THE TRANSFER PERFORMANCE OF ECONOMIC MODELS

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

November 28, 2023

Abstract. Economists often estimate models using data from a particular domain, e.g. estimating risk preferences in a particular subject pool or for a specific class of lotteries. Whether a model's predictions extrapolate well across domains depends on whether the estimated model has captured generalizable structure. We provide a tractable formulation for this "out-of-domain" prediction problem and define the *transfer error* of a model based on how well it performs on data from 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. and use these intervals to compare the transferability of economic models and black box algorithms for predicting certainty equivalents. We find that in this application, the black box algorithms we consider outperform standard economic models when estimated and tested on data from the same domain, but the economic models generalize across domains better than the black-box algorithms do.

We thank Jiafeng Chen, Stefano DellaVigna, Nic Fishman, Andrei Iakovlev, Michael Jordan, Brandon Kaplowitz, Kyohei Okumura, 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, Harvard.
[‡]Department of Economics, MIT.
[#]Stanford Graduate School of Business.
[†]Department of Economics, Northwestern University.
[*]Department of Computer Science, Northwestern University.

1. Introduction

Economic models estimated on data from one context are often used to guide predictions and policy in a range of other contexts. For example, a model of information diffusion estimated on data of microfinance takeup in one Indian village may be used to guide policy decisions about the seeding of microfinance in another village, and a model of risk preferences estimated on willingness-to-pay data for certain lotteries may be used to predict willingness-to-pay for new lotteries. How should the generalizability of a model to new settings be assessed and predicted?

The generalizability of models is a classic concern in economics (Haavelmo, 1944; Pearl and Bareinboim, 2011; Tipton and Olsen, 2018; Chassang and Kapon, 2022), but has a new salience due to the rise of "black-box" machine learning prediction methods. Machine learning methods have been shown to out-predict economic models (Hartford et al., 2016; Plonsky et al., 2019; Hofman et al., 2021) and identify new interpretable regularities that the models do not capture (e.g., Fudenberg and Liang (2019), Peterson et al. 2021, Ludwig and Mullainathan (2023)). At the same time, other papers (e.g. Coveney et al. (2016); Athey (2017); Beery et al. (2018); Manski (2021)) have argued that structured economic models capture regularities that generalize well across domains, and may thus be more reliable for making predictions in new contexts. Whether economic models in fact generalize better is an important empirical question.

Our paper's contribution is twofold. First, we provide a tractable approach for evaluating cross-domain transfer performance based on techniques that generalize conformal inference (e.g. Vovk et al., 2005; Barber et al., 2021; Angelopoulos et al., 2022). In our statistical model, behavior in different economic contexts is governed by different distributions. Unlike previous approaches, we do not restrict these distributions to be close to one another, but instead assume that they are i.i.d. Under this assumption, we derive finite-sample forecast intervals for a large class of measures of transfer performance. These forecasts can be applied to evaluate economic models, regression models, and black box algorithms alike.[1] Second, we use these forecast intervals to compare the generalizability of economic models and black box machine learning methods in a specific economic application (predicting certainty equivalents for lotteries), and find that economic models generalize better.

[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.

Our conceptual framework, described in Section 2, is an extension of 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 drawn from the same distribution. We depart from this framework by supposing that the distribution of the data varies across a set of "domains," but that these domain-specific distributions are themselves drawn i.i.d. from a meta-distribution. As Section 2 explains, our results apply to a large class of measures for the transferability for a model, which we call *transfer errors*. Transfer errors can be used to evaluate the performance of many common empirical techniques, including using a model that is trained on a sample from one domain to predict in a sample from an as-yet unobserved domain, and asking whether a qualitative prediction based on estimated parameters from one sample will generalize to another.

Section 3 shows how to construct forecast intervals with guaranteed coverage probability for any transfer error, using a meta-data set of samples from already observed domains. Our approach is to split the observed domains in the meta-data set into training and test domains, estimate the parameters of the model on the samples from the training domains, and evaluate its transfer error on each of the test domains. Pooling these transfer errors across different choices of training and test domains yields an empirical distribution of transfer errors. We show that for every quantile τ, the interval bounded by the τ-th and $(1-\tau)$-th quantiles of the pooled transfer error is a valid forecast interval for the transfer error on a new, unseen domain. We also relax our i.i.d. sampling assumption, deriving a modified procedure for cases where the distributions in training domains are drawn i.i.d. from one distribution, while the distribution in the target domain is drawn from another. Both procedures (and all other methods described in this paper) are implemented in an `R` package (`transferUQ`), available on Github.[2]

Section 4 applies these procedures to compare the transferability of economic models and black box algorithms in a classic economic problem: predicting certainty equivalents for binary lotteries. The samples correspond to observations from different subject pools, so a model's transfer error describes how well it predicts outcomes in one subject pool when estimated on data from another. We evaluate two models of risk preferences, expected utility

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

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 in this setting? 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 when we restrict the analysis to a subset of our samples involving the same set of lotteries, the resulting forecast intervals are nearly identical across all of the prediction methods, so black box methods do not always transfer worse. Instead, 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). We leave to future work the question of what other properties of the transfer problem are relevant for determining which of black box algorithms and economic models are better suited to prediction.

1.1. **Related Literature.** This paper is situated at the intersection of several literatures in economics, computer science, and statistics. These literatures consider several related but distinct tasks: 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. Our results are most closely related to this third strand.

The first objective, synthesizing results across different domains, is a particular focus of the literature on meta-analysis.[3] 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 of her application. Unlike our approach, the predictive intervals reported in those papers are valid only under a parametric model for the distribution of effects across domains.

[3]See Card and Krueger (1995), Benartzi et al. (2017), DellaVigna and Pope (2019), Hummel and Maedche (2019), Bandiera et al. (2021), Imai et al. (2020) and Vivalt (2020) among others.

There is also a large literature that aims 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).[4] 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. In contrast, 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.

Finally, the literature on external validity studies the extent to which results obtained in one domain hold more generally. This paper does not focus on the generalizability of insights from randomized control trials (e.g. Deaton, 2010; Imbens, 2010) or laboratory experiments (e.g. Levitt and List, 2007; Al-Ubaydli and List, 2015), but instead on a model's generalizability across exchangeable domains.[5] Our use of exchangeability to construct bounds extends work on conformal inference (e.g. Vovk et al., 2005; Barber et al., 2021; Angelopoulos et al., 2022) by replacing the assumption of exchangeable observations with that of exchangeable domains.[6] Section 3.2 relaxes this assumption; our results there connect to the literature on sensitivity analysis (e.g. Aronow and Lee, 2013; Andrews and Oster, 2019; Nie et al., 2021; Sahoo et al., 2022).

Finally, we join a small but growing body of work regarding the relative value of economic models and black box algorithms, and how the two approaches can be combined for better prediction and explanation of social science phenomena (Athey and Imbens, 2016; Fudenberg and Liang, 2019; Agrawal et al., 2020). Several recent papers compare the predictiveness of black box algorithms with that of more structured economic models.[7] While black box methods are often very effective given a large quantity of data from the domain of interest, our results suggest that they may be less effective at transferring predictions across domains.

[4]Our problem corresponds to *homogeneous* domain generalization, where the set of outcomes $\mathcal{Y}$ is constant across domains, in contrast to *heterogenous* domain generalization, where the outcome set potentially varies across domains as well. There is also a related literature on *domain adaptation*, which aims to improve predictions when some data from the target domain is available – see Zhou et al. (2021).

[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).

[6]This also differentiates our work from the out-of-distribution prediction literature in computer science (Shen et al., 2021), which bounds expected transfer error when the test and training distributions are close.

[7]See e.g. Noti et al. (2016), Plonsky et al. (2017), Plonsky et al. (2019), Camerer et al. (2019), Fudenberg and Liang (2019), and Ke et al. (2020).

Hofman et al. (2021) organizes recent work in this area and concludes that more work is needed on the question "how well does a predictive model fit to one data distribution generalize to another?" Our paper takes an important step in this direction.

2. Framework

2.1. **Data generation process.** 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 $S = \{(x_i, y_i)\}_{i=1}^n$. We assume that samples $S_1, S_2, \ldots$ are generated i.i.d. from a 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 $\mathbb{N}$. That is, 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.[8] For example, if the samples are data from experiments conducted in different locations, then the i.i.d. assumption means that each location is drawn independently from a fixed distribution. This assumption rules out predictable patterns in how the joint distribution varies across samples, but allows for arbitrary relationships between the realized distributions; in particular, we do not constrain the distributions P_d, $P_{d'}$ to be close or share a common support over $\mathcal{X}$ or $\mathcal{Y}$.

The analyst has access to *metadata* consisting of n samples

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

A set $\mathbf{T}$ is drawn uniformly over all subsets of $\{1, \ldots, n\}$ of size $r < n$, and the samples $S_{\mathbf{T}} = (S_d)_{d \in \mathbf{T}}$ are used to train the model. We refer to these samples as the *training samples* or *training data*, and call a new sample S_{n+1} on which the model is evaluated the *target sample*. The choice of r, i.e., the number of training samples, depends on what the analyst wants to understand. In many contexts, including 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), economists transfer quantitative conclusions from a single domain to another. In this case $r = 1$, and the relevant question is whether extrapolating from one sample leads to good predictions at the new location. If instead experiments are run at $r > 1$ different locations and the observations

[8]This can be understood as a version of cluster sampling (Liang and Zeger, 1986; Bugni et al., 2023), where our 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.

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 location, and $r > 1$ is appropriate.

We provide a method for constructing a forecast interval for what we call *transfer errors*.

Definition 1 (Transfer Errors). A *transfer error* is any random variable that can be written as a function of the training $S_{\mathbf{T}}$, the target sample S_{n+1}, and potentially an independent noise variable.

To simplify notation, we write these transfer errors as $e_{\mathbf{T},n+1}$, although their values depend on the realization of the full vector $(S_d)_{i=1}^{n+1}$ as well as the realization of the set of samples $\mathbf{T}$ used for training.

This is a large class of variables that can measure many notions of transferability, including the ones we describe below.

2.2. **Model Transfer.** Suppose we use the training samples $S_{\mathbf{T}}$ to select a *prediction rule* $f_{S_{\mathbf{T}}} : \mathcal{X} \to \mathcal{Y}$, e.g., by estimating a parametric model or by fitting a black box algorithm.[9] If this rule is used to predict outcomes in a different sample, how accurate will those predictions be?

Example 1 (Certainty Equivalents). 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 CARA preferences. The firm then uses this estimated model to predict willingness-to-pay from consumers in California for a different set of lotteries. How accurate will those predictions be?

Example 2 (Network Diffusion). 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

[9]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 $f_{S_{\mathbf{T}}} = \rho(S_{\mathbf{T}})$ for the realized prediction rule.

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. How much less accurate will this prediction be than if the economist could re-estimate the structural model on data from this new village?

Fix a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, and define the error of prediction rule f on sample S to be

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

i.e., the average loss when using f to predict y given x. Our leading example of a transfer error (Definition 1) is the raw error of the model when it is estimated on the training samples and used to predict outcomes in the target sample:

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

Any normalization of (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 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.

Alternatively, we might normalize (1) 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

$$\frac{e(f_{S_\mathbf{T}}, S_{n+1})}{\min_{m \in \mathcal{M}} e\left(f^m_{S_{n+1}}, S_{n+1}\right)} \tag{3}$$

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 $\mathcal{M}$.[10] The main advantage of the specifications in (2) and (3) is that the raw error in (1) is very sensitive to how predictable y is given x

[10]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.

in the target sample, which may differ across domains but is not directly related to the transferability of the model.

