THE TRANSFER PERFORMANCE OF ECONOMIC MODELS

ISAIAH ANDREWS[§], DREW FUDENBERG[‡], LIHUA LEI[#], ANNIE LIANG[†], AND CHAOFENG WU[*]

March 4, 2025

Abstract. Economists often use models estimated on data from a particular domain to make predictions in another. We provide a tractable formulation for this "out-of-domain" prediction problem and define the *transfer error* of a model based on its performance in a new domain. We derive finite-sample forecast intervals that are guaranteed to cover realized transfer errors with a user-selected probability when domains are i.i.d. We apply these intervals in an application to compare the transferability of economic models and black box algorithms, finding that black box algorithms outperform economic models for prediction within a domain, but generalize more poorly across domains.

We thank Jiafeng Chen, Stefano DellaVigna, Nic Fishman, Matthew Gentzkow, Andrei Iakovlev, Michael Jordan, Brandon Kaplowitz, Paulo Natenzon, Kyohei Okumura, Pietro Ortoleva, Michael Ostrovsky, Ashesh Rambachan, Bas Sanders, Jesse Shapiro, Eisho Takatsuji, and Davide Viviano for helpful comments. We also thank NSF grants 1851629 and 1951056 for financial support, and the Quest high performance computing facility at Northwestern University for computational resources and staff assistance.

[§]Department of Economics, MIT.
[‡]Department of Economics, MIT.
[#]Stanford Graduate School of Business.
[†]Department of Economics, Northwestern University.
[*]Department of Computer Science, Northwestern University.

1. Introduction

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

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

Our paper provides a conceptual framework that formalizes the "out-of-domain" transfer problem, and allows systematic comparison between economic modeling and machine learning algorithms. Relative to the large body of work on external validity described in Section 1.1, our out-of-domain test considers performance within a class of prediction problems rather than conditioning on a particular realization of the data available to the economist. For example, instead of assessing whether a structural model of information diffusion will transfer well from a specific Indian village to another, we ask whether the structural model transfers well "in general" across Indian villages. Formally, we consider an ex-ante perspective in which the economist is aware of a set of domains that are relevant (such as Indian villages) but does not know which specific transfer problem will realize from this class of possible transfer problems.

Our main theoretical contribution is the construction of finite-sample forecast intervals that characterize how well a model or algorithm performs in new domains.[1] In the main text, we provide intervals that are valid for a benchmark setting where all domains are equally likely to realize for training and testing (corresponding to an assumption that domains are exchangeable) as well as for a setting in which the testing domain is qualitatively different from the training domains. (Appendix R.1 generalizes our approach and results even further.) The definition of "performance" is left to the user, and our results apply for a large class of measures that includes our motivating example of predictive accuracy, as well as others, such as how well a qualitative finding based on parameter estimates generalizes across domains.

Finally, we use our framework to compare the generalizability of economic models and black box machine learning methods when predicting certainty equivalents for lotteries. While economic models of risk preference have been studied extensively from the perspectives of how well they fit economic data (Harless and Camerer, 1994; Hey and Orme, 1994; Bruhin et al., 2010; Bernheim and Sprenger, 2020) and how their predictive performance compares to that of machine learning algorithms (Peysakhovich and Naecker, 2017; Plonsky et al., 2019; Fudenberg et al., 2022), the existing tests have primarily been within the context of a single domain.[2] We offer a new perspective on how well these models transfer across domains, finding that although machine learning algorithms outperform the economic models when trained and tested on (disjoint) data generated under the same conditions, the economic models generalize better. Our analysis suggests the primary reason for this is not that the machine learning algorithms overfit, but rather that economic models of risk preference are more effective at extrapolating from observed cases to new ones.

We now describe the main parts of the paper in more detail. Section 2 describes our conceptual framework, which extends the familiar notion of out-of-sample evaluation to out-of-domain evaluation. In the standard out-of-sample test, a model's free parameters are estimated on a training sample, and the predictions of the estimated model are evaluated on a test sample, where the observations in the training and test samples are disjoint but drawn from the same distribution. We depart from this framework by supposing that the

[1] We use the term "forecast interval," rather than "confidence interval," to reflect the random nature of the target, namely the *realized* (rather than expected, median, etc.) transfer error, but they can also be viewed as confidence intervals for these random targets.

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

distribution of the data varies across a set of "domains," e.g. different subject pools or choice frames.

We adopt the perspective of an external *analyst* who wants to assess the efficacy of a procedure that a *researcher* uses to make predictions in a new domain. For example, the researcher's procedure might be to estimate an economic model on data from one domain and use the estimated model to make predictions in another. Alternatively, the researcher's procedure might involve pooling data across domains for estimation, or instead training and porting a black box algorithm. The procedure's exact performance depends on the domains that are used for estimation—the *training domains*—and the domain that is realized for prediction—the *target domain*. The analyst's goal is to develop a forecast interval for the random performance of this procedure across the many possible realizations of those domains. This forecast interval can then be used to compare procedures for generalization.

Section 3 constructs forecast intervals for the baseline setting where the distributions governing different domains are exchangeable. This means that while there may be ex-post differences between the domains on which the model is estimated and on which it is applied, these differences are not ex-ante known to the analyst. We propose the following protocol: The analyst first collects data across as many domains as possible, henceforth the analyst's *meta-data set*. The analyst then repeatedly splits these into training and test domains, and runs the researcher's procedure for each of these domain realizations. For example, if the researcher's procedure involves transferring an economic model, then the analyst would estimate the researcher's model on the training domains and use it to make predictions in the test domain. Intuitively, the transfer performance between domains in the analyst's meta-data set can serve as a proxy for the transfer performance to the new (unobserved) target domain. Pooling these transfer performances across different choices of training and test domains yields an empirical distribution of transfer errors. We show that the quantiles of this distribution can be used to compute valid finite-sample forecast intervals for the transfer error on the target domain. (Because in most applications, models are estimated on data from a small number of domains, we exclusively consider finite-sample results in this paper.)

Section 4 extends our methods for the scenario where the training domains are known to be systematically different from the target domain. Specifically, we suppose that the training domains are exchangeable, but allow the target domain to be governed by a qualitatively

different distribution. We fully generalize our procedure and results when the analyst either knows or can bound a likelihood ratio relating the target and training distributions. We further develop two partial orders for comparing how well models generalize in environments where the analyst lacks any knowledge at all about the likelihood ratio. Though this ordering is inherently incomplete, we demonstrate that it can be used to compare economic models and black box algorithms in our subsequent application. Each of these procedures (and all other methods described in this paper) are implemented in an `R` package (`transferUQ`), available on Github.[3]

Our statistical framework and results extend the recent literature on conformal inference (e.g. Vovk et al., 2005; Tibshirani et al., 2019; Lei and Candès, 2021) and randomization inference (e.g. Ritzwoller et al., 2024) by replacing the assumption of (weighted) exchangeable observations with that of (weighted) exchangeable domains. Importantly, while conformal inference aims to predict a single data point from other data points, our goal is to generate forecast intervals on a function of multiple data points, involving both observed and unobserved data. Our problem thus involves a multi-dimensional structure that falls outside of the standard framework and requires substantively new arguments.

Section 5 applies our procedures to compare the transferability of economic models and black box algorithms for predicting certainty equivalents for binary lotteries. In this application, we evaluate how well a model (or algorithm) predicts certainty equivalents reported by one subject pool when estimated on data from another. We evaluate two models of risk preferences (expected utility and cumulative prospect theory) and two popular black box machine learning algorithms (random forest and kernel regression). We find that although the black box algorithms outperform the economic models out-of-sample when trained and tested on data from the same domain, the economic models generalize more reliably across domains. Specifically, while the forecast intervals for the black box algorithms and economic models overlap, the forecast intervals for the black box methods are wider, and their upper bounds are substantially higher.

Why do the black boxes perform worse at transfer prediction? A natural explanation, based on intuition from conventional out-of-sample testing, is that black boxes are very flexible and hence learn idiosyncratic details that do not generalize across subject pools. But

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

when we restrict the analysis to a subset of samples involving the same set of lotteries, the economic models and black box algorithms have nearly indistinguishable forecast intervals. Alternatively, when we define the domains so that they involve disjoint lotteries, the improvement of the transfer performance of economic models over black box algorithms is even larger than in our main analysis. This tells us that black box algorithms are not universally worse at transfer, but rather their relative performance depends on specific characteristics of the transfer problem. In particular, black boxes seem to transfer worse when the primary source of variation across samples is a shift in the marginal distribution over features (i.e., which lotteries appear in the sample), rather than a shift in the distribution of outcomes conditional on features (i.e., the distribution of certainty equivalents given fixed lotteries).

1.1. **Related Literature.** Our results apply for a broad class of definitions of transfer error, but our primary motivation is evaluating how well an economic model estimated on data from one domain predicts in another. Hofman et al. (2021) gives an in-depth argument for why this important, calling for more work on the question "how well does a predictive model fit to one data distribution generalize to another?" for social science models. This is exactly what we consider.

Predictive accuracy has long had a central role in experimental economics.[4] While models have often been assessed based on how well they fit data from a given domain, recent papers also examine how well predictions transfer across domains. For example, Külpmann and Kuzmics (2022) estimates various game-theoretic models on 2×2 normal-form games and evaluates their predictive performance on 3×3 normal-form games; Natenzon (2019) estimates discrete choice models on data for four choice menus and evaluates their predictive performance on a fifth menu; and Fudenberg and Karreskog Rehbinder (2024) evaluates the cross-domain predictive accuracy of models of how players learn in the infinitely-repeated prisoner's dilemma, where the domains are various sets of payoff matrices for the stage game. This paper provides a general framework that nests these transfer problems, and develops formal statistical results for assessing transfer performance.

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

Transfer prediction is also an important consideration when comparing economic models with black box algorithms. Several papers have compared the predictive performance of machine learning algorithms with that of more structured economic models in out-of-sample tests (Peysakhovich and Naecker, 2017; Noti et al., 2016; Plonsky et al., 2019; Camerer et al., 2019; Fudenberg and Liang, 2019; Hirasawa et al., 2022; Hsieh et al., 2023), concluding that machine learning algorithms are more predictive. In the application we consider, black box methods turn out to be less effective at transferring predictions across domains. This result joins Gechter et al. (2019)'s finding that structural models deliver better policy recommendations for conditional cash transfer policies in new contexts than black box methods do. We hope that our methods will be used to assess the transferability of economic models in other problems as well, contributing to a more comprehensive picture of how well economic models generalize.

Finally, our theoretical framework and results lie at the intersection of several literatures in economics, computer science, and statistics. These literatures consider several related but distinct objectives: synthesizing evidence across different domains, improving the quality of extrapolation from one domain to another, and quantifying the extent to which insights from one domain generalize to another.

The first objective, synthesizing results across different domains, is a particular focus of the literature on meta-analysis (Card and Krueger, 1995; DellaVigna and Pope, 2019; Bandiera et al., 2021; Imai et al., 2020; Vivalt, 2020). Our goal is instead to assess the cross-domain forecast accuracy of a model. These problems are related, and Meager (2019) and Meager (2022) in particular provide posterior predictive intervals for new domains in the context they consider. Unlike our approach, the predictive intervals reported in those papers are valid only under a parametric model for the distribution of effects across domains and a distributional assumption on the domain-specific effect estimates.

The second objective is to extrapolate results from one domain to another. Within computer science, the literature on domain generalization (Blanchard et al. 2011 and Muandet et al. 2013) develops models that generalize well to new unseen domains (Zhou et al., 2021). Similarly, several papers within economics (e.g., Hotz et al. 2005 and Dehejia et al. 2021) use knowledge about the distribution of covariates to extrapolate out-of-domain. Our focus is not on developing new models or algorithms with good out-of-domain guarantees, but rather

on developing forecast intervals for the out-of-domain performance of models and algorithms that are used in practice.

The third related objective in the literature is to understand the extent to which results obtained in one domain hold more generally, i.e. their external validity. Unlike papers that assess the generalizability of insights from randomized control trials (e.g. Deaton, 2010; Imbens, 2010; DellaVigna and Linos, 2020) or laboratory experiments (e.g. Levitt and List, 2007; Al-Ubaydli and List, 2015), here we consider generalizability across random domains.[5] Section 4 and Appendix R.1 relaxes the exchangeability assumption to allow bounded deviations from exchangeability; our results there connect to the literature on sensitivity analysis (e.g. Rosenbaum, 2005; Aronow and Lee, 2013; Andrews and Oster, 2019; Sahoo et al., 2022).

2. Framework

Consider a fixed procedure for extrapolating predictions across domains, e.g., estimating a structural economic model on data from one domain and using the estimated model to make predictions in another. We adopt the perspective of an external analyst who wants to evaluate the effectiveness of this procedure. The analyst is not focused on extrapolation from one specific domain to another (e.g., from an American dataset to a German dataset), but would rather like to understand whether the procedure generally performs well across a class of transfer tasks (e.g., extrapolating across countries). To this end, the analyst evaluates transfer error from an ex-ante perspective without knowing which domains are used to estimate and evaluate the model. The analyst seeks to construct forecast intervals for the procedure's error when transferring from a (random) set of training domains to a new (random) target domain. This speaks to the question of whether one procedure for extrapolation (such as transferring an estimated economic model) generally performs better than another (such as transferring a trained black box algorithm).

This section proceeds as follows: Section 2.1 describes the extrapolation procedure that the analyst would like to evaluate, and a large class of measures for the procedure's *transfer error*. Section 2.2 formalizes the analyst's problem.

[5]Another set of papers study the generalizability of instrumental variables estimates (e.g. Angrist and Fernández-Val, 2013; Bertanha and Imbens, 2020) and causal effects (e.g. Pearl and Bareinboim, 2014; Park et al., 2023).

2.1. **Transfer errors.** Let $\mathcal{X}$ be a set of covariate vectors and $\mathcal{Y}$ be a set of outcomes. An *observation* is a pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and a *sample* is a set of observations $S = \{(x_i, y_i)\}_{i=1}^m$. We consider samples S_d indexed to *domains* $d = 1, 2, \ldots$, such as in the following examples:

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

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

Example 3 (Different Choice Menus). There is a finite set of goods $a_1, a_2, \ldots, a_m$, and each sample S_d includes observed choices from a different subset of available goods. For example, S_1 might contain all choices from binary menus and S_2 all choices from those menus that include a_1.

For now we take these samples as given; Section 2.2 lays out the underlying statistical model for how these samples are generated, which we will use to prove our results.

A researcher observes samples from some set of training domains $d \in \mathcal{T}$, and uses these samples $S_{\mathcal{T}} \equiv (S_d)_{d \in \mathcal{T}}$ to make predictions in a new (yet unseen) target domain d^*. We will refer to $S_{\mathcal{T}}$ as the *training samples* and to S_{d^*} as the *target sample*.

The number of training domains $r \equiv |\mathcal{T}|$ is a parameter of the research procedure, and should reflect what is done in practice. In economics, it is common to transfer quantitative conclusions from a single domain to another, e.g., for parameter calibration in structural models (Greenwood et al., 1997; McKay et al., 2016; Oswald, 2019) and extrapolation of treatment effect estimates beyond the experimental population (Mogstad and Torgovitsky, 2018; Tipton and Olsen, 2018; Cattaneo et al., 2021; Maeba, 2022). In this case $r = 1$, and the relevant question is whether extrapolating from one sample leads to good predictions in the new domain. If instead data is gathered from $r > 1$ different domains and the observations are aggregated and used to estimate a model (as in the meta-analyses of Meager 2019,

2022), the relevant question may be how well the estimated model on the aggregated data generalizes to a new domain, and $r > 1$ is appropriate.

Our results apply to the following class of measures for transfer performance.

Definition 1. A *transfer error* is any quantity $e_{\mathcal{T},d^*}$ that can be written as a function of the training data $S_{\mathcal{T}}$, the target sample S_{d^*}, and potentially an independent noise variable.

Our leading examples are transfer errors that measure how well a fixed model or algorithm transfers across domains. That is, suppose the training samples $S_{\mathcal{T}}$ are used to select a *prediction rule* $f_{S_{\mathcal{T}}} : \mathcal{X} \to \mathcal{Y}$, e.g., by estimating a parametric model or by fitting a black box algorithm.[6] The accuracy of the prediction rule is evaluated using a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, where

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

denotes the average loss when using f to predict y given x for observations $(x, y) \in S$. One specification of a transfer error is then

$$e_{\mathcal{T},d^*} = e(f_{S_{\mathcal{T}}}, S_{d^*}) \tag{1}$$

i.e., the raw error of the model when it is estimated on the training samples and used to predict outcomes in the target sample.

Example 4 (Transferring Models of Risk Preferences). The covariates $\mathcal{X}$ describe different lotteries, i.e., each covariate vector x includes a description of (say) two possible prizes and their corresponding probabilities. The outcome y is the average willingness-to-pay for this lottery. A firm acquires willingness-to-pay data from consumers in Illinois for a given set of lotteries, and uses this data to estimate a model of risk preferences, e.g., estimating parameter values for an Expected Utility model with CRRA preferences. The firm then uses this estimated model to predict willingness-to-pay from consumers in California for a different set of lotteries. The measure in (1) assesses the accuracy of those predictions.

Any normalization of equation (1) with respect to a function of the target sample is also a transfer error. For example, we might normalize (1) with respect to the in-sample error of

[6]That is, let $\mathcal{S}$ denote the set of all finite sets of finite samples, and let $\mathcal{Y}^{\mathcal{X}}$ be the set of all prediction rules. Then a "model" is a mapping $\rho : \mathcal{S} \to \Delta(\mathcal{Y}^{\mathcal{X}})$ and we write $f_{S_{\mathcal{T}}} = \rho(S_{\mathcal{T}})$ for the realized prediction rule.

the model when trained on the target sample,

$$e_{\mathbf{T},n+1} = \frac{e(f_{S_{\mathbf{T}}}, S_{n+1})}{e(f_{S_{n+1}}, S_{n+1})}. \tag{2}$$

This quantity reveals how much less accurate the model is than if it had been directly trained on the target sample.

Example 5 (The Value of Re-Estimating Diffusion Models). The covariates $\mathcal{X}$ describes the network of relationships across households in a village, and the identity of households which are seeded with information about a microfinance program. The outcome y is the average takeup rate of the program across households. A development economist observes the takeup decisions in a single village in India following an experiment in which certain households are seeded with information about the program. The economist uses this data to estimate a structural model of information diffusion, and then predicts the average takeup rate in a new village using the estimated model. The transfer error in (2) assesses how much less accurate this prediction is compared to if the economist could re-estimate the structural model on data from this new village.

Although we focus on the transfer errors defined in (1) and (2), Definition 1 is substantially broader. Appendix P describes several other specifications of transfer errors, including the stability of parameters and errors in counterfactual predictions.

2.2. **The analyst's problem.** We now consider the perspective of an external analyst, who would like to evaluate the transfer guarantees of the procedure described above. Rather than assessing the transfer error $e_{\mathcal{T},d^*}$ for a specific set of training domains $\mathcal{T}$ and target domain d^*, the analyst considers a random version of this quantity, where the samples used for training and evaluation of the model are not yet known. In Example 4, this corresponds to an analyst who is interested in how well the CRRA model transfers across arbitrary locations, as opposed to from one specific location to another.

Formally, the analyst has access to *metadata* consisting of n samples

$$\mathbf{M} = \{S_1, \ldots, S_d, \ldots S_n\}.$$

We assume that $n > r$; that is, the analyst can collect a larger number of samples than were used by the researcher. The analyst models the researcher's set $\mathcal{T}$ of r training domains

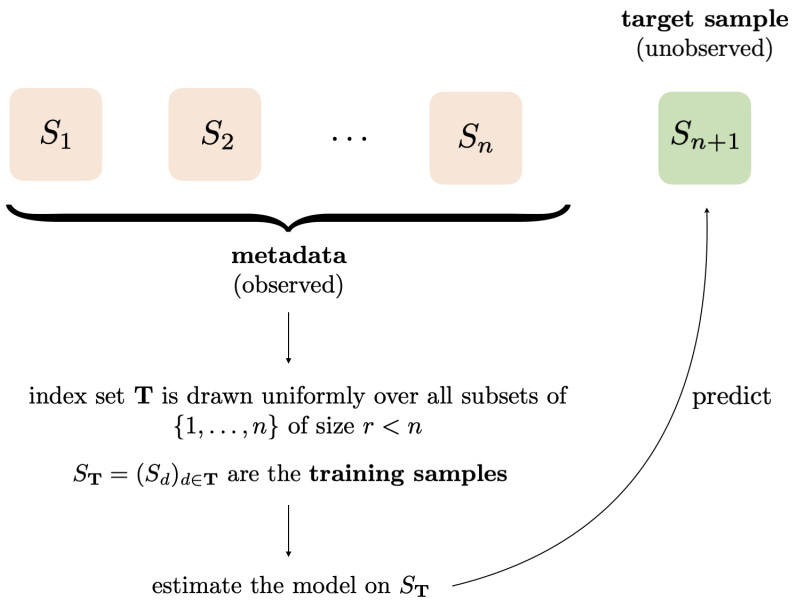


FIGURE 1. *This figure depicts the transfer error $e_{\mathbf{T},n+1}$ as defined according to (1). It measures the prediction error of a model estimated on the samples $S_{\mathbf{T}}$ and evaluated on the sample S_{n+1}, where $S_{\mathbf{T}}$ consists of r random training samples from the metadata, and S_{n+1} is an unobserved sample from a new domain.*

as drawn uniformly over all subsets of $\{1,\ldots,n\}$ of size r. We use $\mathbf{T}$ to mean the random variable whose realization is $\mathcal{T}$, so that $S_{\mathbf{T}} = (S_d)_{d\in\mathbf{T}}$ is the researcher's (random) vector of training samples. The target domain (on which predictions will be made) is a final sample S_{n+1}, which, unlike the metadata, is not observed by the analyst. The quantity of interest is $e_{\mathbf{T},n+1}$, i.e., the random transfer error when the researcher extrapolates predictions from $S_{\mathbf{T}}$ to S_{n+1}. Figure 1 depicts this transfer error for the model transfer specification of (1).

