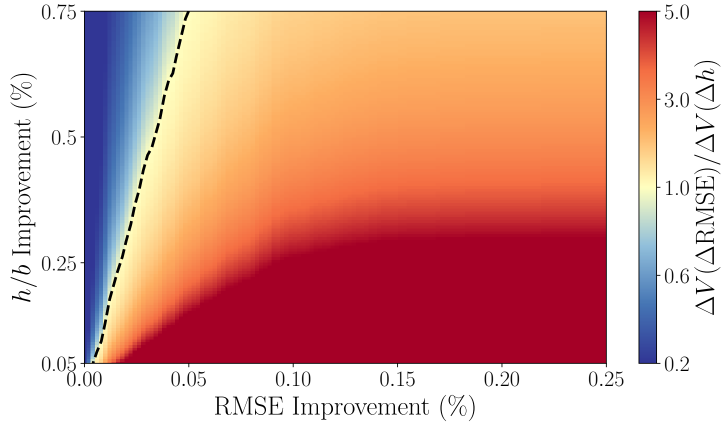


Figure 5: Ratio of welfare gains ($\beta = 0.25$, $\alpha = 0.01$) from prediction improvement (x-axis: RMSE reduction at $\frac{h}{b} = 2$) versus harm reduction (y-axis: reduction in $\frac{h}{b}$). Heatmap values are truncated at 0.2 and 5.0.

capacity expansion would otherwise be the clear priority. This setting was absent from prior work; `rvp` allowed us to derive these conclusions empirically, without new theoretical analysis.

How much more should we be willing to pay for reducing harm versus improving prediction? A social planner may also consider reducing the harm caused by misallocation, for example, mitigating lock-in effects through better timing of interventions [Wunsch, 2016]. How does the value of such improvements compare to investing in better prediction? Even without cost information on either lever, we can compare their relative welfare impacts. Figure 5 shows the ratio of welfare gains from prediction improvement versus reductions in $\frac{h}{b}$. This ratio indicates how much more a planner should be willing to pay for one lever over the other. In this low-capacity regime ($\alpha = 0.01$), even small prediction improvements generate large relative gains. A 4% RMSE reduction matches the welfare gain of halving misallocation harm. For a social planner, this comparison offers useful information. Halving the relative harm of an intervention is difficult. It likely requires restructuring program timing, duration, or delivery. A small RMSE improvement, by contrast, may be far more achievable.

6 Case Study: Poverty Targeting in Ethiopia

A second, complementary case study is the targeting of anti-poverty programs to poor households. The poverty targeting case study differs from the public employment services case study in two key ways: (1) the costs of all policy levers are well-known, allowing the planner to optimize directly for welfare given a budget constraint, and (2) the treatment intensity is continuous, allowing the planner to optimize over the trade-off between improving treatment and increasing capacity (as well as improving prediction).

Background. We focus on proxy means tests for the targeting of cash transfer programs. Direct cash transfer programs are a globally ubiquitous approach to distributing aid, with 1.36 billion people receiving cash transfers from 962 programs globally in 2020 [Gentilini, 2022]. Proxy means tests (PMTs) — which are used to target cash transfer programs in at least fifty low- and middle-income countries [Barrientos, 2018] — use a simple predictive model to infer household

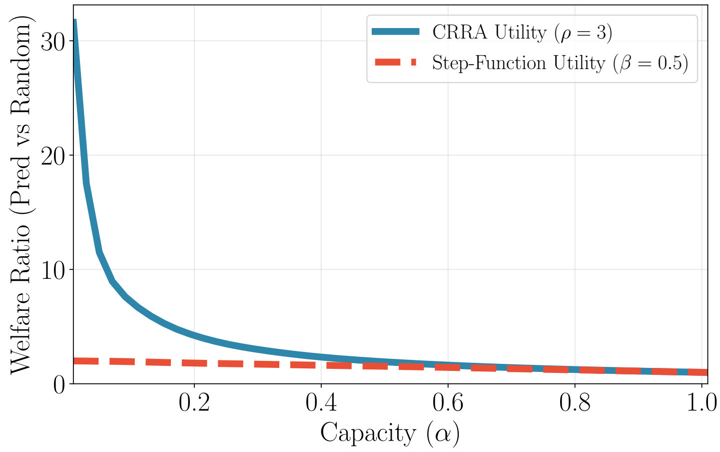


Figure 6: Ratio of welfare impacts under predictive targeting to welfare impacts under random targeting, as a function of program capacity (keeping benefits fixed at $100 USD PPP per household).