2.3. **Parameter Transfer.** When a model has interpretable parameters, we may also 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 3 (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 well does this estimate proxy for the effectiveness of the same job training program when implemented at another location?

Example 4 (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.[11] To assess parameter variation, first fix a distance metric $d(\theta, \theta')$ (e.g., Euclidean distance) to assess how different two parameter vectors θ and θ' are. Then the transfer error

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

tells us how far apart 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

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

[11] If there are ties, break them arbitrarily

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

2.4. **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 5 (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 6 (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 7 (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 8 (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.

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.

3. Theoretical Results

3.1. **Main Results.** Our goal is to develop forecast intervals for any transfer error $e_{\mathbf{T},n+1}$ (Definition 1). That is, we will provide 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. In many applications only a limited number of domains will be observed, so our focus is on finite-sample results.

For any choice of training samples $\mathcal{T} \subseteq \{1, \ldots, n\}$ and test sample $d \in \{1, \ldots, n\} \backslash \mathcal{T}$ from $\mathbf{M}$, let $e_{\mathcal{T},d}^{\mathbf{M}}$ be the (observed) transfer error from samples in $\mathcal{T}$ to sample d.[12] Further, let $\mathbb{T}_{s,t}$ denote the set of all vectors of length s that consist of distinct elements from $\{1, \ldots, t\}$, so that (for example) $\mathbb{T}_{r+1,n}$ corresponds to all possible choices of r training samples and a single test sample from the metadata $\{1, \ldots, n\}$. Then

$$F_{\mathbf{M}} = \frac{(n-r-1)!}{n!} \sum_{(\mathcal{T},d) \in \mathbb{T}_{r+1,n}} \delta_{e_{\mathcal{T},d}^{\mathbf{M}}} \tag{4}$$

is the empirical distribution of transfer errors in the pooled sample $\left\{e_{\mathcal{T},d}^{\mathbf{M}} : (\mathcal{T}, d) \in \mathbb{T}_{r+1,n}\right\}$ as we vary the samples in the metadata that are used for training and testing. (Throughout we use δ to denote the Dirac measure).

[12] For example, if we use the specification of transfer error in (1), then $e_{\mathcal{T},d}^{\mathbf{M}}$ is the raw error of the model estimated on samples $S_{\mathcal{T}}$ and used for prediction on sample 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 $\bar{e}_\tau^{\mathbf{M}} \equiv \overline{Q}_\tau(F_{\mathbf{M}})$ and $\underline{e}_\tau^{\mathbf{M}} \equiv \underline{Q}_{1-\tau}(F_{\mathbf{M}})$ be the τth upper quantile and $(1-\tau)$th lower quantile of the empirical distribution of transfer errors in the pooled sample.

These quantiles can be used to construct a valid forecast interval for the transfer error on the target sample:

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

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{e}_\tau^{\mathbf{M}}\right) \geq \tau\left(\frac{n-r}{n+1}\right), \tag{5}$$

and

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

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-$\left(2\tau\left(\frac{n-r}{n+1}\right) - 1\right)$ forecast interval for $e_{\mathbf{T},n+1}$.

The parameters r and τ are choice variables. The size of τ influences the width of the forecast interval, where larger choices of τ lead to wider forecast intervals with higher confidence guarantees. The choice of r determines how many samples in the meta-data are used for training versus testing. Larger choices of r mean that the model will be estimated on a larger quantity of data, but we will have fewer samples on which to evaluate the performance of the estimated model.

The next result shows that the guarantees in Proposition 1 are tight to $O(1/n)$.

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*

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

and

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

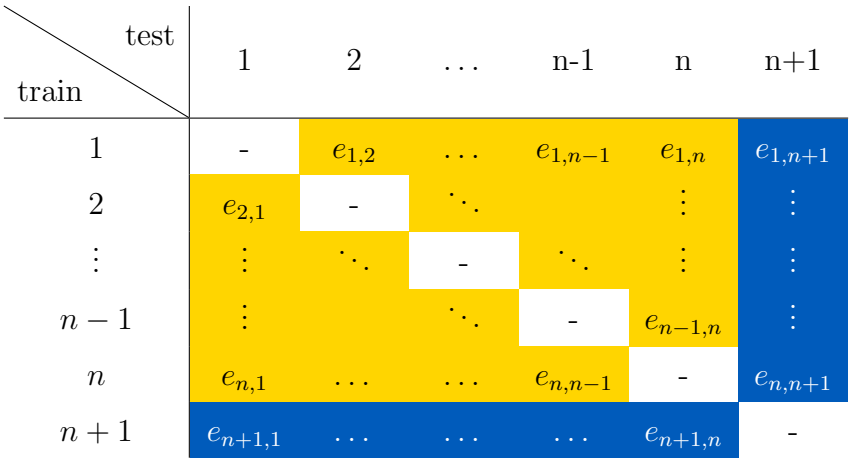


FIGURE 1. *Matrix of transfer errors when training on one domain (row) and testing on another (column).*

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

$$\left\{e^{\mathbf{M}}_{\mathcal{T},d} : (\mathcal{T}, d) \in \mathbb{T}_{r+1,n+1}\right\}. \tag{6}$$

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

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq e^*_\tau \mid \{S_d\}_{d=1}^{n+1}\right) \geq \tau. \tag{7}$$

In the case $r = 1$ where precisely one sample is used for training, the set of pooled errors (6) is the shaded cells in Figure 1 (either yellow or blue), and the inequality in (7) says that the probability that the value of a randomly drawn cell falls below the τth upper quantile of cells is at least τ.

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

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

Applying the law of iterated expectations (with respect to the sample $\{S_i\}_{i=1}^{n+1}$) yields the one-sided forecast interval in (5), and a similar argument yields the two-sided forecast interval.

3.2. **Relaxing the i.i.d. Assumption.** 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, if the samples in the metadata are from experiments run at different locations, the i.i.d. assumption fails if there is selection bias over where experiments are run.[13] We now relax this assumption 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. 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}.

We again construct a forecast interval for $e_{\mathbf{T},n+1}$ using the pooled sample of transfer errors across samples in the metadata, $\left\{e^{\mathbf{M}}_{\mathcal{T},d} : (\mathcal{T}, d) \in \mathbb{T}_{r+1,n}\right\}$, giving different probabilities to each $e^{\mathbf{M}}_{T,d}$ instead of uniform weights. Under the i.i.d. assumption, each sample in the metadata is equally representative of the training and target distributions. When we relax that assumption, then 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

$$W_d = \frac{(n-r-1)!}{(n-1)!} \frac{\omega(S_d)}{\sum_{j=1}^{n} \omega(S_j)}. \tag{8}$$

To interpret this quantity, consider an alternative data-generating process for the metadata (mimicking the larger environment) 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

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

[13] One way this selection bias can arise is if the observed sites were chosen based on characteristics which are correlated with effect sizes, as Allcott (2015) found in the Opower energy conservation experiments.

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[14]

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

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

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

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 (4).

Definition 5 (Quantiles of $F^{\omega}_{\mathbf{M}}$). For any likelihood ratio $\omega(\cdot)$ and quantile $\tau \in (0,1)$, define $\bar{e}^{\mathbf{M},\omega}_{\tau} = \overline{Q}_{\tau}(F^{\omega}_{\mathbf{M}})$ and $\underline{e}^{\mathbf{M},\omega}_{\tau} = \underline{Q}_{1-\tau}(F^{\omega}_{\mathbf{M}})$ 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

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

This result strictly generalizes Proposition 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.

If the analyst does not know ω, but can uniformly upper and lower bound it across samples, the following result applies.

[14]This is known as weighted exchangeability; see Tibshirani et al. (2019).

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.

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

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

and

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

When there is no natural bound for w, it can still be possible in some cases to compare the transferability of two models $i = 1, 2$. Let $\left[\underline{e}_{1,\tau}^{\mathbf{M},\omega}, \bar{e}_{1,\tau}^{\mathbf{M},\omega}\right]$ and $\left[\underline{e}_{2,\tau}^{\mathbf{M},\omega}, \bar{e}_{2,\tau}^{\mathbf{M},\omega}\right]$ denote the respective forecast intervals. For each model i, let

$$\bar{e}_{i,\tau}^{\mathbf{M}}(\Gamma) = \sup_{\omega \in \mathcal{W}(\Gamma)} \bar{e}_{i,\tau}^{\mathbf{M},\omega} \tag{9}$$

be the worst-case upper bound across likelihood ratios in $\mathcal{W}(\Gamma)$. As shown in Appendix P this quantity can be computed in $O(n^{r+1})$ time.

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

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

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

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

Many decision rules will not be comparable under either of these definitions, but both of these orders do have bite in our application. The even stronger requirement that $\bar{e}_{1,\tau}^{\mathbf{M},\omega} \leq \underline{e}_{2,\tau}^{\mathbf{M},\omega}$,

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.[15]

3.3. **Discussion. What is a domain?** The specified domains determine the transfer question the analyst is interested in and the content of the i.i.d. sampling assumption. Suppose, for instance, that the meta-data consists of experimental results from multiple papers, where each paper reports results from experiments at multiple sites. If each site is treated as a separate domain, then i.i.d. sampling corresponds to drawing further sites, some from the papers already observed and others from as-yet-unobserved papers. Alternatively, if each paper is a separate domain, then i.i.d. sampling corresponds to drawing new papers, each with its own sites.

Evaluating transfer performance versus choosing a cross-domain prediction rule. This paper provides forecast intervals for the transfer errors of a given economic model or black box algorithm. A complementary question is how to identify models and algorithms that lead to better transfer performance. We do not pursue that question here, but note that many of the approaches that have been proposed (such as distributionally robust optimization (Rahimian and Mehrotra, 2019)), can be evaluated using our methods.

Number of Domains Versus Observations. The meta-data involve both a finite number of observed domains and a finite number of observations per domain. These two sources of finiteness enter into our results in different ways. Increasing the number of observations per domain changes the distribution of $e_{\mathbf{T},n+1}$: In the limit of infinitely many observations per domain, 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, the error $e_{\mathbf{T},n+1}$ measures how well an imperfectly estimated model transfers. In contrast, increasing the number of observed domains (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.

[15]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, which seems unlikely.

4. Application

To illustrate our methods, we evaluate the transferability of predictions of certainty equivalents for binary lotteries, where the domains correspond to different subject pools. Section 4.1 describes our metadata, and Section 4.2 describes the decision rules we consider. Section 4.3 conducts "within-domain" out-of-sample tests, where the training and test data are drawn from the same domain. Section 4.4 constructs forecast intervals for three different kinds of transfer error.

4.1. **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. In Online Appendix Q.2, we repeat our analysis with each paper treated as a separate domain, and show that the results are qualitatively similar.[16] Our samples range in size from 72 observations to 8906 observations, with an average of 2752.7 observations per sample.[17] We convert all prizes to dollars using purchasing power parity exchange rates (from OECD 2023) in the year of the paper's publication

Within each sample, observations take the form $(z_1, z_2, p; y)$, where z_1 and z_2 denote the possible prizes of the lottery (and we adopt the convention that $|z_1| > |z_2|$), p is the probability of z_1, and y is the reported certainty equivalent by a given subject. Thus our feature space is $\mathcal{X} = \mathbb{R} \times \mathbb{R} \times [0, 1]$, the outcome space is $\mathcal{Y} = \mathbb{R}$, and a prediction rule is any mapping from binary lotteries into predictions of the reported certainty equivalent. We use squared-error loss $\ell(y, y') = (y - y')^2$ to evaluate the error of the prediction, but for ease of interpretation we report results in terms of root-mean-squared error, which puts the errors in the same units as the prizes.[18] Since different subjects report different certainty equivalents for the same lottery, the best achievable error is generally bounded away from zero.