The analyst's goal is to develop forecast intervals for the transfer error $e_{\mathbf{T},n+1}$, i.e., interval-valued functions of the meta-data $\mathbf{M}$ which cover $e_{\mathbf{T},n+1}$ with the prescribed probability, regardless of the distribution μ that governs samples across domains. Section 3 provides these intervals for a baseline setting in which domains are exchangeable, and Section 4 generalizes them to a setting in which the observed domains are systematically different from the target domain.

3. Baseline Setting: Exchangeability

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

Assumption 1. *There is a fixed (but unknown) meta-distribution $\mu \in \Delta(\mathcal{P} \times \mathbb{N})$ over joint distributions $\mathcal{P} \equiv \Delta(\mathcal{X} \times \mathcal{Y})$ and sample sizes $\mathbb{N}$, where each sample S_d is generated by first drawing a distribution and sample size $(P_d, m_d) \sim \mu$, and then independently drawing m_d observations (x, y) from P_d.*[7]

In Examples 1-3, this assumption implies that (from the analyst's perspective) the subject pools, choice frames, or choice menus differentiating the samples are themselves drawn i.i.d. from a fixed distribution. Assumption 1 is standard in conformal inference (Vovk et al., 2005), permutation testing Romano (1990), and randomization inference (Ritzwoller et al., 2024).[8] In contrast, the literature on external validity (see Section 1.1) typically assumes that the distributions governing behavior in different domains are close in some distance metric (Adjaho and Christensen, 2022), share a common support over $\mathcal{X}$ or $\mathcal{Y}$ (Sahoo et al., 2022; Lei et al., 2023), or can be estimated using background covariates (Tipton and Olsen, 2018). Relative to these assumptions, our approach has the advantage of allowing for arbitrary and unknown relationships between the realized distributions governing domains, but it rules out ex-ante predictable patterns in how the joint distribution varies across samples (such as time trends).

Section 3.1 presents our forecast intervals for this setting, and Section 3.2 proves the validity and tightness of these intervals.

3.1. Baseline procedure.

Thanks to Assumption 1, the observed samples in the metadata, $\{S_1, \ldots, S_n\}$, can act as surrogates for the unseen target sample, S_{n+1}. As before, let $e^{\mathbf{M}}_{\mathcal{T},d}$ denote the (observed) transfer error from any selection of training samples $\mathcal{T} \subseteq \{1, \ldots, n\}$

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

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

to any surrogate target sample $d \in \{1, \ldots, n\} \backslash \mathcal{T}$ from the metadata (where we now make the dependence of this quantity on $\mathbf{M}$ explicit). We use $\mathbb{T}_{r+1,n}$ to denote the set of $\frac{n!}{(n-r-1)!}$ unique pairs $(\mathcal{T}, d)$ that can be constructed in this way. Then

$$F_{\mathbf{M}} = \frac{(n-r-1)!}{n!} \sum_{(\mathcal{T},d) \in \mathbb{T}_{r+1,n}} \delta_{e^{\mathbf{M}}_{\mathcal{T},d}} \tag{3}$$

is the empirical distribution of transfer errors in the pooled sample $\left\{e^{\mathbf{M}}_{\mathcal{T},d} : (\mathcal{T}, d) \in \mathbb{T}_{r+1,n}\right\}$ as we vary which samples in the metadata are used for training and testing. (Throughout δ denotes the Dirac measure). In the case where $r = 1$, so that a single sample is used for training, the observed transfer errors can be represented as a matrix as depicted in Figure 2, and $F_{\mathbf{M}}$ is their empirical distribution.

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

FIGURE 2. *$e_{d,d'}$ is the transfer error from sample S_d to $S_{d'}$.*

Definition 2 (Upper and Lower Quantiles). For any distribution P let $\overline{Q}_\tau(P) = \inf\{b : P((-\infty, b]) \geq \tau\}$ and $\underline{Q}_\tau(P) = \sup\{b : P([b, \infty)) \geq 1 - \tau\}$ denote the upper and lower τth quantiles, respectively.

These quantiles coincide for continuously distributed variables with connected support.

Definition 3 (Quantiles of $F_{\mathbf{M}}$). For any $\tau \in (0, 1)$, let $\overline{e}^{\mathbf{M}}_\tau \equiv \overline{Q}_\tau(F_{\mathbf{M}})$ and $\underline{e}^{\mathbf{M}}_\tau \equiv \underline{Q}_{1-\tau}(F_{\mathbf{M}})$ be the τth upper quantile and $(1-\tau)$th lower quantile of the empirical distribution of transfer errors in the pooled sample.

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

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

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

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{e}_\tau^{\mathbf{M}}\right) \geq \tau \left(\frac{n-r}{n+1}\right), \tag{4}$$

and

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

Thus $\left(-\infty, \bar{e}_\tau^{\mathbf{M}}\right]$ is a level-$\left(\frac{\tau(n-r)}{n+1}\right)$ one-sided forecast interval for $e_{\mathbf{T},n+1}$, and $\left[\underline{e}_\tau^{\mathbf{M}}, \bar{e}_\tau^{\mathbf{M}}\right]$ is a level-$\left((2\tau-1)\left(\frac{n-r}{n+1}\right)\right)$ forecast interval for $e_{\mathbf{T},n+1}$.

Parameter τ influences the width of the forecast interval, where larger choices of τ lead to wider forecast intervals with higher confidence guarantees. Parameter r determines how many samples in the meta-data are used for training versus testing. As discussed in Section 2.1, r is determined by the the research procedure under evaluation.[9]

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

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

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

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

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

and

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

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

$$\left\{e_{\mathcal{T},d}^{\mathbf{M}} : (\mathcal{T}, d) \in \mathbb{T}_{r+1,n+1}\right\}. \tag{5}$$

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

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq e_{\tau}^{*} \mid \{S_d\}_{d=1}^{n+1}\right) \geq \tau. \tag{6}$$

In the case $r = 1$ where precisely one sample is used for training, the set of pooled errors (5) is the shaded cells in Figure 3 (either yellow or blue), and the inequality in (6) says that the probability that the value of a randomly drawn cell falls below the τth upper quantile of cells is at least τ.

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

FIGURE 3. *Transfer errors when training on one domain (row) and testing on another (column).*

The analyst does not observe the target sample S_{n+1}, and so does not know e_{τ}^{*}. We instead use $\bar{e}_{\tau}^{\mathbf{M}}$, the τth upper quantile of the pooled sample of errors when transferring across samples in $\mathbf{M}$, to construct the forecast intervals. In Figure 3, the probability that $e_{\mathbf{T},n+1} \leq \bar{e}_{\tau}^{\mathbf{M}}$ is the probability that the value of a randomly drawn shaded cell (yellow or blue) falls below the τth quantile of the yellow cells. By a straightforward counting argument,

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

Applying the law of iterated expectations (with respect to the sample $\{S_i\}_{i=1}^{n+1}$) yields the one-sided forecast interval in (4). The proof for the two-sided forecast interval follows a similar logic but is more involved, see Appendix A.3 for details.

4. Beyond Exchangeability

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

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

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

The following subsections consider successively weaker assumptions about what the analyst knows about the likelihood ratio ω. Sections 4.1 and 4.2 respectively generalize our previous results when the analyst either knows ω or can bound it. Section 4.3 provides two ways of comparing the generalizability of models in environments where the analyst knows nothing about ω.

4.1. The analyst knows the likelihood ratio ω.

We again construct a forecast interval for $e_{\mathbf{T},n+1}$ using the pooled sample of transfer errors across samples in the metadata, that is, $\left\{e^{\mathbf{M}}_{\mathcal{T},d} : (\mathcal{T}, d) \in \mathbb{T}_{r+1,n}\right\}$. Different from the previous section, we no longer assign

uniform weights to each $e^{\mathbf{M}}_{T,d}$. Intuitively, under our previous i.i.d. assumption, each sample in the metadata was equally representative of the training and target distributions, but in this relaxed model, whether a sample S_d is more representative of the training or testing distribution depends on its relative likelihood under ν and μ.

A crucial quantity is the following:

Definition 4. For every domain $d \in \{1, \ldots, n\}$, define

$$W_d = \frac{(n-r-1)!}{(n-1)!} \frac{\omega(S_d)}{\sum_{j=1}^{n} \omega(S_j)}. \tag{7}$$

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

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

denote the permutations that specify $(t_1, \ldots, t_r)$ for training and d as the target. Then conditional on a realization of the metadata $(S_1, \ldots, S_n)$, the probability that $(S_{t_i})_{i=1}^{r}$ are the training samples and S_d is the test sample is[10]

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

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

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

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

be the weighted empirical distribution of transfer errors, where each sample d is weighted according to W_d. When the two meta-distributions μ and ν are identical as in our main model, then $W_d \equiv (n-r-1)!/n!$ for every domain d, so the distribution $F^{\omega}_{\mathbf{M}}$ is simply $F_{\mathbf{M}}$ as defined in (3).

Definition 5 (Quantiles of $F^{\omega}_{\mathbf{M}}$). For any likelihood ratio $\omega(\cdot)$ and quantile $\tau \in (0,1)$, define $\bar{e}^{\mathbf{M},\omega}_{\tau} = \overline{Q}_{\tau}\left(F^{\omega}_{\mathbf{M}}\right)$ and $\underline{e}^{\mathbf{M},\omega}_{\tau} = \underline{Q}_{1-\tau}\left(F^{\omega}_{\mathbf{M}}\right)$ to be, respectively, the τth upper quantile and $(1-\tau)$th lower quantile of the weighted distribution of transfer errors in the pooled sample.

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

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

and

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

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

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

and

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

This result strictly generalizes Proposition 1 and Claim 1, since when $w(\cdot)$ is the identity then $\bar{e}^{\mathbf{M},\omega}_{\tau} = \bar{e}^{\mathbf{M}}_{\tau}$ and $\underline{e}^{\mathbf{M},\omega}_{\tau} = \underline{e}^{\mathbf{M}}_{\tau}$, and the bounds in this theorem reduce to those given in Proposition 1.

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

Definition 6 (Bounded Likelihood-Ratios). For any $\Gamma \geq 1$, let $\mathcal{W}(\Gamma)$ be the class of density ratios that satisfy $\omega(S) \in [\Gamma^{-1}, \Gamma]$ for all samples S.

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

$$\overline{e}_\tau^{\mathbf{M}}(\Gamma) = \sup_{\omega \in \mathcal{W}(\Gamma)} \overline{e}_\tau^{\mathbf{M},\omega}, \quad \underline{e}_\tau^{\mathbf{M}}(\Gamma) = \inf_{\omega \in \mathcal{W}(\Gamma)} \overline{e}_\tau^{\mathbf{M},\omega} \tag{8}$$

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

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

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

and

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

4.3. **The analyst knows nothing about** ω**.** Finally, we provide two ways for comparing the transferability of two models $i = 1, 2$ when the analyst cannot bound ω. We do not provide formal results about these orders, but show that they have bite in our subsequent application (see Section 5).

Let $\overline{e}_{i,\tau}^{\mathbf{M}}(\Gamma)$ and $\underline{e}_{i,\tau}^{\mathbf{M}}(\Gamma)$ again denote the worst case bounds for model i, as defined in (8).

Definition 7 (Worst-Case Dominance). Say that model 1 worst-case-upper-dominates model 2 at the τ-th quantile if

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

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

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

Definition 8 (Everywhere Dominance). Say that model 1 everywhere-upper-dominates model 2 at the τ-th quantile if

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

Many decision rules will not be comparable under either of these definitions, but we show they are empirically relevant in our application. The even stronger requirement that

$\overline{e}^{\mathbf{M},\omega}_{1,\tau} \leq \underline{e}^{\mathbf{M},\omega}_{2,\tau}$, i.e., that the upper bound of model 1's forecast interval is smaller than the lower bound of model 2's forecast interval, is likely too stringent to be useful in practice.[11]

5. Application

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

Section 5.1 describes our metadata, and Section 5.2 describes the decision rules we consider. Section 5.3 conducts "within-domain" out-of-sample tests, where the training and test data are drawn from the same domain. Section 5.4 compares transfer performance across domains by constructing forecast intervals for three different definitions of transfer error.

5.1. **Data.** Our metadata consists of samples of certainty equivalents from 44 subject pools, which we treat as the domains. These data are drawn from 14 papers in experimental economics, with twelve papers contributing one sample each, one paper contributing two, and a final paper (a study of risk preferences across countries) contributing 30 samples. Our samples range in size from 72 observations to 8906 observations, with an average of 2752.7 observations per sample.[12] Besides the difference in subject pools, these samples may differ in other details, such as whether the lotteries were restricted to the gain domain. We convert all prizes to dollars using purchasing power parity exchange rates (from OECD 2023) in the year of the paper's publication

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

[12]Online Appendix S.1 describes our data sources in more detail.

Within each sample, observations take the form $(z_1, z_2, p; y)$, where z_1 and z_2 denote the possible prizes of the lottery (and we adopt the convention that $|z_1| > |z_2|$), p is the probability of z_1, and y is the reported certainty equivalent by a given subject. Thus our feature space is $\mathcal{X} = \mathbb{R} \times \mathbb{R} \times [0, 1]$, the outcome space is $\mathcal{Y} = \mathbb{R}$, and a prediction rule is any mapping from binary lotteries into predictions of the reported certainty equivalent. We use squared-error loss $\ell(y, y') = (y - y')^2$ to evaluate the error of the prediction, but for ease of interpretation we report results in terms of root-mean-squared error, which puts the errors in the same units as the prizes.[13] Since different subjects report different certainty equivalents for the same lottery, the best achievable error is generally bounded away from zero.

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

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

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

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

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

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

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

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

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

where

$$v(z) = \begin{cases} z^{\alpha} & \text{if } z \geq 0 \\ -(-z)^{\beta} & \text{if } z < 0 \end{cases} \tag{9}$$

is a value function for money, and

$$w(p) = \frac{\delta p^{\gamma}}{\delta p^{\gamma} + (1-p)^{\gamma}} \tag{10}$$

is a probability weighting function.

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

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

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

We also evaluate restricted specifications of CPT that have appeared elsewhere in the literature: CPT with free parameters α and β (setting $\delta = \gamma = 1$) describes an expected utility decision-maker whose utility function is as given in (9); CPT with free parameters α, β and γ (setting $\delta = 1$) is the specification used in Karmarkar (1978); and CPT with free parameters δ and γ (setting $\alpha = \beta = 1$) describes a risk-neutral CPT agent whose utility function over money is $u(z) = z$ but who exhibits nonlinear probability weighting. Additionally, we include CPT with the single free parameter γ (setting $\alpha = \beta = \delta = 1$), which Fudenberg et al. (2023) found to be an especially effective one-parameter specification.

Black Box Algorithms. We consider two popular machine learning algorithms. First, we train a *random forest* (RF), which is an ensemble learning method consisting of a collection of decision trees.[14] Second, we train a *kernelized ridge regression* model (KR), which modifies OLS to weight observations at nearby covariate vectors more heavily, and additionally places a penalty term on the size of the coefficients. Specifically, we use the radial basis function kernel $\kappa(x, \tilde{x}) = e^{-\gamma\|x-\tilde{x}\|_2^2}$ to assess the similarity between covariate vectors x and $\tilde{x}$. Given

[14]A decision tree recursively partitions the input space, and learns a constant prediction for each partition element. The random forest algorithm collects the output of the individual decision trees, and returns their average as the prediction. Each decision tree is trained with a sample (of equal size to our training data) drawn with replacement from the actual training data. At each decision node, the tree splits the training samples into two groups using a True/False question about the value of some feature, where the split is chosen to greedily minimize mean squared error.

training data $\{(x_i, y_i)\}_{i=1}^N$, the estimated weight vector is $\vec{w} = (\mathbb{K} + \lambda I_N)^{-1}\vec{y}$, where $\mathbb{K}$ is the $N \times N$ matrix whose (i,j)-th entry is $\kappa(x_i, x_j)$, I_N is the $N \times N$ identity matrix, and $\vec{y} = (y_1, \ldots, y_N)'$ is the vector of observed outcomes in the training data. The estimated prediction rule is $\sigma(x) = \sum_{i=1}^N w_i \kappa(x, x_i)$.

There are at least two approaches for cross-validating hyper-parameters such as the size of the trees in the random forest algorithm. First, when there are multiple training domains one can cross-validate across them; we use this in Appendix S.6. Second, one can cross-validate across observations within the training domains. Since we are interested in cross-domain performance, rather than within-domain performance, it is not guaranteed that this will improve performance, and indeed we find that choosing the hyper-parameters via within-domain cross-validation leads to worse transfer performance than using default values. Thus in our main analysis with a single training domain, we set all hyper-parameters to default values.[15]

Discussion. There is no established definition of what constitutes an economic model versus a black box algorithm, but one way of distinguishing between the two approaches is whether the prediction method is tailored to a general application or a general-purpose method of prediction. EU and CPT model the risk preferences of economic agents; we would not expect these models to predict well if we changed our problem to image classification. In contrast, random forest algorithms and kernel regression have been successfully applied across a wide array of prediction problems. In this sense, EU and CPT are economic models, while RF and KR are not. Our approach and results can, however, equally be applied to evaluate prediction methods that are a hybrid of the two approaches. For example, Plonsky et al. (2019) and Hsieh et al. (2023) consider black box algorithms whose inputs are based on prior economic theory. We leave investigation of the transfer performance of such methods to future work.

We note finally that although black box algorithms are traditionally perceived as more flexible than economic models, whether this is in fact the case is something that has to be determined case-by-case. In particular, Fudenberg et al. (2023) shows that although CPT

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

uses only four parameters, it imposes very few restrictions on mappings from binary lotteries to certainty equivalents.

5.3. **Within-domain performance.** We first evaluate how these models perform when trained and evaluated on data from the same subject pool. We compute the tenfold cross-validated out-of-sample error for each decision rule in each of the 44 domains.[16] The two black box methods (random forest and kernel regression) each achieve lower cross-validated error than EU and CPT in 38 of the 44 domains, although the improvement is not large. To obtain a simple summary statistic for the comparison between the economic models and black boxes, we normalize each economic model's error (in each domain) by the random forest error. Table 1 averages this ratio across domains and shows that on average, the cross-validated errors of the economic models are slightly larger than the random forest error. That is, the CPT error is on average 1.06 times the random forest error, and the EU error is on average 1.21 times the random forest error.[17]

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

TABLE 1. *Average ratio of out-of-sample errors relative to random forest.*

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

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

[17]The numbers in Table 1 are very similar if we normalized by the kernel regression error instead.

boxes are confined to the domains on which they were trained. We next separate these explanations by evaluating the transfer performance of the models.

5.4. **Transfer error.** We use the results in Section 3.2 to construct forecast intervals for the two specifications of transfer error defined in (1) and(2), which we will subsequently call *raw transfer error* and *transfer shortfall* respectively. We also consider another normalization of the raw transfer error with respect to a proxy for the best achievable error on the target sample. Let $m \in \mathcal{M}$ index a set of models that each prescribe rules f^m for mapping data to prediction rules. Then *transfer shortfall*

$$\frac{e(f_{S_{\mathbf{T}}}, S_{n+1})}{\min_{m \in \mathcal{M}} e\left(f^m_{S_{n+1}}, S_{n+1}\right)} \tag{11}$$

reveals how much lower the accuracy of the transferred model $f_{S_{\mathbf{T}}}$ is compared to the best in-sample accuracy using a model from $\mathcal{M}$.[18] One advantage of this specification relative to (1) is that the raw error is very sensitive to the predictability of y given x in the target sample, which may differ across domains but is not directly related to the model's transferability.

In our meta-data there are $n = 44$ domains, and we choose $r = 1$ of these to use as the training domain which corresponds to the question, "If the researcher draws one domain at random, and then tries to generalize to another domain, how well will they do?" Figure 4 displays two-sided forecast intervals for transfer performance, transfer deterioration, and transfer shortfall (where R includes all decision rules shown in the figure). These forecast intervals use $\tau = 0.95$, so the upper bound of the forecast interval is the 95th percentile of the pooled transfer errors (across choices of the training and test domains), and the lower bound of the forecast interval is the 5th percentile of the pooled transfer errors. (See Table 5 in Appendix S.3 for the exact numbers.) Applying Proposition 1, these are 86% forecast intervals. Choosing larger τ results in wider forecast intervals that have higher coverage levels, and we report some of these alternative forecast intervals in Online Appendix S.4, including a 96% forecast interval.