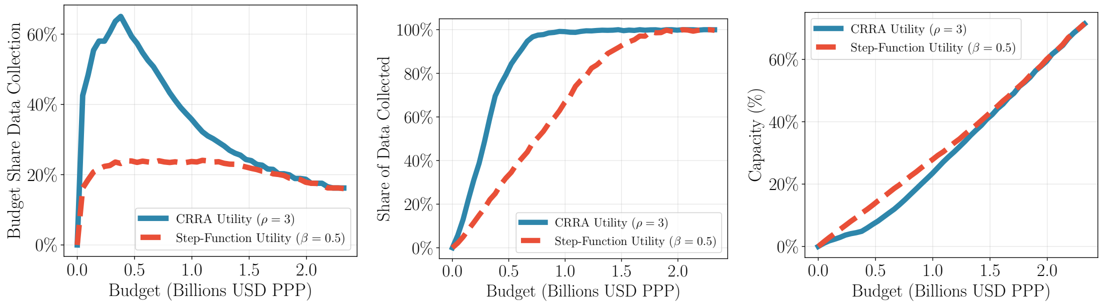
poverty, where households with the highest-estimated poverty probability are eligible for cash transfer programs Grosh et al. [1995], Brown et al. [2018]. While the accuracy of proxy means tests for identifying poor households has been assessed in a number of contexts Brown et al. [2018], Schnitzer and Stoeffler [2024], McBride and Nichols [2018], the trade-offs between investing in prediction and allocation in the PMT setting have not previously been analyzed. We use `rvp` to conduct such an analysis for the first time, using publicly available survey data from 4,694 Ethiopian households in 2015. See Appendix B for experimental details and additional background.

Problem setup. The PMT model relies on survey data collected in short "poverty scorecard" surveys, which are enumerated in the field. The model is trained to estimate household poverty on a subset of households for which both the poverty scorecard predictors and a measure of household poverty are collected, and is deployed to predict poverty for all households for which scorecards (features) have been collected. The poorest-estimated households are eligible for program benefits (up to the capacity threshold). We simulate such a predictive approach using household survey data from Ethiopia's 2015 Living Standards Measurement survey, which includes both poverty scorecard questions and a measure of total household consumption expenditures (serving as the "ground truth" measure of poverty).

What is the value of prediction for targeting cash transfers? We again begin by assessing the welfare gains associated with predictive accuracy for targeting in the cash transfer setting. The welfare gains from predictive targeting depend on how we quantify the impact of a cash transfer as a function of pre-transfer poverty. We evaluate two utility functions, and show how the value of prediction depends on the choice of utility function.

Our first utility function is a step function, as in Section 5, where welfare gains are proportional to the size of the transfer if a targeted household is poor (which we define as households below the average consumption level $\bar{w}$): $u(w, a) = b \cdot a$ if $w \leqslant \bar{w}$, and 0 otherwise.

Here, w is pre-transfer household consumption expenditures and b is the size of the cash transfer delivered to eligible households. The step utility function corresponds closely to standard evaluation metrics such as accuracy, precision, and recall for measuring the accuracy of a targeting rule. Within households that are among the poorest half, prioritizing poorer



(a) (b) (c)

Figure 7: **(a)** Welfare-optimizing share of budget allocated to data collection vs. transfers, as a function of program budget; **(b)** optimal share of data collected as a function of program budget; **(c)** optimal program capacity as a function of program budget.

households has no additional benefit, and providing transfers to richer households has no benefit at all.

An alternative utility function models positive effects for cash transfers to any household, but larger impacts for transfers to poorer households. The constant relative risk aversion (CRRA) utility function [Hanna and Olken, 2018] is standard in development economics, assuming diminishing impacts of cash transfers as a function of household welfare:

$$u(w,a) = \frac{(w+ba)^{1-\rho} - w^{1-\rho}}{1-\rho}, \qquad \rho > 0,\ \rho \neq 1$$

The curvature of the CRRA utility function is governed by ρ, with higher values of ρ corresponding to larger utility gains from cash transfers to poorer households (relative to cash transfers to richer households). Values of ρ from 2-4 are standard in the literature Gandelman and Hernandez-Murillo [2015]; following Hanna and Olken [2018] we use a value of $\rho = 3$ in our empirical simulations.

We evaluate a simple allocation policy that ranks households by predicted consumption $p(x)$ and allocates to those predicted poorest up to capacity $\pi(x) = \mathbf{1}\{p(x) \leqslant t_\alpha\}$. Figure 6 shows the ratio of welfare under predictive versus random targeting, varying capacity (i.e. the share of households eligible for cash transfers) with a fixed benefit size of \$100 USD PPP (purchasing power parity) — corresponding to an 8% increase in yearly consumption for the average household in Ethiopia in 2015. The value of predictive targeting is substantially higher under CRRA utility than step function utility, and highest at constrained capacity levels. Similar patterns hold when varying benefit size at fixed capacity (Figure 11).

6.1 What is the value of improving prediction when designing a new program? (Q1)

Because the poverty scorecard surveys used to estimate household poverty in proxy means tests are expensive to conduct — with a median per-survey cost of \$13 USD PPP reported in the literature [Aiken et al., 2023] — such PMT registry databases often fail to cover all households in a country, or even all households in the poorest parts of a country. A recent review of PMT registries in low- and middle-income countries found a median registry coverage of 21% of households (ranging from less than 1% in Belize to over 99% in Argentina) [Grosh et al., 2022].

Households without PMT scorecard data collected are by default ineligible for aid programs targeted with a proxy means test, effectively dramatically eroding predictive accuracy of the targeting model at deployment time.

The administrators of PMT-targeted cash transfer programs thus face an allocation design problem: with a fixed budget to invest, should they prioritize investing in data collection to improve targeting accuracy (by reducing the errors of exclusion arising from households not having PMT data collected), or invest in increasing the cash transfer budget dispersed by the program?

