Testing Conditional Stochastic Dominance at Target Points*

Federico A. Bugni
Department of Economics
Northwestern University
federico.bugni@northwestern.edu

Ivan A. Canay
Department of Economics
Northwestern University
iacanay@northwestern.edu

Deborah Kim
Department of Economics
University of Warwick
deborah.kim@warwick.ac.uk

August 19, 2026

Abstract

This paper introduces a test for conditional stochastic dominance between two distributions at prespecified values of a conditioning covariate, referred to as target points. The test uses a one-sided Kolmogorov–Smirnov statistic computed from induced order statistics, the outcomes attached to the conditioning observations closest to the target point, and compares it to a critical value that, given the number of neighbors, requires no resampling, kernel smoothing, or parametric assumptions. The same procedure applies whether the outcomes are continuous, discrete, or mixed, and requires only continuity of the conditional distributions in the conditioning variable. We establish asymptotic validity under two frameworks: one in which the number of neighbors is held fixed, where the induced order statistics converge to independent draws from the conditional distributions at the target point; and one in which it grows with the sample size, where we obtain an explicit rate that accommodates both an estimated target point and a data-dependent choice of the number of neighbors. We connect the test to permutation-based inference, provide a refined critical value for discrete outcomes, propose a rule for selecting the tuning parameters, and illustrate the procedure in two empirical applications whose recorded outcomes exhibit mass points. Monte Carlo simulations confirm its strong finite-sample performance.

KEYWORDS: stochastic dominance, regression discontinuity design, induced order statistics, rank tests, permutation tests.

JEL classification codes: C12, C14.

*We thank Xinran Li, David Kaplan, Matias Cattaneo, and participants at several conferences and seminars for helpful comments and discussion.

1 Introduction

The concept of stochastic dominance has long been central to many areas of applied research. This paper studies a conditional version of the problem: testing conditional stochastic dominance (CSD) at one or more prespecified values of a conditioning covariate, which we call target points. Comparisons of this kind arise whenever the question of interest concerns a particular value of the conditioning variable rather than its entire range. Examples include the outcome distributions on either side of the cutoff in a regression discontinuity design (RDD), and the comparison of two groups at a fixed quantile of a covariate such as skill or experience.

Two empirical applications, developed in Section C.3, motivate and illustrate the procedure. In the first, we ask whether retirement shifts the distribution of household expenditure upward, following Battistin et al. (2009) and Shen and Zhang (2016). This application uses a regression discontinuity in the age of the household head, with the known retirement-age cutoff as the target point. In the second, we ask whether small classes stochastically dominate regular classes in student achievement at the median level of teacher experience, using the Project STAR data as in Qu and Yoon (2015). Here, the target point is a quantile of the conditioning variable and must be estimated. In both applications, the recorded outcomes exhibit mass points, making atomlessness of the conditional outcome distributions an unattractive maintained assumption. Existing procedures for CSD at a target point typically impose this assumption, often together with smoothness conditions, and treat the target point as known. Our procedure accommodates continuous, discrete, and mixed outcomes and allows the target point to be either known or estimated.

There is a long literature on inference for unconditional stochastic dominance, which compares entire marginal distributions; see Whang (2019) for a textbook treatment. A related and equally active literature studies inference for conditional stochastic dominance, comparing subgroups defined by covariates; see Whang (2019, Chapter 4) for a review. As that review makes clear, this literature addresses two distinct problems: dominance uniformly over a range of conditioning values, and dominance at one or more specific target points. Our paper concerns the latter, a problem that recurs across applied work; see, among others, Shen (2019), Qu and Yoon (2019), Barcellos et al. (2023), and Li and Sunder (2024). We review both branches and position our contribution relative to them after describing the test.

The primary goal of this paper is to test whether the conditional cumulative distribution function (CDF) of one variable stochastically dominates that of the other at specific values of a conditioning variable. Formally, we consider the null hypothesis

$$H_0 : F_Y(t|z) \leq F_X(t|z) \quad \text{for all } (t,z) \in \mathbf{R} \times \mathcal{Z} \ ,$$

against the alternative that there exists some (t,z) for which the reversed inequality holds strictly. Here, $F_Y(\cdot|z)$ and $F_X(\cdot|z)$ represent the conditional CDFs of the random variables Y and X, respectively, given $Z = z$. Importantly, we focus on situations where the set of target values,

$\mathcal{Z}$, is not the entire support of Z, but rather a finite collection of points (including the case of a singleton), which may be known or estimated from the data. To address this testing problem, we propose a novel procedure that leverages induced order statistics based on independent samples from Y and X. Our test statistic, a Kolmogorov–Smirnov-type measure, captures the maximal deviation between the empirical CDFs of the two samples, conditional on observations near the target points. Crucially, the critical value we propose is derived in a deterministic, non-data-dependent manner once the tuning parameters are accounted for, ensuring computational simplicity.