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

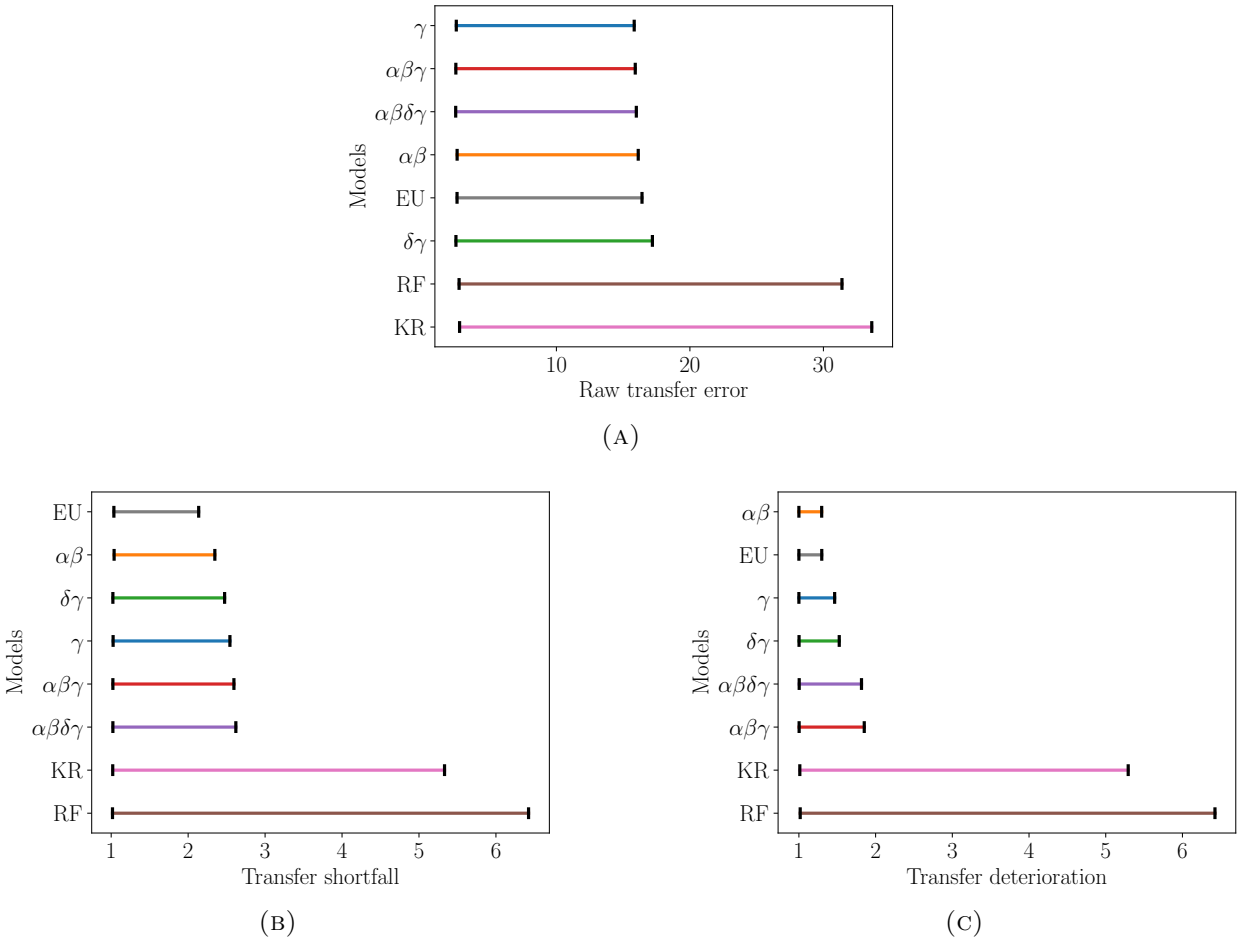


FIGURE 4. *86% (n=44, $\tau = 0.95$) forecast intervals for (a) raw transfer error, (b) transfer shortfall (with $\mathcal{R}$ consisting of the decision rules shown in the figure), and (c) transfer deterioration.*

Our main takeaway from Figure 4 is that although the prediction methods we consider are very similar from the perspective of within-domain prediction, they have very different out-of-domain implications. Panel (a) of Figure 4 shows that the black box forecast intervals for raw transfer error have upper bounds that are roughly twice those of the economic models. Panel (b) shows that the contrast between the economic models and the black boxes is even larger for transfer shortfall, which removes the common variation across models that emerges from variation in the predictability of the different target samples. Thus, although the economic models and the black box models select prediction rules that are close for the purposes of prediction in the training domain, they sometimes have very different performances in the test domain, and the prediction rules selected by the economic models generalize substantially

better. Panel (c) of Figure 4, which reports transfer deterioration, shows that it is less important to re-estimate the economic models on new target domains than to retrain the black-box algorithms.

All of the forecast intervals overlap for each of the three measures. This is not surprising, as variation in the transfer errors due to the random selection of training and target domains cannot be eliminated even with data from many domains. We expect the black box intervals and the economic model intervals to overlap so long as the economic model errors on "upper tail" training and target domain pairs exceed the black box errors on "lower tail" training and target domain pairs. Section Q provides confidence intervals for different population quantities, including quantiles of the transfer error distribution and the expected transfer error, whose width we do expect to vanish as the number of domains grow large. There, we find similar conclusions with regards to the relative performance of the black box algorithms and economic models.

The appendix provides several robustness checks and complementary analyses. Online Appendix S.4 plots the τ-th percentile of pooled transfer errors as τ varies, demonstrating that forecast intervals constructed using other choices of τ (besides $\tau = 0.95$) would look similar to those shown in the main text. Online Appendix S.5 provides 86% forecast intervals for the ratio of the raw CPT transfer error to the raw random forest transfer error, and finds that the random forest error is sometimes much higher than the CPT error, but is rarely much lower. Online Appendix S.6 considers an alternative choice for the number of training domains, setting $r = 3$ instead of $r = 1$. While the results are similar, the contrast between the economic models and black boxes is not as large, suggesting that the relative performance of the black boxes improves given a larger number of training domains. Online Appendix S.2 provides forecast intervals when each of the 14 papers is treated as a different domain; once again the black box methods transfer worse than the economic models do.

We next use our theoretical results from Section 4 to study the consequences of relaxing the i.i.d. assumption in our comparison of CPT$(\alpha, \beta, \delta, \gamma)$ and RF. Since the main differences observed above concerned the upper bounds of our forecast intervals, we limit attention to $\tau \geq 0.5$, and compare the methods in terms of worst-case and everywhere upper-dominance with respect to all three measures of transfer performance. These results are summarized in Table 2.

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

TABLE 2. *Comparison between CPT and RF in terms of worst-case and everywhere upper-dominance. Each cell gives the range of τ at which CPT dominates RF.*

Table 2 shows that CPT worst-case-upper-dominates RF at all quantiles $\tau \geq 0.5$ and for all three transfer error measures. Hence, our finding that the upper tail of transfer errors is larger for RF than for CPT is robust to relaxing the assumption that the training and test domains are drawn from the same distribution, provided that we are comfortable comparing the upper bound for one method to the upper bound for the other. In Appendix S.8, we provide a more detailed view of worst-case-upper-dominance by plotting $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$ as functions of τ and Γ, respectively.

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

5.5. **Do black boxes transfer poorly because they are too flexible?** One tempting explanation of why the black box algorithms transfer less well is that they may overfit to idiosyncratic details of the training samples that do not generalize across subject pools. For example, suppose some subject pools tend to value lotteries depending on the specific digits they contain.[19] This regularity could not be captured by the economic models, because they do not include parameters for individual digits, but could be learned by a random forest algorithm. If so, the random forest would have better within-domain prediction for those subject pools, but worse transfer performance (assuming this regularity does not generalize across subject pools).

While the flexibility of black box algorithms is likely an important determinant of their transfer performance, a second analysis shows that this cannot be a complete explanation

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

of our result. One of the papers we use is based on samples of certainty equivalents from 30 countries (l'Haridon and Vieider, 2019). Of the 30 samples from this paper, 29 samples report certainty equivalents for the same 28 lotteries, and the remaining sample reports certainty equivalents for 24 of those lotteries. We repeat our analysis using these 30 samples as the metadata, and find that the forecast intervals for raw transfer error are indistinguishable across the prediction methods (Panel (a) of Figure 5). There is some separation between the forecast intervals for the remaining two measures, but in both cases the CPT and random forest forecast intervals are substantially more similar than in the original data. If overfitting were the main explanation of our previous results, we would expect the black box algorithms to overfit here as well.

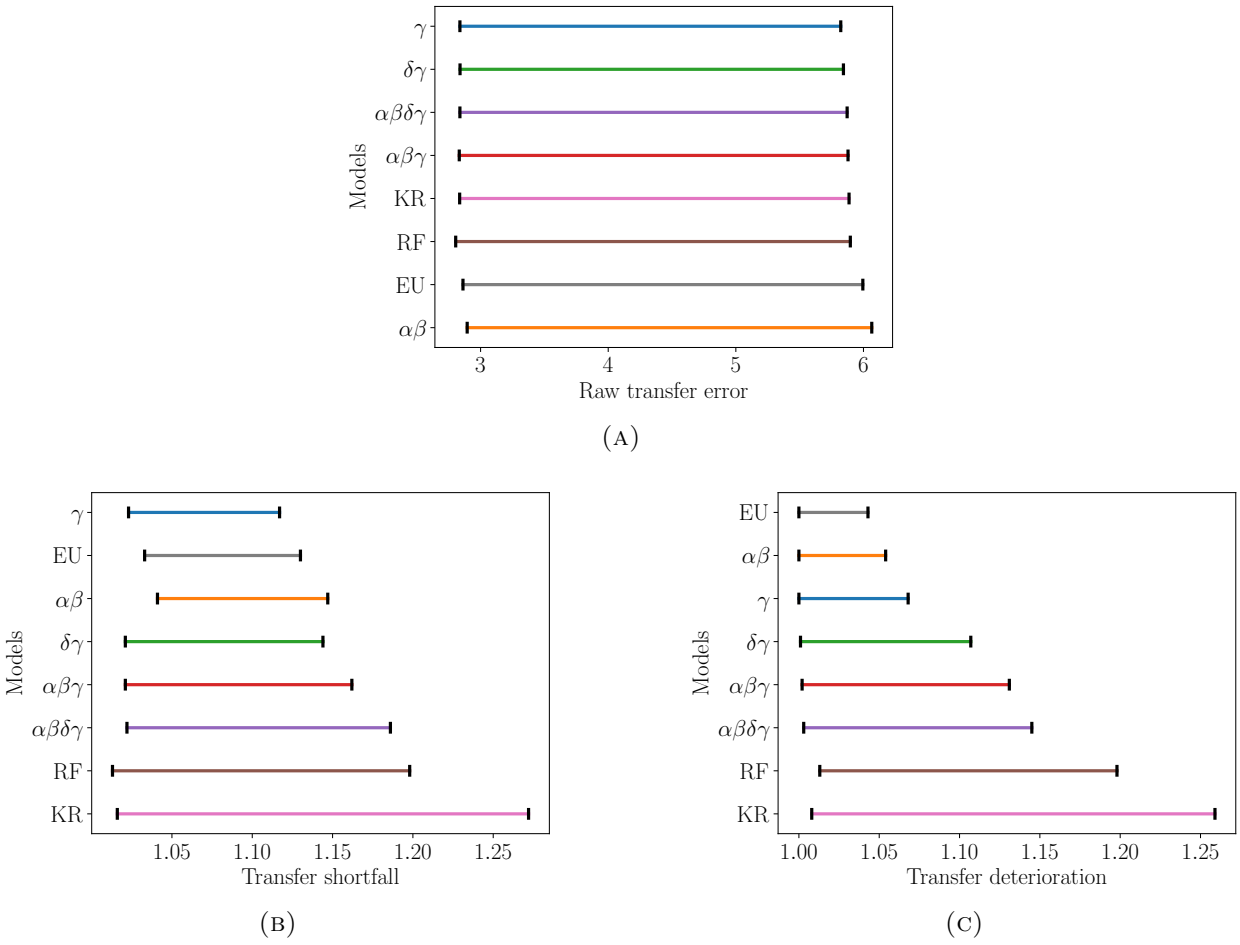


FIGURE 5. *This figure reports 84% (n=30, τ=0.95) forecast intervals when we transfer across subject pools in the l'Haridon and Vieider (2019) data.*

In contrast, we find that the economic models again outperform the black box algorithms when we consider a different definition of domains for the l'Haridon and Vieider (2019) data. Specifically, we aggregate all observations for the 24 lotteries that are shared in all 30 samples, and split these observations into 24 samples, where each sample includes all reported certainty equivalents (across subject pools) for a given lottery. For this new definition of domains, our transfer measures evaluate how well a model estimated on data from certain lotteries predict certainty equivalents for other lotteries. Figure 6 reports 83-level confidence intervals, and we find that the economic models transfer substantially better than the black box algorithms. In fact, isolating the difference across domains to be differences across lotteries exaggerates the relative value of economic models even relative to our original Figure 4 (which uses a definition of domains that combines several sources of variation). For consistency, Figure 6 reports confidence intervals for $r = 1$ (corresponding to training on one lottery and predicting on another), but we show in Appendix S.7 that the qualitative features of this figure hold also for $r = 3$ and $r = 5$.

Taken together, our empirical results suggest that the crucial difference between economic models and black box algorithms isn't that one is more flexible and hence more inclined to overfit, but rather that economic models perform better in certain kinds of transfer tasks.[20] The next section discusses more formally one potential explanation for the difference in the relative performance of economic models and black box algorithms in these two transfer tasks.

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

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

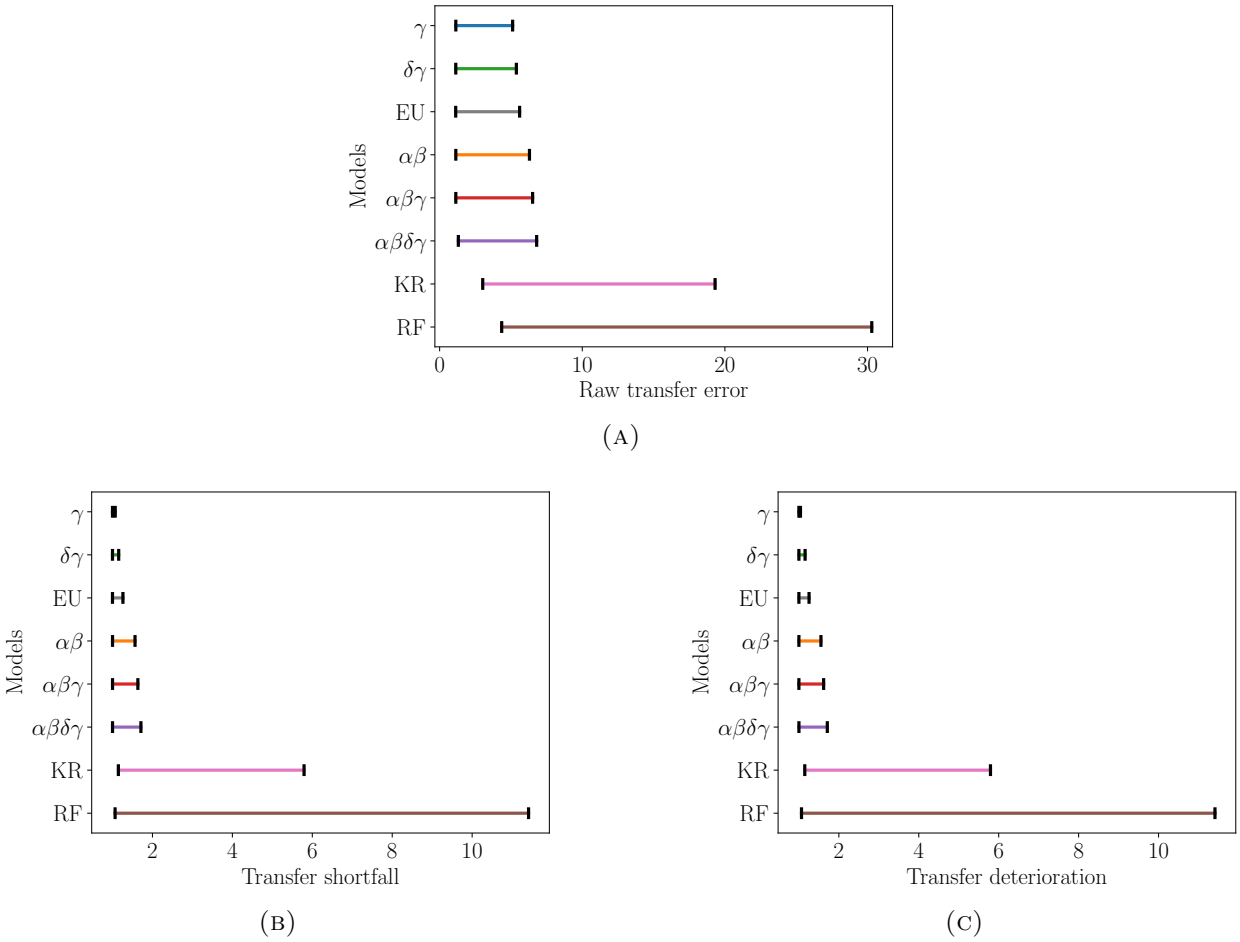


FIGURE 6. *This figure reports 83% (n=24, τ=0.95) forecast intervals when we transfer across lotteries in the l'Haridon and Vieider (2019) data.*

Our findings in Figure 5 suggest that black boxes do as well as economic models at transfer prediction when the marginal distribution over features P_X is held constant across samples. Intuitively, when the relevant feature vectors are the same in every sample, a black box algorithm can perform well by simply memorizing a prediction for each of these feature vectors. In contrast, when the set of lotteries varies across samples, then good transfer prediction necessarily involves extrapolation, and an algorithm that hasn't identified the right structure for relating behavior across lotteries will fail to generalize. Since economic models of risk preferences are intended to relate an individual's preferences across lotteries that permits extrapolation of this form, our empirical results suggest that they do so effectively.

For a simple, stylized, example of this contrast, consider three domains with degenerate distributions over observations. In domain 1, the distribution is degenerate at the lottery

$(z_1, z_2, p) = (10, 0, 1/2)$ and certainty equivalent $y = 3$. In domain 2, the distribution is degenerate at the lottery $(z_1, z_2, p) = (10, 0, 1/2)$ and certainty equivalent $y = 4$. In domain 3, the distribution is degenerate at a new lottery $(z_1, z_2, p) = (20, 10, 1/10)$ and certainty equivalent $y = 11$. Suppose EU and a decision tree are both trained on a sample from domain 1. The CRRA parameter $\eta \approx 0.64$ perfectly fits the observation $(10, 0, 1/2; 3)$, as does the trivial decision tree that predicts $y = 3$ for all lotteries. The estimated EU model and decision tree are equivalent for predicting observations in domain 2: both predict $y = 3$ and achieve a mean-squared error of 1. But their errors are very different on domain 3: the EU prediction for the new lottery is approximately 10.8 with a mean-squared error of approximately 0.05, while the decision tree's prediction is 3 with a mean-squared error of 64.

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

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

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

sample $\mathbf{M}_{\mathcal{T}}$ and the test sample S_d. We evaluate two prediction methods: OLS and a random forest algorithm. Table 3 reports tenfold cross-validated errors for each of these feature sets and prediction methods. As a benchmark, we also consider the best possible constant prediction.

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

TABLE 3. *Cross-Validated MSE*

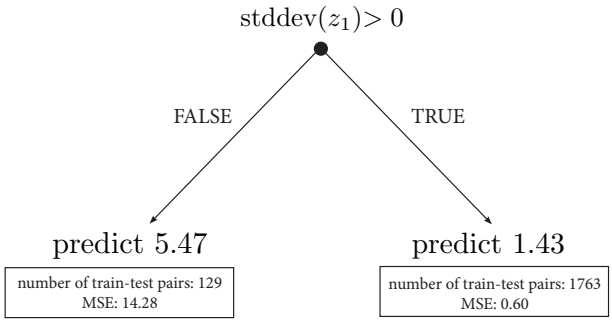


FIGURE 7. *Best 1-split decision tree based on training and test features.*

The best constant prediction achieves a mean-squared error of 2.57, which can be more than halved using features of the training set alone. Using features of both the training and test sets, the random forest algorithm reduces error to 30% of the constant model. Crucially, the random forest algorithm is permitted to learn nonlinear combinations of the input features, and thus discover relationships between the training and test lotteries that are relevant to the relative performance of the black box and the economic model.

The random forest algorithm is too opaque to deliver insight into how it achieves these better predictions, but we can obtain some understanding by examining the best 1-split decision tree, shown in Figure 7 below. This decision tree achieves a cross-validated MSE of 1.75, reducing the error of the constant model by 32%. It partitions the set of (train,test) domain pairs into two groups depending on whether the standard deviation of z_1 (the larger

prize) in the training set of lotteries exceeds zero. There are three domains in which the prizes (z_1, z_2) are held constant across all training lotteries (although the probabilities vary). In the 129 transfer prediction tasks where one of these three domains is used for training, the decision tree predicts the ratio of the random forest error to the CPT error to be 5.47. For all other transfer prediction tasks, the decision tree predicts 1.43.

This finding reinforces our intuition that the relative performance of the black boxes and economic models is driven in part by whether the training sample covers the relevant part of the feature space. When the training observations concentrate on an unrepresentative part of the feature space (such as all lotteries that share a common pair of prize outcomes), then the black boxes transfer much more poorly than economic models.

Our results also clarify a contrast between transfer performance and classical out-of-sample performance. In out-of-sample testing, the marginal distribution on $\mathcal{X}$ is the same for the training and test samples, so the set of training lotteries is likely to be representative of the set of test lotteries as long as the training sample is sufficiently large. When test and training samples are governed by distributions with different marginals on $\mathcal{X}$, the set of training lotteries can be unrepresentative of the set of test lotteries regardless of the number of training observations. Training on observations pooled across many domains alleviates the potential unrepresentativeness of the training data, but the number of domains needed will depend on properties of the distribution: An environment where each domain puts weight on exactly one lottery that is itself sampled i.i.d. may be difficult for black-box algorithms,[21] while an environment where the marginal distribution is degenerate on the same lottery in all domains may be easier. There is no analog in out-of-sample testing for the role played by variation in the marginal distribution on $\mathcal{X}$ across domains. Moving beyond our specific application, we expect this variation to be an important determinant of the relative transfer performance of black box algorithms and economic models in general.

6. Conclusion

Our measures of transfer error quantify how well a model's performance on one domain extrapolates to other domains. We applied these measures to show that the predictions

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

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

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

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

References

Abdellaoui, M., P. Klibanoff, and L. Placido (2015): "Experiments on compound risk in relation to simple risk and to ambiguity," *Management Science*, 61, 1306–1322.

Adjaho, C. and T. Christensen (2022): "Externally valid treatment choice," *arXiv preprint arXiv:2205.05561*, 1.

Al-Ubaydli, O. and J. A. List (2015): "On the Generalizability of Experimental Results in Economics," in *Handbook of Experimental Economic Methodology*, Oxford University Press.

Alfaro-Urena, A., B. Faber, C. Gaubert, I. Manelici, and J. P. Vasquez (2023): "Responsible Sourcing? Theory and Evidence from Costa Rica," Working Paper.

Allcott, H. (2015): "Site Selection Bias in Program Evaluation," *Quarterly Journal of Economics*, 130, 1117–1165.

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

Andrews, I. and E. Oster (2019): "A simple approximation for evaluating external validity bias," *Economics Letters*, 178, 58–62.

Angrist, J. D. and I. Fernández-Val (2013): *ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework*, Cambridge University Press, vol. 3 of *Econometric Society Monographs*, 401–434.

Aronow, P. M. and D. K. Lee (2013): "Interval estimation of population means under unknown but bounded probabilities of sample selection," *Biometrika*, 100, 235–240.

Athey, S. (2017): "Beyond prediction: Using big data for policy problems," *Science*, 355, 483–485.

