RESEARCH STATEMENT

YUEHAO BAI

I work in econometrics, the branch of economics that uses statistical methods to analyze economic data. As an econometrician, I develop methods to learn the causal effects of policies from experimental and observational data. My research focuses on three related areas:

1. Design and analysis of randomized experiments ([3], [4], [5], [6], [7], [8], [9], [10], [11], [13]). In particular, how should we design experiments to estimate causal effects as precisely as possible for a given sample size? Given such designs, how should we conduct inference while properly accounting for statistical uncertainty?
2. Identification and inference with multi-valued treatments and instrumental variables ([12], [16], [17]). When individuals choose among multiple policy programs, how do conclusions depend on assumptions about their choices? How can we assess the credibility of these assumptions? How can we analyze causal effects while allowing these assumptions to remain flexible?
3. Inference for partially identified economic models ([2], [14], [18]). In many applications, including some in the first two areas, causal effects are only partially identified: even without statistical uncertainty, we cannot pinpoint a single value for the effect but can only conclude it lies in an interval. To conduct inference in these settings, how can we leverage features of the model, especially linearity?

My research is methodological and closely motivated by challenges in empirical economic research. The methods I propose are directly applicable to empirical research in economics and have also been applied in other fields such as statistics and the medical sciences. For my contributions to the literature on randomized experiments, the *Journal of Political Economy Microeconomics* invited me, along with Azeem Shaikh and Max Tabord-Meehan, to write a comprehensive survey article on the analysis of randomized experiments [11]. To implement the methods in [3], [4], [5], [6], [7], [8], [9], [10], [11], and [13], we developed and maintain an R package

on the Comprehensive R Archive Network (CRAN) [19]. We are also developing software for the methods in [14], [16], [17], and [18].

DESIGN AND ANALYSIS OF RANDOMIZED EXPERIMENTS

Randomized controlled trials (RCTs) are widely used in economics and related fields to study the effects of interventions (usually called treatments). In economics, a treatment could be an educational or job training program, cash transfers, malaria treatment, and so on. Because RCTs are often expensive to run, researchers want to estimate causal effects as precisely as possible for a given sample size. A key decision for precision is how to stratify (group) units according to their baseline variables (covariates) so that treated and untreated units are more comparable. More than 2000 RCTs on the AEA RCT Registry are stratified. My research studies a class of stratified designs for achieving maximum precision. These designs are called highly stratified designs and include matched-pair designs as an important case, where units are matched according to their baseline covariates and then one unit in each pair is assigned to treatment at random.

In "Optimality of Matched-Pair Designs in Randomized Controlled Trials" [4], published in the *American Economic Review*, I establish a general finite-sample optimality result for matched-pair designs (and highly stratified designs in general). Matched-pair designs have been used at least once by 56% of the researchers interviewed by [20], yet before my paper [4], it was not known whether they lead to better precision of estimators and how units should be paired. I show that a particular matched-pair design is optimal in the sense that it minimizes the mean-squared error of the difference-in-means estimator for estimating the average treatment effect (ATE) among all stratifications in which half of each stratum is treated. When the treated fraction is not one half, the optimal design is still a highly stratified design. The optimal design matches the units according to an index function that summarizes how the covariates predict a particular weighted average of the potential outcomes. The result does not restrict the heterogeneity of treatment effects, makes no parametric assumption about functional forms, and holds exactly at each fixed sample size. To my knowledge, it is the first optimality result for matched-pair designs (and highly stratified designs in general) at this level of generality.

The index function is generally unknown, but the optimal design can be characterized explicitly in an important special case: when the index function is monotonic in a scalar covariate. For example, it is reasonable to assume a higher baseline test score predicts a higher endline test score on average. In such settings, the optimal design simply matches the units according to their baseline outcomes. I then conduct extensive simulation studies based on data from ten RCTs to compare the performance of different stratifications in more complicated settings and recommend particular matched-pair designs for practice.

Despite their optimality for estimation, matched-pair designs pose a serious challenge for inference: standard inference procedures do not correctly quantify statistical uncertainty because treatment status is no longer independent across units. Indeed, if one unit in a pair is treated, then the other unit is necessarily untreated, so treatment status is perfectly negatively correlated within a pair. The difficulty for inference was one of the main reasons that matched-pair designs were not adopted as widely as they should have been.