Our contributions in this paper are both methodological and theoretical. First, we introduce a novel test for CSD at target points. The test is built from induced order statistics: it orders the conditioning observations by their distance to the target point and compares the outcomes attached to the q_y and q_x closest observations in each of two independent samples through a one-sided Kolmogorov–Smirnov statistic. Like nonparametric methods generally, the procedure requires two tuning parameters: the numbers of neighbors q_y and q_x, which play a role analogous to a bandwidth. What distinguishes the test is its scope. The same procedure applies without modification whether Y and X are continuous, discrete, or mixed, and it requires only that the conditional CDFs be continuous in z at each target point, rather than differentiable or otherwise smooth. It uses no kernel weighting and imposes no parametric structure on the conditional distributions.

Second, we establish the asymptotic validity of our test under two alternative asymptotic frameworks. In the first framework, the numbers of effective "local" observations, q_y and q_x, are held fixed as the sample size $n \to \infty$. In this setting, we show that the test statistic converges to a limiting experiment in which the induced order statistics behave as independent draws from the true conditional CDFs at the target point. This convergence yields a feasible critical value that remains valid without relying on resampling methods such as the bootstrap. Our regularity conditions in this fixed-q framework are mild: the conditional CDFs at the target point may contain finitely many discontinuities; Y and X may be continuous, discrete, or mixed; and we require only that the conditional CDFs $F_Y(t \mid z)$ and $F_X(t \mid z)$ be continuous in z at the target point, uniformly in t. This is a weaker condition than the smoothness assumptions, such as twice differentiability in z, typically imposed in nonparametric methods. In the second framework, the numbers of effective observations q_y and q_x are allowed to diverge as $n \to \infty$. Building on recent results on convergence rates for induced order statistics (Bugni et al., 2025), we derive an explicit, finite-sample bound on the rate at which the rejection probability of our test approaches its nominal level as a function of q_y, q_x, and n. We establish this result at a level of generality that mirrors how the test is used in practice: the conditioning point may be estimated from the data, and the tuning parameters (q_y, q_x) may themselves be chosen in a data-dependent manner, with the known-point and deterministic-q results each following as a special case. This second framework complements the fixed-q results by demonstrating that the test continues to control size when the number of effective observations increases, and it provides the theoretical basis for the data-dependent rule we propose for selecting q_y and q_x. A feature common to both frameworks is that the target point need not be known in advance: we allow it to be estimated from the data, for example, when it is

a quantile of the conditioning variable, and show that this does not affect the asymptotic validity of the test. This size control can be conservative when the outcomes are discrete or mixed, in which case the achieved level may lie strictly below α. As we discuss next, a refined critical value mitigates this when Y and X have few support points.

Third, we show that the proposed critical value aligns with the one obtained from a permutation-based approach when the random variables Y and X are both continuous, thus establishing a natural connection between our method and the broader literature on permutation-based inference. To the best of our knowledge, this result provides the first formal justification for the validity of permutation-based inference in testing unconditional stochastic dominance. We demonstrate that the critical value of our test cannot be improved when both Y and X are continuous. This connection is not merely technical: it situates our test alongside design-based approaches to treatment-effect inference, where permutation and randomization arguments are standard, and it explains why a data-independent critical value can deliver finite-sample control in our setting. We develop the connection in Section 5.3, where we also relate it to the design-based analysis of Caughey et al. (2023). However, this result on validity of permutation tests does not extend to the case when either Y or X is discretely distributed. For this latter case, we introduce a *refined* critical value, which is typically smaller than the default one we propose and only a function of the support points for Y and X. This refinement enhances power relative to the default critical value, though it comes at the cost of increased computational complexity.

Finally, we examine the finite-sample performance of our test through Monte Carlo simulations and present two empirical application to illustrate its implementation. The Monte Carlo simulations and the empirical applications suggest that the data-dependent rule we propose for selecting the key tuning parameters performs well in practice.

Related literature

Our work contributes to the extensive literature on inference for unconditional stochastic dominance. Early contributions include Hodges (1958) and McFadden (1989), with seminal subsequent developments in Anderson (1996), Davidson and Duclos (2000), Abadie (2002), Barrett and Donald (2003), Linton et al. (2005), Linton et al. (2010), Davidson and Duclos (2013), and Lok and Tabri (2021), among others. Whang (2019) provides a comprehensive textbook treatment of this literature. This literature studies hypotheses of stochastic dominance between marginal distributions and predominantly relies on asymptotic approximations and resampling methods. As explained above, our paper instead studies conditional stochastic dominance at specific target points. Nevertheless, our paper is connected to the unconditional literature because, in the limit experiment underlying our procedure, the conditional stochastic dominance testing problem reduces to a finite-sample unconditional one.