We use `rvp` to simulate a planner designing a new aid program with a fixed budget. With no data yet collected for program targeting, how much of the budget should be invested in PMT data collection vs. cash transfers? We assume a fixed benefit size, so the only investment in allocation available to the planner is increasing capacity (we relax this constraint in Section 6.2). We simulate aid programs with budgets up to $2.7 billion PPP per year.[4] We then identify the welfare-optimizing choice of investments in data collection for targeting vs. distribution of cash transfers.

Figure 7 shows that the optimal investment in data collection is higher under CRRA utility than the more egalitarian step function. For both programs, the optimal budget allocations for limited-budget programs (under USD $1.5 billion PPP) are a mix of data collection and expanding capacity; for large-budget programs collecting data for all households is the welfare-optimizing approach.

6.2 What is the value of improving prediction when expanding an existing program? (Q1)

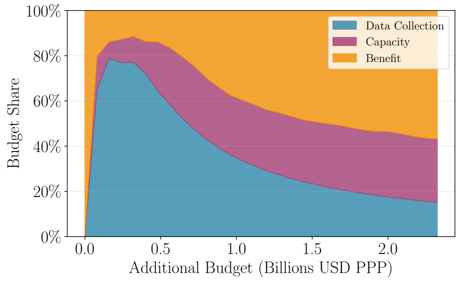
Section 6.1 assumed a fixed transfer size of $100 USD PPP to all beneficiaries, but a planner may have the option to choose between increasing program allocations on the extensive margin by increasing capacity, or on the intensive margin by increasing the benefits allocated to eligible households. Consider a planner who has already made initial investments in data collection, capacity, and benefit sizes. Given an increase in available budget, where should improvements be prioritized?

We consider a typical real-world cash transfer program, with 20% data coverage to begin with, 5% capacity, and a yearly cash transfer to eligible households of $100 PPP.

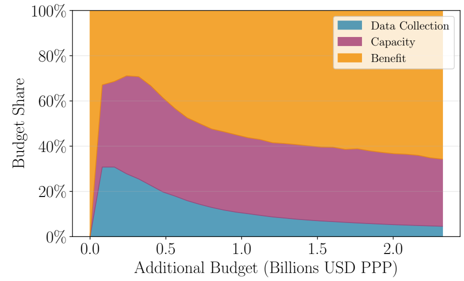
What if the existing data collection coverage was much smaller (1%) or much larger (70%)? What if the existing capacity was much smaller (0.5%) or much larger (40%)? What if the benefit size was much smaller ($10 per household) or much larger ($1,000 per household)? `rvp` answers this question by optimizing over policy levers, given the planner's baseline. Figure 8 shows the best incremental investments under CRRA utility for varying initial conditions.

The best allocation of resources to each policy lever is very dependent on the original investments, demonstrated by the variation across subplots in Figure 8. When there is little data already labeled, data collection is prioritized, but the share of data labeled does not make a large difference in how the transfer budget should be allocated between the intensive and extensive margin. Data collection to improve predictive accuracy is also more important when there is high initial capacity and large initial transfer sizes.

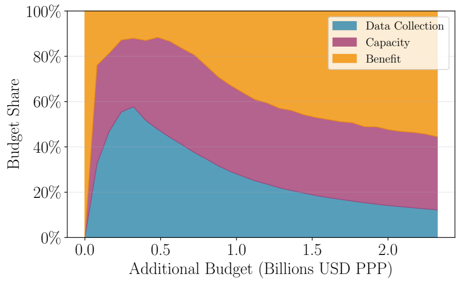
[4] This maximum budget corresponds to 100% capacity for a 100 per year cash transfer program covering Ethiopia's 27.3 million households if no budget were allocated fro data collection.



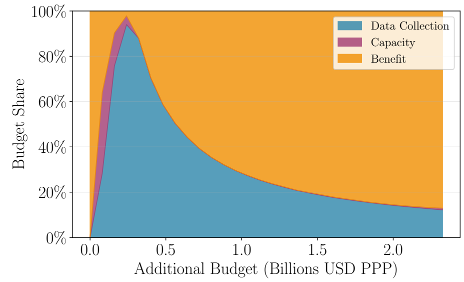
(a) Low initial data share



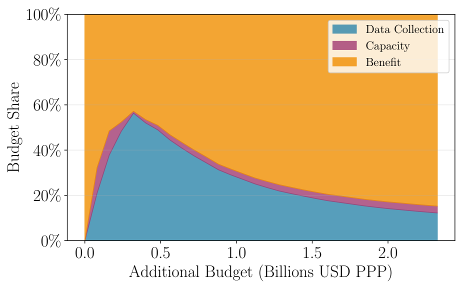
(b) High initial data share



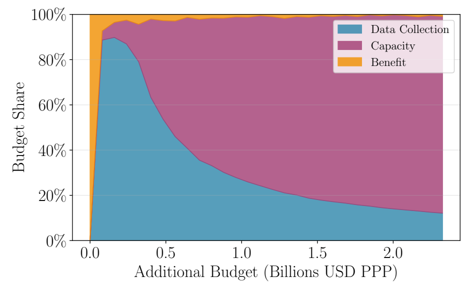
(c) Low initial capacity



(d) High initial capacity



(e) Low initial transfer



(f) High initial transfer

Figure 8: Optimal marginal investments in data collection, capacity, and benefit size as a function of additional budget, varying initial conditions. (a)-(b): Low (1%) vs. high (70%) initial data share, holding capacity (5%) and transfer ($100 PPP) fixed. (c)-(d): Low (0.5%) vs. high (40%) initial capacity, holding data (20%) and transfer ($100 PPP) fixed. (e)-(f): Low ($10 PPP) vs. high ($1000 PPP) initial transfer, holding data (20%) and capacity (5%) fixed.