4.2. **Models and black boxes.** We consider two parametric economic models of certainty equivalents.

[16] In both cases, the domains sometime combines observations from different experimental treatments. In Etchart-Vincent and l'Haridon (2011), we pool reported certainty equivalents across three payment conditions: real losses, hypothetical-losses, and losses-from-initial-endowment.

[17] Online Appendix Q.1 describes our data sources in more detail.

[18] 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 $g(x) = \sqrt{x}$.

Expected Utility. First we consider an expected utility agent with a CRRA utility function parameterized by $\eta \geq 0$. 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.

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

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

where

$$v(z) = \begin{cases} z^\alpha & \text{if } z \geq 0 \\ -(-z)^\beta & \text{if } z < 0 \end{cases} \tag{10}$$

is a value function for money, and

$$w(p) = \frac{\delta p^\gamma}{\delta p^\gamma + (1-p)^\gamma} \tag{11}$$

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 (10); 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.[19] 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 Boxes. Finally, we consider two machine learning algorithms. First, we train a *random forest*, which is an ensemble learning method consisting of a collection of decision trees.[20] Second, we train a *kernelized ridge regression* model, 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 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, in settings with multiple training domains one can cross-validate across training domains. This procedure is not relevant to our analysis in Section 4.4, which considers transfer from a single training domain to a single test domain, but we use it in Appendix Q.7 when we consider multiple training domains. 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

[19]See Fehr-Duda and Epper (2012) for further discussion of these different parametric forms, and some non-nested versions that have been used in the literature.

[20]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.

values. Thus in our main analysis with a single training domain, we set all hyper-parameters to default values.[21]

4.3. **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.[22] 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. Figure 2 reports the CDF of tenfold cross-validated errors for the random forest, kernel regression, EU, and CPT.[23]

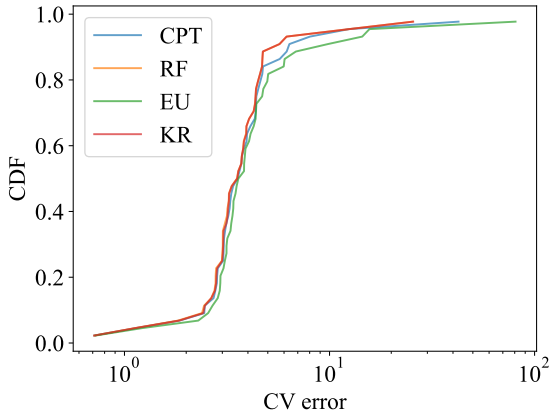


FIGURE 2. *CDF of Cross-Validated Errors. (The CDFs for kernel regression and random forest overlap.)*

To obtain simpler summary statistics 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: 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.[24]

[21]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.
[22]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.
[23]Online Appendix Q.3 shows that the CDFs for in-sample errors are likewise very close.
[24]The numbers in Table 1 are very similar if we normalized by the kernel regression error instead.

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

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

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

4.4. **Transfer error.** We use the results in Section 3 to construct forecast intervals for the three specifications of transfer error defined in (1), (2), and (3), which we will subsequently call *raw transfer error*, *transfer deterioration*, and *transfer shortfall*.

In our meta-data there are $n = 44$ domains, and we choose $r = 1$ of these to use as the training domain. This choice of r corresponds to the question, "If I draw one domain at random, and then try to generalize to another domain, how well do I do?"

Figure 3 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.[25] Applying Proposition 1, these are 81% 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 Q.5, including a 91% forecast interval.

[25]See Table 5 in Appendix Q.4 for the exact numbers.

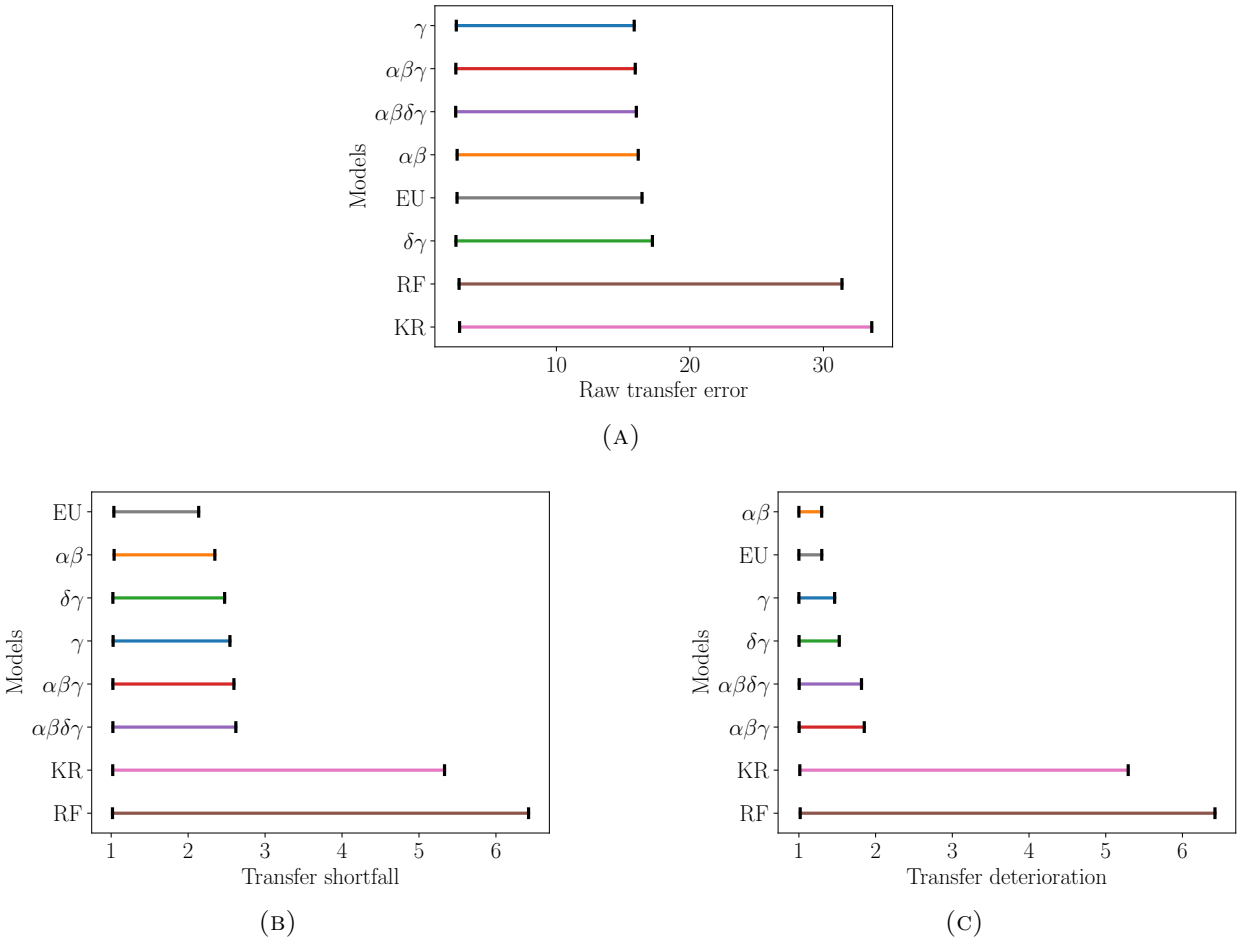


FIGURE 3. *81% (n=44, $\tau = 0.95$) forecast intervals for (a) raw transfer error, (b) transfer shortfall (with $\mathcal{R}$ consisting of the decision rules shown in the figure), and (c) transfer deterioration.*

Our main takeaway from Figure 3 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 3 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 3, 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 5 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 Q.5 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 Q.6 provides 81% 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 Q.7 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 Q.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 3.2 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 the 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 for a given degree of relaxation we are comfortable comparing the upper bound for one method to the upper bound for the other. In Appendix Q.8, we provide a more detailed view of worst-case-upper-dominance by plotting $\bar{e}^{\mathbf{M}}_{\tau}(\Gamma)$ 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 $\tau \geq 0.647$.

4.5. **Do black boxes transfer poorly because they are too flexible?** One tempting explanation of our empirical findings is that because the black boxes are more flexible than the economic models, they can learn idiosyncratic details that do not generalize across subject pools, such as that some subject pools tend to value lotteries depending on specific digits they contain.[26] This would lead the black boxes to have better within-domain prediction for those subject pools, but worse transfer performance if the regularity does not generalize across subject pools.

While the flexibility of black box algorithms is likely an important determinant of their transfer performance, there are at least two reasons this cannot be a complete explanation of our results. First, the flexibility gap between the black boxes and economic models is not

[26]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. Such a 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.

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).

Second, black boxes do not always transfer worse. One of the papers we use is based on samples of certainty equivalents from 30 countries (l'Haridon and Vieider, 2019). Crucially, of the 30 samples from this paper, 29 samples report certainty equivalents for the same 27 lotteries, and the remaining sample reports certainty equivalents for 23 of those lotteries. We repeat our analysis using these 30 samples as the domains, and find that the forecast intervals for raw transfer error are indistinguishable across the prediction methods (Panel (a) of Figure 4). 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 more similar than in the original data.