In "Inference in Experiments with Matched Pairs" [3], published in the *Journal of the American Statistical Association*, my coauthors and I show that under matched-pair designs, conventional inference procedures such as the two-sample t-test and the "matched pairs" t-test are asymptotically conservative in the sense that the limiting rejection probabilities under the null hypothesis are usually strictly smaller than the nominal level. Therefore, standard errors from these tests tend to be larger than necessary, making it harder to detect a nonzero ATE. We then propose an adjusted t-test and an accompanying randomization test that are asymptotically exact in the sense that the limiting rejection probabilities under the null hypothesis are equal to the nominal level. Through simulation studies and an RCT that we conducted on the Amazon Mechanical Turk platform, we find that our methods can dramatically reduce standard errors and increase statistical power to detect a nonzero ATE.

The papers discussed above focus on the ATE and abstract from several complications that commonly arise in RCTs: treatment may be multi-valued or assigned at the cluster level; compliance with treatment assignments may be imperfect; units may drop out between the baseline and endline surveys; and additional covariates not used for matching may be available for estimation. My coauthors and I address these challenges in a series of papers.

In "Inference for Matched Tuples and Fully Blocked Factorial Designs" [8], published in *Quantitative Economics*, we propose designs in the spirit of matched pairs for settings with multiple treatment arms. We call these designs matched tuples designs and adapt them to factorial settings, in which multiple binary treatments are jointly randomized. As in [3], we provide inference procedures for treatment effects that are asymptotically exact. We also show that with a multi-valued treatment, conventional tests based on block fixed effects may not control size.

In economics, treatment is often assigned at the level of a cluster, such as a school, a village, or a city. In "Inference in Cluster Randomized Trials with Matched Pairs" [9], published in the *Journal of Econometrics*, we propose designs and inference procedures for settings in which the clusters are matched according to their baseline covariates. An important distinction from the rest of the literature is that we carefully distinguish between two parameters of interest in cluster randomized trials: equally-weighted and size-weighted average treatment effects. The distinction matters because, for some policy questions, larger clusters carry greater weight.

In "Inference in Experiments with Matched Pairs and Imperfect Compliance" [10], published in the *Journal of Business & Economic Statistics*, we study estimation and inference under matched-pair designs with imperfect compliance, meaning that units may not take up the treatment they are assigned to. In such settings, we focus on the local average treatment effect (LATE), a common parameter that summarizes the effect of the treatment for the subgroup of compliers.

In "Revisiting the Analysis of Matched-Pair and Stratified Experiments in the Presence of Attrition" [6], published in the *Journal of Applied Econometrics*, we study settings in which units may drop out between the baseline and endline surveys. We show that several popular estimators in fact estimate different parameters, some of which are not causal, and clarify a misconception in the literature that attrition precludes the use of matched-pair designs.

In practice, researchers often observe additional covariates that are not used for matching and may hope to further increase precision by adjusting for them. In "Covariate Adjustment in Experiments with Matched Pairs" [7], published in the *Journal of Econometrics*, we study a large class of covariate adjustments based on the "doubly robust" moment condition. We find that linear adjustments without pair

fixed effects may harm precision, but adding pair fixed effects avoids this problem. Either way, inference must be adjusted to account for the matched-pair design. We then study adjustments with many covariates based on machine learning methods such as LASSO.

The parallel structure in the papers above suggests that these problems may be analyzed in a common framework that had not yet been developed in the literature. We provide such a synthesis in "On the Efficiency of Highly Stratified Experiments" [13], published in the *Annals of Statistics*. The point of departure is a set of just-identified moment equations. This framework accommodates many common parameters of interest in causal inference, including the ATE, LATE, quantile treatment effect, weighted treatment effect, and others. We first derive the limiting variance of the naive plug-in estimator based on the moment equations under highly stratified designs. Under arbitrary designs, this naive estimator is generally not efficient in the sense that some other estimators have lower limiting variances. Under highly stratified designs, however, we show that its limiting variance coincides with the semiparametric efficiency bound, a bound that characterizes the lowest limiting variance attainable by any regular estimator.

This result strongly suggests that highly stratified designs are optimal for estimating any parameter in this framework, with the caveat that existing results on efficiency bounds assume treatment status is independent. As mentioned above, treatment status is dependent under highly stratified designs, so we further establish a common efficiency bound for a large class of experimental designs under weak conditions on large-sample convergence. The paper therefore settles a long-standing question in the literature by showing highly stratified designs are indeed efficient "by design."