7 Conclusion

Building on past theoretical work on the relative value of prediction in allocation problems, this paper introduces an empirical framework for reasoning about investing in improved prediction relative to other policy levers available to the implementers of systems that allocate scarce social goods. We show the value of this framework in two empirical examples, and provide the `rvp` software toolkit that generalizes our framework to a variety of prediction-allocation settings. We hope that this work will allow researchers and practioners to apply similar methods to new prediction-allocation problems, and to reason about improving prediction as part of the broader set of design choices available in such settings.

Acknowledgements

UFA acknowledges the support by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education. We are thankful to Chris Haysa and Sendhil Mullainathan for insightful comments and discussions.

References

E. Achterhold, M. Mühlböck, N. Steiber, and C. Kern. Fairness in Algorithmic Profiling: The Amas Case. *Minds and Machines*, 35(1):9, 2025.

E. Aiken, T. Ohlenburg, and J. Blumenstock. Moving targets: When does a poverty prediction model need to be updated? In *Proceedings of the 6th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, pages 117–117, 2023.

E. Aiken, A. Ashraf, J. Blumenstock, R. Guiteras, and A. M. Mobarak. Scalable Targeting of Social Protection: When Do Algorithms Out-Perform Surveys and Community Knowledge? Technical report, National Bureau of Economic Research, 2025.

V. Alatas, A. Banerjee, R. Hanna, B. A. Olken, and J. Tobias. Targeting the Poor: Evidence from a Field Experiment in Indonesia. *American Economic Review*, 102(4):1206–1240, 2012.

D. Allhutter, F. Cech, F. Fischer, G. Grill, and A. Mager. Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics are Made Effective. *Frontiers in Big Data*, 3:5, 2020.

M. Arntz and R. A. Wilke. Unemployment Duration in Germany: Individual and Regional Determinants of Local Job Finding, Migration and Subsidized Employment. *Regional Studies*, 43(1):43–61, 2009.

S. Athey and S. Wager. Policy Learning With Observational Data. *Econometrica*, 89(1):133–161, 2021.

G. Azmat, M. Güell, and A. Manning. Gender Gaps in Unemployment Rates in OECD Countries. *Journal of Labor Economics*, 24(1):1–37, 2006.

R. L. Bach, C. Kern, H. Mautner, and F. Kreuter. The impact of modeling decisions in statistical profiling. *Data & Policy*, 5:e32, 2023.

A. Banerjee, R. Hanna, B. A. Olken, and S. Sumarto. The (lack of) distortionary effects of proxy-means tests: Results from a nationwide experiment in Indonesia. *Journal of Public Economics Plus*, 1:100001, 2020.

A. Barrientos. Social assistance in low and middle income countries 2000-2015. *Research Handbook on Poverty and Inequality*, 2018.

S. G. Bishu and M. G. Alkadry. A systematic review of the gender pay gap and factors that predict it. *Administration & Society*, 49(1):65–104, 2017.

C. Brown, M. Ravallion, and D. Van de Walle. A poor means test? Econometric targeting in Africa. *Journal of Development Economics*, 134:109–124, 2018.

S. Casacuberta and M. Hardt. Good Allocations from Bad Estimates. *arXiv preprint arXiv:2601.05597*, 2026.

B. Cockx, M. Lechner, and J. Bollens. Priority to unemployed immigrants? A causal machine learning evaluation of training in Belgium. *Labour Economics*, 80:102306, 2023. ISSN 0927-5371.

A. Coston, A. Mishler, E. H. Kennedy, and A. Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 582–593, New York, NY, USA, 2020. Association for Computing Machinery.

A. Cotter, H. Jiang, M. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019.

S. Desiere and L. Struyven. Using Artificial Intelligence to classify Jobseekers: The Accuracy-Equity Trade-off. *Journal of Social Policy*, 50(2):367–385, 2021.

S. Desiere, K. Langenbucher, and L. Struyven. Statistical profiling in public employment services: An international comparison. OECD Social, Employment and Migration Working Papers 224, OECD Publishing, Feb 2019.

A. N. Elmachtoub and P. Grigas. Smart "Predict, then Optimize". *Management Science*, 68(1): 9–26, 2022.

U. Fischer-Abaigar, C. Kern, N. Barda, and F. Kreuter. Bridging the gap: Towards an expanded toolkit for ai-driven decision-making in the public sector. *Government Information Quarterly*, 41(4):101976, 2024. ISSN 0740-624X.

U. Fischer-Abaigar, C. Kern, and J. C. Perdomo. The Value of Prediction in Identifying the Worst-Off. In *International Conference on Machine Learning*, pages 17239–17261. PMLR, 2025.