Our paper lies within the literature on inference for conditional stochastic dominance. As

explained in Whang (2019, Chapter 4), this literature considers two distinct testing problems. The first concerns whether conditional stochastic dominance holds over a range or the entire support of the conditioning variable. The second concerns whether it holds at one or more specific points in that support. As explained earlier, our paper studies the latter problem. There is an extensive literature on the former, including Delgado and Escanciano (2013), Gonzalo and Olmo (2014), Chang et al. (2015), Linton et al. (2016), Andrews and Shi (2017), Linton et al. (2023), and Wu and Kaplan (2025), among others. These procedures test restrictions that hold almost surely over a range of conditioning values and therefore do not directly provide inference for the conditional distributions at an isolated target point.

Our paper lies squarely within the literature on testing stochastic dominance at specific target points. A growing body of related work develops methods for distributional comparisons in this setting, including Donald et al. (2012), Qu and Yoon (2015), Shen and Zhang (2016), Shen (2019), Goldman and Kaplan (2018, Sections 6.1-6.2), and Qu and Yoon (2019). As mentioned earlier, the focus on particular target points may be dictated by the research design, as with the cutoff in an RDD, or by the substantive economic question, as with the median or another prespecified quantile of interest. Our method requires weaker assumptions than those developed in the previous studies. First, all existing methods assume that the conditional outcome distributions are continuous, and some impose additional smoothness conditions on the corresponding density functions. Our method, by contrast, accommodates continuous, discrete, and mixed data without requiring any modification to the inferential procedure. We view this flexibility with respect to the distributional form of the data as an important advantage of our approach. Second, some methods, such as that of Qu and Yoon (2019), are based on conditional quantiles and require the quantile indices to be bounded away from the tails of the distribution. We impose no such restriction. Finally, all existing methods treat the target point as fixed and known, whereas our method also permits it to be estimated, as occurs, for example, when the target point is defined as a quantile of the distribution of the conditioning variable. These differences are not merely incremental refinements. To our knowledge, none of the existing methods covers the two-sample problem at a target point while allowing one or both conditional outcome distributions to contain atoms.

The remainder of the paper is organized as follows. Section 2 defines the testing problem and introduces notation. Section 3 presents the induced order statistics that form the basis of our procedure and introduces the proposed test. Section 4 establishes its asymptotic validity under two alternative asymptotic frameworks. Section 5 discusses several refinements and extensions, including results on rates of convergence, a data-dependent rule for selecting the tuning parameters, and a refinement of the test that increases power when Y and X are discrete. Importantly, Section 5 also provides formal results on the validity of permutation tests for stochastic dominance with continuously distributed random variables, offering novel arguments for the validity of permutation-based inference in this setting. Section 6 evaluates the finite-sample performance of our test through Monte Carlo simulations. Finally, Section 7 concludes. All proofs are contained in the Appendix.

2 Testing problem

Let (Y, Z) and (X, Z) denote random pairs, each supported on $\mathbf{R} \times \mathbf{R}$. Let P_Y and P_X denote the joint distributions of (Y, Z) and (X, Z), respectively. For $t \in \mathbf{R}$ and z in the support of Z, define the conditional distribution functions

$$F_Y(t|z) := P_Y\{Y \leq t \mid Z = z\} \qquad \text{and} \qquad F_X(t|z) := P_X\{X \leq t \mid Z = z\} \ .$$

We are interested in testing the null hypothesis:

$$H_0 : F_Y(t|z) \leq F_X(t|z) \text{ for all } (t, z) \in \mathbf{R} \times \mathcal{Z} \tag{2.1}$$

versus the alternative hypothesis:

$$H_1 : F_Y(t|z) > F_X(t|z) \text{ for some } (t, z) \in \mathbf{R} \times \mathcal{Z} \ ,$$

where $\mathcal{Z} = \{z_1, \ldots, z_L\}$ is a finite set of target values of the conditioning variable. When $L = 1$, we write $\mathcal{Z} = \{z_0\}$. The case where $L = 1$ and Z is a continuously distributed random variable is both simpler and particularly relevant in empirical applications. To minimize notational clutter, we focus on this case for most of the remainder of the paper, unless otherwise stated, with Section C.1 addressing the case where $L > 1$.

To test this hypothesis, we assume the analyst observes two independent samples:

$$\{(Y_i, Z_i) : 1 \leq i \leq n_y\} \qquad \text{and} \qquad \{(X_j, Z_j) : 1 \leq j \leq n_x\} \ . \tag{2.2}$$

We refer to the first sample as the Y-sample, which consists of n_y i.i.d. draws from P_Y, and to the second sample as the X-sample, which consists of n_x i.i.d. draws from P_X. A leading example, and the one that motivates much of our analysis, is the sharp RDD. There, an outcome $\tilde{Y}$ is observed together with a running variable $\tilde{Z}$, treatment is assigned to units on one side of a known cutoff z_0, and interest centers on comparing the outcome distributions just above and just below z_0. The two samples in (2.2) arise by splitting the data at the cutoff,

$$\begin{aligned} \{(Y_i, Z_i) : 1 \leq i \leq n_y\} &= \{(\tilde{Y}_i, \tilde{Z}_i) : \tilde{Z}_i \leq z_0\}, \\ \{(X_j, Z_j) : 1 \leq j \leq n_x\} &= \{(\tilde{Y}_j, \tilde{Z}_j) : \tilde{Z}_j > z_0\} \ , \end{aligned}$$

so that testing (2.1) at the single target point z_0 asks whether one outcome distribution conditionally dominates the other at the threshold. In this design, z_0 is a boundary point of the support of Z within each subsample; our conditions accommodate this case, and we return to it in the Monte Carlo simulations of Section 6 and the empirical application of Section C.3.