In "Why Randomize? Minimax Optimality under Permutation Invariance" [5], published in the *Journal of Econometrics*, I answer the bigger-picture question why researchers should randomize. Clearly, a pure Bayesian would not randomize but rather assign treatment status deterministically to minimize the Bayesian risk. I show, however, that randomization is needed if the decision criterion is minimax instead of Bayesian. I further prove that without any covariate, the difference-in-means estimator together with complete randomization is minimax optimal, and

with a discrete covariate, the block-size weighted difference-in-means estimator together with stratified block randomization is minimax optimal.

Looking ahead, I plan to continue working on empirically relevant econometric problems in RCTs. For example, in the working paper "A New Design-Based Variance Estimator for Finely Stratified Experiments" [15], we propose a new variance estimator for highly stratified designs that is upward biased in a design-based framework. We plan to use an approach based on graph Laplacians to study a general class of designs inspired by stratification. I am also studying econometric issues in RCTs with bipartite networks and RCTs with group formation. These projects, together with the survey article [11] and the R package "sreg: Stratified Randomized Experiments" [19], aim to provide a comprehensive econometric toolkit for designing and analyzing RCTs.

IDENTIFICATION AND INFERENCE WITH MULTI-VALUED TREATMENTS AND INSTRUMENTAL VARIABLES

As discussed in the previous section, in many real-world economic applications, treatment is multi-valued rather than binary. This is the case in both experimental and observational settings: individuals may be randomized into one of many job-training programs, or they may choose to attend one of many colleges. In addition, treatment is often endogenous in the sense that units self-select into different treatment choices. A standard approach to address treatment endogeneity is to use an instrumental variable that shifts the treatment and is exogenous in the sense of being randomly assigned or "as good as" randomly assigned. Even when a high-quality instrument exists, researchers need to impose assumptions on choice behavior (potential treatments) to evaluate treatment effects. With a binary treatment and a binary instrument, a standard assumption is that the potential treatment is monotonic in the instrument; under this assumption, the LATE can be identified, although it may or may not be policy relevant.

Multi-valued treatments and multi-valued instruments present both opportunities and challenges. They allow researchers to impose a wider range of assumptions on potential treatments and to study a richer class of causal parameters, but most policy-relevant causal parameters are only partially identified: even without statistical uncertainty, the population distribution of the observed data is consistent with

many possible values of the parameter. In a series of papers, my coauthors and I study how identification and inference for causal parameters interact with these assumptions on potential treatments.

In "On the Identifying Power of Generalized Monotonicity for Average Treatment Effects" [12], published in the *Journal of the American Statistical Association*, we study which assumptions on potential treatments have identifying power for the ATEs. Here, having identifying power means that the identified sets for the ATEs can strictly shrink after imposing the assumptions. We study an assumption that we call "generalized monotonicity," meaning that each value of the treatment has a corresponding value of the instrument that maximally encourages individuals towards it: if an individual doesn't choose this treatment value at this instrument value, then they never choose this treatment value. Generalized monotonicity nests the classical monotonicity assumption in the binary setting as well as many other assumptions that have been employed in economics and statistics to evaluate multi-valued treatments.

We maintain instrument exogeneity and show the striking result that as long as the model does not directly restrict the potential outcomes, imposing generalized monotonicity will either not change the identified sets for ATEs or make them empty. The same result holds, moreover, even if the assumption is strictly stronger than generalized monotonicity. The results imply that many assumptions employed in the literature have no identifying power for the ATEs. Therefore, in order to help identify treatment effects, researchers need to impose assumptions that violate generalized monotonicity, impose shape restrictions on potential outcomes directly, or move away from ATEs to distributional or conditional treatment effects.

Indeed, researchers are often interested in treatment effects conditional on "generalized principal strata," sets defined by values of potential treatments and potential outcomes. These parameters include pairwise LATEs, unconditional parameters such as average and distribution effects, and many other more complex parameters. Most of them are partially identified and the form of the identified sets is unknown except in a few special cases. Therefore, for many parameters there had been no methods to construct confidence intervals that control size uniformly over a reasonably large class of distributions. Further complicating the problem, researchers

usually want to impose additional assumptions on potential treatments and potential outcomes when studying these parameters.