Bandiera, O., G. Fischer, A. Prat, and E. Ytsma (2021): "Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments," *American Economic Review: Insights*, 3, 435–54.

Banerjee, A., A. G. Chandrasekhar, S. Dalpath, E. Duflo, J. Floretta, M. O. Jackson, H. Kannan, F. Loza, A. Sankar, A. Schrimpf, and M. Shrestha (2024): "Selecting the Most Effective Nudge: Evidence from a Large-Scale Experiment on Immunization," *Econometrica*, forthcoming.

Bates, S., A. Angelopoulos, L. Lei, J. Malik, and M. Jordan (2021): "Distribution-free, risk-controlling prediction sets," *Journal of the ACM*, 68, 1–34.

Beery, S., G. Van Horn, and P. Perona (2018): "Recognition in Terra Incognita," in *Proceedings of the European Conference on Computer Vision*.

Bentkus, V. (2004): "On Hoeffding's inequalities," *The Annals of Probability*, 32, 1650–1673.

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

Bertanha, M. and G. W. Imbens (2020): "External Validity in Fuzzy Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, 38, 593–612.

Blanchard, G., G. Lee, and C. Scott (2011): "Generalizing from Several Related Classification Tasks to a New Unlabeled Sample," in *Advances in Neural Information Processing Systems*, ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, Curran Associates, Inc., vol. 24.

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

Bruhin, A., H. Fehr-Duda, and T. Epper (2010): "Risk and rationality: Uncovering heterogeneity in probability distortion," *Econometrica*, 78, 1375–1412.

Bugni, F., I. Canay, A. Shaikh, and M. Tabord-Meehan (2023): "Inference for Cluster Randomized Experiments with Non-ignorable Cluster Sizes," Working Paper.

Camerer, C. F., G. Nave, and A. Smith (2019): "Dynamic unstructured bargaining with private information: theory, experiment, and outcome prediction via machine learning," *Management Science*, 65, 1867–1890.

Card, D. and A. B. Krueger (1995): "Time-Series Minimum-Wage Studies: A Meta-analysis," *The American Economic Review*, 85, 238–243.

Cattaneo, M. D., L. Keele, R. Titiunik, and G. Vazquez-Bare (2021): "Extrapolating treatment effects in multi-cutoff regression discontinuity designs," *Journal of the American Statistical Association*, 116, 1941–1952.

Charnes, A. and W. W. Cooper (1962): "Programming with linear fractional functionals," *Naval Research logistics quarterly*, 9, 181–186.

Chetverikov, D., Z. Liao, and V. Chernozhukov (2021): "On cross-validated lasso in high dimensions," *The Annals of Statistics*, 49, 1300–1317.

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

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

Deaton, A. (2010): "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, 48, 424–55.

Dehejia, R., C. Pop-Eleches, and C. Samii (2021): "From Local to Global: External Validity in a Fertility Natural Experiment," *Journal of Business & Economic Statistics*, 39, 217–243.

DellaVigna, S. and E. Linos (2020): "RCTs to Scale: Comprehensive Evidence from Two Nudge Units," Working Paper.

DellaVigna, S. and D. Pope (2019): "Stability of Experimental Results: Forecasts and Evidence," Working Paper.

Einav, L., A. Finkelstein, I. Pascu, and M. R. Cullen (2012): "How General Are Risk Preferences? Choices under Uncertainty in Different Domains," *American Economic Review*, 102, 2606–38.

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

Fan, Y., D. V. Budescu, and E. Diecidue (2019): "Decisions with compound lotteries." *Decision*, 6, 109.

Fehr-Duda, H., A. Bruhin, T. Epper, and R. Schubert (2010): "Rationality on the rise: Why relative risk aversion increases with stake size," *Journal of Risk and Uncertainty*, 40, 147–180.

Fortin, N. M., A. Hill, and J. Huang (2014): "Superstition in the Housing Market," *Economic Inquiry*, 52, 974–993.

Fudenberg, D., W. Gao, and A. Liang (2023): "How Flexible is that Functional Form? Quantifying the Restrictiveness of Theories," *Review of Economics and Statistics,* forthcoming.

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

Fudenberg, D., J. Kleinberg, A. Liang, and S. Mullainathan (2022): "Measuring the Completeness of Economic Models," 130, 956–990.

Fudenberg, D. and A. Liang (2019): "Predicting and Understanding Initial Play," *American Economic Review*, 109, 4112–4141.

Gechter, M., C. Samii, R. Dehejia, and C. Pop-Eleches (2019): "Evaluating Ex Ante Counterfactual Predictions Using Ex Post Causal Inference," Working Paper.

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

Greenwood, J., Z. Hercowitz, and P. Krusell (1997): "Long-Run Implications of Investment-Specific Technological Change," *The American Economic Review*, 87, 342–362.

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

Harless, D. W. and C. F. Camerer (1994): "The Predictive Utility of Generalized Expected Utility Theories," *Econometrica*, 62, 1251–1289.

Hartford, J. S., J. R. Wright, and K. Leyton-Brown (2016): "Deep Learning for Predicting Human Strategic Behavior," in *Advances in Neural Information Processing Systems*, ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Curran Associates, Inc., vol. 29.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009): *The elements of statistical learning: data mining, inference, and prediction*, vol. 2, Springer.

Hey, J. D. and C. Orme (1994): "Investigating Generalizations of Expected Utility Theory Using Experimental Data," *Econometrica*, 62, 1291–1326.

Hirasawa, T., M. Kandori, and A. Matsushita (2022): "Using Big Data and Machine Learning to Uncover How Players Choose Mixed Strategies," Working Paper.

Hoeffding, W. (1963): "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*, 58, 13–30.

Hofman, J. M., D. J. Watts, S. Athey, F. Garip, T. L. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M. J. Salganik, S. Vazire, A. Vespignani, and T. Yarkoni (2021): "Integrating explanation and prediction in computational social science," *Nature*, 595, 181–188.

Hotz, V. J., G. W. Imbens, and J. H. Mortimer (2005): "Predicting the efficacy of future training programs using past experiences at other locations," *Journal of Econometrics*, 125, 241–270.

Hsieh, S.-L., S. Ke, C. Zhao, and Z. Wang (2023): "A Logit Neural-Network Utility Model," Working Paper.

Imai, T., T. A. Rutter, and C. F. Camerer (2020): "Meta-Analysis of Present-Bias Estimation using Convex Time Budgets," *The Economic Journal*, 131, 1788–1814.

Imbens, G. W. (2010): "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, 48, 399–423.

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

Külpmann, P. and C. Kuzmics (2022): "Comparing Theories of One-Shot Play Out of Treatment," .

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

Lefebvre, M., F. M. Vieider, and M. C. Villeval (2010): "Incentive effects on risk attitude in small probability prospects," *Economics Letters*, 109, 115–120.

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

Lei, L., R. Sahoo, and S. Wager (2023): "Policy learning under biased sample selection," *arXiv preprint arXiv:2304.11735*.

Levitt, S. D. and J. A. List (2007): "Viewpoint: On the Generalizability of Lab Behaviour to the Field," *The Canadian Journal of Economics*, 40, 347–370.

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

Liang, K.-Y. and S. L. Zeger (1986): "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13–22.

Ludwig, J. and S. Mullainathan (2023): "Machine Learning as a Tool for Hypothesis Generation," Working Paper.

Maeba, K. (2022): "Extrapolation of Treatment Effect Estimates Across Contexts and Policies: An Application to Cash Transfer Experiments," Working Paper.

Manski, C. F. (2021): "Econometrics for Decision Making: Building Foundations Sketched by Haavelmo and Wald," *Econometrica*, 89, 2827–2853.

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

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

McKay, A., E. Nakamura, and J. Steinsson (2016): "The Power of Forward Guidance Revisited," *American Economic Review*, 106, 3133–58.

Meager, R. (2019): "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments," *American Economic Journal: Applied Economics*, 11, 57–91.

——— (2022): "Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature," *American Economic Review*, 112, 1818–47.

Mogstad, M. and A. Torgovitsky (2018): "Identification and extrapolation of causal effects with instrumental variables," *Annual Review of Economics*, 10, 577–613.

Muandet, K., D. Balduzzi, and B. Schölkopf (2013): "Domain Generalization via Invariant Feature Representation," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, I–10–I–18.

Murad, Z., M. Sefton, and C. Starmer (2016): "How do risk attitudes affect measured confidence?" *Journal of Risk and Uncertainty*, 52, 21–46.

Murphy, K. (2012): *Machine Learning: a Probabilistic Perspective*, MIT Press.

Natenzon, P. (2019): "Random Choice and Learning," *Journal of Political Economy*, 127, 419–457.

Noti, G., E. Levi, Y. Kolumbus, and A. Daniely (2016): "Behavior-Based Machine-Learning: A Hybrid Approach for Predicting Human Decision Making," *CoRR*, abs/1611.10228.

OECD (2023): "Purchasing Power Parities," Accessed on 10 November, 2022.

Oswald, F. (2019): "The effect of homeownership on the option value of regional migration," *Quantitative Economics*, 10, 1453–1493.

Pan, S. J. and Q. Yang (2010): "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, 22, 1345–1359.

Park, S., A. Tafti, and G. Shmueli (2023): "Transporting Causal Effects Across Populations Using Structural Causal Modeling: The Example of Work-From-Home Productivity," *Information Systems Research*.

Pathak, P. and P. Shi (2013): "Simulating Alternative School Choice Options in Boston - Main Report," Working Paper.

Pearl, J. and E. Bareinboim (2014): "External Validity: From Do-Calculus to Transportability Across Populations," *Statistical Science*, 29, 579 – 595.

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

Peterson, J. C., D. D. Bourgin, M. Agrawal, D. Reichman, and T. L. Griffiths (2021): "Using large-scale experiments and machine learning to discover theories of human decision-making," *Science*, 372, 1209–1214.

Peysakhovich, A. and J. Naecker (2017): "Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity," *Journal of Economic Behavior & Organization*, 133, 373–384.

Plonsky, O., R. Apel, E. Ert, M. Tennenholtz, D. Bourgin, J. Peterson, and ...and I. Erev (2019): "Predicting human decisions with behavioral theories and machine learning," *CoRR*, abs/1904.06866.

Ritzwoller, D. M., J. P. Romano, and A. M. Shaikh (2024): "Randomization Inference: Theory and Applications," *arXiv preprint arXiv:2406.09521*.

Romano, J. P. (1990): "On the behavior of randomization tests without a group invariance assumption," *Journal of the American Statistical Association*, 85, 686–692.

Rosenbaum, P. R. (2005): "Sensitivity analysis in observational studies," *Encyclopedia of statistics in behavioral science*.

Sahoo, R., L. Lei, and S. Wager (2022): "Learning from a biased sample," *arXiv preprint arXiv:2209.01754*.

Sutter, M., M. G. Kocher, D. Glätzle-Rützler, and S. T. Trautmann (2013): "Impatience and uncertainty: Experimental decisions predict adolescents' field behavior," *American Economic Review*, 103, 510–31.

Tibshirani, R. J., R. Foygel Barber, E. Candes, and A. Ramdas (2019): "Conformal prediction under covariate shift," *Advances in neural information processing systems*, 32.

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

Vivalt, E. (2020): "How Much Can We Generalize From Impact Evaluations?" *Journal of the European Economic Association*, 18, 3045–3089.

Vovk, V., A. Gammerman, and G. Shafer (2005): *Algorithmic learning in a random world*, vol. 29, Springer.

Zhou, K., Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy (2021): "Domain Generalization: A Survey," Working Paper.

Appendix A. Proofs

A.1. **Notation.** Throughout let $\mathcal{N} \equiv \{1, \ldots, n\}$. The set $\mathbb{T}_{r,n}$ consists of all vectors of length r with distinct values in $\mathcal{N}$. For any $(d_1, \ldots, d_{r+1}) \in \mathbb{T}_{r+1,n+1}$, let $f(d_1, \ldots, d_{r+1}) = e_{(d_1,\ldots,d_r),d_{r+1}}$ denote the transfer error from training samples $S_{d_1}, \ldots, S_{d_r}$ to test sample $S_{d_{r+1}}$.

A.2. **Proofs of Proposition 1 and Claim 1.** These are special cases of Theorem 1 with $\Gamma = 1$. To avoid repetition, we will only prove Theorem 1.

A.3. **Proof of Theorem 1 and Corollary 1.** We start by proving a simple lemma.

Lemma A.1. *Let $H = (1-\pi)F + \pi G$ be a mixture of two distributions F and G. Further let Z_H be a draw from H and Z_F a draw from F. Then, for any $0 \leq \tau_1 < \tau_2 \leq 1$,*

$$(1-\pi)(\tau_2 - \tau_1) \leq \mathbb{P}\left(Z_H \in [\underline{Q}_{\tau_1}(F), \overline{Q}_{\tau_2}(F)]\right) \leq (1-\pi)\mathbb{P}\left(Z_F \in [\underline{Q}_{\tau_1}(F), \overline{Q}_{\tau_2}(F)]\right) + \pi$$

Proof. Let Z_G be a draw from G, and let W be a binary random variable, independent of Z_F and Z_G, with $\mathbb{E}[W] = \pi$. Then $Z_H \stackrel{d}{=} (1-W)Z_F + WZ_G$ and

$$\begin{aligned}
&\mathbb{P}\left(Z_H \in [\underline{Q}_{\tau_1}(F), \overline{Q}_{\tau_2}(F)]\right) \\
&= (1-\pi)\mathbb{P}\left(Z_F \in [\underline{Q}_{\tau_1}(F), \overline{Q}_{\tau_2}(F)]\right) + \pi\mathbb{P}\left(Z_G \in [\underline{Q}_{\tau_1}(F), \overline{Q}_{\tau_2}(F)]\right) \\
&\in (1-\pi)\mathbb{P}\left(Z_F \in [\underline{Q}_{\tau_1}(F), \overline{Q}_{\tau_2}(F)]\right) + [0, \pi].
\end{aligned}$$

By definition of upper and lower quantiles, $\mathbb{P}\left(Z_F \in [\underline{Q}_{\tau_1}(F), \overline{Q}_{\tau_2}(F)]\right) \geq \tau_2 - \tau_1$. The result then follows. □

Proof of Theorem 1. Throughout the proof we condition on the unordered samples $\{S_1, \ldots, S_{n+1}\}$ and denote by $\{S_{(1)}, \ldots, S_{(n+1)}\}$ any typical realization. Let $\mathcal{F}$ denote the sigma-field generated by the unordered set $\{S_{(1)}, \ldots, S_{(n+1)}\}$. Under the assumed data-generating process,

$$e_{(d_1,\ldots,d_r),n+1} \mid \mathcal{F} \stackrel{d}{=} f(\boldsymbol{\pi}^w(d_1), \ldots, \boldsymbol{\pi}^w(d_r), \boldsymbol{\pi}^w(n+1)), \quad \forall (d_1, \ldots, d_r) \in \mathbb{T}_{r,n}.$$

where $\boldsymbol{\pi}^w$ is a random permutation on $\{1, \ldots, n+1\}$ distributed according to

$$\mathbb{P}(\boldsymbol{\pi}^w = \pi) = \frac{1}{n!} \frac{w(S_{\pi(n+1)})}{\sum_{j=1}^{n+1} \omega(S_j)}. \tag{A.1}$$

On the other hand, $\mathbf{T} \stackrel{d}{=} (\boldsymbol{\pi}^n(1), \ldots, \boldsymbol{\pi}^n(r))$ where $\boldsymbol{\pi}^n$ denotes a uniform random permutation on $\{1, \ldots, n\}$, so $e_{\mathbf{T},n+1} \mid \mathcal{F} \stackrel{d}{=} f(\boldsymbol{\pi}^w \circ \boldsymbol{\pi}^n(1), \ldots, \boldsymbol{\pi}^w \circ \boldsymbol{\pi}^n(r), \boldsymbol{\pi}^w(n+1))$. By (A.1), we have

$$e_{\mathbf{T},n+1} \mid \mathcal{F} \stackrel{d}{=} f(\boldsymbol{\pi}^w(1), \ldots, \boldsymbol{\pi}^w(r), \boldsymbol{\pi}^w(n+1)). \tag{A.2}$$

For any $(d_1, \ldots, d_r, k) \in \mathbb{T}_{r+1,n+1}$, let

$$W'_k = \mathbb{P}((\boldsymbol{\pi}^w(1), \ldots, \boldsymbol{\pi}^w(r), \boldsymbol{\pi}^w(n+1)) = (d_1, \ldots, d_r, k)) = \frac{(n-r)!}{n!} \frac{w(S_k)}{\sum_{j=1}^{n+1} \omega(S_j)}.$$

Thus,

$$e_{\mathbf{T},n+1} \mid \mathcal{F} \sim \sum_{(d_1,\ldots,d_r,k) \in \mathbb{T}_{r+1,n+1}} W'_k \cdot \delta_{f(d_1,\ldots,d_r,k)}. \tag{A.3}$$

Because $\mathbb{P}(Z \leq \bar{Q}_\tau(F)) \geq \tau$,

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{Q}_\tau \left(\sum_{(d_1,\ldots,d_r,k) \in \mathbb{T}_{r+1,n+1}} W'_k \cdot \delta_{f(d_1,\ldots,d_r,k)} \right) \mid \mathcal{F} \right) \geq \tau. \tag{A.4}$$

Let

$$\Omega_{n+1} = \sum_{(d_1,\ldots,d_r,k) \in \mathbb{T}_{r+1,n+1} \setminus \mathbb{T}_{r+1,n}} W'_k. \tag{A.5}$$

By definition, the element $n+1$ must belong to every tuple $(d_1, \ldots, d_r, k) \in \mathbb{T}_{r+1,n+1} \setminus \mathbb{T}_{r+1,n}$. Thus, W'_{n+1} shows up $|\mathbb{T}_{r+1,n+1} \setminus \mathbb{T}_{r+1,n}|/(r+1)$ times in the sum (A.5). By symmetry, each of the other W_k's shows up $|\mathbb{T}_{r+1,n+1} \setminus \mathbb{T}_{r+1,n}|r/(r+1)n$ times. Since

$$|\mathbb{T}_{r+1,n+1} \setminus \mathbb{T}_{r+1,n}| = |\mathbb{T}_{r+1,n+1}| - |\mathbb{T}_{r+1,n}| = \frac{(n+1)!}{(n-r)!} - \frac{n!}{(n-r-1)!} = (r+1)\frac{n!}{(n-r)!},$$

we obtain that

$$\begin{aligned} \Omega_{n+1} &= \frac{n!}{(n-r)!} W'_{n+1} + \frac{r(n-1)!}{(n-r)!} \sum_{k=1}^{n} W'_k \\ &= \frac{r}{n} + \frac{(n-1)!}{(n-r-1)!} W'_{n+1} = \frac{r}{n} + \frac{n-r}{n} \frac{w(S_{n+1})}{\sum_{j=1}^{n+1} \omega(S_j)}. \end{aligned} \tag{A.6}$$

where the second to last equality uses $\sum_{k=1}^{n+1} W'_k = \frac{(n-r)!}{n!}$. By (A.6) and (7),

$$W_k = \frac{W'_k}{1 - \Omega_{n+1}}.$$

Thus, the distribution in (A.3) can be written as the following mixture distribution

$$\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}} W'_k \cdot \delta_{f(d_1,\ldots,d_r,k)} = (1-\Omega_{n+1})F^{\omega}_{\mathbf{M}} + \Omega_{n+1}\cdot G$$

where $G = \sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}\setminus\mathbb{T}_{r+1,n}} (W'_k/\Omega_{n+1})\cdot\delta_{f(d_1,\ldots,d_r,k)}$. By Lemma A.1 with $\tau_1 = 0, \tau_2 = \tau$, and $F = F^{\omega}_{\mathbf{M}}$, we have

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{Q}_\tau(F^{\omega}_{\mathbf{M}}) \mid \mathcal{F}\right) \geq \tau(1-\Omega_{n+1}).$$

Moreover, when $f(d_1,\ldots,d_r,k)$ are mutually distinct,

$$\mathbb{P}\left(Z \leq \bar{Q}_\tau(F^{\omega}_{\mathbf{M}})\right) \leq 1-\tau+\max_k W_k,$$

where Z is the draw from the distribution inside $\bar{Q}_\tau$ and

$$\max_k W_k = \frac{(n-r-1)!}{(n-1)!}\frac{\max_{k\leq n}\omega(S_k)}{\sum_{j=1}^n \omega(S_j)}.$$

Thus, Lemma A.1 implies

$$\begin{aligned}
&\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{Q}_\tau(F^{\omega}_{\mathbf{M}}) \mid \mathcal{F}\right) \\
&\leq \left(\tau + \frac{(n-r-1)!}{(n-1)!}\frac{\max_{k\leq n}\omega(S_k)}{\sum_{j=1}^n \omega(S_j)}\right)(1-\Omega_{n+1}) + \Omega_{n+1} \\
&= 1 - \left(1-\tau - \frac{(n-r-1)!}{(n-1)!}\frac{\max_{k\leq n}\omega(S_k)}{\sum_{j=1}^n \omega(S_j)}\right)(1-\Omega_{n+1}).
\end{aligned}$$

The result then follows by the iterated law of expectation and (A.6), which implies

$$1-\Omega_{n+1} = \frac{n-r}{n}\frac{\sum_{j=1}^n \omega(S_j)}{\sum_{j=1}^{n+1}\omega(S_j)}.$$

The two-sided guarantee can be similarly obtained by Lemma A.1 with $\tau_1 = 1-\tau$ and $\tau_2 = \tau$.

To prove Corollary 1, we note that

$$\frac{\omega(S_{n+1})}{\sum_{j=1}^{n+1}\omega(S_j)} \leq \frac{\Gamma}{n\Gamma^{-1}+\Gamma} = \frac{1}{n\Gamma^{-2}+1}.$$

Thus,

$$1 - \Omega_{n+1} = \frac{n-r}{n}\left(1 - \frac{\omega(S_{n+1})}{\sum_{j=1}^{n+1}\omega(S_j)}\right) \geq \frac{n-r}{n}\frac{n\Gamma^{-2}}{n\Gamma^{-2}+1} = \frac{n-r}{n+\Gamma^2}.$$

□

Online appendix to the paper

The Transfer Performance of Economic Models

Isaiah Andrews Drew Fudenberg Lihua Lei Annie Liang Chaofeng Wu

March 4, 2025

Appendix P. Other Transfer Problems

Although we have focused on specifications of transfer error that evaluate how well a model transfers from one domain to another, our results apply for the substantially broader class of random variables given in Definition 1. We discuss below other interesting specifications of $e_{\mathcal{T},d^*}$ and what they might measure.