Independence of the samples implies that the data-generating distribution is the product mea-

sure

$$P := P_Y \otimes P_X \ .$$

All probability and expectation statements in what follows are taken with respect to P unless otherwise noted. We denote by $\mathbf{P}$ the class of data-generating distributions P satisfying the regularity conditions imposed later, and let

$$\mathbf{P}_0 := \{P \in \mathbf{P} : (2.1) \text{ holds}\} \tag{2.3}$$

denote the subset of distributions $P \in \mathbf{P}$ satisfying the null hypothesis in (2.1).

Remark 1. *Our analysis assumes two independent samples. If instead one observes i.i.d. draws*

$$\{(Y_i, X_i, Z_i) : 1 \le i \le n\} \ ,$$

the nearest-neighbor step can still be applied by keeping the observations whose values of Z_i are closest to z_0. The resulting local experiment, however, consists of draws from the joint conditional law $\mathcal{L}\big((Y, X) \mid Z = z_0\big)$ rather than independent draws from P_Y and P_X, and since the null hypothesis in (2.1) only restricts the two marginal conditional distributions, the within-pair dependence is left unrestricted. This substantially changes the formal analysis and would likely lead to a different critical value, which we leave for future work. ■

3 Test based on induced ordered statistics

Let the observed data be those given in (2.2). Let (q_y, q_x) be two small positive integers (relative to (n_y, n_x)) and consider the point $z_0 \in \mathcal{Z}$. The test we propose is based on the following two samples:

- The q_y values of $\{Y_i : 1 \le i \le n_y\}$ associated with the q_y values of $\{Z_i : 1 \le i \le n_y\}$ closest to z_0, and
- The q_x values of $\{X_j : 1 \le j \le n_x\}$ associated with the q_x values of $\{Z_j : 1 \le j \le n_x\}$ closest to z_0.

To define these samples formally, we introduce g-order statistics for the conditioning variable Z, where $g(Z) := |Z - z_0|$; see Reiss (1989, Section 2.1) and Kaufmann and Reiss (1992). For any two values z, z' in the support of Z, we define the ordering $\le_g$ as follows:

$$z \le_g z' \quad \text{if and only if} \quad g(z) \le g(z') \ .$$

This defines a g-ordering on the support of Z. The g-order statistics $Z_{g,(i)}$ are then the values of

Z ordered according to this criterion:

$$Z_{g,(1)} \leq_g Z_{g,(2)} \leq_g \cdots \leq_g Z_{g,(n)} \ .$$

If there are ties in the g-ordering, they can be resolved arbitrarily—for example, by relying on the original sample order.

We then take the values of $\{Y_i : 1 \leq i \leq n_y\}$ associated with the q_y smallest g-ordered statistics of Z in the Y-sample, denoted by

$$Y_{n_y,[1]}, Y_{n_y,[2]}, \ldots, Y_{n_y,[q_y]} \ . \tag{3.1}$$

That is, $Y_{n_y,[j]} = Y_k$ if $Z_{g,(j)} = Z_k$ for $k = 1, \ldots, q_y$. Similarly, we take the values of $\{X_i : 1 \leq i \leq n_x\}$ associated with the q_x smallest g-ordered statistics of Z in the X-sample, denoted by

$$X_{n_x,[1]}, X_{n_x,[2]}, \ldots, X_{n_x,[q_x]} \ . \tag{3.2}$$

The random variables in (3.1) and (3.2) are referred to as *induced order statistics* or *concomitants of order statistics*; see David and Galambos (1974); Bhattacharya (1974); Canay and Kamat (2018). Intuitively, we view these samples as independent samples of Y and X, conditional on Z being 'close' to z_0. A key feature of our test is that it relies solely on these induced order statistics. Letting $n := \min\{n_y, n_x\}$ and $q := q_y + q_x$, the effective (pooled) sample is

$$S_n = (S_{n,1}, \ldots, S_{n,q}) := (Y_{n_y,[1]}, \ldots, Y_{n_y,[q_y]}, X_{n_x,[1]}, \ldots, X_{n_x,[q_x]}) \ . \tag{3.3}$$

It is also important to note that the first q_y elements of S_n are associated with the Y-sample, while the remaining q_x elements come from the X-sample.