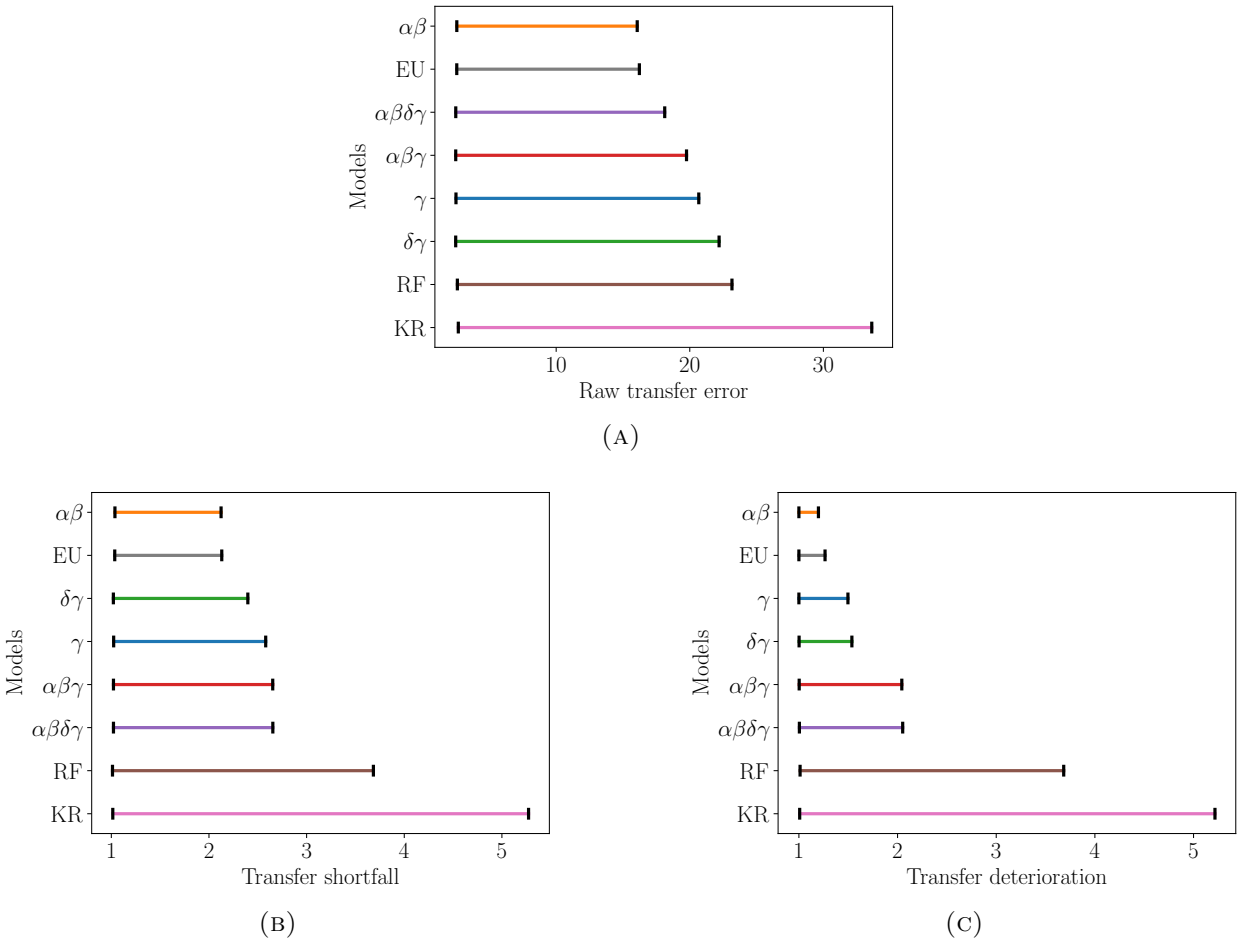


FIGURE 4. *78% (n=30, τ=0.95) forecast intervals using samples in l'Haridon and Vieider (2019).*

These observations show that flexible prediction methods do not always transfer poorly, although they do perform poorly in certain kinds of transfer prediction tasks. The next section explores one potential explanation.

4.6. **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*).

Our findings in Figure 4 suggest that black boxes do as well as economic models at transfer prediction when the marginal distribution over features P_X is unchanging across samples. Intuitively, when the relevant feature vectors are held constant across samples, 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.

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

4.7. **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
- 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*

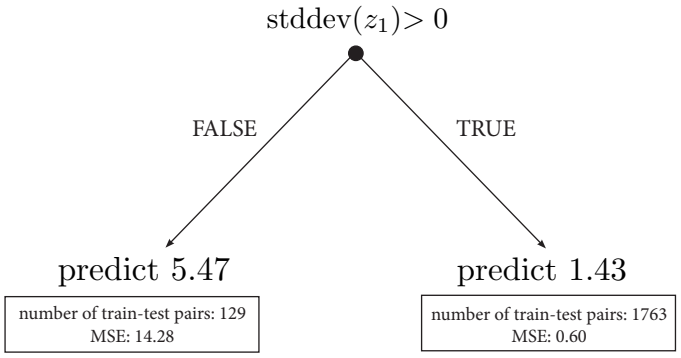


FIGURE 5. *Best 1-split decision tree based on training and test features.*

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 5 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,[27] 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.

5. 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 5.2 provides confidence intervals for quantiles of the transfer error distribution, and Section 5.3 provides a confidence interval for the expected transfer error. Since these quantities can be perfectly recovered given data from an infinite number of domains, we expect the lengths of these intervals to vanish as the number of observed domains grows large, unlike the forecast intervals from Section 3.

5.1. **Preliminary Lemma.**

We start by establishing a bound that will be useful in the subsequent construction of confidence intervals. Let

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

[27] 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.

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

Definition 9. 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

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

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

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

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

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

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

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

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

$$\mathbb{P}(U_\beta \leq x) \leq B_{n,r+1}(x, \beta) \qquad \mathbb{P}(U_\beta \geq x) = \mathbb{P}(1 - U_\beta \leq 1 - x) \leq B_{n,r+1}(1-x, 1-\beta). \quad (12)$$

Definition 10. For any quantile $\beta \in (0,1)$ and confidence level $1 - \alpha \in (0,1)$, let $\hat{u}^+_\beta(\alpha) = \inf\{u : B_{n,r+1}(u;\beta) \geq \alpha\}$ and $\hat{u}^-_\beta(\alpha) = \sup\{u : B_{n,r+1}(1-u; 1-\beta) \geq \alpha\}$. Further define $\hat{q}^{\mathrm{L}}_\beta(\alpha) \equiv \min\left\{q : \varphi(q, \mathbf{M}) \geq \hat{u}^+_\beta(\alpha)\right\}$ and $\hat{q}^{\mathrm{U}}_\beta(\alpha) \equiv \max\left\{q : \varphi(q, \mathbf{M}) \leq \hat{u}^-_r(\alpha)\right\}$.

Since $B_{n,r+1}(u;\cdot)$ is right-continuous in u, it follows from (12) that $\mathbb{P}(U_\beta < \hat{u}^+_\beta(\alpha)) \leq \alpha$ and $\mathbb{P}(U_\beta > \hat{u}^-_\beta(\alpha)) \leq \alpha$. Since $\varphi(q, \mathbf{M})$ is monotonically increasing in q, the event $\{U_\beta < \hat{u}^+_\beta(\alpha)\}$ is equivalent to $\{q_\beta < \hat{q}^{\mathrm{L}}_\beta(\alpha)\}$, while $\{U_\beta > \hat{u}^-_\beta(\alpha)\}$ is equivalent to $\{q_\beta > \hat{q}^{\mathrm{U}}_\beta(\alpha)\}$. This yields:

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

Figure 6 applies Proposition 2 to construct two-sided 81% confidence interval for the median raw transfer error, median transfer shortfall, and median transfer deterioration. As in Figure 3, these confidence intervals are substantially wider for the black box algorithms, and have higher upper bounds.

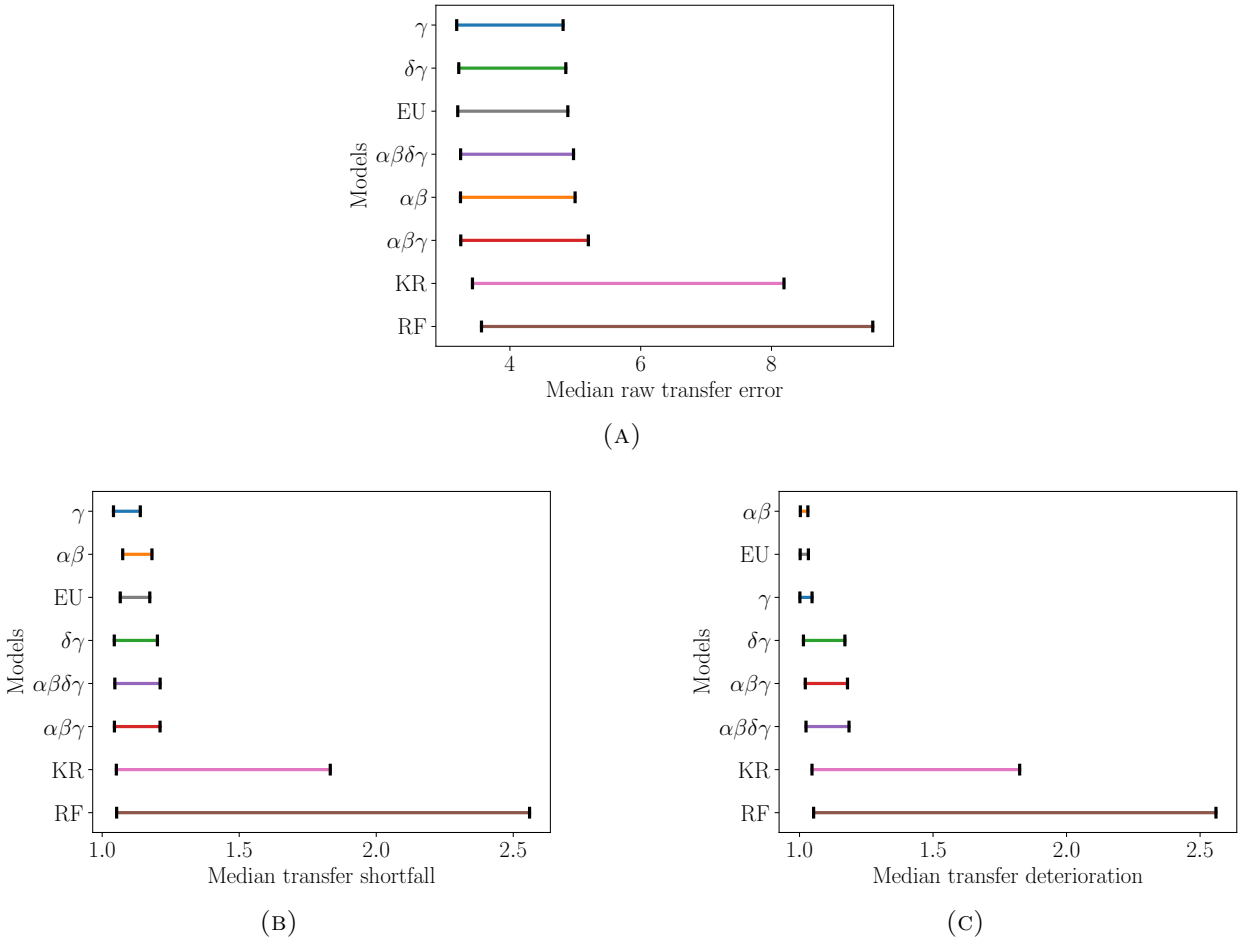


FIGURE 6. *81% confidence intervals for the median of (a) raw transfer error, (b) transfer shortfall, and (c) transfer deterioration.*

5.3. **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 1 implies that $\mathbb{P}(U \leq x) \leq B_{n,r+1}(x,\mu)$ and $\mathbb{P}(U \geq x) \leq B_{n,r+1}(1-x,1-\mu)$ for all $x \in \mathbb{R}$.

Definition 11. For any confidence guarantee $1-\alpha \in (0,1)$, let $\hat{\mu}^+(\alpha) = \sup\{\mu : B_{n,r+1}(U;\mu) \geq \alpha\}$ and $\hat{\mu}^-(\alpha) = \inf\{\mu : B_{n,r+1}(1-U;1-\mu) \geq \alpha\}$.

It follows from (12) that $\mathbb{P}(U < \hat{u}^+(\alpha)) \leq \alpha$ and $\mathbb{P}(U > \hat{u}^-(\alpha)) \leq \alpha$, which implies:

Proposition 3. *If $e_{\mathcal{T},d} \in [0,1]$ almost surely, then $\mathbb{P}\left(\mu \leq \hat{\mu}^+(\alpha)\right) \geq 1-\alpha$ and $\mathbb{P}\left(\mu \in [\hat{\mu}^-(\alpha/2), \hat{\mu}^+(\alpha/2)]\right) \geq 1-\alpha$.*

Figure 7 applies this result to construct two-sided 81% confidence intervals for the transfer errors we considered in Section 4.4. 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.