P.1. **Parameter Transfer.** When a model has interpretable parameters, we may be interested in whether the parameter values estimated on the training data will be a good proxy for the best-fitting parameters in the target sample.

Example 6 (Effectiveness of a Job Training Program). An economist has estimated the effectiveness of a job training program using a data set from one location (as in Hotz et al. (2005)). How similar would the estimate be if the program were implemented at another location?

Example 7 (Loss Aversion). An economist observes on a data set of choice over lotteries that "losses loom larger than gains," specifically that the loss aversion parameter in Prospect Theory has a value larger than 1. If the economist were to elicit choices over a different set of lotteries, would this qualitative conclusion continue to hold?

Consider any model that can be defined as a set $\mathcal{F}_\Theta = \{f_\theta\}_{\theta\in\Theta}$ of prediction rules $f_\theta : \mathcal{X} \to \mathcal{Y}$, which depend continuously on a parameter θ in a compact parameter space Θ. Given any training data $S_{\mathbf{T}}$, let $\hat{\theta}(S_{\mathbf{T}}) = \text{arginf}_{\theta\in\Theta} \sum_{d\in\mathbf{T}} \frac{|S_d|}{\sum_{d\in\mathbf{T}}|S_d|} \sum_{d\in\mathbf{T}} e(f_\theta, S_d)$ be the parameter value that minimizes a weighted sum of the errors across the samples in the training data, and let $f_{\hat{\theta}(S_{\mathbf{T}})}$ denote the corresponding prediction rule.[22] To assess parameter variation, first

[22]If there are ties, break them arbitrarily

fix a distance metric $d(\theta, \theta')$ (e.g., Euclidean distance) to assess how different two parameter vectors θ and θ' are. Then the transfer error

$$e_{\mathbf{T},n+1} = d\left(\hat{\theta}(S_{\mathbf{T}}), \hat{\theta}(S_{n+1})\right)$$

measures how far the estimated parameters on the training data are from the best-fitting parameters on the target sample.

We can also assess how well a qualitative prediction that is based on the estimated parameters will transfer to the target sample (e.g., a prediction that some coefficient is positive). Let A denote any event that can be described as a function of the parameter θ. Then

$$e_{\mathbf{T},n+1} = \begin{cases} 1 & \text{if } \mathbb{1}\left(\hat{\theta}(S_{\mathbf{T}}) \in A\right) = \mathbb{1}\left(\hat{\theta}(S_{n+1}) \in A\right) \\ 0 & \text{otherwise} \end{cases}$$

is a transfer error which tells us whether the prediction about A based on the training samples also holds in the target sample.

P.2. **Other Estimation Procedures.** In the examples above, a model is trained on r training samples and used to predict properties of a target sample. Our results apply also for other training procedures. To avoid introducing extensive notation, we describe these procedures informally.

Example 8 (Transfer Learning). In *transfer learning* problems in computer science (see e.g., Pan and Yang (2010)), some observations from the target sample are available in addition to the training samples $S_{\mathbf{T}}$. The model or algorithm is trained on these observations jointly, with some specification of how to weight the target sample observations relative to the other training data. The performance of a model estimated in this way is another transfer error.

Example 9 (Transfer of Specific Parameters). While some economic parameters are viewed as constant across domains, other parameters may be viewed as domain-specific. For example, spatial models of trade often have structural parameters (e.g., the elasticity of demand substitution between goods produced in different countries) whose values are set using estimates from another paper, and "fundamentals" (e.g., productivity in each country), which are re-estimated on each sample (see for example Alfaro-Urena et al., 2023). The performance of a model that is estimated and evaluated in this way is a transfer error.

Example 10 (Using Cross-Validation to Tune Parameters). Our framework can also accommodate training procedures in which cross-validation is used to tune select model parameters. For example, black box algorithms often have a complexity parameter (e.g., the penalization parameter in LASSO or the depth of decision trees in a random forest algorithm). One way of choosing the size of this parameter is based on out-of-sample fit (Hastie et al., 2009; Chetverikov et al., 2021). In our setting, this means holding out one of the training samples to use for testing, training the algorithm on the remaining $r-1$ training samples, and evaluating fit on the remaining test sample. The chosen complexity parameter is the one that yields the lowest average error across the r possible choices of the test sample. Fixing this value for the complexity parameter, the algorithm is then fit to the entire training data. The performance of such an algorithm on the target sample is a transfer error.

Example 11 (Counterfactual Predictions). One way that economic models are used is to form predictions for outcomes under policy changes that have yet to be implemented. For instance, McFadden (1974) predicted the demand impacts of the then-new BART rapid transit system in the San Franciso Bay Area, and Pathak and Shi (2013) predicted demand for schools under changes to the Boston school choice system. One can generalize our framework to cover the case where each sample S_d is instead a pair of two observations, $S_d = (S_d^0, S_d^1)$. The pre-intervention samples $(S_1^0, \ldots, S_{n+1}^0)$ are drawn i.i.d. from one distribution, while the post-intervention samples $(S_1^1, \ldots, S_{n+1}^1)$ are drawn i.i.d. from another. In this more general setting, a transfer error is any function of the training pairs $\{(S_d^0, S_d^1)\}_{d \in \mathbf{T}}$, the target pair (S_{n+1}^0, S_{n+1}^1), and potentially an independent noise variable.[23]

Appendix Q. Extensions and further results

Our main results focus on forecasting realized transfer errors, which is useful when we want to know the range of plausible errors in transferring a given model to a new domain. We now complement those results with procedures for inference focused on population quantities: Section Q.2 provides confidence intervals for quantiles of the transfer error distribution, and Section Q.3 provides a confidence interval for the expected transfer error. Since these quantities can be perfectly recovered given data from an infinite number of domains, we

[23]Our theoretical results generalize completely for transfer errors defined in this way; the main limitation is the difficulty of obtaining sufficiently many pre- and post-intervention pairs. We mention this potential application in the case that such data does eventually become available.

expect the lengths of these intervals to vanish as the number of observed domains grows large, unlike the forecast intervals from Section 3.2.

Q.1. **Preliminary Lemma.** We start by establishing a bound that will be useful in the subsequent construction of confidence intervals. Let

$$U = \frac{(n-k)!}{n!} \sum_{(i_1,\ldots,i_k)\in\mathbb{T}_{r+1,n}} \phi(Z_{i_1},\ldots,Z_{i_k})$$

be an arbitrary U-statistic of degree k with a bounded (and potentially asymmetric) kernel ϕ that takes values in $[0,1]$.

Definition Q.1. For every $n, k \in \mathbb{Z}_+$ and $x, y \in \mathbb{R}$, define

$$B_{n,k}(x;y) \equiv \min\left\{ b^1_{n,k}(x;y), b^2_{n,k}(x;y), b^3_{n,k}(x;y) \right\}$$

where

$$b^1_{n,k}(x;y) \equiv \exp\left\{ -\lceil n/k \rceil \left(x\wedge y \log\left(\frac{x\wedge y}{y}\right) + (1-x\wedge y)\log\left(\frac{1-x\wedge y}{1-y}\right)\right)\right\}$$

$$b^2_{n,k}(x;y) \equiv e\cdot \mathbb{P}\left(\text{Binom}(\lceil n/k\rceil; y) \le \lceil \lceil n/k\rceil \cdot x\rceil\right)$$

$$b^3_{n,k}(x;y) \equiv \min_{\lambda>0} \frac{n\lambda}{k}\left(x - \frac{\lambda}{\lambda + kG(\lambda)} y\right)$$

Lemma Q.1. *If $\phi(Z_1,\ldots,Z_k) \in [0,1]$ almost surely, then $P(U \le x) \le B_{n,k}(x;\mathbb{E}(U))$ for every $x \in [0,1]$.*

Q.2. **Quantiles of transfer error.** Let F denote the CDF of $e_{\mathbf{T},n+1}$, which we assume is continuous. This section builds a confidence interval for the β-th quantile of F, denoted q_β.

For arbitrary $q \in \mathbb{R}$ and realized metadata $\mathbf{M} = \{S_1,\ldots,S_n\}$, define

$$\varphi(q,\mathbf{M}) = \frac{(n-r-1)!}{n!} \sum_{(d_1,\ldots,d_{r+1})\in\mathbb{T}_{r+1,n}} \mathbb{I}(e_{(d_1,\ldots,d_r),d_{r+1}}) \le q)$$

where $\mathbb{I}(\cdot)$ is the indicator function, recalling that $e_{(d_1,\ldots,d_r),d_{r+1}}$ denotes the observed transfer error from samples $(S_{d_1},\ldots,S_{d_r})$ to sample $S_{d_{r+1}}$. This is the fraction of observed transfer errors in the metadata (from r training samples to one test sample) that are less than q. Then $U_\beta \equiv \varphi(q_\beta,\mathbf{M})$ is a U-statistic where by definition, $\mathbb{E}[U_\beta] = \beta$. Lemma Q.1 then

implies

$$\mathbb{P}(U_\beta \le x) \le B_{n,r+1}(x,\beta) \qquad \mathbb{P}(U_\beta \ge x) = \mathbb{P}(1-U_\beta \le 1-x) \le B_{n,r+1}(1-x,1-\beta). \quad \text{(Q.1)}$$

Definition Q.2. For any quantile $\beta \in (0,1)$ and confidence level $1-\alpha \in (0,1)$, let $\hat{u}_\beta^+(\alpha) = \inf\{u : B_{n,r+1}(u;\beta) \ge \alpha\}$ and $\hat{u}_\beta^-(\alpha) = \sup\{u : B_{n,r+1}(1-u;1-\beta) \ge \alpha\}$. Further define $\hat{q}_\beta^{\mathrm{L}}(\alpha) \equiv \min\left\{q : \varphi(q,\mathbf{M}) \ge \hat{u}_\beta^+(\alpha)\right\}$ and $\hat{q}_\beta^{\mathrm{U}}(\alpha) \equiv \max\left\{q : \varphi(q,\mathbf{M}) \le \hat{u}_r^-(\alpha)\right\}$.

Since $B_{n,r+1}(u;\cdot)$ is right-continuous in u, it follows from (Q.1) that $\mathbb{P}(U_\beta < \hat{u}_\beta^+(\alpha)) \le \alpha$ and $\mathbb{P}(U_\beta > \hat{u}_\beta^-(\alpha)) \le \alpha$. Since $\varphi(q,\mathbf{M})$ is monotonically increasing in q, the event $\{U_\beta < \hat{u}_\beta^+(\alpha)\}$ is equivalent to $\{q_\beta < \hat{q}_\beta^{\mathrm{L}}(\alpha)\}$, while $\{U_\beta > \hat{u}_\beta^-(\alpha)\}$ is equivalent to $\{q_\beta > \hat{q}_\beta^{\mathrm{U}}(\alpha)\}$. This yields:

Proposition Q.1. *For any quantile $\beta \in (0,1)$ and confidence level $1-\alpha \in (0,1)$, $P(q_\beta \le \hat{q}_\beta^U(\alpha)) \ge 1-\alpha$ and $\mathbb{P}\left(q_\beta \in \left[\hat{q}_\beta^L(\alpha/2), \hat{q}_\beta^U(\alpha/2)\right]\right) \ge 1-\alpha$.*

Figure 8 applies Proposition Q.1 to construct two-sided 81% confidence interval for the median raw transfer error, median transfer shortfall, and median transfer deterioration. As in Figure 4, these confidence intervals are substantially wider for the black box algorithms, and have higher upper bounds.

Q.3. **Expected transfer error.** This section constructs confidence intervals for the expected transfer error, $\mu \equiv \mathbb{E}(e_{\mathbf{T},n+1})$, under the assumption that transfer errors are uniformly bounded (in which case it is without loss to set $e_{\mathbf{T},n+1} \in [0,1]$). Define the U-statistic

$$U = \frac{(n-r-1)!}{n!} \sum_{(d_1,\ldots,d_{r+1}) \in \mathbb{T}_{r+1,n}} e_{(d_1,\ldots,d_r),d_{r+1}}.$$

Because $\mathbb{E}[U] = \mu$, Lemma Q.1 implies that $\mathbb{P}(U \le x) \le B_{n,r+1}(x,\mu)$ and $\mathbb{P}(U \ge x) \le B_{n,r+1}(1-x,1-\mu)$ for all $x \in \mathbb{R}$.

Definition Q.3. For any confidence guarantee $1-\alpha \in (0,1)$, let $\hat{\mu}^+(\alpha) = \sup\{\mu : B_{n,r+1}(U;\mu) \ge \alpha\}$ and $\hat{\mu}^-(\alpha) = \inf\{\mu : B_{n,r+1}(1-U;1-\mu) \ge \alpha\}$.

It follows from (Q.1) that $\mathbb{P}(U < \hat{u}^+(\alpha)) \le \alpha$ and $\mathbb{P}(U > \hat{u}^-(\alpha)) \le \alpha$, which implies:

Proposition Q.2. *If $e_{\mathcal{T},d} \in [0,1]$ almost surely, then $\mathbb{P}\left(\mu \le \hat{\mu}^+(\alpha)\right) \ge 1-\alpha$ and $\mathbb{P}\left(\mu \in [\hat{\mu}^-(\alpha/2), \hat{\mu}^+(\alpha/2)]\right) \ge 1-\alpha$.*

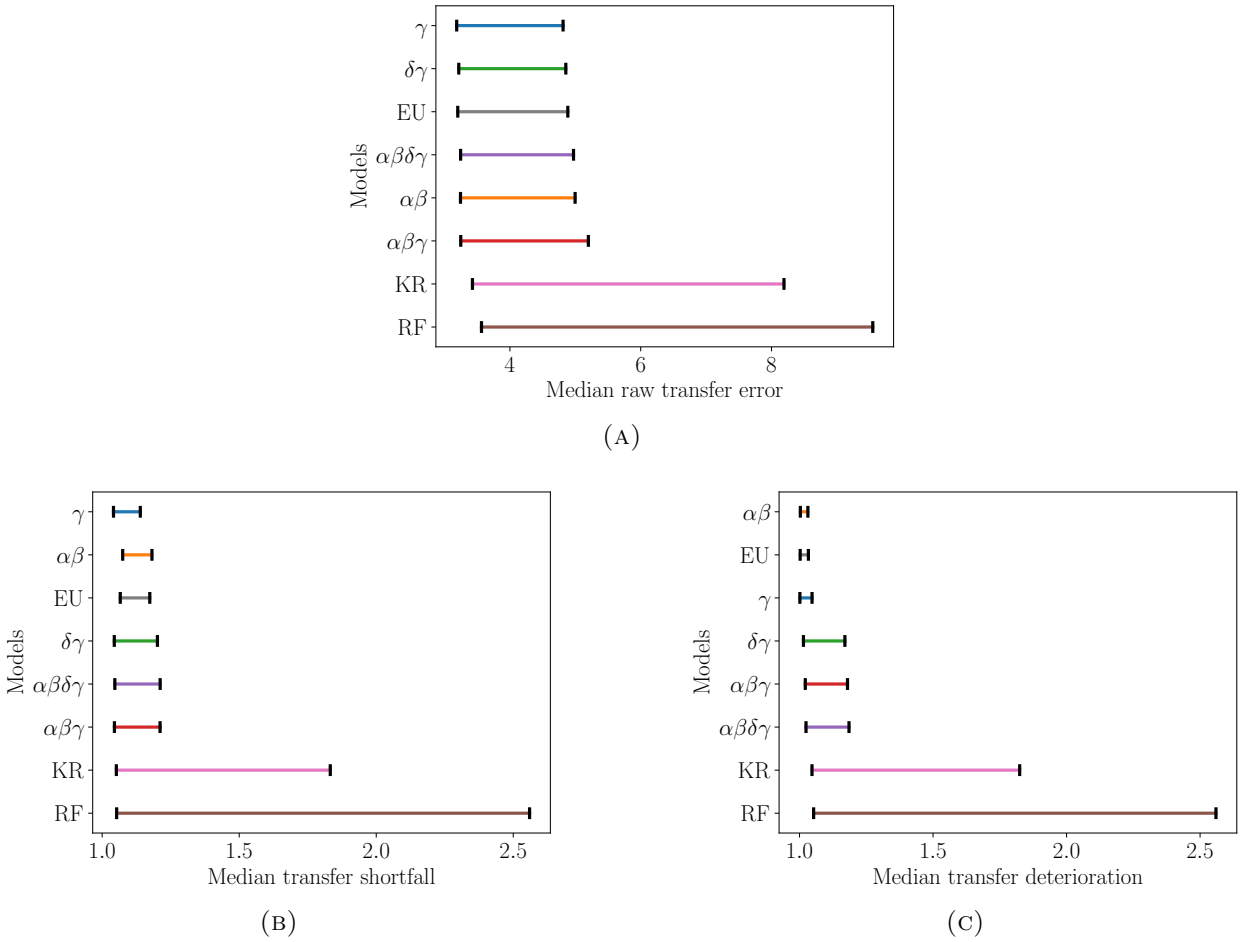


FIGURE 8. *81% confidence intervals for the median of (a) raw transfer error, (b) transfer shortfall, and (c) transfer deterioration.*

Figure 9 applies this result to construct two-sided 81% confidence intervals for the transfer errors we considered in Section 5.4. Since transfer shortfall and transfer deterioration are not bounded, we report instead confidence intervals for the expectation of their inverses $\frac{\min_{m\in\mathcal{M}} e\left(f^m_{S_{n+1}},S_{n+1}\right)}{e(f_{S_\mathbf{T}},S_{n+1})}$ and $\frac{e(f_{S_{n+1}},S_{n+1})}{e(f_{S_\mathbf{T}},S_{n+1})}$; lower values for these measures correspond to worse transfer performance. We again find that the confidence intervals for the black box algorithms are qualitatively worse than those for the economic models.

Q.4. **Proof of Lemma Q.1.** Hoeffding (1963) shows that $P(U \leq x) \leq b^1_{n,k}(x, \mathbb{E}(U))$, and Bates et al. (2021) shows that $P(U \leq x) \leq b^2_{n,k}(x, \mathbb{E}(U))$. We now show that if $x \in [0, 1]$ then $P(U \leq x) \leq b^3_{n,k}(x, \mathbb{E}(U))$. To do this, we use a series of intermediate results to extend

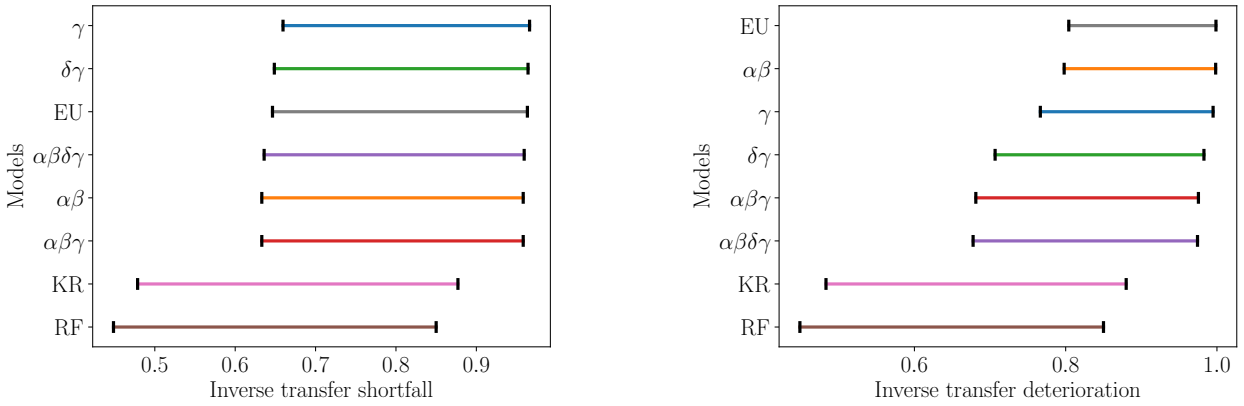


FIGURE 9. *81% forecast intervals for (a) expected inverse transfer shortfall, (b) expected inverse transfer deterioration.*

a result of Bates et al. (2021) on U-statistics of degree 2 with bounded kernels to U-statistics with bounded kernels for any order $k \geq 2$.

Let $Z_1, \ldots, Z_n$ be i.i.d. random variables and $\phi : \mathbb{R}^k \to [0, 1]$ be a bounded function. Then a U-statistic of degree k is defined as

$$U = \frac{(n-k)!}{n!} \sum_{i_1,\ldots,i_k} \phi(Z_{i_1}, \ldots, Z_{i_k}), \tag{Q.2}$$

where $\sum_{i_1,\ldots,i_k}$ denotes the sum over all k-tuples in $\mathcal{N}$ with mutually distinct elements. The average of Z_i is a special case of (Q.2) with $k = 1$ and $\phi(z) = z$.

Let $m = \lfloor n/k \rfloor$ and $\boldsymbol{\pi}^n : \mathcal{N} \mapsto \mathcal{N}$ be a uniformly random permutation. For each permutation π, define

$$W_\pi = \frac{1}{m} \sum_{j=1}^{m} \phi\left(Z_{\pi((j-1)k+1)}, \ldots, Z_{\pi(jk)}\right).$$

Note that the summands in W_π are independent given π. Then $U = \mathbb{E}_{\boldsymbol{\pi}^n}[W_{\boldsymbol{\pi}^n}]$, where $\mathbb{E}_{\boldsymbol{\pi}^n}$ denotes the expectation with respect to $\boldsymbol{\pi}^n$ when conditioning on $Z_1, \ldots, Z_n$. By Jensen's inequality, for any convex function ψ, $\mathbb{E}[\psi(U)] = \mathbb{E}[\psi(\mathbb{E}_{\boldsymbol{\pi}^n}[W_{\boldsymbol{\pi}^n}])] \leq \mathbb{E}[\mathbb{E}_{\boldsymbol{\pi}^n}\psi(W_{\boldsymbol{\pi}^n})] = \mathbb{E}_{\boldsymbol{\pi}^n}[\mathbb{E}\psi(W_{\boldsymbol{\pi}^n})]$. Since W_π has identical distributions for all π,

$$\mathbb{E}[\psi(U)] \leq \mathbb{E}[\psi(W_{\mathbf{id}})] \tag{Q.3}$$

where **id** is the permutation that maps each element to itself.