Having defined the induced order statistics, we now define our test statistic as

$$T(S_n) = \sup_{t \in \mathbf{R}} \left(\hat{F}_{n,Y}(t) - \hat{F}_{n,X}(t) \right) = \max_{k \in \{1,\ldots,q_y\}} \left(\hat{F}_{n,Y}(S_{n,k}) - \hat{F}_{n,X}(S_{n,k}) \right) \ , \tag{3.4}$$

where the empirical CDFs are

$$\hat{F}_{n,Y}(t) := \frac{1}{q_y} \sum_{j=1}^{q_y} I\{S_{n,j} \leq t\} \quad \text{and} \quad \hat{F}_{n,X}(t) := \frac{1}{q_x} \sum_{j=1}^{q_x} I\{S_{n,q_y+j} \leq t\} \ .$$

The test statistic in (3.4) is a two-sample one-sided Kolmogorov–Smirnov (KS) test statistic, see Hajek et al. (1999, p. 99), and the test we propose rejects the null hypothesis in (2.1) when $T(S_n)$ exceeds a critical value, defined next.

To introduce the critical value, let $\alpha \in (0, 1)$ be given and $\{U_j : 1 \leq j \leq q\}$ be a sequence of

uniform random variables i.i.d., that is, $U_j \sim U[0,1]$. Define $\Delta_{q_y,q}(u)$ as

$$\Delta_{q_y,q}(u) := \frac{1}{q_y}\sum_{j=1}^{q_y} I\{U_j \le u\} - \frac{1}{q-q_y}\sum_{j=q_y+1}^{q} I\{U_j \le u\} \ , \tag{3.5}$$

and

$$c_\alpha(q_y,q) := \inf_{x\in\mathbf{R}}\left\{P\left\{\sup_{u\in(0,1)} \Delta_{q_y,q}(u) \le x\right\} \ge 1-\alpha\right\} \ . \tag{3.6}$$

The test we propose for the null hypothesis in (2.1) is

$$\phi(S_n) := I\{T(S_n) > c_\alpha(q_y,q)\} \ . \tag{3.7}$$

We reiterate that (3.7) corresponds to our test with a single target point ($L = 1$). Section C.1 extends this framework to the general case with multiple target points ($L > 1$). Furthermore, in addition to our default critical value in (3.6), we provide a refined critical value tailored for discrete Y and X with finite support in Section 5.2.

The choice of the two-sample one-sided KS statistic in (3.4) is crucial for accommodating discrete or mixed outcomes. Alternative statistics entail a familiar power trade-off: the KS statistic is naturally sensitive to pronounced positive peaks in $F_Y(t|z_0) - F_X(t|z_0)$ over a relatively narrow range, whereas integrated statistics such as Cramér–von Mises may have greater power against smaller but more diffuse departures, with no uniform ranking between them. In the continuous case, the Cramér–von Mises analogue remains asymptotically valid; with discrete or mixed outcomes, however, the analogous Cramér–von Mises and Anderson–Darling tests need not control size. This is why our default procedure relies on the KS statistic; see the discussion following Theorem 2 and Section C.2.

Remark 2. *From a computational perspective, a notable feature of the test $\phi(S_n)$ is that the supremum over $\mathbf{R}$ in (3.4) can be replaced by a maximum over the set $\{1,\dots,q_y\}$. This simplification arises because the KS test statistic increases only when evaluated at points corresponding to the Y-sample, which are the first q_y elements in the vector S_n. Consequently, the supremum in $c_\alpha(q_y,q)$ can also be replaced with a maximum over $\{1,\dots,q_y\}$. This allows us to compute the critical value $c_\alpha(q_y,q)$ with arbitrary accuracy through simulation.* ■

Remark 3. *When both Y and X are continuous, the test $\phi(S_n)$ is essentially exact; see Lemma 1. By contrast, in the discrete or mixed case, the test may be conservative because the default critical value in (3.6) is an upper bound on, rather than the relevant quantile. Consequently, the achieved level may remain strictly below α even asymptotically. We revisit this issue after Theorem 2 and, in Section 5.2, develop a refined critical value that is less conservative for discrete outcomes.* ■

Remark 4. *The values of q_y and q_x are tuning parameters chosen by the researcher. Section 4 shows that the test is asymptotically valid when (q_y,q_x) are either held fixed or allowed to diverge slowly with the sample size. In Section 5.1, we develop data-dependent guidelines for selecting*

(q_y, q_x) *that are compatible with these asymptotic requirements, and we assess their practical performance through simulations in Section 6 and in the empirical applications in Section C.3.* ■

Remark 5. *The RDD introduced in Section 2 is a running example for our test. With observed sample* $\{(\tilde{Y}_i, \tilde{Z}_i) : 1 \leq i \leq n\}$ *and cutoff* z_0*, the two samples in (2.2) are obtained by splitting the data at* z_0*, and the test compares the induced order statistics on either side of the threshold. The cutoff* z_0 *is a boundary point of the support of* Z *within each subsample, a case our assumptions accommodate.* ■