N. Gandelman and R. Hernandez-Murillo. Risk Aversion at the Country Level. *Review*, 97(1): 53–66, 2015.

U. Gentilini. Cash Transfers in Pandemic Times: Evidence, Practices, and Implications from the Largest Scale up in History. *World Bank*, 2022.

D. Goller, M. Lechner, T. Pongratz, and J. Wolff. Active labor market policies for the long-term unemployed: New evidence from causal machine learning. *Labour Economics*, 94:102729, 2025. ISSN 0927-5371.

M. Grosh, J. L. Baker, et al. Proxy means tests for targeting social programs. *Living standards measurement study (LSMS) working paper*, 118:1–49, 1995.

M. Grosh, P. Leite, M. Wai-Poi, and E. Tesliuc. *Revisiting Targeting in Social Assistance: A New Look at Old Dilemmas*. World Bank Publications, 2022.

R. Hanna and B. A. Olken. Universal Basic Incomes versus Targeted Transfers: Anti-Poverty Programs in Developing Countries. *Journal of Economic Perspectives*, 32(4):201–26, 2018.

L. Hu, I. R. L. Navon, O. Reingold, and C. Yang. Omnipredictors for Constrained Optimization. In *International Conference on Machine Learning*, pages 13497–13527. PMLR, 2023.

S. Jain, K. Creel, and A. Wilson. Scarce Resource Allocations That Rely on Machine Learning Should Be Randomized. In *International Conference on Machine Learning*, pages 21148–21169. PMLR, 2024.

S. Jain, M. Wang, K. Creel, and A. Wilson. Allocation multiplicity: Evaluating the promises of the rashomon set. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, page 2040–2055. Association for Computing Machinery, 2025.

Á. F. Junquera and C. Kern. From rules to forests: rule-based versus statistical models for jobseeker profiling. *Journal for Labour Market Research*, 59(1):1–27, 2025.

N. Kallus. Balanced Policy Evaluation and Learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8909–8920, 2018.

C. Kern, R. Bach, H. Mautner, and F. Kreuter. When Small Decisions Have Big Impact: Fairness Implications of Algorithmic Profiling Schemes. *ACM J. Responsib. Comput.*, 1(4), Nov. 2024.

C. Kern, U. Fischer-Abaigar, J. Schweisthal, D. Frauen, R. Ghani, S. Feuerriegel, M. van der Schaar, and F. Kreuter. Algorithms for reliable decision-making need causal reasoning. *Nature Computational Science*, 5(5):356–360, 2025.

T. Kitagawa and A. Tetenov. Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica*, 86(2):591–616, 2018.

J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer. Prediction Policy Problems. *American Economic Review*, 105(5):491–95, May 2015.

J. Körtner and R. Bach. Inequality-Averse Outcome-Based Matching. https://osf.io/preprints/osf/yrn4d, 2023.

J. Körtner and G. Bonoli. Predictive Algorithms in the Delivery of Public Employment Services. In *Handbook of Labour Market Policy in Advanced Democracies*, pages 387–398. Edward Elgar Publishing, 2023.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116 (10):4156–4165, 2019.

H. Le, C. Voloshin, and Y. Yue. Batch Policy Learning under Constraints. In *International Conference on Machine Learning*, volume 97, pages 3703–3712. PMLR, 2019.

L. T. Liu, I. D. Raji, A. Zhou, L. Guerdan, J. Hullman, D. Malinsky, B. Wilder, S. Zhang, H. Adam, A. Coston, B. Laufer, E. Nwankwo, M. Zanger-Tishler, E. Ben-Michael, S. Barocas, A. Feller, M. Gerchick, T. Gillis, S. Guha, D. Ho, L. Hu, K. Imai, S. Kapoor, J. Loftus, R. Nabi, A. Narayanan, B. Recht, J. C. Perdomo, M. Salganik, M. Sendak, A. Tolbert, B. Ustun, S. Venkatasubramanian, A. Wang, and A. Wilson. Bridging Prediction and Intervention Problems in Social Systems. *arXiv preprint arXiv:2507.05216*, 2026.

A. Loxha and M. Morgandi. Profiling the unemployed: a review of OECD experiences and implications for emerging economies. *Social Protection and labor discussion paper*, SP 1424, 2014.

C. F. Manski. Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, 72(4): 1221–1246, 2004.

T. Mashiat, P. J. Fowler, and S. Das. Who Pays the RENT? implications of Spatial Inequality for Prediction-Based Allocation Policies. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2):1686–1697, 2025.

L. McBride and A. Nichols. Retooling Poverty Targeting Using Out-Of-Sample Validation and Machine Learning. *The World Bank Economic Review*, 32(3):531–550, 2018.

A. Narayan and N. Yoshida. Proxy Means Tests for Targeting Welfare Benefits in Sri Lanka. *South Asia Poverty Reduction and Economic Management*, 2005.

A. Noriega-Campero, B. Garcia-Bulle, L. F. Cantu, M. A. Bakker, L. Tejerina, and A. Pentland. Algorithmic targeting of social policies: fairness, accuracy, and distributed governance. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 241–251. Association for Computing Machinery, 2020.

A. Osikominu. Quick Job Entry or Long-Term Human Capital Development? The Dynamic Effects of Alternative Training Schemes. *Review of Economic Studies*, 80(1):313–342, 2013.