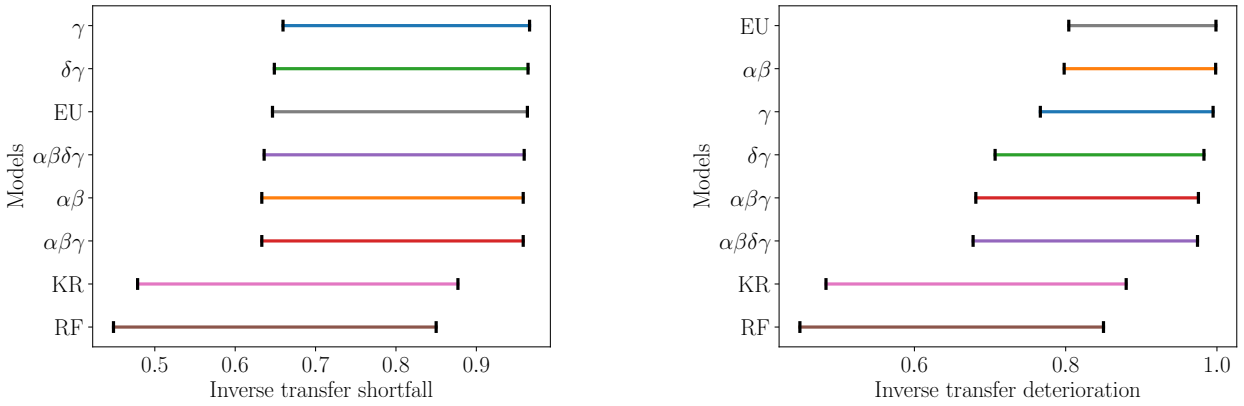


FIGURE 7. *81% forecast intervals for (a) expected inverse transfer shortfall, (b) expected inverse transfer deterioration.*

6. Conclusion

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

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.

Agrawal, M., J. C. Peterson, and T. L. Griffiths (2020): "Scaling up psychology via Scientific Regret Minimization," *Proceedings of the National Academy of Sciences*, 117, 8825–8835.

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.

Angelopoulos, A. N., S. Bates, A. Fisch, L. Lei, and T. Schuster (2022): "Conformal risk control," *arXiv preprint arXiv:2208.02814*.

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.

Athey, S. and G. Imbens (2016): "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360.

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.

Barber, R. F., E. J. Candes, A. Ramdas, and R. J. Tibshirani (2021): "Predictive inference with the jackknife+," *The Annals of Statistics*, 49, 486–507.

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*.

Benartzi, S., J. Beshears, K. L. Milkman, C. R. Sunstein, R. H. Thaler, M. Shankar, W. Tucker-Ray, W. J. Congdon, and S. Galing (2017): "Should Governments Invest More in Nudging?" *Psychological Science*, 28, 1041–1055.

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.

Chassang, S. and S. Kapon (2022): "Designing Randomized Controlled Trials with External Validity in Mind," Working Paper.

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

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.

Fehr-Duda, H. and T. Epper (2012): "Probability and Risk: Foundations and Economic Implication of Probability-Dependent Risk Preferences," *Annual Review of Economics*, 4, 567–593.

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., 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.

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.

Haavelmo, T. (1944): "The Probability Approach in Econometrics," *Econometrica*, 12, iii–115.

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

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.

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.

Hummel, D. and A. Maedche (2019): "How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies," *Journal of Behavioral and Experimental Economics*, 80, 47–58.

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.

Ke, S., C. Zhao, Z. Wang, and S.-L. Hsieh (2020): "Behavioral Neural Networks," Working Paper.

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.

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.

Nie, X., G. Imbens, and S. Wager (2021): "Covariate Balancing Sensitivity Analysis for Extrapolating Randomized Trials across Locations," Working Paper.

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 (2011): "Transportability across studies: A formal approach," Tech. rep., UCLA Dept of Computer Science.

——— (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.

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.

Plonsky, O., I. Erev, T. Hazan, and M. Tennenholtz (2017): "Psychological forest: Predicting human behavior," *AAAI Conference on Artificial Intelligence*, 31, 656–662.

Rahimian, H. and S. Mehrotra (2019): "Distributionally Robust Optimization: A Review," Working Paper.

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

Shen, Z., J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui (2021): "Towards Out-Of-Distribution Generalization: A Survey," *CoRR*, abs/2108.13624.

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.** Since $\mathbf{T}$ is a random subset of $\{1, \ldots, n\}$ that is independent of $\{S_1, \ldots, S_{n+1}\}$, $\mathbb{P}\left(e_{\mathbf{T},n+1} \in A\right) = \mathbb{P}(e_{(1,\ldots,r),n+1} \in A)$ for any event A that is independent of $\mathbf{T}$. Thus, it suffices to prove Proposition 1 by replacing $\mathbf{T}$ with $(1, \ldots, r)$. We start by proving the one-sided guarantee:

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

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. Ranging over all possible choices of r (ordered) training environments and a single test environment from $\{1, \ldots, n+1\}$ yields the multiset of transfer errors[28]

$$\mathcal{C} = \left\{f(d_1, \ldots, d_{r+1}) : (d_1, \ldots, d_{r+1}) \in \mathbb{T}_{r+1,n+1}\right\}, \tag{A.1}$$

which has size $(n+1)!/(n-r)!$. The subset of these transfer errors that don't use sample k for either training or testing is

$$\mathcal{C}_{-k} = \left(f(d_1, \ldots, d_{r+1}) : (d_1, \ldots, d_{r+1}) \in \mathbb{T}_{r+1,n+1}, d_j \neq k, j = 1, \ldots, r+1\right),$$

which has size $n!/(n-r-1)!$. For any $\tau \in [0,1]$, let $\bar{E}_\tau^*$ denote the $\lceil \tau n!/(n-r-1)! \rceil$-th smallest element in $\mathcal{C}$, and let $\bar{E}_{k,\tau}$ denote the $\lceil \tau n!/(n-r-1)! \rceil$-th smallest element in $\mathcal{C}_{-k}$. Since $\mathcal{C}_{-k} \subseteq \mathcal{C}$,

$$\bar{E}_{k,\tau} \geq \bar{E}_\tau^* \text{ for all } k \text{ and } \tau. \tag{A.2}$$

[28]A multiset is a set-like, unordered collection where repeated values are multiply counted.

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

$$\left(e_{(1,\ldots,r),n+1}, \bar{e}_\tau\right) \mid \mathcal{F} \stackrel{d}{=} \left(f(\boldsymbol{\pi}^*(1), \ldots, \boldsymbol{\pi}^*(r), \boldsymbol{\pi}^*(n+1)), \bar{E}_{\boldsymbol{\pi}^*(n+1),\tau}\right)$$

where $\boldsymbol{\pi}^*$ denotes a random permutation drawn from the uniform distribution over all permutations on $\{1, \ldots, n+1\}$. Thus

$$\mathbb{P}\left(e_{(1,\ldots,r),n+1} \leq \bar{e}_\tau \mid \mathcal{F}\right) = \mathbb{P}_{\boldsymbol{\pi}^*}\left(f(\boldsymbol{\pi}^*(1), \ldots, \boldsymbol{\pi}^*(r), \boldsymbol{\pi}^*(n+1)) \leq \bar{E}_{\boldsymbol{\pi}^*(n+1),\tau}\right). \tag{A.3}$$

It follows that

$$\begin{aligned}
\mathbb{P}\left(e_{(1,\ldots,r),n+1} \leq \bar{e}_\tau \mid \mathcal{F}\right) &\geq \mathbb{P}_{\boldsymbol{\pi}^*}\left(f(\boldsymbol{\pi}^*(1), \ldots, \boldsymbol{\pi}^*(r), \boldsymbol{\pi}^*(n+1)) \leq \bar{E}^*_\tau\right) && \text{by (A.2) and (A.3)} \\
&= \frac{(n-r)!}{(n+1)!} \sum_{(d_1,\ldots,d_{r+1}) \in \mathbb{T}_{r+1,n+1}} \mathbb{I}(f(d_1, \ldots, d_{r+1}) \leq \bar{E}^*_\tau) \\
&= \frac{(n-r)!}{(n+1)!} \sum_{f \in \mathcal{C}} \mathbb{I}(f \leq \bar{E}^*_\tau) \\
&\geq \frac{(n-r)!}{(n+1)!} \left\lceil \tau \frac{n!}{(n-r-1)!} \right\rceil \geq \tau \left(1 - \frac{r+1}{n+1}\right).
\end{aligned}$$

The one-sided guarantee then follows from the law of iterated expectation.

Turning to the two-sided guarantee, we first notice that the one-sided guarantee implies

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

and, by symmetry,

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

Thus,

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

This completes the proof of Proposition 1.

To prove Claim 1, define $\bar{E}'_\tau$ to be the $\lfloor (1-\tau)n!/(n-r-1)! + 1 \rfloor$-th largest element in $\mathcal{C}$ as defined in (A.1). Since $|\mathcal{C}_{-k}| = n!/(n-r-1)!$, the quantity $\bar{E}_{k,\tau}$ is also the $\lfloor (1-\tau)n!/(n-r-1)! + 1 \rfloor$-th largest element in $\mathcal{C}_{-k}$. So $\bar{E}_{k,\tau} \leq \bar{E}'_\tau$ for any k and τ.

By (A.3), $\mathbb{P}\left(e_{(1,\ldots,r),n+1} \leq \bar{e}_\tau \mid \mathcal{F}\right) \leq \mathbb{P}_{\boldsymbol{\pi}^*}\left(f(\boldsymbol{\pi}^*(1), \ldots, \boldsymbol{\pi}^*(r), \boldsymbol{\pi}^*(n+1)) \leq \bar{E}'_\tau\right)$

$$= \frac{(n-r)!}{(n+1)!} \sum_{(d_1,\ldots,d_{r+1})\in\mathbb{T}_{r+1,n+1}} \mathbb{I}(f(d_1,\ldots,d_{r+1}) \le \bar{E}'_\tau)$$
$$= \frac{(n-r)!}{(n+1)!} \sum_{f\in\mathcal{C}} \mathbb{I}(f \le \bar{E}'_\tau).$$

Claim 1 assumes that $\mathcal{C}$ has no ties almost surely, so $\sum_{f\in\mathcal{C}} \mathbb{I}(f > \bar{E}'_\tau) = \left\lfloor (1-\tau)\frac{n!}{(n-r-1)!} \right\rfloor$. Thus

$$\begin{aligned}
\mathbb{P}\left(e_{(1,\ldots,r),n+1} \le \bar{e}_\tau \mid \mathcal{F}\right) &\le \frac{(n-r)!}{(n+1)!}\left(|\mathcal{C}| - \left\lfloor (1-\tau)\frac{n!}{(n-r-1)!}\right\rfloor\right) \\
&\le \frac{(n-r)!}{(n+1)!}\left(\frac{(n+1)!}{(n-r)!} - (1-\tau)\frac{n!}{(n-r-1)!} + 1\right) \\
&= \tau + (1-\tau)\frac{r+1}{n+1} + \frac{(n-r)!}{(n+1)!}.
\end{aligned}$$

The two-sided guarantee follows by symmetry.