Recalling that Hoeffding's inequality is derived from the moment-generating function $\psi(z) = e^{\lambda z}$ (Hoeffding, 1963), and the Bentkus inequality is derived from the piecewise linear function $\psi(z) = (z-t)_+$ (Bentkus, 2004), the following tail inequalities for U-statistics are a direct consequence of (Q.3).

Proposition Q.3. *Let U be a U-statistic of order k with a bounded kernel $\phi \in [0,1]$ in the form of* (Q.2) *and $m = \lfloor n/k \rfloor$. Then*

(1) (Hoeffding inequality for U-statistics, Section 5 of Hoeffding 1963)

$$\mathbb{P}(U \leq x) \leq \exp\left\{-m h_1\left(x \wedge \mathbb{E}[U]; \mathbb{E}[U]\right)\right\},$$

where

$$h_1(y;\mu) = y \log\left(\frac{y}{\mu}\right) + (1-y)\log\left(\frac{1-y}{1-\mu}\right).$$

(2) (Bentkus inequality for U-statistics, modified from Bentkus 2004)

$$\mathbb{P}\left(U \leq x\right) \leq e\mathbb{P}\left(\mathrm{Bin}\left(m; \mathbb{E}[U]\right) \leq \lceil mx \rceil\right).$$

Other concentration inequalities can be derived from the leave-one-out property. Write $U(Z_1, \ldots, Z_n)$ for U and let $U_i = \inf_{z_i} U(Z_1, \ldots, Z_{i-1}, z_i, Z_{i+1}, \ldots, Z_n)$. Note that U_i is independent of Z_i. Since $\phi(\cdot) \geq 0$, we have $0 \leq U - U_i \leq \frac{(n-k)!}{n!} \sum_{j=1}^{k} \sum_{i_1,\ldots,i_k, i_j = i} \phi(Z_{i_1}, \ldots, Z_{i_k})$ so $\frac{n}{k}(U - U_i) \leq 1$ and

$$\begin{aligned}
\sum_{i=1}^{n} (U - U_i)^2 &\leq \frac{((n-k)!)^2}{(n!)^2} \sum_{i=1}^{n} \left(\sum_{j=1}^{k} \sum_{i_1,\ldots,i_k, i_j = i} \phi(Z_{i_1}, \ldots, Z_{i_k}) \right)^2 \\
&\overset{(i)}{\leq} \frac{k(n-k)!}{n \cdot n!} \sum_{j=1}^{k} \sum_{i=1}^{n} \sum_{i_1,\ldots,i_k, i_j = i} \phi(Z_{i_1}, \ldots, Z_{i_k})^2 \\
&\overset{(ii)}{\leq} \frac{k(n-k)!}{n \cdot n!} \sum_{j=1}^{k} \sum_{i=1}^{n} \sum_{i_1,\ldots,i_k, i_j = i} \phi(Z_{i_1}, \ldots, Z_{i_k}) \\
&= \frac{k^2}{n} U,
\end{aligned}$$

where (i) applies the Cauchy-Schwarz inequality and (ii) uses the fact that $\phi(\cdot) \leq 1$. If we let $W = (n/k)U$ and $W_i = (n/k)U_i$, then $W - W_i \leq 1, \quad \sum_{i=1}^{n} (W - W_i)^2 \leq kW$. This implies

that W as a function of $Z_1, \ldots, Z_n$ satisfies the assumptions for the claim (34) in Theorem 13 of Maurer (2006) with constant $a = k$.[24]

Proposition Q.4 (Theorem 13, Maurer 2006)**.** *Let $G(\lambda) = (e^\lambda - \lambda - 1)/\lambda$. Then for any $\lambda > 0$,*

$$\log \mathbb{E}[e^{\lambda(\mathbb{E}[W]-W)}] \leq \frac{k\lambda G(\lambda)}{\lambda + kG(\lambda)}\mathbb{E}[W].$$

This further implies that for any $x \in (0, \mathbb{E}[U])$,

$$\mathbb{P}\left(U \leq x\right) \leq \exp\left\{\min_{\lambda>0} \frac{n\lambda}{k}\left(x - \frac{\lambda}{\lambda + kG(\lambda)}\mathbb{E}[U]\right)\right\}.$$

Putting Proposition Q.3 and Proposition Q.4 together yields Lemma Q.1.

Appendix R. Supplementary Material to Section 4

R.1. **A more general distribution shift model.** In this subsection we discuss a more general distribution shift model that allows $S_1, \ldots, S_n$ to have non-identical distributions. Let p denote the joint density of $S_1, \ldots, S_{n+1}$ (with respect to a dominating measure). Let $\boldsymbol{\pi}^p$ be a random permutation on $\{1, \ldots, n+1\}$ such that, for any realization $\{s_1, \ldots, s_{n+1}\}$ of $\{S_1, \ldots, S_{n+1}\}$,

$$\begin{aligned}&\mathbb{P}\left(\boldsymbol{\pi}^p(1) = d_1, \ldots, \boldsymbol{\pi}^p(n+1) = d_{n+1} \mid \{S_{(1)}, \ldots, S_{(n+1)}\} = \{s_1, \ldots, s_{n+1}\}\right)\\ &= \frac{p(s_{d_1}, \ldots, s_{d_{n+1}})}{\sum_{(d'_1,\ldots,d'_{n+1})\in\mathbb{T}_{n+1,n+1}} p(s_{d'_1}, \ldots, s_{d'_{n+1}})}.\end{aligned}$$

Then,

$$\begin{aligned}&(s_{\boldsymbol{\pi}^p(1)}, \ldots, s_{\boldsymbol{\pi}^p(n)}) \mid \{S_{(1)}, \ldots, S_{(n+1)}\} = \{s_1, \ldots, s_{n+1}\}\\ &\overset{d}{=} (S_1, \ldots, S_{n+1}) \mid \{S_{(1)}, \ldots, S_{(n+1)}\} = \{s_1, \ldots, s_{n+1}\}.\end{aligned}$$

Again let $\mathcal{F}$ denote the sigma-field generated by the unordered set $\{S_{(1)}, \ldots, S_{(n+1)}\}$.

Definition R.1. For any $\Gamma \geq 1$, let $\mathcal{P}(\Gamma; r)$ be the class of distributions on $(S_1, \ldots, S_{n+1})$ with

$$\frac{(n+1)!}{(n-r)!}\mathbb{P}\left(\boldsymbol{\pi}^p(1) = d_1, \ldots, \boldsymbol{\pi}^p(r) = d_r, \boldsymbol{\pi}^p(n+1) = k \mid \mathcal{F}\right) \in \left[\Gamma^{-1}, \Gamma\right] \text{ almost surely,}$$

[24]Theorem 13 of Maurer (2006) states a weaker result that $\log \mathbb{E}[e^{\lambda(\mathbb{E}[W]-W)}] \leq \frac{k\mathbb{E}[W]}{2}\lambda^2$. The stronger version stated here can be found in the second last display in the proof of Theorem 13 of Maurer (2006).

for any $(d_1, \ldots, d_r, k) \in \mathbb{T}_{n+1}$.

Above, $(n-r)!/(n+1)!$ is the probability under a uniform permutation and thus the LHS can be interpreted as the density ratio between $\boldsymbol{\pi}^p$ and a uniform permutation, which measures the deviation from exchangeability.

By (A.1), when $\nu \in \mathcal{W}(\Gamma)$, the joint density p satisfies

$$\begin{aligned}&\frac{(n+1)!}{(n-r)!}\mathbb{P}(\boldsymbol{\pi}^p(1) = d_1, \ldots, \boldsymbol{\pi}^p(r) = d_r, \boldsymbol{\pi}^p(n+1) = k)\\ &= \frac{(n+1)\omega(S_k)}{\sum_{j=1}^{n+1}\omega(S_j)} \le \frac{(n+1)\Gamma}{n\Gamma^{-1}+\Gamma} = \frac{(n+1)\Gamma^2}{n+\Gamma^2}.\end{aligned}$$

Thus,

$$p \in \mathcal{P}\left(\frac{(n+1)\Gamma^2}{n+\Gamma^2}; r\right) \subset \mathcal{P}\left(\Gamma^2; r\right).$$

Next, we derive forecast intervals akin to Corollary 1.

Theorem R.1. *Suppose the joint distribution of $(S_1, \ldots, S_{n+1})$ lies in $\mathcal{P}(\Gamma; r)$. Then*

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \le \bar{e}^{\mathbf{M}}_{1-(1-\tau)/\Gamma}\right) \ge \tau\left(1 - \frac{(r+1)\Gamma}{n+1}\right),$$

and

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

Proof. For notational convenience, for any $(d_1, \ldots, d_r, k) \in \mathbb{T}_{r+1,n+1}$, let

$$A_{d_1,\ldots,d_r,k} = \mathbb{P}\left(\boldsymbol{\pi}^p(1) = d_1, \ldots, \boldsymbol{\pi}^p(r) = d_r, \boldsymbol{\pi}^p(n+1) = k \mid \mathcal{F}\right).$$

Again, we condition on the unordered samples $S_1, \ldots, S_{n+1}$ and denote by $S_{(1)}, \ldots, S_{(n+1)}$ a typical realization. By similar arguments used to show (A.3), we have

$$e_{\mathbf{T},n+1} \mid \mathcal{F} \sim \sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}} A_{d_1,\ldots,d_r,k} \cdot \delta_{f(d_1,\ldots,d_r,k)}. \tag{R.1}$$

Thus,

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \le \bar{Q}_\tau\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}} A_{d_1,\ldots,d_r,k} \cdot \delta_{f(d_1,\ldots,d_r,k)}\right) \mid \mathcal{F}\right) \ge \tau.$$

Let

$$\Omega_{n+1} = \sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}\setminus\mathbb{T}_{r+1,n}} A_{d_1,\ldots,d_r,k}.$$

Then we can rewrite the distribution in (R.1) as a mixture distribution

$$\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}} A_{d_1,\ldots,d_r,k}\cdot\delta_{f(d_1,\ldots,d_r,k)} = (1-\Omega_{n+1})\cdot F + \Omega_{n+1}\cdot G,$$

where

$$F = \sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n}} \frac{A_{d_1,\ldots,d_r,k}}{1-\Omega_{n+1}}\cdot\delta_{f(d_1,\ldots,d_r,k)}, \quad G = \sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}\setminus\mathbb{T}_{r+1,n}} \frac{A_{d_1,\ldots,d_r,k}}{\Omega_{n+1}}\cdot\delta_{f(d_1,\ldots,d_r,k)}.$$

By Lemma A.1 with $\tau_1 = 0, \tau_2 = \tau$, we have that

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \le \bar{Q}_\tau\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n}} A_{d_1,\ldots,d_r,k}\cdot\delta_{f(d_1,\ldots,d_r,k)}\right) \mid \mathcal{F}\right) \ge \tau(1-\Omega_{n+1}).$$

By definition,

$$A_{d_1,\ldots,d_r,k} \in \frac{(n-r)!}{(n+1)!}\cdot\left[\Gamma^{-1},\Gamma\right].$$

The largest possible value for $\bar{Q}_\tau\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n}} A_{d_1,\ldots,d_r,k}\cdot\delta_{f(d_1,\ldots,d_r,k)}\right)$ is achieved when $A_{d_1,\ldots,d_r,k} = \Gamma(n-r)!/(n+1)!$ for the largest values of $f(d_1,\ldots,d_r,k)$. Thus,

$$\bar{Q}_\tau\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n}} A_{d_1,\ldots,d_r,k}\cdot\delta_{f(d_1,\ldots,d_r,k)}\right) \le \bar{Q}_{\tau'}\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n}} \delta_{f(d_1,\ldots,d_r,k)}\right) = \bar{e}^{\mathbf{M}}_{\tau'},$$

where $\Gamma(1-\tau') = 1-\tau$. Clearly, $\tau' = 1-(1-\tau)/\Gamma$. Thus,

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \le \bar{e}^{\mathbf{M}}_{\tau'} \mid \mathcal{F}\right) \ge \tau(1-\Omega_{n+1}).$$

Moreover,

$$\Omega_{n+1} \le \Gamma\frac{(n-r)!}{(n+1)!}\cdot(|\mathbb{T}_{r+1,n+1}| - |\mathbb{T}_{r+1,n}|) = \frac{(r+1)\Gamma}{n+1}.$$

Thus, the result for the one-sided interval is proved. The result for the two-sided interval can be proved similarly by Lemma A.1 with $\tau_1 = 1-\tau, \tau_2 = \tau$ and by noting that the smallest possible value for $\underline{Q}_\tau\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n}} A_{d_1,\ldots,d_r,k}\cdot\delta_{f(d_1,\ldots,d_r,k)}\right)$ is achieved when $A_{d_1,\ldots,d_r,k} = \Gamma(n-r)!/(n+1)!$ for the smallest values of $f(d_1,\ldots,d_r,k)$, and hence

$$\underline{Q}_\tau\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n}} A_{d_1,\ldots,d_r,k}\cdot\delta_{f(d_1,\ldots,d_r,k)}\right) \ge \underline{Q}_{1-\tau'}\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n}} \delta_{f(d_1,\ldots,d_r,k)}\right) = \underline{e}^{\mathbf{M}}_{\tau'}.$$

□

R.2. **Algorithm for evaluating worst-case-upper-dominance.** We provide an algorithm that computes $\bar{e}_\tau(\Gamma)$ with a single τ in $O(rn^{r+1}\log n)$ time and computes $\bar{e}_\tau(\Gamma)$ for all $\tau \in (0,1)$ in $O(rn^{r+1}\log n + n^{r+2})$ time. First, sort the elements in $\{f(d_1,\ldots,d_{r+1}) : (d_1,\ldots,d_{r+1}) \in \mathbb{T}_{r+1,n}\}$ as

$$f_{(1)} \le f_{(2)} \le \ldots \le f_{(|\mathbb{T}_{r+1,n}|)},$$

where

$$f_{(j)} = f(d^{(j)}), \quad d^{(j)} = (d_1^{(j)},\ldots,d_{r+1}^{(j)}) \in \mathbb{T}_{r+1,n}.$$

Let $\psi^{(j)} \in \{0,1\}^n$ with

$$\psi_i^{(j)} = I\left(d_{r+1}^{(j)} = i\right).$$

Further define the cumulative sum of $\psi^{(j)}$ as

$$\Psi^{(j)} = \sum_{\ell=1}^{j} \psi^{(\ell)}.$$

Let $w = (\omega(S_1),\ldots,\omega(S_n))^T$ and $\mathbf{1}_n = (1,1,\ldots,1)^T$. By (7), for each j,

$$f_{(j)} \ge \bar{e}_\tau^{\mathbf{M},\omega} \iff \frac{(n-r-1)!}{(n-1)!}\frac{w^T\Psi^{(j)}}{w^T\mathbf{1}_n} \ge \tau.$$

Therefore,

$$\bar{e}_\tau^{\mathbf{M},\omega} = f_{J^\omega}, \quad \text{where } J_\tau^\omega = \min\left\{j : \frac{w^T\Psi^{(j)}}{w^T\mathbf{1}_n} \ge \tau\frac{(n-1)!}{(n-r-1)!}\right\}.$$

By definition, the set of w generated by all $\omega \in \mathcal{W}(\Gamma)$ is $[\Gamma^{-1},\Gamma]^n$. Thus,

$$\bar{e}_\tau^{\mathbf{M}}(\Gamma) = f_{J_\tau(\Gamma)}, \quad \text{where } J_\tau(\Gamma) = \min\left\{j : \min_{w\in[\Gamma^{-1},\Gamma]^n}\frac{w^T\Psi^{(j)}}{w^T\mathbf{1}_n} \ge \tau\frac{(n-1)!}{(n-r-1)!}\right\}. \tag{R.2}$$

Via some algebra, we can further simplify the expression of $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$.

Theorem R.2. *Let $\bar{\Psi}_k^{(j)}$ be the average of the k-smallest coordinates of $\Psi^{(j)}$ and*

$$Q_j(\Gamma) = \frac{j}{n} + \min_{k\in\mathcal{N}}\frac{\bar{\Psi}_k^{(j)} - \frac{j}{n}}{1+\frac{n}{k(\Gamma^2-1)}}.$$

Then $Q_j(\Gamma)$ is strictly increasing in both j and Γ. Moreover, $\bar{e}_\tau^{\mathbf{M}}(\Gamma) = f_{(J_\tau(\Gamma))}$, where

$$J_\tau(\Gamma) = \min\left\{j \ge \tau\frac{n!}{(n-r-1)!} : Q_j(\Gamma) \ge \tau\frac{(n-1)!}{(n-r-1)!}\right\}.$$

Proof. First, we prove that

$$\min_{w\in[\Gamma^{-1},\Gamma]^n} \frac{w^T\Psi^{(j)}}{w^T\mathbf{1}_n} = \min_{w\in\{\Gamma^{-1},\Gamma\}^n} \frac{w^T\Psi^{(j)}}{w^T\mathbf{1}_n}. \tag{R.3}$$

Let $g_j(w) = w^T\Psi^{(j)}/w^T\mathbf{1}_n$. Then g_j is continuous and bounded on the closed set $[\Gamma^{-1},\Gamma]^n$ and thus the minimum can be achieved. Let

$$w^{(j)}(\Gamma) = \underset{w:g_j(w)=\min_{w\in[\Gamma^{-1},\Gamma]^n} g_j(w)}{\operatorname{argmin}} \sum_{i=1}^n \min\left\{|w_i-\Gamma|, |w_i-\Gamma^{-1}|\right\}.$$

Suppose there exists $i\in\mathcal{N}$ such that $w_i^{(j)}(\Gamma)\in(\Gamma^{-1},\Gamma)$. Then

$$g_j(w_i,w_{-i}) = \frac{\Psi_i^{(j)}w_i + \Psi_{-i}^{(j)T}w_{-i}}{w_i+\mathbf{1}_{n-1}^T w_{-i}} = \Psi_i^{(j)} + \frac{\Psi_{-i}^{(j)T}w_{-i} - \Psi_i^{(j)}\cdot\mathbf{1}_{n-1}^T w_{-i}}{w_i+\mathbf{1}_{n-1}^T w_{-i}},$$

where $\Psi_{-i}^{(j)}$ and w_{-i} are the leave-i-th-entry subvectors of $\Psi^{(j)}$ and w. Clearly, g_j is a monotone function of w_i for any given w_{-i}. Since $w^{(j)}(\Gamma)$ is a minimizer and $w_i^{(j)}(\Gamma)\in(\Gamma^{-1},\Gamma)$, we must have $\Psi_{-i}^{(j)T}w_{-i} - \Psi_i^{(j)}\cdot\mathbf{1}_{n-1}^T w_{-i} = 0$. Define $\widetilde{w}^{(j)}(\Gamma)$ with

$$\widetilde{w}_i^{(j)}(\Gamma) = \Gamma, \quad \widetilde{w}_{-i}^{(j)}(\Gamma) = w_{-i}^{(j)}(\Gamma).$$

Then

$$g_j(\widetilde{w}^{(j)}(\Gamma)) = g_j(w^{(j)}(\Gamma)) = \min_{w\in[\Gamma^{-1},\Gamma]^n} g_j(w),$$

while

$$\sum_{i=1}^n \min\left\{|\widetilde{w}_i^{(j)}(\Gamma)-\Gamma|, |\widetilde{w}_i^{(j)}(\Gamma)-\Gamma^{-1}|\right\} < \sum_{i=1}^n \min\left\{|w_i^{(j)}(\Gamma)-\Gamma|, |w_i^{(j)}(\Gamma)-\Gamma^{-1}|\right\}.$$

This contradicts the definition of $w^{(j)}(\Gamma)$, so $w^{(j)}(\Gamma)\in\{\Gamma^{-1},\Gamma\}^n$, which completes the proof of (R.3).

For any $w\in\{\Gamma^{-1},\Gamma\}^n$ with $|\{i: w_i=\Gamma\}| = k$, the Fréchet-Hoeffding inequality implies that Γ's are allocated to the k smallest entries of $\Psi^{(j)}$. Thus,

$$\min_{w\in\{\Gamma^{-1},\Gamma\}^n} \frac{w^T\Psi^{(j)}}{w^T\mathbf{1}_n} = \min_{k\in\mathcal{N}\cup\{0\}} \frac{\Gamma k\bar{\Psi}_k^{(j)} + \Gamma^{-1}(\mathbf{1}_n^T\Psi_i^{(j)} - k\bar{\Psi}_k^{(j)})}{\Gamma k + \Gamma^{-1}(n-k)}.$$

By definition, $\mathbf{1}_n^T\Psi_i^{(j)} = j$. Then for each k, the above expression can be simplified as

$$\frac{\Gamma k\bar{\Psi}_k^{(j)} + \Gamma^{-1}(\mathbf{1}_n^T\Psi_i^{(j)} - k\bar{\Psi}_k^{(j)})}{\Gamma k + \Gamma^{-1}(n-k)} = \frac{\Gamma k\bar{\Psi}_k^{(j)} + \Gamma^{-1}(j - k\bar{\Psi}_k^{(j)})}{\Gamma k + \Gamma^{-1}(n-k)}$$

$$= \frac{(\Gamma - \Gamma^{-1})k\bar{\Psi}_k^{(j)} + \Gamma^{-1}j}{(\Gamma - \Gamma^{-1})k + \Gamma^{-1}n} = \frac{j}{n} + \frac{(\Gamma - \Gamma^{-1})k\left(\bar{\Psi}_k^{(j)} - \frac{j}{n}\right)}{(\Gamma - \Gamma^{-1})k + \Gamma^{-1}n}$$
$$= \frac{j}{n} + \frac{\bar{\Psi}_k^{(j)} - \frac{j}{n}}{1 + \frac{n}{k(\Gamma^2-1)}}.$$

The above expression is j/n for both $k = n$ and $k = 0$, so we can remove 0 from the minimum, and thus

$$\min_{w \in [\Gamma^{-1}, \Gamma]^n} \frac{w^T \Psi^{(j)}}{w^T \mathbf{1}_n} = Q_j(\Gamma).$$

By (R.2),

$$\bar{e}_\tau^{\mathbf{M}}(\Gamma) = \min\left\{ j : Q_j(\Gamma) \geq \tau \frac{(n-1)!}{(n-r-1)!} \right\}.$$

Finally, we can restrict to $j \geq \tau n!/(n-r-1)!$ because $Q_j(\Gamma) \leq \frac{j}{n}$ by taking $k = n$. □

Since $Q_j(\Gamma)$ is increasing in j, $J_\tau(\Gamma)$ can be found via binary search with iteration complexity $O(\log n^{r+1}) = O(r \log n)$. Each iteration costs at most $O(n)$ operations to sort the entries of $\Psi^{(j)}$ based on the ordered version of $\Psi^{(j-1)}$, since there is only entry updated, and $O(n)$ additional operations to compute $Q_j(\Gamma)$. Thus, the overall computational overhead after obtaining $(f_{(1)}, \ldots, f_{(|\mathbb{T}_{r+1,n}|)})$ is just $O(rn \log n)$, which is much smaller than the cost of sorting f-values $O(n^{r+1} \log n^{r+1}) = O(rn^{r+1} \log n)$.