A. Osikominu. The dynamics of training programs for the unemployed. *IZA World of Labor*, 2021.

J. C. Perdomo. The Relative Value of Prediction in Algorithmic Decision Making. In *International Conference on Machine Learning*, pages 40439–40460. PMLR, 2024.

J. C. Perdomo, T. Britton, M. Hardt, and R. Abebe. Difficult Lessons on Social Prediction from Wisconsin Public Schools. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, page 2682–2704. Association for Computing Machinery, 2025.

P. Premand and P. Schnitzer. Efficiency, Legitimacy, and Impacts of Targeting Methods: Evidence from an experiment in Niger. *The World Bank Economic Review*, 35(4):892–920, 2021.

K. Ren, Y. Byun, and B. Wilder. Decision-Focused Evaluation of Worst-Case Distribution Shift. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, pages 3076–3093. PMLR, 2024.

H. Rudolph and M. Müntnich. "Profiling" zur Vermeidung von Langzeitarbeitslosigkeit. *Mitteilungen aus der Arbeitsmarkt-und Berufsforschung*, 34(4):530–553, 2001.

A. Schmucker and P. vom Berge. Sample of Integrated LabourMarket Biographies Regional File (SIAB-R) 1975–2021, 2023a. FDZ-Datenreport, 07/2023 (en), Nürnberg.

A. Schmucker and P. vom Berge. Factually anonymous version of the Sample of Integrated Labour Market Biographies (SIAB-Regionalfile) – Version 7521 v1. Research Data Centre of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB), 2023b. URL 10.5164/IAB.SIAB-R7521.de.en.v1.

P. Schnitzer and Q. Stoeffler. Targeting Social Safety Nets: Evidence from Nine Programs in the Sahel. *The Journal of Development Studies*, 60(4):574–595, 2024.

V. Sharma and B. Wilder. Comparing Targeting Strategies for Maximizing Social Welfare with Limited Resources. In *The Thirteenth International Conference on Learning Representations, ICLR*, 2025.

A. Shirali, R. Abebe, and M. Hardt. Allocation Requires Prediction Only if Inequality is Low. In *International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

S. Wager and S. Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

A. Wang, S. Kapoor, S. Barocas, and A. Narayanan. Against Predictive Optimization: On the Legitimacy of Decision-making Algorithms That Optimize Predictive Accuracy. *ACM J. Responsib. Comput.*, 1(1), Mar. 2024.

B. Wilder and P. Welle. Learning treatment effects while treating those in need. In *Proceedings of the 26th ACM Conference on Economics and Computation*, EC '25, page 448–473. Association for Computing Machinery, 2025.

C. Wunsch. How to minimize lock-in effects of programs for unemployed workers. *IZA World of Labor*, page 288, 2016.

S. Zezulka and K. Genin. From the Fair Distribution of Predictions to the Fair Distribution of Social Goods: Evaluating the Impact of Fair Machine Learning on Long-Term Unemployment. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1984–2006, 2024.

E. Zschirnt and D. Ruedin. Ethnic Discrimination in Hiring Decisions: A Meta-Analysis of Correspondence Tests 1990–2015. *Journal of Ethnic and Migration Studies*, 42(7):1115–1134, 2016.

A Identifying Long-Term Unemployment in Germany

A.1 Background

Algorithmic profiling of job seekers in the delivery of support measures are debated and implemented by Public Employment Services (PES) in various countries [Loxha and Morgandi, 2014, Körtner and Bonoli, 2023]. In Germany, the introduction of algorithmic decision-support tools has been discussed since the early 2000s [Rudolph and Müntnich, 2001]. A common narrative in these discussions is the hope to improve efficiency over rule- or case-worker-based profiling workflows, supported by comparative studies that report higher accuracy of algorithmic profiling relative to human case-workers in predicting long-term unemployment risks [Desiere et al., 2019, Junquera and Kern, 2025]. Yet, German PES currently rely on case-worker-based profiling, motivating our case study as an illustrative investigation of how different profiling regimes could support employment services in Germany.

The hope of increased efficiency is contrasted by concerns of disparate impact of prediction-based profiling as historical disadvantages of minorities are often deeply embedded in labour market data. In the German context, this includes long-standing disadvantages and discrimination based on gender and ethnicity [Azmat et al., 2006, Zschirnt and Ruedin, 2016]. Kern et al. [2024] show how profiling models in the German context can incur considerable error disparities based on nationality. Similar concerns are raised for the AMAS model in Austria by Allhutter et al. [2020] and Achterhold et al. [2025], and for a profiling system used by the Flemish PES Desiere and Struyven [2021]. Zezulka and Genin [2024] study the link between LTU risk predictions and targeting decisions with data from the Swiss PES. In the targeting context, Cockx et al. [2023] and Goller et al. [2025] exploit effect heterogeneity of active labour market programs (ALMPs) to find optimal allocations of programs to job seekers. Körtner and Bach [2023] demonstrate how inequality-averse allocation of ALMPs to job seekers could mitigate historical differences in group-specific unemployment risks. Yet, a common thread in these works is that most parameters of the allocation problem are treated as given, i.e, they present modeling solutions given fixed, and often insufficient, data rather than asking how (subgroup) outcomes could be improved with an expanded set of actions, such as additional data collection or increased screening capacity.