Remark 6. *In many applications, the conditioning point* z_0 *is not a fixed value but a feature of the distribution of* Z *that the researcher estimates from the same data—the leading example being the median, or another quantile, of* Z*. The results below, including the RDD setting discussed in Remark 5, are stated to formally account for this case: they allow* z_0 *to be replaced by a consistent estimator* $\hat{z}_n$ *that depends on the* Z*-observations alone, so our conclusions apply directly to such settings.* ■

4 Asymptotic framework and formal results

In this section, we derive the asymptotic properties of the test in (3.7) using two alternative asymptotic frameworks. The first one requires $q := q_y + q_x$ to be fixed as $n \to \infty$ and represents a finite sample situation where the effective number of observations used by the test is too small to credibly invoke approximations that require a "large" value of q. The second framework requires $q \to \infty$ slowly as $n \to \infty$, and represents a finite sample situation where the effective number of observations used by the test is large enough to invoke approximations for "large" q. Throughout, we allow the conditioning point z_0 to be either known or estimated from the data, as anticipated in Remark 6. Let $\mathcal{A}$ denote the σ-algebra generated by the Z-observations from both samples, and let $\hat{z}_n$ be the estimator of z_0 mentioned above. To avoid clutter, in the main text we continue to write S_n for the induced order statistics in (3.3), with the understanding that they are constructed using the nearest neighbors of either z_0 or $\hat{z}_n$, depending on the context; the appendix makes this distinction explicit.

4.1 Asymptotic results for fixed q

In this section, we examine the asymptotic properties of the test in (3.7) within a framework where $q := q_y + q_x$ is fixed and $n := \min\{n_y, n_x\} \to \infty$. We first derive the asymptotic properties of induced order statistics in (3.3), and then present our main theorem.

We start by deriving a result on the induced order statistics collected in the vector S_n in (3.3). To do so, we make the following assumptions. We state the assumptions for all $z \in \mathcal{Z}$ so that they also apply to the multiple-target case considered in Section C.1.

Assumption 1. *For each $z \in \mathcal{Z}$, there exists an $\mathcal{A}$-measurable estimator $\hat{z}_n$ such that $\hat{z}_n \xrightarrow{p} z$.*

Assumption 2. *For any $\varepsilon > 0$ and $z \in \mathcal{Z}$, $P\{Z \in (z-\varepsilon, z+\varepsilon)\} > 0$.*

Assumption 3. *For any $z \in \mathcal{Z}$ and any sequence $\{z_k\}$ in the support of Z such that $z_k \to z$,*

$$\sup_{t \in \mathbf{R}} \left| F_Y(t|z_k) - F_Y(t|z) \right| \to 0 \quad \textit{and} \quad \sup_{t \in \mathbf{R}} \left| F_X(t|z_k) - F_X(t|z) \right| \to 0 \ .$$

The $\mathcal{A}$-measurability requirement in Assumption 1 restricts $\hat{z}_n$ to be a function of the Z-observations alone—as with the empirical median or any empirical quantile of Z—and rules out estimators that depend on the outcomes Y or X. Taking $\hat{z}_n := z_0$ is allowed for, so this assumption nests the known-z_0 case as a special case. Assumption 2 requires that the distribution of Z is locally dense at each of the points in $\mathcal{Z}$. This includes the case where Z has a mass point at $z \in \mathcal{Z}$. Assumption 3 is a smoothness assumption required to guarantee that conditioning on observations close to z is informative about the distribution conditional on $Z = z$.

Theorem 1. *Let Assumptions 1–3 hold. Then,*

$$S_n \xrightarrow{d} S = (S_1, \ldots, S_q) \ , \tag{4.1}$$

where for any $s := (s_1, \ldots, s_q) \in \mathbf{R}^q$, the random vector S satisfies

$$P\{S \leq s\} = \prod_{j=1}^{q_y} F_Y(s_j|z_0) \cdot \prod_{j=q_y+1}^{q} F_X(s_j|z_0) \ .$$

Theorem 1 is a special case of Theorem 4 in the appendix with $L = 1$, which generalizes Canay and Kamat (2018, Theorem 4.1) in two directions: first, it accommodates multiple conditioning values ($L > 1$); and second, it allows the conditioning point z_0 to be estimated by $\hat{z}_n$ rather than known, in which case S_n is understood to be constructed using neighbors of $\hat{z}_n$. It establishes that the limiting distribution of the induced order statistics in the vector S_n is such that the elements of the vector, denoted by S, are mutually independent. Specifically, the first q_y elements of this vector follow the distribution $F_Y(\cdot|z_0)$, while the remaining q_x elements follow $F_X(\cdot|z_0)$. The proof leverages the fact that the induced order statistics S_n in (3.3) are conditionally independent given $(Z_1, \ldots, Z_n)$, with conditional CDFs

$$F_Y(\cdot|Z_{n_y,(1)}), \ldots, F_Y(\cdot|Z_{n_y,(q_y)}), F_X(\cdot|Z_{n_x,(1)}), \ldots, F_X(\cdot|Z_{n_x,(q_x)}) \ .$$

The result then follows by showing that $Z_{n_y,(j)} \xrightarrow{p} z_0$ and $Z_{n_x,(j)} \xrightarrow{p} z_0$ for all $j \in \{1, \ldots, q\}$, and invoking standard properties of weak convergence. Theorem 1 plays a crucial role in our asymptotic validity result presented in Theorem 2.