In some cases, we want to compute $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$ for all $\tau \in [0, 1]$ at once. The following result links $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$ to an induced distribution on the f's.

Corollary R.1. *For any $\Gamma \geq 1$, let μ_Γ be a weighted measure with*

$$\mu_\Gamma = \sum_{j=1}^{|\mathbb{T}_{r+1,n}|} \frac{(n-r-1)!}{(n-1)!}(Q_j(\Gamma) - Q_{j-1}(\Gamma)) \cdot \delta_{f_{(j)}},$$

where $Q_0(\Gamma) = 0$. Then $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$ is the τ-th quantile of μ_Γ.

Since the ordering takes $O(rn^{r+1} \log n)$ time and computing each $Q_j(\Gamma)$ takes $O(n)$ time, the total computational cost to compute $\bar{e}_\tau^{\mathbf{M}}(\Gamma)$ for all $\tau \in [0, 1]$ is $O(rn^{r+1} \log n + n^{r+2})$.

R.3. **Algorithm for evaluating everywhere dominance.** Let $f_{(j),1}$ and $f_{(j),2}$ be the j-th largest transfer errors for method 1 and 2, respectively. Similarly, the count vectors for two methods are denoted by $\Psi^{(j),1}$ and $\Psi^{(j),2}$. Then method 1 does NOT everywhere-upper-dominate method 2 at the τ-th quantile if and only if there exists $j_1, j_2 \in \{1, \ldots, |\mathbb{T}_{r+1,n}|\}$

and $W \in [0,\infty)^n$ such that

$$f_{(j_1),1} > f_{(j_2),2}, \quad \frac{(n-r-1)!}{(n-1)!}\frac{w^T\Psi^{(j_1-1),1}}{w^T\mathbf{1}_n} < \tau \le \frac{(n-r-1)!}{(n-1)!}\frac{w^T\Psi^{(j_2),2}}{w^T\mathbf{1}_n}. \tag{R.4}$$

Above $\Psi^{(0),1} = (0,0,\ldots,0)^T$.

To avoid pairwise comparisons, which incur $O(n^{2(r+1)})$ computation, we can check (R.4) by only focusing on $j_1 = m(j), j_2 = j$ where

$$m(j) = \min\{j' : f_{(j'),1} > f_{(j),2}\}.$$

It is easy to see that (R.4) holds for some pair $(j_1, j_2) \in \{1,\ldots,|\mathbb{T}_{r+1,n}|\}^2$ if and only if it holds for $(m(j), j)$ for some $j \in \{1,\ldots,|\mathbb{T}_{r+1,n}|\}$. For any given j, (R.4) reduces to

$$\frac{(n-r-1)!}{(n-1)!}\frac{w^T\Psi^{(m(j)-1),1}}{w^T\mathbf{1}_n} < \tau \le \frac{(n-r-1)!}{(n-1)!}\frac{w^T\Psi^{(j),2}}{w^T\mathbf{1}_n}, \quad w \in [0,\infty)^n.$$

This is equivalent to solving the following linear fractional programming problem and then checking if the objective is below τ:

$$\min \frac{w^T a^{(j)}}{w^T\mathbf{1}_n}, \quad \text{s.t.,} \ \frac{w^T b^{(j)}}{w^T\mathbf{1}_n} \ge \tau, \quad w \in [0,\infty)^n,$$

where

$$a^{(j)} = \Psi^{(m(j)),1} \cdot \frac{(n-r-1)!}{n-1)!}, \quad b^{(j)} = \Psi^{(j),2} \cdot \frac{(n-r-1)!}{(n-1)!}.$$

We can apply the Charnes-Cooper transformation (Charnes and Cooper, 1962) by introducing $v = w/w^T\mathbf{1}_n$ to transform it into a linear programming problem:

$$\min v^T a^{(j)}, \quad \text{s.t.,} \ v^T b^{(j)} \ge \tau, v^T\mathbf{1}_n = 1, v \in [0,\infty)^n. \tag{R.5}$$

Solving these $O(n^{r+1})$ LP problems can be accelerated by the following two observations:

(1) Using the same argument as in the last step of the proof of Theorem R.2, we can restrict

$$j \ge \tau\frac{n!}{(n-r-1)!}.$$

(2) When $a_i^{(j)} \ge b_i^{(j)}$ for every $i \in \mathcal{N}$, then the objective of (R.5) can never be below τ.

Appendix S. Supplementary material for Section 5

S.1. **Description of data.** We briefly describe the individual samples in our meta-data.

TABLE 4

Source of Data	# Obs	# Subj	# Lottery	Country	Gains Only
Abdellaoui et al. (2015)	801	89	3	France	Y
Fan et al. (2019)	4750	125	19	US	Y
Bouchouicha and Vieider (2017)	3162	94	66	UK	N
Sutter et al. (2013)	661	661	4	Austria	Y
Etchart-Vincent and l'Haridon (2011)	3036	46	20	France	N
Fehr-Duda et al. (2010)	8560	153	56	China	N
Lefebvre et al. (2010)	72	72	2	France	Y
Halevy (2007)	366	122	2	Canada	Y
Anderhub et al. (2001)	183	61	1	Israel	Y
Murad et al. (2016)	2131	86	25	UK	Y
Dean and Ortoleva (2019)	1032	179	3	US	Y
Bernheim and Sprenger (2020)	1071	153	7	US	Y
Bruhin et al. (2010)	8906	179	50	Switzerland	N
Bruhin et al. (2010)	4669	118	40	Switzerland	N
l'Haridon and Vieider (2019)	1708	61	28	Australia	N
l'Haridon and Vieider (2019)	2548	95	28	Belgium	N
l'Haridon and Vieider (2019)	2350	84	28	Brazil	N
l'Haridon and Vieider (2019)	2240	80	28	Cambodia	N
l'Haridon and Vieider (2019)	2687	96	28	Chile	N
l'Haridon and Vieider (2019)	5711	204	28	China	N
l'Haridon and Vieider (2019)	3072	128	24	Colombia	N
l'Haridon and Vieider (2019)	2968	106	28	Costa Rica	N
l'Haridon and Vieider (2019)	2770	99	28	Czech Republic	N
l'Haridon and Vieider (2019)	3906	140	28	Ethiopia	N
l'Haridon and Vieider (2019)	2604	93	28	France	N
l'Haridon and Vieider (2019)	3639	130	28	Germany	N
l'Haridon and Vieider (2019)	2352	84	28	Guatemala	N
l'Haridon and Vieider (2019)	2492	89	28	India	N
l'Haridon and Vieider (2019)	2352	84	28	Japan	N
l'Haridon and Vieider (2019)	2716	97	28	Kyrgyzstan	N
l'Haridon and Vieider (2019)	1791	64	28	Malaysia	N
l'Haridon and Vieider (2019)	3360	120	28	Nicaragua	N
l'Haridon and Vieider (2019)	5638	202	28	Nigeria	N

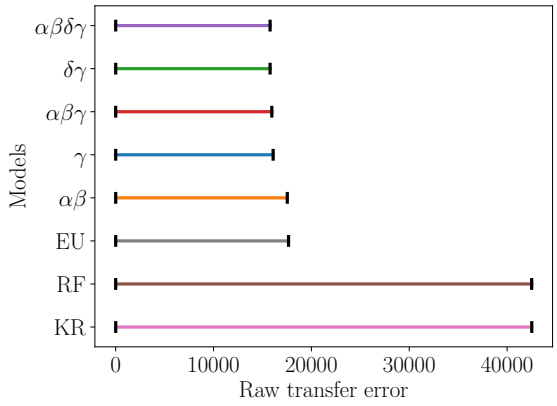
l'Haridon and Vieider (2019)	2660	95	28	Peru	N
l'Haridon and Vieider (2019)	2491	89	28	Poland	N
l'Haridon and Vieider (2019)	1959	70	28	Russia	N
l'Haridon and Vieider (2019)	1819	65	28	Saudi Arabia	N
l'Haridon and Vieider (2019)	1988	71	28	South Africa	N
l'Haridon and Vieider (2019)	2240	80	28	Spain	N
l'Haridon and Vieider (2019)	2212	79	28	Thailand	N
l'Haridon and Vieider (2019)	2070	74	28	Tunisia	N
l'Haridon and Vieider (2019)	2240	80	28	UK	N
l'Haridon and Vieider (2019)	2701	97	28	US	N
l'Haridon and Vieider (2019)	2436	87	28	Vietnam	N

S.2. **Papers as domains.** We now consider an alternative definition of domains, with each of the 14 papers representing a different domain. This changes the content of the i.i.d. assumption imposed in Section 3.2, where we now assume that samples are i.i.d. across papers, but may be dependent across subject pools within the same paper. We repeat our main analysis and report 78% two-sided forecast intervals in Figure 10. These intervals are qualitatively similar to those reported in Figure 4.

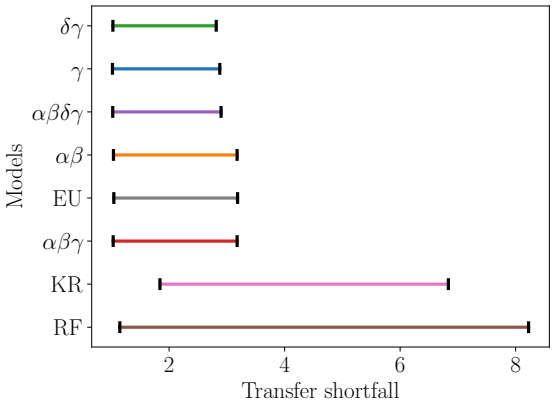
S.3. **Supplementary tables and figures for main analysis.** Table 5 reports the forecast intervals that are depicted in Figure 4.

Model	Raw Transfer Error	Transfer Shortfall	Transfer Deterioration
CPT variants			
γ	[2.50,15.83]	[1.03,2.54]	[1.00,1.47]
α, β	[2.56,16.13]	[1.04,2.35]	[1.00,1.30]
δ, γ	[2.48,17.19]	[1.02,2.47]	[1.00,1.53]
α, β, γ	[2.47,15.91]	[1.02,2.60]	[1.00,1.85]
$\alpha, \beta, \delta, \gamma$	[2.46,15.99]	[1.02,2.62]	[1.00,1.82]
EU models			
EU	[2.56,16.41]	[1.04,2.14]	[1.00,1.30]
ML algorithms			
Random Forest	[2.71,31.39]	[1.02,6.42]	[1.02,6.42]
Kernel Regression	[2.75,33.62]	[1.02,5.33]	[1.01,5.29]

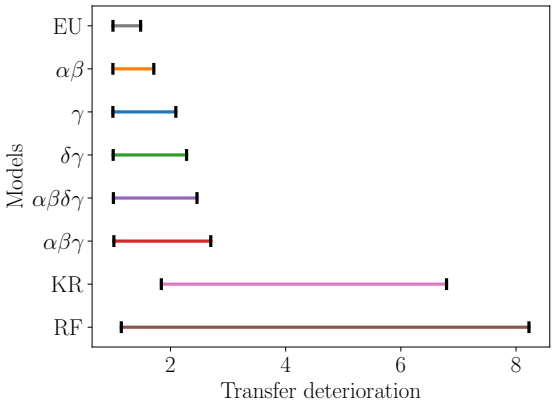
TABLE 5. *86% (n=44, $\tau = 0.95$) forecast intervals*



(A) *Raw transfer error*



(B) *Transfer shortfall*



(C) *Transfer deterioration*

FIGURE 10. *78% (n=14, $\tau = 0.95$) forecast intervals for each of the three measures, treating each paper as a separate domain.*

S.4. **Alternative forecast intervals.** In this section, we report alternative forecast intervals for our three measures. Table 6 constructs 96% two-sided forecast intervals (setting $\tau = 1$),[25] and Table 7 reports 91% one-sided forecast intervals (setting $\tau = 0.95$). All of the forecast intervals are qualitatively similar to the 86% two-sided forecast intervals reported in the main text.

Finally, Figure 11 plots the τ-th percentile of the pooled transfer errors as τ varies. The figure shows that the qualitative conclusions we have drawn about the relative performance of black boxes and economic models are not specific to any choice of τ.[26] In fact, in Panels

[25]The lower bounds of these intervals are the minimum transfer error (among the pooled transfer errors) and the upper bounds are the maximum transfer error.

[26]To improve readability, we remove extreme numbers by truncating $\tau \in [5, 95]$, and show results only for the $\alpha\beta\gamma\delta$ specification of the CPT model.

Model	Raw Transfer Error	Transfer Shortfall	Transfer Deterioration
CPT main variants			
γ	[0.81,23104.96]	[1.01,7.31]	[1.00,7.22]
α, β	[0.71,19999.41]	[1.00,5.28]	[1.00,5.27]
δ, γ	[0.71,23052.76]	[1.00,7.25]	[1.00,7.18]
α, β, γ	[0.71,28122.26]	[1.00,5.65]	[1.00,5.60]
$\alpha, \beta, \delta, \gamma$	[0.71,27959.10]	[1.00,6.01]	[1.00,5.95]
EU models			
EU	[0.72,22787.99]	[1.00,4.44]	[1.00,1.75]
ML algorithms			
Random Forest	[0.96,42520.49]	[1.01,33.17]	[1.01,33.17]
Kernel Regression	[1.01,42519.23]	[1.01,6.835]	[1.00,6.79]

TABLE 6. *96% (n=44, $\tau = 1$) two-sided forecast intervals*

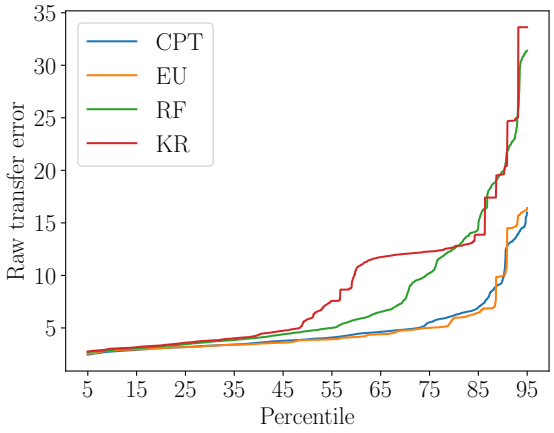
Model	Raw Transfer Error	Transfer Shortfall	Transfer Deterioration
CPT main variants			
γ	[0,15.83]	[1,2.54]	[1,1.47]
α, β	[0,16.13]	[1,2.35]	[1,1.30]
δ, γ	[0,17.19]	[1,2.47]	[1,1.53]
α, β, γ	[0,15.91]	[1,2.60]	[1,1.85]
$\alpha, \beta, \delta, \gamma$	[0,15.99]	[1,2.62]	[1,1.82]
EU models			
EU	[0,16.41]	[1,2.14]	[1,1.30]
ML algorithms			
Random Forest	[0,31.39]	[1,6.42]	[1,6.42]
Kernel Regression	[0,33.62]	[1,5.33]	[1,5.29]

TABLE 7. *91% (n=44, $\tau = 0.95$) one-sided forecast intervals*

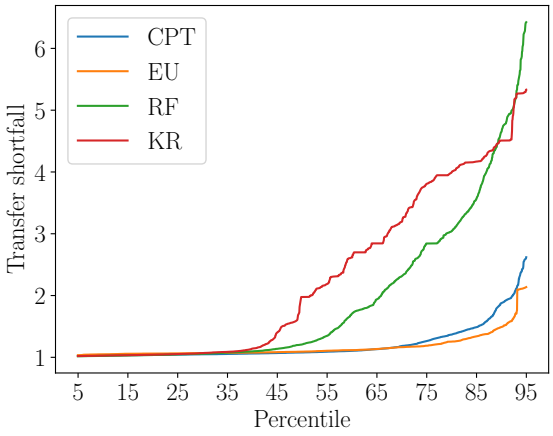
(a) and (c), the black box curves lie everywhere above the CPT and EU curves, so both the lower and upper bounds of the black boxes' forecast intervals are higher than those of the economic models for every choice of τ.

S.5. **Forecast intervals for the ratio of raw CPT and RF transfer errors.** Let $e_{\mathcal{T},d}$ be the ratio of the raw random forest transfer error to the raw CPT transfer error (i.e., using the specification in (1)), henceforth the *transfer error ratio*.

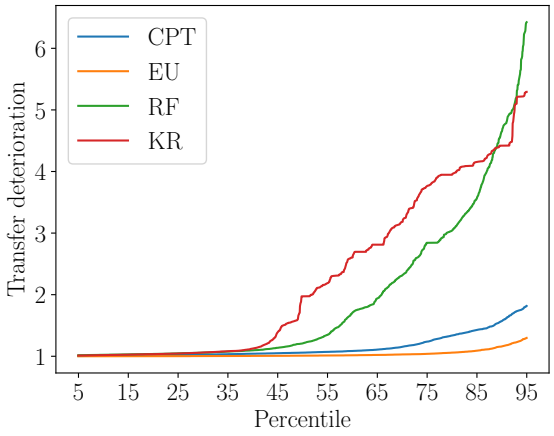
Panel (a) of Figure 12 reports 86% two-sided forecast intervals for the raw transfer error ratio for each CPT specification. The lower bound for each CPT model is approximately 0.9, while the upper bound is as large as 4.5. Panel (b) of the figure is a histogram of raw



(A) *Raw transfer error*



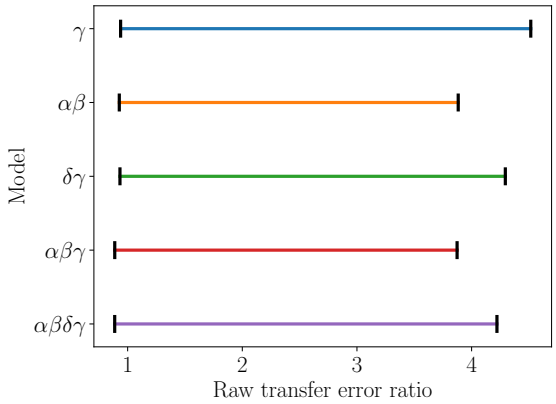
(B) *Transfer shortfall*



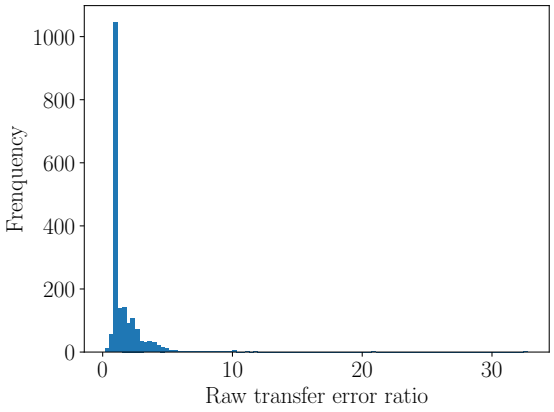
(C) *Transfer deterioration*

FIGURE 11. *Error percentiles from 5 to 95 (truncated for readability).*

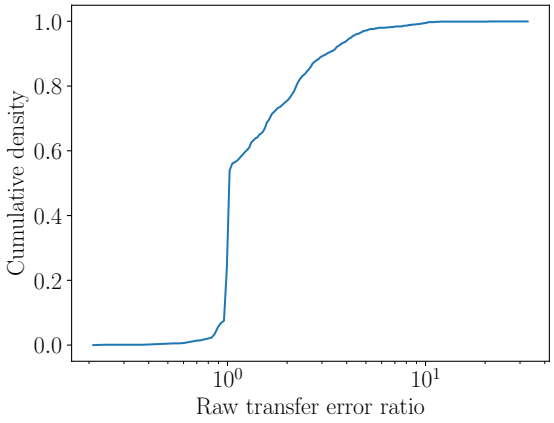
transfer error ratios for the 4-parameter CPT model when the training domains $\mathcal{T}$ and the target domains d are drawn uniformly at random from the set of domains in the meta-data. This distribution has a large cluster of ratios around 1 (i.e., raw CPT transfer errors are similar to the raw random forest errors) and a long right tail of ratios achieving a max value of 32.8 (i.e., the random forest error can be up to 32 times as large as the CPT error). The cumulative distribution function of $e_{\mathcal{T},d}$, reported in Panel (c) of Figure 12, shows that the random forest algorithm outperforms CPT in approximately 35% of $(\mathcal{T}, d)$ pairs, although CPT rarely has a much worse raw transfer error than the random forest and is sometimes much better.



(A) *86% (n=44, $\tau = 0.95$) Forecast intervals*



(B) *Density*



(C) *CDF*

FIGURE 12. *Forecast intervals, density, and cdf for the ratio of the raw random forest transfer error to the raw CPT transfer error.*

S.6. **Alternative Choice of r.** Here we consider an alternative choice for the number of training domains, setting $r = 3$ instead of $r = 1$.

This corresponds to randomly choosing 3 of the 44 domains to be the training domains, finding the best prediction rule for this pooled data, and using the estimated prediction rule to predict the remaining 41 samples. For this analysis we use domain cross-validation to select tuning parameters for the black box algorithms, as described in Example 10.