A.2 Data

We make use of a comprehensive dataset on German jobseekers. The Sample of Integrated Labour Market Biographies (SIAB) is provided via a Scientific Use File from the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) [Schmucker and vom Berge, 2023b]. The SIAB is a 2% sample of all governmental labor market records in Germany, comprising approximately 60 million entries spanning 1970 to 2021. It contains information on employment status, jobseeker demographics, employment and benefits histories, and participation in active labor market programs. Further details on its content, sampling procedure, and anonymization can be found in the accompanying data report [Schmucker and vom Berge, 2023a].

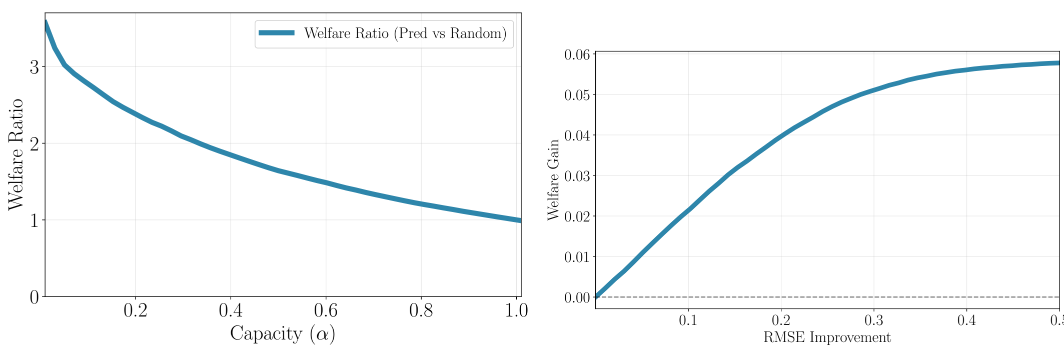
A.3 Predictive Modeling

We construct a prediction task for long-term unemployment, predicting the number of months a jobseeker will remain unemployed at entry into unemployment. Following Bach et al. [2023], Kern et al. [2024], Fischer-Abaigar et al. [2025], we use the same spell aggregation procedure to construct covariates capturing sociodemographic information, labor market and benefits history, and characteristics of the last job held. This leads to 56 numerical and 24 categorical variables; the latter are one-hot encoded. To simulate deployment conditions in real employment offices, we use records from 2010 and 2011 for training, data from 2012 for validation, and data from 2015 for testing. We train a gradient boosting model (CatBoost) on this temporal split. All policy lever comparisons in this paper are conducted on the test set ($n = 86,692$).

A.4 Replication

We provide full replication code for the policy lever comparisons using `rvp`. Due to data sensitivity, the underlying records are not publicly available but can be applied for at the Research Data Centre (FDZ) of the Institute for Employment Research (IAB).

A.5 Additional Analysis



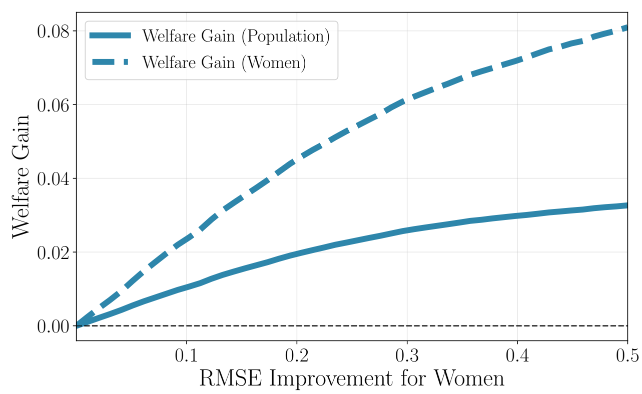
(a) Value of Prediction

(b) Improvements in Prediction

Figure 9: **(a)** Welfare ratio of prediction-based targeting relative to random allocation across different baseline capacities α. **(b)** Welfare gains from improving prediction ($\alpha = 0.1$, $\beta = 0.15$).

Value of Prediction for Women. Evidence suggests that certain demographic groups, such as women or non-citizens, face distinct labor market challenges, including discrimination and greater vulnerability to long-term unemployment [Arntz and Wilke, 2009, Bishu and Alkadry, 2017]. Concerns have also been raised about how statistical profiling tools may reinforce existing inequalities for these groups [Allhutter et al., 2020, Kern et al., 2024]. An employment agency might therefore ask what would be gained from improving prediction specifically for women.

Figure 10 shows welfare gains from prediction improvements for women (setting $\eta(x) = \eta \cdot \mathbf{1}\{\text{female}\}$). We evaluate these gains under two welfare functions; one that averages utility across the full population, and one that averages only over women. A policymaker might primarily aim to improve allocation for women motivated by evidence of labor market discrimination or concerns about algorithmic fairness. Under such a subgroup-specific objective, welfare gains rise more with prediction quality, making prediction improvements relatively more valuable than aggregate welfare alone would suggest. The tradeoffs we study in the following sections are shaped by these choices, how utility is specified and what outcomes are prioritized.