In "Inference for Treatment Effects Conditional on Generalized Principal Strata using Instrumental Variables" [17], revision requested by the *Review of Economic Studies*, we provide a unified method to conduct inference for all of these parameters while allowing researchers to impose flexible assumptions on potential treatments and potential outcomes. The key insight is that a parameter value is consistent with the distribution of the data if and only if a particular linear program is feasible. As a result, confidence intervals can be computed through test inversion over a grid of values. Our method can also be readily applied to test whether the model is correctly specified; that is, whether the assumptions on potential treatments and potential outcomes are consistent with the data. Such specification tests are important for evaluating the credibility of these assumptions.

In a leading special case, the sharp testable implications of the model can be obtained analytically. In the working paper "Sharp Testable Implications of Encouragement Designs" [16], we study arguably the most natural extension of the monotonicity assumption from the binary setting to multi-valued settings. The assumption requires that each value of the instrument increases the appeal of only one treatment value, thus leading to an encouragement design. We show through a novel constructive argument that the distribution of the data is consistent with an encouragement design if and only if a set of conditional moment inequalities hold. This constructive proof allows us to obtain the sharp testable implications analytically with a general-valued outcome. Based on our results, the encouragement design assumption can then be tested using off-the-shelf methods for testing moment inequalities.

Looking ahead, I plan to continue working on econometric problems involving multi-valued treatments and instruments. One empirically important setting is factorial experiments, where my coauthors and I plan to impose restrictions motivated by economic theory and study how they affect the identification of treatment effects. I also plan to study mediation analysis using the econometric tools developed in [18], discussed below.

INFERENCE FOR PARTIALLY IDENTIFIED ECONOMIC MODELS

The previous two sections show that in many economic models, the parameters of interest are partially identified. In a series of papers, my coauthors and I study how to conduct inference in these models. A unified theme is that the methods apply to high-dimensional settings that are prevalent in practice.

In "A Two-Step Method for Testing Many Moment Inequalities" [2], published in the *Journal of Business & Economic Statistics*, we study inference with many moment inequalities. A leading example is entry games in industrial organization. We show that the two-step procedure of [21] is asymptotically valid even in high-dimensional settings where the number of inequalities may grow exponentially in the sample size. The procedure remains computationally feasible in such settings, and our result provides a theoretical justification for its use.

In the working paper "Inference for Linear Systems with Unknown Coefficients" [18], we study inference for linear systems. In contrast to some prior work, we allow all parameters in the system of equations, including the slope coefficients, to be unknown. Many economic problems have this structure. Using Farkas' lemma, we first characterize the closure of the null hypothesis in terms of a union of inequalities. Motivated by this characterization, we then propose a test based on sample splitting. The test is asymptotically valid under minimal assumptions on the geometry of the linear system.

Looking ahead, I plan to continue working on inference for partially identified models. A particularly interesting direction is to study models where some constraints may be quadratic or, more generally, convex, and to examine whether extensions of Farkas' lemma can be used to develop inference procedures similar to those in [18] for more complicated problems.

REFERENCES

[1] Y. Bai, H. Ho, G. A. Pouliot, and J. Shea, "Inference for Support Vector Regression under l1 Regularization," *AEA Papers and Proceedings*, vol. 111, pp. 611–615, May 2021, doi: 10.1257/pandp.20211035.

[2] Y. Bai, A. Santos, and A. M. Shaikh, "A Two-Step Method for Testing Many Moment Inequalities," *Journal of Business & Economic Statistics*, vol. 40, no. 3, pp. 1070–1080, July 2022, doi: 10.1080/07350015.2021.1897016.

[3] Y. Bai, J. P. Romano, and A. M. Shaikh, "Inference in Experiments With Matched Pairs," *Journal of the American Statistical Association*, vol. 117, no. 540, pp. 1726–1737, Oct. 2022, doi: 10.1080/01621459.2021.1883437.

[4] Y. Bai, "Optimality of Matched-Pair Designs in Randomized Controlled Trials," *American Economic Review*, vol. 112, no. 12, pp. 3911–3940, Dec. 2022, doi: 10.1257/aer.20201856.

[5] Y. Bai, "Why randomize? Minimax optimality under permutation invariance," *Journal of Econometrics*, vol. 232, no. 2, pp. 565–575, Feb. 2023, doi: 10.1016/j.jeconom.2021.10.009.