In addition to Assumptions 1 to 3, we also require that, conditional on $Z = z$, the random variables Y and X have distributions with a finite number of discontinuity points. To state this

assumption formally, let $\mathcal{D}_Y(z)$ and $\mathcal{D}_X(z)$ denote the sets of discontinuity points of the CDFs of $Y|Z=z$ and $X|Z=z$, respectively.

Assumption 4. *For any $z \in \mathcal{Z}$, $|\mathcal{D}_Y(z)|$ and $|\mathcal{D}_X(z)|$ are finite.*

It is important to note that Assumption 4 allows both Y and X to be continuous, discrete, or mixed random variables. However, it excludes cases where these variables have countably many discontinuities conditional on $z \in \mathcal{Z}$. We also point out that Theorem 1 does not require Assumption 4.

We now formalize our main result in Theorem 2, which shows that the test defined in (3.7) is asymptotically level α under the assumptions we just introduced. Below, we denote by $E_P[\cdot]$ the expected value with respect to the distribution $P \in \mathbf{P}$.

Theorem 2. *Let $\mathbf{P}$ be the space of distributions that satisfy Assumptions 1 to 4, and let $\mathbf{P}_0$ be as in (2.3). Let $\alpha \in (0,1)$ be given, T be the KS test statistic in (3.4), and $\phi(\cdot)$ be the test in (3.7). Then,*

$$\limsup_{n\to\infty} E_P[\phi(S_n)] \leq \alpha \qquad \text{for all } P \in \mathbf{P}_0 \ . \tag{4.2}$$

Theorem 2 establishes the asymptotic validity of the test in (3.7). There are three main reasons why the inequality in (4.2) may be strict, resulting in the limiting rejection probability strictly below α. First, when the inequality in (2.1) is strict, the test may reject with probability below α, with the magnitude depending on the distance of P from the least-favorable boundary of $\mathbf{P}_0$. Second, in cases where the support of Y and X take finitely many points, the critical value defined in (3.6) may be a strict upper bound for the desired quantile, as discussed further in Section 5.2. Finally, the test statistic $\Delta_{q_y,q}(u)$ in (3.5) is discretely distributed, taking only a limited number of distinct values. Consequently, the achieved significance level

$$\bar{\alpha} := P\left\{ \sup_{u\in(0,1)} \Delta_{q_y,q}(u) > c_\alpha(q_y,q) \right\} \tag{4.3}$$

satisfies $\bar{\alpha} \leq \alpha$ by definition, but may be strictly less than α. Whether $\bar{\alpha} = \alpha$ occurs or not depends on whether the critical value $c_\alpha(q_y,q)$ aligns exactly with one of the discrete jumps in the CDF of $\sup_{u\in(0,1)} \Delta_{q_y,q}(u)$, which depends on α, q_y, and q.

We derive Theorem 2 by linking the weak convergence of the induced order statistics S_n to the limit variable S in (4.1) with the finite-sample validity of the test $\phi(S)$ in the limit experiment. This connection becomes nontrivial when the data are not continuously distributed. The KS statistic plays a central role in addressing these challenges. First, weak convergence of S_n to S does not by itself imply convergence of the rejection probability, because the test is discontinuous at configurations involving ties. Our proof constructs an almost-sure representation under which $S_n \to S$ and shows that, with probability approaching one, all pairwise order relations, including ties, agree between the two vectors. On this event, the empirical CDF comparisons and therefore

$\phi(S_n)$ and $\phi(S)$ coincide. This yields convergence of the rejection probability to that in the limit experiment. Second, the structure of the KS statistic implies that our test controls size in the limit experiment over our class of null distributions—including those that are discrete or mixed—thereby establishing asymptotic validity. By contrast, as shown in Section C.2 in the appendix, analogous results generally fail when using the one-sided versions of the Cramér–von Mises or Anderson–Darling statistics, which are commonly used in the stochastic dominance testing. Nonetheless, in the special case where the data are continuously distributed, we show in Theorem 6 that the Cramér–von Mises-based test remains valid.