Figure 13 is the analog of Figure 4. Again we choose $\tau = 0.95$, thus constructing forecast intervals whose lower bounds are the 5% percentile of pooled transfer errors, and whose upper bounds are the 95% percentile of pooled transfer errors. Applying Proposition 1, these are 82% forecast intervals. The most notable change is that the random forest forecast

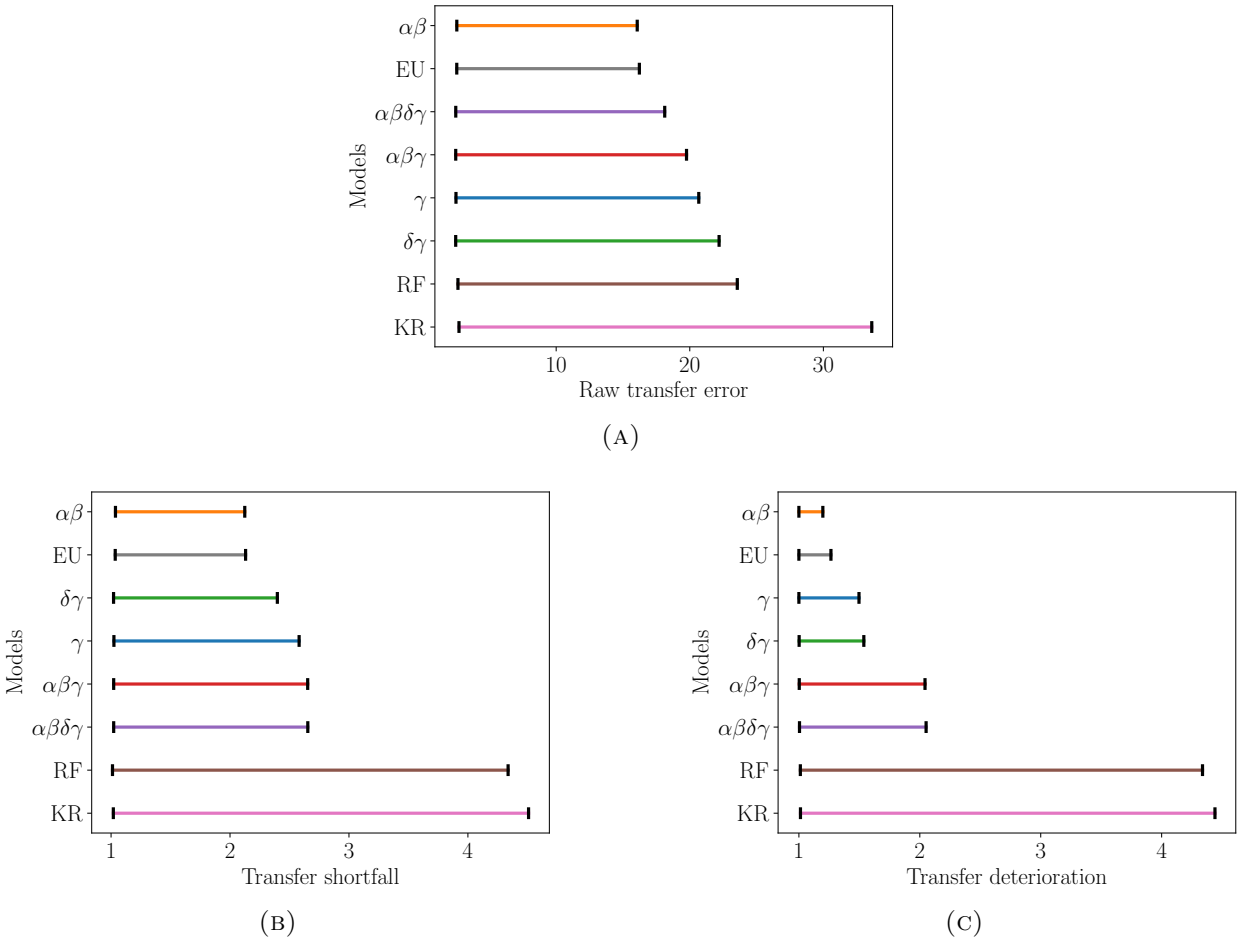


FIGURE 13. *82% (n=44, $\tau = 0.95$) forecast intervals for (a) raw transfer error, (b) transfer shortfall, and (c) transfer deterioration, with the choice of $r = 3$.*

interval shrinks considerably, which suggests that the raw transfer error of the random forest algorithm becomes less variable when it is trained on more domains. Otherwise, all of the qualitative statements in the main text for $r = 1$ continue to hold. In particular, as with $r = 1$, we find that the forecast intervals for all three of our measures have higher lower and upper bounds for the black box algorithms than for the CPT specifications.

S.7. **Supplementary Material to Section 5.5.** Here we consider an alternative choice of for the number of training samples, setting $r = 3$ and $r = 5$ instead $r = 1$. Recalling that each sample includes the observations associated with a unique lottery, this corresponds to randomly choosing three (or five) of the 24 lotteries for training, finding the best prediction

rule for this pooled data, and using the estimated prediction rule to predict certainty equivalents for the remaining 21 (or 19) lotteries. We use domain cross-validation to select tuning parameters for the black box algorithms, as described in Example 10.

Figure 14 and Figure 15 are the analog of Figure 6, with $r = 3$ and $r = 5$ respectively. We again choose $\tau = 0.95$, thus constructing forecast intervals whose lower bounds are the 5% percentile of pooled transfer errors, and whose upper bounds are the 95% percentile of pooed transfer errors. Applying Proposition 1, these are 76% for $r = 3$ and 68% for $r = 5$ forecast intervals. The most notable change is that the forecast intervals shrink for all of the prediction methods, which suggests that the raw transfer error becomes less variable when it is trained on more lotteries. Otherwise, all of the qualitative statements in the main text for $r = 1$ continue to hold, and in particular the economic models continue to transfer better than the black box algorithms do.

S.8. **More details on worst-case dominance.** Figures 16 and 17 compare the worst case upper bound of the forecast intervals for CPT and RF for our three transfer measures as either γ or τ varies. In each case the dominance relation is clear.

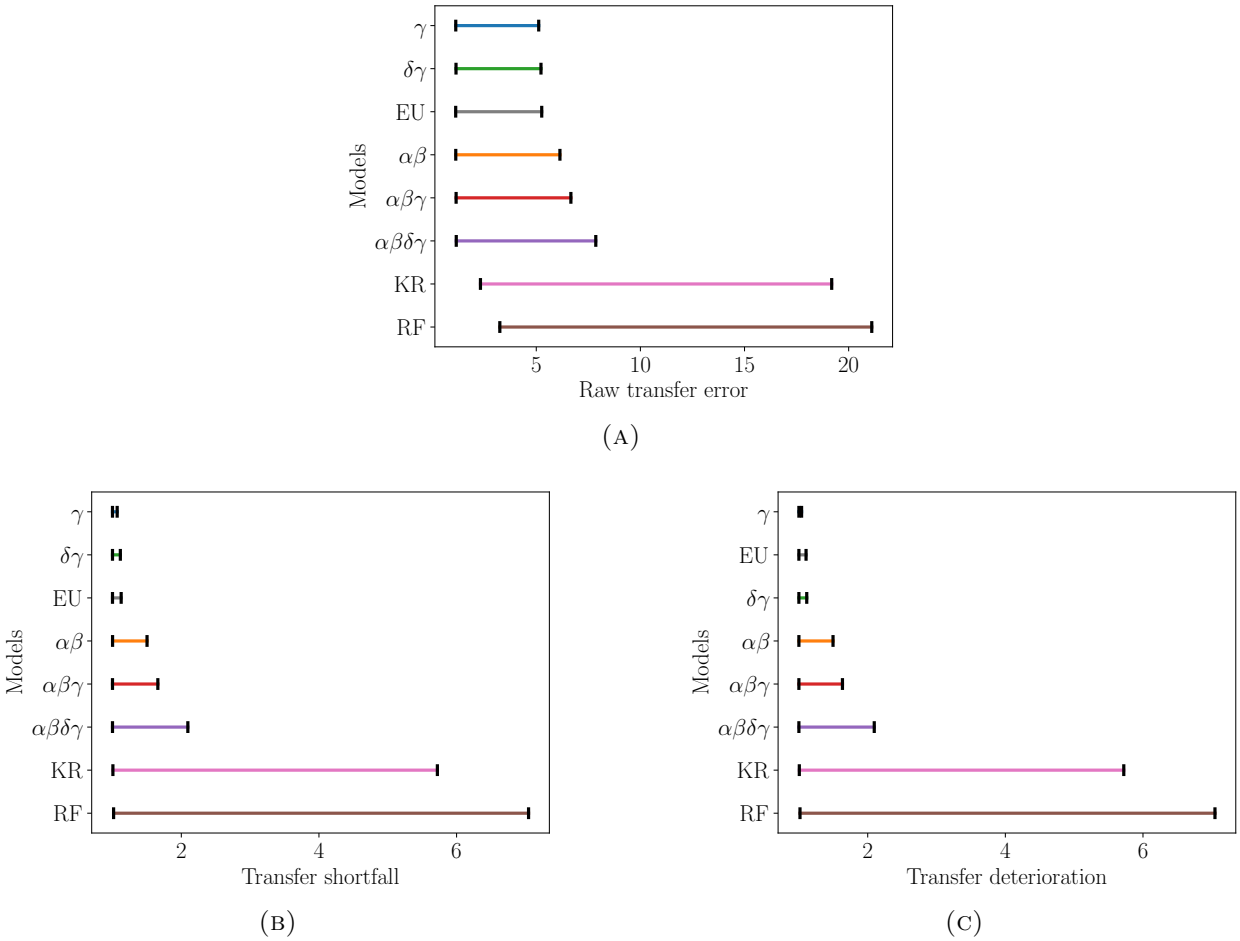


FIGURE 14. *76% (n=24, τ=0.95) forecast intervals using common lotteries in l'Haridon and Vieider (2019), with the choice of* $r = 3$.

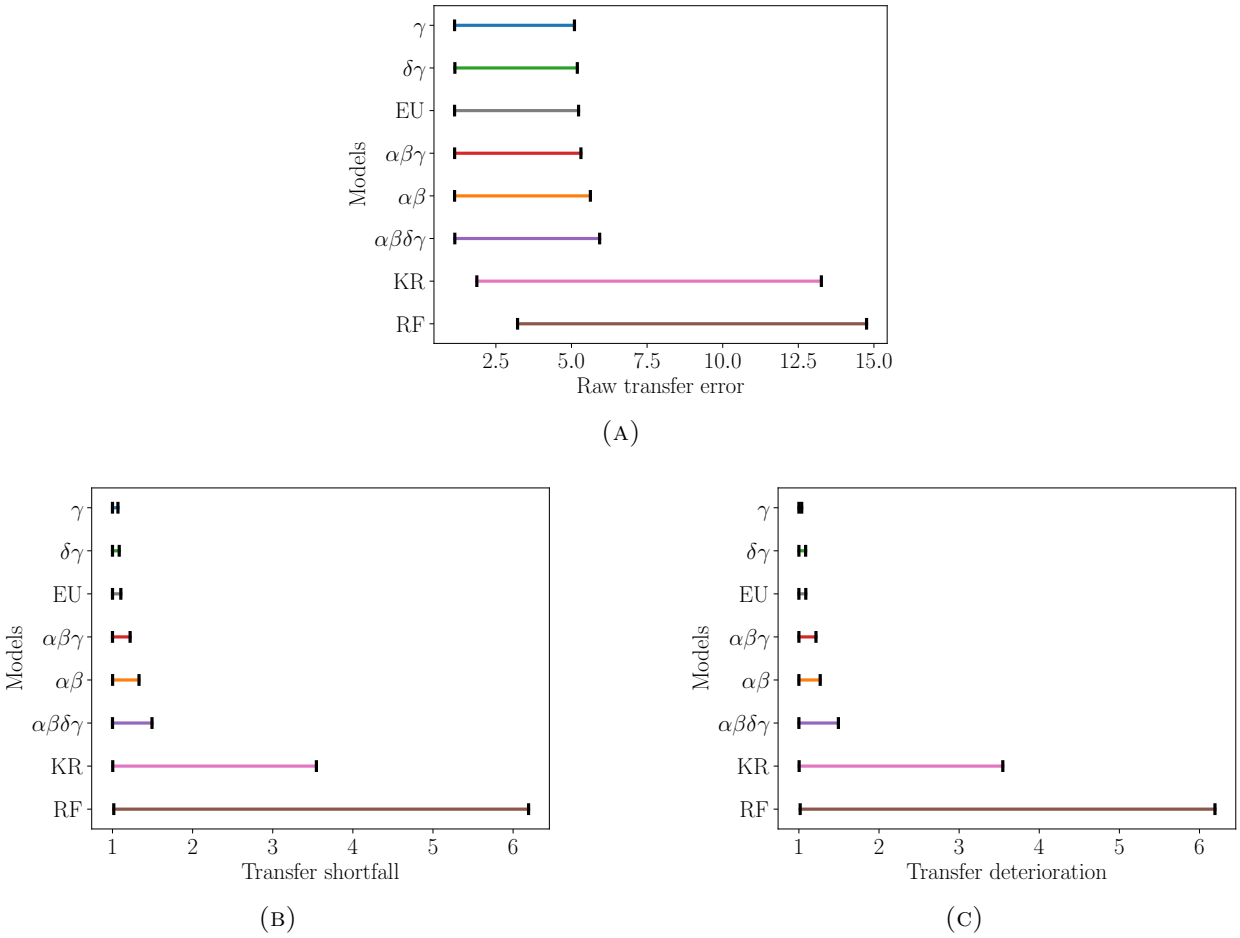


FIGURE 15. *68% (n=24, τ=0.95) forecast intervals using common lotteries in l'Haridon and Vieider (2019), with the choice of $r = 5$.*

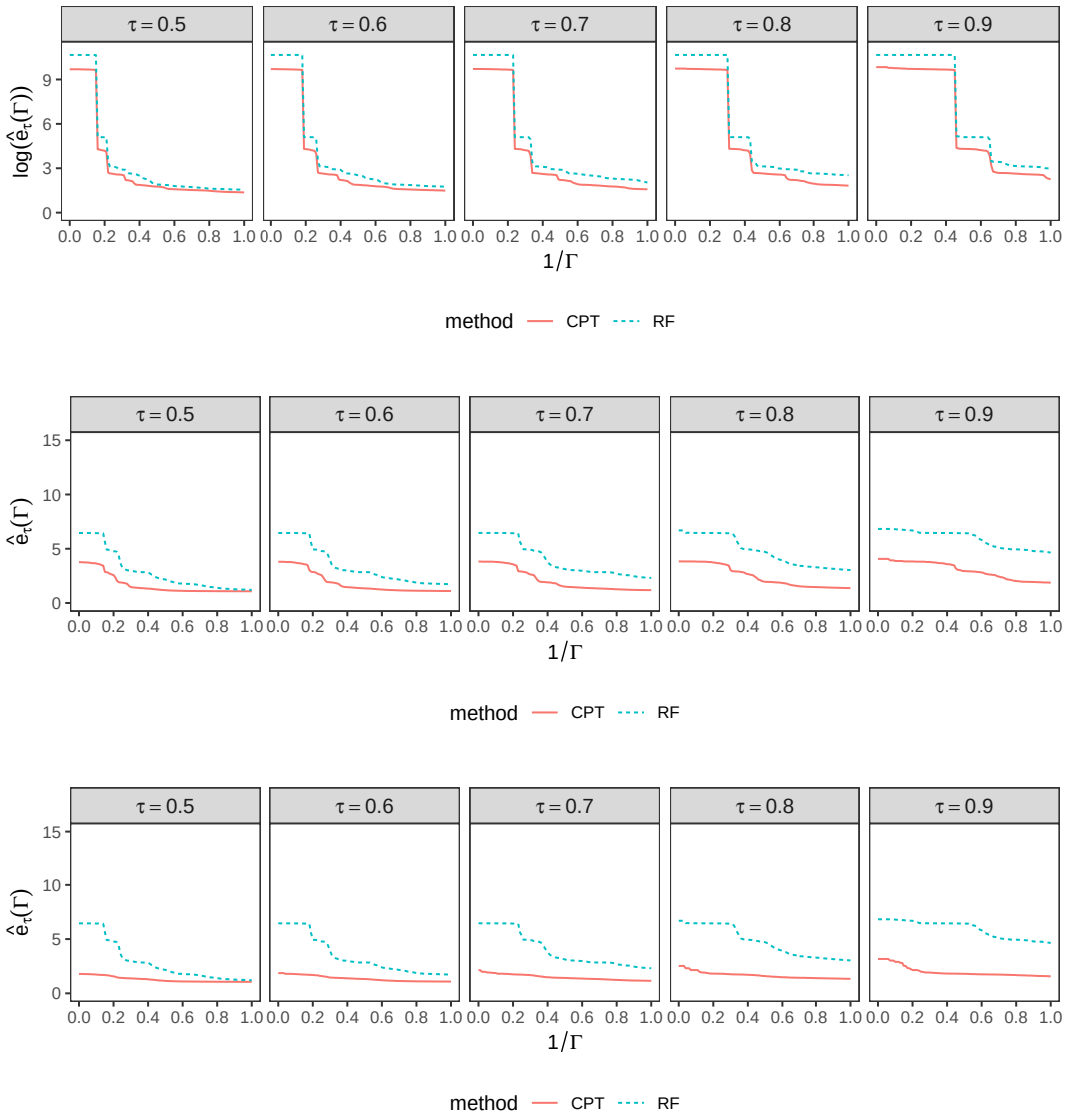


FIGURE 16. *The worst case upper prediction bound* $\hat{e}_\tau(\Gamma)$ *(as defined in (8)) for (a) raw transfer error, (b) transfer shortfall, and (c) transfer deterioration of CPT and RF as a function of* $\Gamma \in [1, \infty)$*, discretized at* $100/i (i = 0, 1, \ldots, 100)$*, at different quantiles* $\tau \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

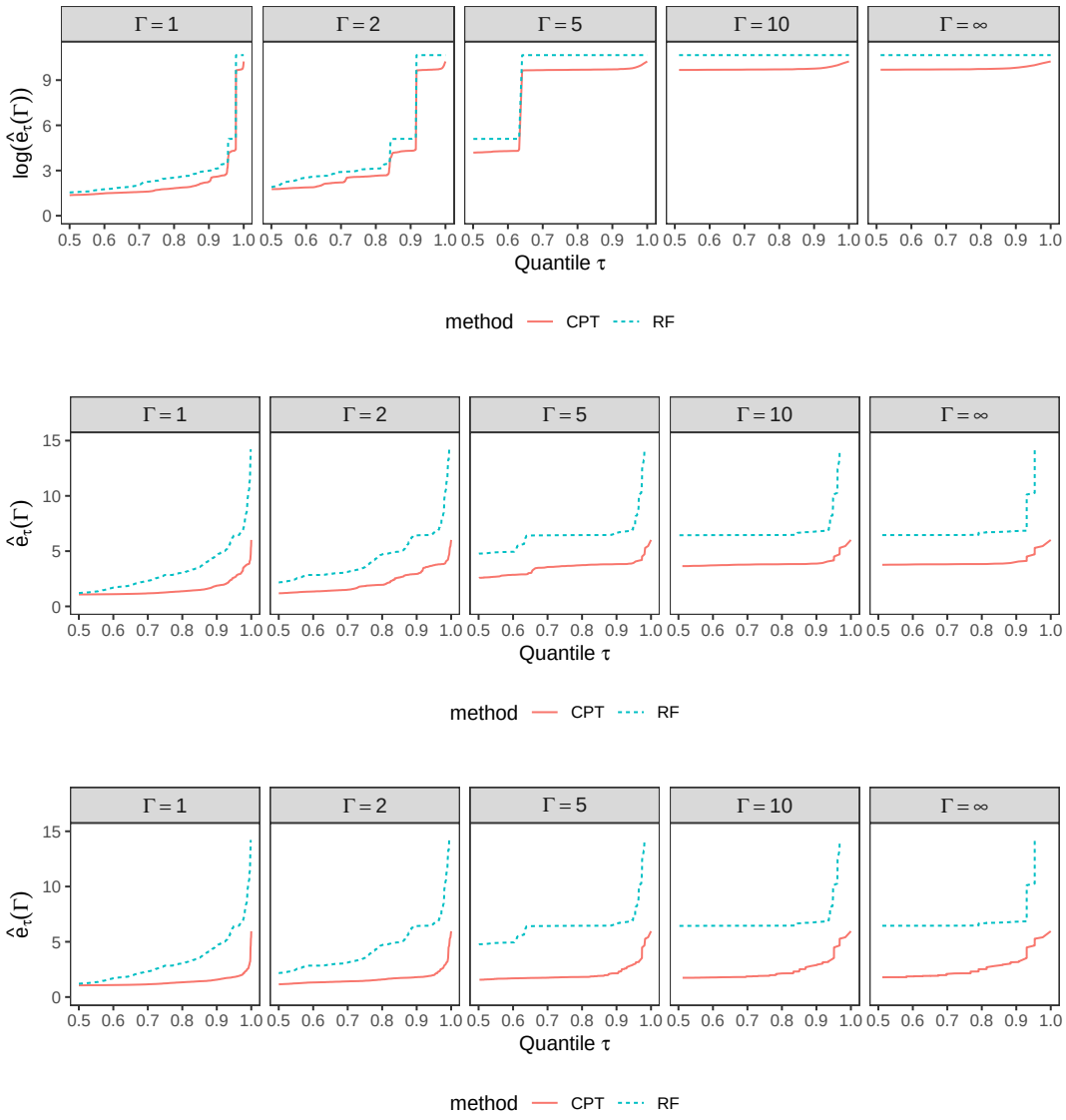


FIGURE 17. *The worst case upper prediction bound* $\hat{e}_\tau(\Gamma)$ *(as defined in (8)) for (a) raw transfer error, (b) transfer shortfall, and (c) transfer deterioration of CPT and RF as a function of* $\tau \in [0.5, 1]$ *without discretization for* $\Gamma \in \{1, 2, 5, 10, \infty\}$.