[6] Y. Bai, M. H. Hsieh, J. Liu, and M. Tabord-Meehan, "Revisiting the analysis of matched-pair and stratified experiments in the presence of attrition," *Journal of Applied Econometrics*, vol. 39, no. 2, pp. 256–268, 2024, doi: 10.1002/jae.3025.

[7] Y. Bai, L. Jiang, J. P. Romano, A. M. Shaikh, and Y. Zhang, "Covariate adjustment in experiments with matched pairs," *Journal of Econometrics*, vol. 241, no. 1, p. 105740, Apr. 2024, doi: 10.1016/j.jeconom.2024.105740.

[8] Y. Bai, J. Liu, and M. Tabord-Meehan, "Inference for Matched Tuples and Fully Blocked Factorial Designs," *Quantitative Economics*, vol. 15, no. 2, pp. 279–330, 2024, doi: 10.3982/QE2354.

[9] Y. Bai, J. Liu, A. M. Shaikh, and M. Tabord-Meehan, "Inference in cluster randomized trials with matched pairs," *Journal of Econometrics*, vol. 245, no. 1, p. 105873, Oct. 2024, doi: 10.1016/j.jeconom.2024.105873.

[10] Y. Bai, H. Guo, A. M. Shaikh, and M. Tabord-Meehan, "Inference in Experiments with Matched Pairs and Imperfect Compliance," *Journal of Business & Economic Statistics*, vol. 43, no. 3, pp. 627–642, July 2025, doi: 10.1080/07350015.2024.2416972.

[11] Y. Bai, A. M. Shaikh, and M. Tabord-Meehan, "A Primer on the Analysis of Randomized Experiments and a Survey of some Recent Advances," *Journal of Political Economy Microeconomics*, 2026, doi: 10.48550/arXiv.2405.03910.

[12] Y. Bai, S. Huang, S. Moon, A. M. Shaikh, and E. J. Vytlacil, "On the Identifying Power of Generalized Monotonicity for Average Treatment Effects," *Journal of the American Statistical Association*, vol. 0, no. 0, pp. 1–8, Apr. 2026, doi: 10.1080/01621459.2026.2620141.

[13] Y. Bai, J. Liu, A. M. Shaikh, and M. Tabord-Meehan, "On the Efficiency of Highly Stratified Experiments." Accessed: Mar. 31, 2026. [Online]. Available: http://arxiv.org/abs/2307.15181

[14] Y. Bai, A. Santos, and A. M. Shaikh, "On Testing Systems of Linear Inequalities with Known Coefficients." 2022.

[15] Y. Bai, X. Huang, J. P. Romano, A. M. Shaikh, and M. Tabord-Meehan, "A New Design-Based Variance Estimator for Finely Stratified Experiments." Accessed: Mar. 26, 2025. [Online]. Available: http://arxiv.org/abs/2503.10851

[16] Y. Bai, S. Huang, and M. Tabord-Meehan, "Sharp Testable Implications of Encouragement Designs." Accessed: Nov. 24, 2025. [Online]. Available: http://arxiv.org/abs/2411.09808

[17] Y. Bai, S. Huang, S. Moon, A. Santos, A. M. Shaikh, and E. J. Vytlacil, "Inference for Treatment Effects Conditional on Generalized Principal Strata using Instrumental Variables." Accessed: Nov. 24, 2025. [Online]. Available: http://arxiv.org/abs/2411.05220

[18] Y. Bai, K. Ponomarev, A. Santos, A. M. Shaikh, M. Tabord-Meehan, and A. Torgovitsky, "Inference for Linear Systems with Unknown Coefficients." 2026.

[19] J. Trifonov, Y. Bai, A. Shaikh, and M. Tabord-Meehan, "sreg: Stratified Randomized Experiments." Accessed: Sept. 02, 2025. [Online]. Available: https://cran.r-project.org/web/packages/sreg/index.html

[20] M. Bruhn and D. McKenzie, "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, vol. 1, no. 4, pp. 200–232, Oct. 2009, doi: 10.1257/app.1.4.200.

[21] J. P. Romano, A. M. Shaikh, and M. Wolf, "A Practical Two-Step Method for Testing Moment Inequalities," *Econometrica*, vol. 82, no. 5, pp. 1979–2002, 2014, Accessed: Apr. 14, 2023. [Online]. Available: https://www.jstor.org/stable/24029299