(a) Improvements for Women

Figure 10: Welfare gains from improving prediction for women ($\alpha = 0.1$, $\beta = 0.15$). Welfare is evaluated under two objectives: averaged over the full population, and averaged over women only.

B Poverty Targeting

B.1 Background

Proxy means tests are one of the most popular approaches for targeting cash transfer programs and other social protection programs in low-income contexts where comprehensive adminsitrative data on incomes or consumption expenditures are unavailable. They are currently used to target social protection programs in at least fifty low- and middle-income countries [Barrientos, 2018]. The accuracy of proxy means tests in practice for identifying poor households have been studied in many studies across Africa Brown et al. [2018], Schnitzer and Stoeffler [2024], McBride and Nichols [2018], Asia Narayan and Yoshida [2005], and Latin America Noriega-Campero et al. [2020]. Other work has assessed the fairness of PMTs relative to other targeting approaches Noriega-Campero et al. [2020], susceptibility to intentional data misreporting and collusion Banerjee et al. [2020], and acceptability to data subjects Premand and Schnitzer [2021], Alatas et al. [2012], but little previous work has addressed the benefit of investing in PMT coverage and accuracy in comparison relative to investment in allocating resources to the cash transfer programs and other social programs targeted with PMTs. Aiken et al. [2025] provide an initial framework for such cost-benefit trade-offs using data on targeting cash transfers in Bangladesh, we expand this framework in this paper to account for additional policy levers and data availability constraints.

B.2 Data

The data for our simulations of cash transfer program are from Ethiopia's 2015 Living Standards Measurement Survey. The survey data covers 4,954 nationally representative households. We use data from 4,694 households that have complete information on consumption expenditures, asset posession, housing quality, and household demographics. We convert consumption expenditures data to yearly total consumption expenditures measured in USD PPP using the PPP exchange rate for 2015 from the World Bank.

B.3 Predictive Modeling

The target of our predictive model is log-transformed total household consumption expenditures. The features are 54 variables commonly included in proxy means tests in low-income countries, including:

- Asset possession, such as as whether the household owns radio, TV, satellite dish, sofa, bicycle, car, or bed
- Housing quality, such as how many rooms the dwelling has, the construction material of the roof and floor, and the primary materials used for lighting and cooking
- Household demographics, such as the number of members, number of children, gender and age of the household head, and education and literacy levels of the household head
- The region of residence of the household

All continuous feature variables are winsorized at the 99th percentile and normalized to a 0-1 range. All categorical variables are one hot encoded.

We train the predictive model on a random 75% of households and produce predictions on the remaining 25%. We repeat all our experiments for poverty targeting over 100 random train-test splits and report the average result. Our predictive model is a LASSO regression with regularization strength determined via three-fold cross validation.

B.4 Simulating test-time data availability

Several of our experiments involve simulating lower test-time data availability – scenarios where for some households in the test set feature data is not available. We simulate this by removing predictions for these households from the test set; households without predictions are ineligible to be targeted regardless of program capacity. In some of our simulations, the share of households with data availability can be lower than the program capacity—in this setting, households without data are targeted at random (along with all households with data available) to fill the program capacity constraint.

B.5 Replication

All code for the poverty targeting simulations is available, including cleaning the survey data to prepare it for predictive modeling. The survey data can be downloaded from the LSMS program at `https://microdata.worldbank.org/index.php/catalog/2783`.

B.6 Additional Analysis

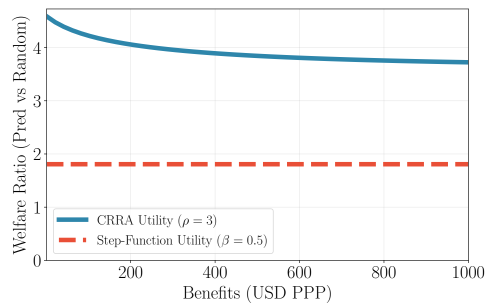


Figure 11: Ratio of welfare impacts under predictive targeting to welfare impacts under random targeting, as a function of benefit size (keeping program capacity fixed at 20%).

C `rvp` Toolkit

Table 1 summarizes the mapping between the formal objects defined in the main text and their software implementations. Full documentation, installation instructions, and worked examples are available here.

Table 1: Mapping from conceptual components to `rvp` classes.

Concept	**Class**	**Important Parameters**
Distribution $\mathcal{D}$	`AllocationData`	`df`, `predictions_col`, `ground_truth_col`
Utility $u(w, a)$	`CRRAUtility` `PartitionedUtility`	`rho`, `b` `thresholds`, `values`
Constraint α	`CoverageConstraint`	`max_coverage`, `population_size`
Policy π	`RankingPolicy`	`ascending`
Problem	`AllocationProblem`	`data`, `utility`, `constraint`, `policy`

Levers ℓ	`PredictionImprovementLever`
	`ExpandCoverageLever`
	`DataLabelingLever`
	`CRRABenefitLever`
		`error_reduction`, `covariate_mask`
		`coverage_increase`, `marginal_cost_per_person`
		`label_share`, `cost_per_label`
		`new_benefit`

Comparison	`optimize_budget_allocation`
	`LeverComparison`
	`plot_welfare_curve`
		(Q1) Budget optimization
		(Q2) Equivalent cost, welfare difference
		(Q3) Welfare curves, heatmaps