A.3. **Proof of Theorem 1.** Again let $\mathcal{F}$ denote the sigma-field generated by the unordered set $\{S_1,\ldots,S_{(n+1)}\}$. Under the assumed data-generating process,

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

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

$$\mathbb{P}(\boldsymbol{\pi}^w = \pi) = \frac{1}{n!}\frac{w(S_{\pi(n+1)})}{\sum_{j=1}^{n+1}\omega(S_j)}. \tag{A.4}$$

On the other hand, $\mathbf{T} \overset{d}{=} (\boldsymbol{\pi}^n(1),\ldots,\boldsymbol{\pi}^n(r))$ where $\boldsymbol{\pi}^n$ denotes a uniform random permutation on $\{1,\ldots,n\}$, so $e_{\mathbf{T},n+1} \mid \mathcal{F} \overset{d}{=} f(\boldsymbol{\pi}^w\circ\boldsymbol{\pi}^n(1),\ldots,\boldsymbol{\pi}^w\circ\boldsymbol{\pi}^n(r),\boldsymbol{\pi}^w(n+1))$. By (A.4), we have

$$e_{\mathbf{T},n+1} \mid \mathcal{F} \overset{d}{=} f(\boldsymbol{\pi}^w(1),\ldots,\boldsymbol{\pi}^w(r),\boldsymbol{\pi}^w(n+1)). \tag{A.5}$$

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

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

Thus,

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

Note that for any random variable $Z \sim F$, $\mathbb{P}(Z \leq \bar{Q}_\tau(F)) \geq \tau$. Then

$$\mathbb{P}\left(e_{\mathbf{T},n+1} \leq \bar{Q}_\tau\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n+1}} W'_k \cdot \delta_{f(d_1,\ldots,d_r,k)}\right) \mid \mathcal{F}\right) \geq \tau. \tag{A.6}$$

The value $f(d_1,\ldots,d_r,k)$ can be observed iff $d_1,\ldots,d_r,k$ are all not equal to $n+1$. The above unobserved upper confidence bound can be replaced by setting all unobserved f-values to ∞, i.e.,

$$\bar{Q}_\tau\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n}} W'_k \cdot \delta_{f(d_1,\ldots,d_r,k)} + \Omega_{n+1}\cdot\delta_\infty\right).$$

where

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

Via a simple counting argument, we obtain that

$$\begin{aligned}\Omega_{n+1} &= \frac{n!}{(n-r)!}W'_{n+1} + \frac{r(n-1)!}{(n-r)!}\sum_{k=1}^{n} W'_k \\ &= \frac{r}{n} + \frac{(n-1)!}{(n-r-1)!}W'_{n+1} = \frac{r}{n} + \frac{n-r}{n}\frac{w(S_{n+1})}{\sum_{j=1}^{n+1}\omega(S_j)}.\end{aligned} \tag{A.7}$$

where the second to last equality uses $\sum_{k=1}^{n+1} W'_k = \frac{(n-r)!}{n!}$. By (A.7) and (8), $W_k = \frac{W'_k}{1-\Omega_{n+1}}$, so

$$\begin{aligned}&\bar{Q}_\tau\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n}} W'_k \cdot \delta_{f(d_1,\ldots,d_r,k)} + \Omega_{n+1}\cdot\delta_\infty\right) \\ &= \bar{Q}_{\frac{\tau}{1-\Omega_{n+1}}}\left(\sum_{(d_1,\ldots,d_r,k)\in\mathbb{T}_{r+1,n}} W_k \cdot \delta_{f(d_1,\ldots,d_r,k)}\right),\end{aligned}$$

where the RHS is ∞ if $\tau \leq \Omega_{n+1}$.

Replacing τ by $\tau(1-\Omega_{n+1})$, (A.6) implies that

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

The theorem now follows from the law of iterated expectation.

A.4. **Proof of Lemma 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

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

$$U = \frac{(n-k)!}{n!} \sum_{i_1,\ldots,i_k} \phi(Z_{i_1}, \ldots, Z_{i_k}), \tag{A.8}$$

where $\sum_{i_1,\ldots,i_k}$ denotes the sum over all k-tuples in $\mathcal{N}$ with mutually distinct elements. The average of Z_i is a special case of (A.8) with $k = 1$ and $\phi(z) = z$.

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

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

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

$$\mathbb{E}[\psi(U)] \leq \mathbb{E}[\psi(W_{\mathbf{id}})] \tag{A.9}$$

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 (A.9).

Proposition A.1. *Let U be a U-statistic of order k with a bounded kernel $\phi \in [0,1]$ in the form of* (A.8) *and $m = \lfloor n/k \rfloor$. Then*

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

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

where

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

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

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

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 $a = k$.[29]

Proposition A.2 (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 A.1 and Proposition A.2 together yields Lemma 1.

[29]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).

Online appendix to the paper

The Transfer Performance of Economic Models

Isaiah Andrews Drew Fudenberg Lihua Lei Annie Liang Chaofeng Wu

November 28, 2023

Appendix P. Supplementary Material to Section 3.2

P.1. **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. Recall the definition of $\mathcal{C}$ in (A.1). First, sort the elements in $\mathcal{C}$ as

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

where

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

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

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

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

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

Let $w = (\omega(S_1), \ldots, \omega(S_n))^T$ and $\mathbf{1}_n = (1,1,\ldots,1)^T$. By (8), for each j,

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

Therefore,

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

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

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

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

Theorem P.1. *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

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

Proof. First, we prove that

$$\min_{w\in[\Gamma^{-1},\Gamma]^n} \frac{w^T\Psi^{(j)}}{w^T\mathbf{1}_n} = \min_{w\in\{\Gamma^{-1},\Gamma\}^n} \frac{w^T\Psi^{(j)}}{w^T\mathbf{1}_n}. \tag{P.2}$$

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

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

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

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

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

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

Then

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

while

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

This contradicts the definition of $w^{(j)}(\Gamma)$, so $w^{(j)}(\Gamma)\in\{\Gamma^{-1},\Gamma\}^n$, which completes the proof of (P.2).

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

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

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

$$\begin{aligned}
&\frac{\Gamma k \bar{\Psi}_k^{(j)} + \Gamma^{-1}(\mathbf{1}_n^T \Psi_i^{(j)} - k\bar{\Psi}_k^{(j)})}{\Gamma k + \Gamma^{-1}(n-k)} = \frac{\Gamma k \bar{\Psi}_k^{(j)} + \Gamma^{-1}(j - k\bar{\Psi}_k^{(j)})}{\Gamma k + \Gamma^{-1}(n-k)} \\
&= \frac{(\Gamma - \Gamma^{-1}) k \bar{\Psi}_k^{(j)} + \Gamma^{-1} j}{(\Gamma - \Gamma^{-1})k + \Gamma^{-1} n} = \frac{j}{n} + \frac{(\Gamma - \Gamma^{-1}) k \left(\bar{\Psi}_k^{(j)} - \frac{j}{n} \right)}{(\Gamma - \Gamma^{-1})k + \Gamma^{-1} n} \\
&= \frac{j}{n} + \frac{\bar{\Psi}_k^{(j)} - \frac{j}{n}}{1 + \frac{n}{k(\Gamma^2 - 1)}}.
\end{aligned}$$

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

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

By (P.1),

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

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 P.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 $O(rn^{r+1}\log n + n^{r+2})$.

P.2. **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

$$f_{(j_1),1} > f_{(j_2),2}, \quad \frac{(n-r-1)!}{(n-1)!}\frac{w^T\Psi^{(j_1-1),1}}{w^T\mathbf{1}_n} < \tau \le \frac{(n-r-1)!}{(n-1)!}\frac{w^T\Psi^{(j_2),2}}{w^T\mathbf{1}_n}. \tag{P.3}$$

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

To avoid pairwise comparisons, which incur $O(n^{2(r+1)})$ computation, we can check (P.3) 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 (P.3) 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, (P.3) 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:

$$\min v^T a^{(j)}, \quad \text{s.t.,} \ v^T b^{(j)} \ge \tau, v^T\mathbf{1}_n = 1, v \in [0,\infty)^n. \tag{P.4}$$

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 P.1, we can restrict

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

(2) When $a_i^{(j)} \geq b_i^{(j)}$ for every $i \in \mathcal{N}$, then the objective of (P.4) can never be below τ.

Appendix Q. Supplementary material for Section 4

Q.1. **Description of data.** We briefly describe the individual samples in our meta-data. There are 44 domains in total.

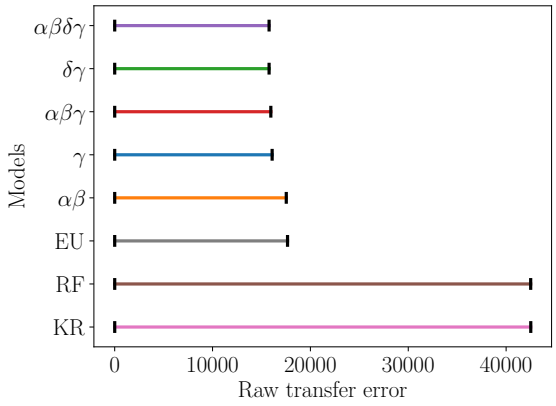
Table 4

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	27	Australia	N
l'Haridon and Vieider (2019)	2548	95	27	Belgium	N
l'Haridon and Vieider (2019)	2350	84	27	Brazil	N
l'Haridon and Vieider (2019)	2240	80	27	Cambodia	N
l'Haridon and Vieider (2019)	2687	96	27	Chile	N
l'Haridon and Vieider (2019)	5711	204	27	China	N
l'Haridon and Vieider (2019)	3072	128	23	Colombia	N
l'Haridon and Vieider (2019)	2968	106	27	Costa Rica	N
l'Haridon and Vieider (2019)	2770	99	27	Czech Republic	N
l'Haridon and Vieider (2019)	3906	140	27	Ethiopia	N
l'Haridon and Vieider (2019)	2604	93	27	France	N

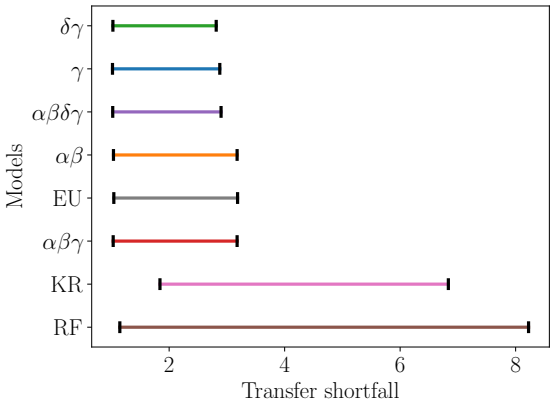
l'Haridon and Vieider (2019)	3639	130	27	Germany	N
l'Haridon and Vieider (2019)	2352	84	27	Guatemala	N
l'Haridon and Vieider (2019)	2492	89	27	India	N
l'Haridon and Vieider (2019)	2352	84	27	Japan	N
l'Haridon and Vieider (2019)	2716	97	27	Kyrgyzstan	N
l'Haridon and Vieider (2019)	1791	64	27	Malaysia	N
l'Haridon and Vieider (2019)	3360	120	27	Nicaragua	N
l'Haridon and Vieider (2019)	5638	202	27	Nigeria	N
l'Haridon and Vieider (2019)	2660	95	27	Peru	N
l'Haridon and Vieider (2019)	2491	89	27	Poland	N
l'Haridon and Vieider (2019)	1959	70	27	Russia	N
l'Haridon and Vieider (2019)	1819	65	27	Saudi Arabia	N
l'Haridon and Vieider (2019)	1988	71	27	South Africa	N
l'Haridon and Vieider (2019)	2240	80	27	Spain	N
l'Haridon and Vieider (2019)	2212	79	27	Thailand	N
l'Haridon and Vieider (2019)	2070	74	27	Tunisia	N
l'Haridon and Vieider (2019)	2240	80	27	UK	N
l'Haridon and Vieider (2019)	2701	97	27	US	N
l'Haridon and Vieider (2019)	2436	87	27	Vietnam	N

Q.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, 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 65% two-sided forecast intervals in Figure 8. These intervals are qualitatively similar to those reported in Figure 3.

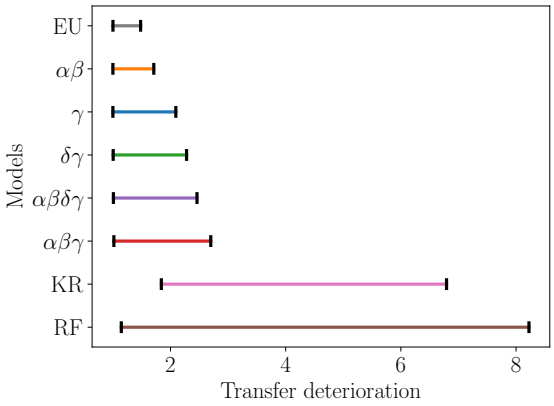
Q.3. **In-sample errors.** Figure 9 displays the CDFs of the in-sample errors of EU, CPT, the random forest algorithm, and kernel regression. As in Figure 2 (which reported out-of-sample errors), these curves are nearly indistinguishable.



(A) *Raw transfer error*



(B) *Transfer shortfall*



(C) *Transfer deterioration*

FIGURE 8. *65% (n=14, $\tau = 0.95$) forecast intervals for each of the three measures, treating each paper as a separate domain.*

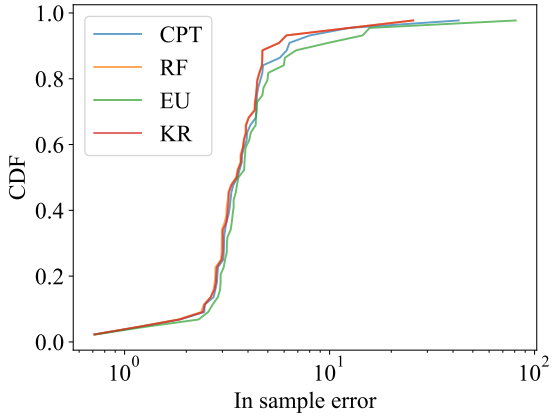


FIGURE 9. *CDF of in-sample errors*

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

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. *81% (n=44, $\tau = 0.95$) forecast intervals*

Q.5. **Alternative forecast intervals.** In this section, we report alternative forecast intervals for our three measures. Table 6 constructs 91% two-sided forecast intervals (setting $\tau = 1$),[30] and Table 7 reports 91% one-sided forecast intervals (setting $\tau = 0.95$). All of the forecast intervals are qualitatively similar to the 81% two-sided forecast intervals reported in the main text.

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. *91% (n=44, $\tau = 1$) two-sided forecast intervals*

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

Model	Raw Transfer Error	Transfer Shortfall	Transfer Deterioration
CPT main variants			
γ	[0,15.83]	[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]

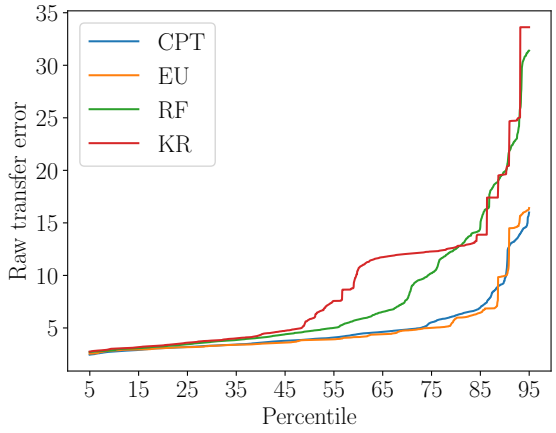
TABLE 7. *91% (n=44, $\tau = 0.95$) one-sided forecast intervals*

Finally, Figure 10 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.[31] In fact, in Panels (a) and (c), the black box curves lie everywhere above the CPT and EU curves, so both the lower and upper bounds of the black boxes' forecast intervals are higher than those of the economic models for every choice of τ.

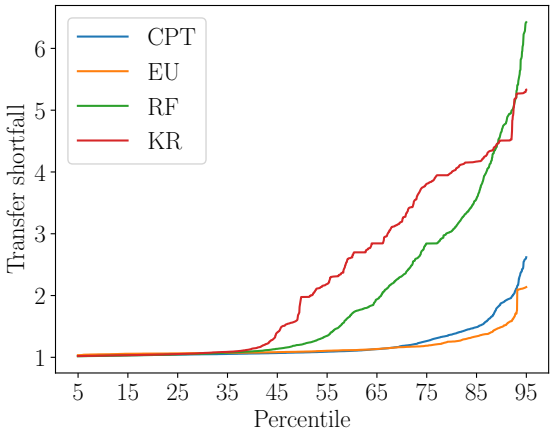
Q.6. **Forecast intervals for the ratio of raw CPT and RF transfer errors.** Let $e_{\mathcal{T},d}$ be the ratio of the raw random forest transfer error to the raw CPT transfer error (i.e., using the specification in (1)), henceforth the *transfer error ratio*.

Panel (a) of Figure 11 reports 81% two-sided forecast intervals for the raw transfer error ratio for each CPT specification. The lower bound for each CPT model is approximately 0.9, while the upper bound is as large as 4.5. Panel (b) of the figure is a histogram of raw transfer error ratios for the 4-parameter CPT model when the training domains $\mathcal{T}$ and the target domains d are drawn uniformly at random from the set of domains in the meta-data. This distribution has a large cluster 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 11, shows that the random forest algorithm outperforms CPT in approximately 35% of $(\mathcal{T}, d)$ pairs, although

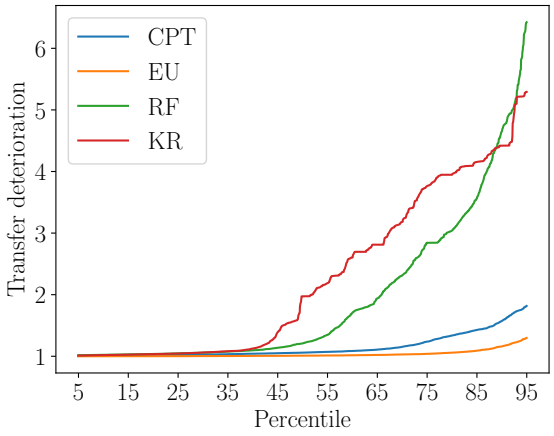
[31]To improve readability, we remove extreme numbers by truncating $\tau \in [5, 95]$, and show results only for the $\alpha\beta\gamma\delta$ specification of the CPT model.



(A) *Raw transfer error*



(B) *Transfer shortfall*



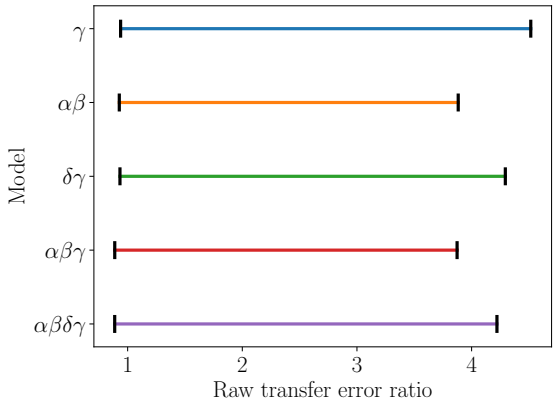
(C) *Transfer deterioration*

FIGURE 10. *Error percentiles from 5 to 95 (truncated for readability).*

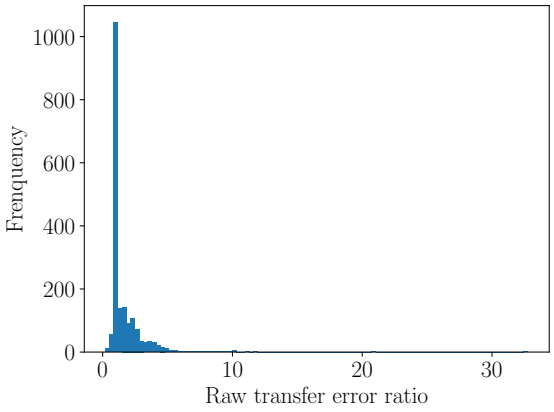
CPT rarely has a much worse raw transfer error than the random forest and is sometimes much better.

Q.7. **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 7.

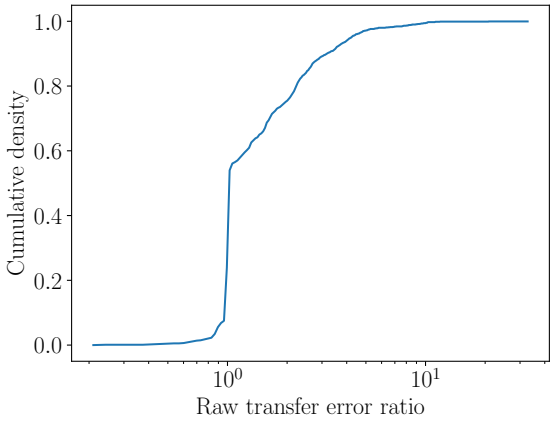
Figure 12 is the analog of Figure 3. Again we choose $\tau = 0.95$, thus constructing forecast intervals whose lower bounds are the 5% percentile of pooled transfer errors, and whose



(A) *81% (n=44, $\tau = 0.95$) Forecast intervals*



(B) *Density*



(C) *CDF*

FIGURE 11. *Forecast intervals, density, and cdf for the ratio of the raw random forest transfer error to the raw CPT transfer error.*

upper bounds are the 95% percentile of pooled transfer errors. Applying Proposition 1, these are 73% forecast intervals. The most notable change is that the random forest forecast interval shrinks considerably, which suggests that the raw transfer error of the random forest algorithm becomes less variable when it is trained on more domains. Otherwise, all of the qualitative statements in the main text for $r = 1$ continue to hold. In particular, as with $r = 1$, we find that the forecast intervals for all three of our measures have higher lower and upper bounds for the black box algorithms than for the CPT specifications.

Q.8. **More details on worst-case dominance.** Figures 13 and 14 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.

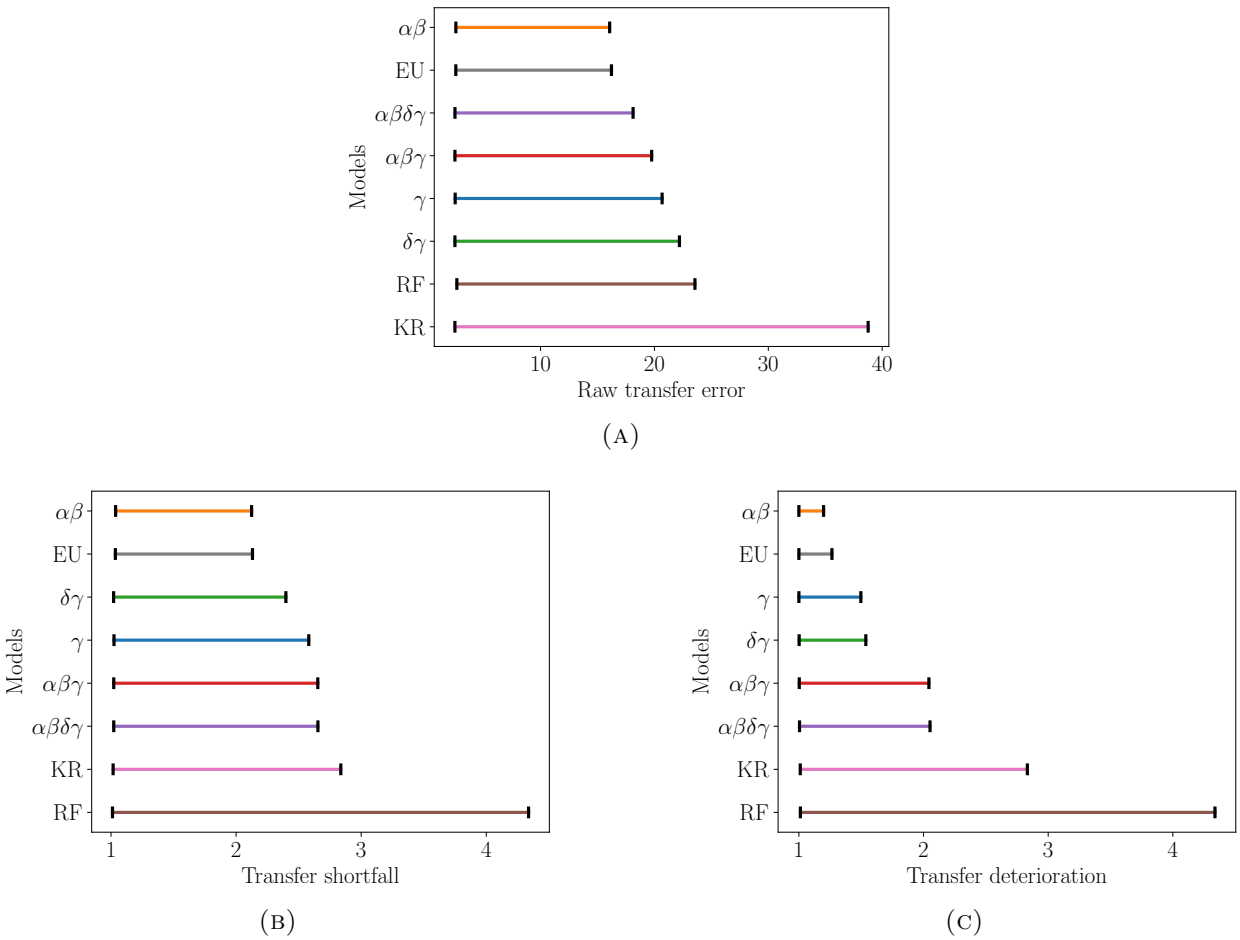


FIGURE 12. *73% (n=44, $\tau = 0.95$) forecast intervals for (a) raw transfer error, (b) transfer shortfall, and (c) transfer deterioration, with the choice of $r = 3$.*

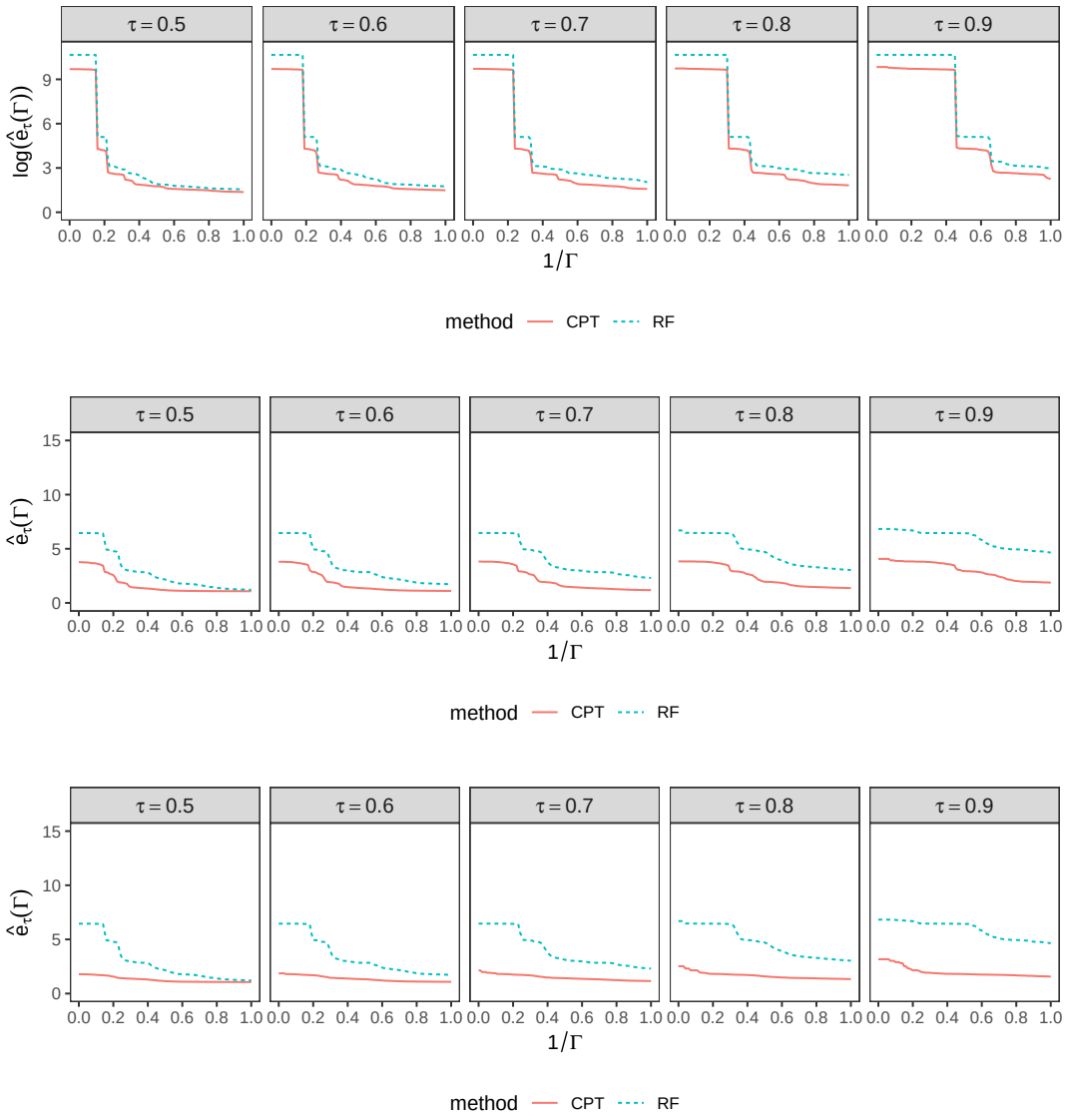


FIGURE 13. *The worst case upper prediction bound* $\hat{e}_\tau(\Gamma)$ *(as defined in (9)) for (a) raw transfer error, (b) transfer shortfall, and (c) transfer deterioration of CPT and RF as a function of* $\Gamma \in [1, \infty)$*, discretized at* $100/i (i = 0, 1, \ldots, 100)$*, at different quantiles* $\tau \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

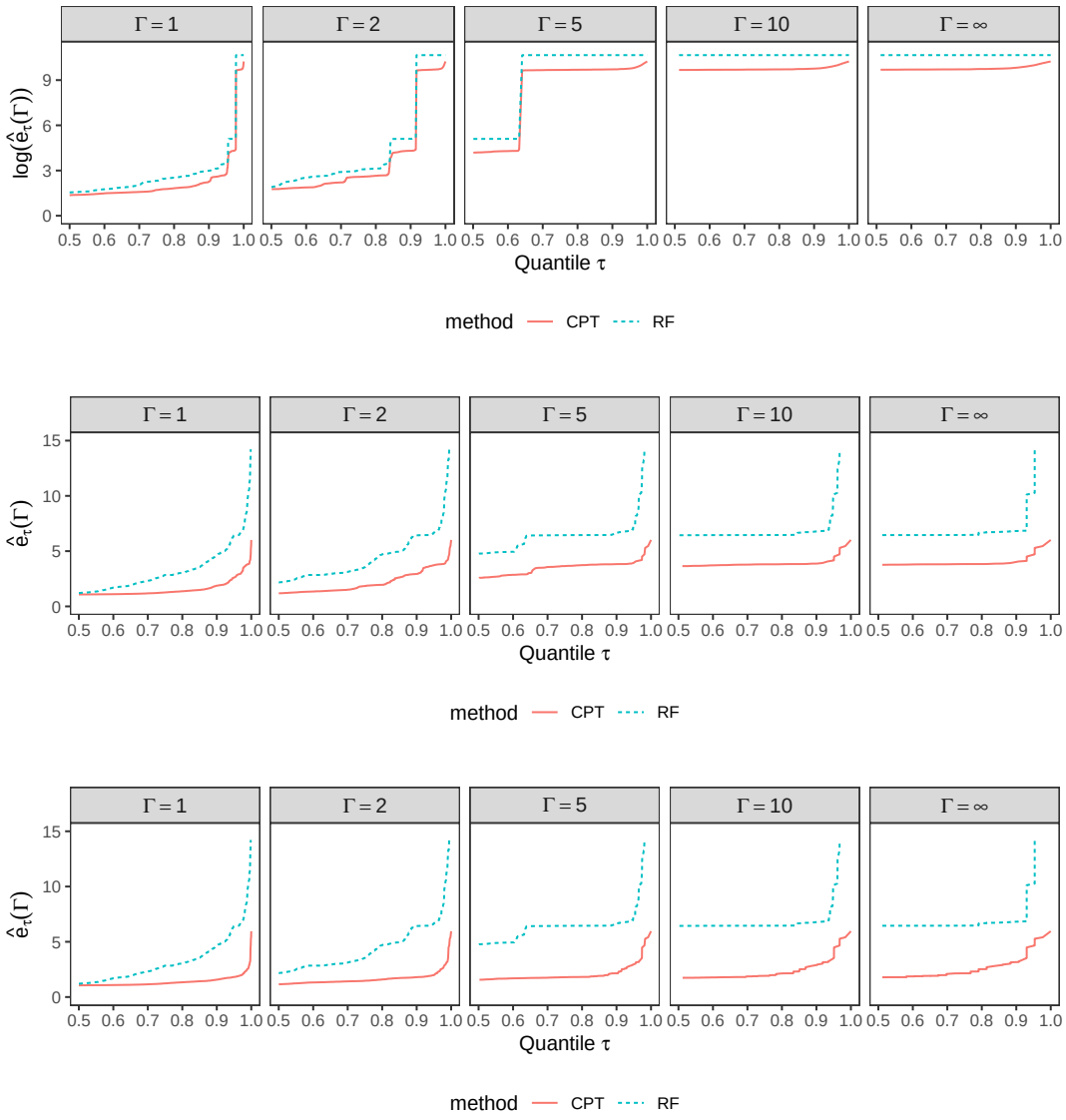


FIGURE 14. *The worst case upper prediction bound* $\hat{e}_\tau(\Gamma)$ *(as defined in (9)) for (a) raw transfer error, (b) transfer shortfall, and (c) transfer deterioration of CPT and RF as a function of* $\tau \in [0.5, 1]$ *without discretization for* $\Gamma \in \{1, 2, 5, 10, \infty\}$.