Two features are worth highlighting. First, the conclusion of Theorem 2 continues to hold when the conditioning point is estimated. The argument relies on only two properties of the neighbor selection underlying S_n: that it is measurable with respect to the Z-observations, so that conditionally on them the selected outcomes remain independent with conditional CDFs $F_Y(\cdot|Z_i)$ and $F_X(\cdot|Z_i)$; and that the selected Z-values converge in probability to z_0. Both properties survive when the neighborhoods are centered at an estimator $\hat{z}_n$: the first because $\hat{z}_n$ is $\mathcal{A}$-measurable, as required by Assumption 1, and the second because consistency of $\hat{z}_n$, combined with Assumption 2, places q observations in any fixed neighborhood of z_0 with probability approaching one. As a result, replacing z_0 by $\hat{z}_n$ leaves the limiting behavior of the test, and hence its asymptotic validity, unchanged. Second, the fixed-q framework cannot, in general, deliver consistency against all fixed alternatives. As $n \to \infty$, the selected observations become increasingly well localized around z_0, but the limiting experiment continues to contain only $q_y + q_x$ observations. Hence, under a fixed alternative, the rejection probability converges to the power of the test in the corresponding fixed-q ideal experiment, which is generally strictly below one. The relevant sample sizes for power in this framework are therefore q_y and q_x, rather than the original sample sizes n_y and n_x. Consistency is obtained in the large-q framework of Section 4.2; see Theorem 3.

Remark 7. *Goldman and Kaplan (2018) develop an inference framework for the two-sample one-sided hypothesis test in (2.1), interpreting it as a multiple-testing procedure and with applications to the RDD. As in this section, they adopt an asymptotic framework with fixed q and construct critical values by simulating i.i.d. $U(0,1)$ random variables given a test statistic. Our approach to stochastic dominance testing, however, differs from theirs in several important respects. First, we focus on the KS statistic, whereas they advocate for a different statistic based on the so-called Dirichlet approach, which they argue may offer power advantages. This distinction allows us to accommodate discrete or mixed distributions, while their analysis is restricted to continuously distributed data; see Appendix C.2 for details.*[1] *Second, we also derive results under an asymptotic framework where $q \to \infty$ as $n \to \infty$. This extension allows us to obtain data-dependent rules for choosing the tuning parameters that satisfy the required rate-of-convergence conditions, and we establish the asymptotic validity of our test when these data-dependent rules are used. Third, we allow the conditioning*

[1] Although Goldman and Kaplan (2018, p. 146) state that their test remains asymptotically valid—albeit conservative—for discrete data, they do not provide a formal proof. By contrast, Theorem 2 establishes this property for the KS statistic. Interestingly, asymptotic validity does not extend to other test statistics: Appendix C.2 shows that it fails for tests based on the Cramér–von Mises and Anderson–Darling statistics.

point z_0 to be estimated from the data (e.g., a quantile of Z) rather than treated as known, and show that this leaves the validity of our test unaffected. ■

4.2 Asymptotic results for large q

In this section, we examine the asymptotic properties of the test in (3.7) within a framework where $q := q_y + q_x \to \infty$ as $n \to \infty$. Motivated by the practical considerations discussed in Section 5.1, we do so at a level of generality that accommodates two features of how the test is used: the conditioning point may be estimated from the data, and the tuning parameters (q_y, q_x) may be data dependent. The known-point, deterministic-q version of our result then follows as a special case; see Remark 8. The assumptions in this section are stronger than those in Section 4.1. This is natural: the weaker assumptions of Section 4.1 ensure convergence to the limit experiment but are silent about its rate, whereas the results below deliver explicit rates and therefore require local smoothness and local mass conditions at z_0.

We first introduce the required notation. Let

$$\mathcal{A} := \sigma\big(\{Z_i : 1 \le i \le n_y\} \cup \{Z_j : 1 \le j \le n_x\}\big)$$

denote the σ-algebra generated by the Z observations from both samples. Given an estimator $\hat{z}_n$ of z_0 and (possibly random) positive integers (q_y, q_x), define $\hat{S}_n$ exactly as S_n in (3.3), with the q_y and q_x nearest neighbors in each sample selected by their distance to $\hat{z}_n$ rather than to z_0, and with distance ties broken by any $\mathcal{A}$-measurable rule. For $W \in \{Y, X\}$ and z such that the conditional distribution is well defined, we denote by $P_{W,z}$ the conditional distribution of W given $Z = z$, whose CDF is $F_W(\cdot|z)$; the target laws are P_{Y,z_0} and P_{X,z_0}. The ideal local experiment is defined conditionally on (q_y, q_x): let

$$S^\circ := S^\circ(q_y, q_x) := (Y_1^\circ, \dots, Y_{q_y}^\circ, X_1^\circ, \dots, X_{q_x}^\circ)\ , \tag{4.4}$$

where, conditional on (q_y, q_x), the components are mutually independent and independent of the data, with $Y_i^\circ \sim P_{Y,z_0}$ and $X_j^\circ \sim P_{X,z_0}$. When (q_y, q_x) is deterministic, S° has the same distribution as the limit vector S in Theorem 1. Since the dimension of $\hat{S}_n$ is random, the objects we compare are the joint laws of $(q_y, q_x, \hat{S}_n)$ and (q_y, q_x, S°); when (q_y, q_x) is deterministic this reduces to comparing $\mathcal{L}(\hat{S}_n)$ and $\mathcal{L}(S^\circ)$. Finally, for probability measures P and Q on $\mathbf{R}$ with densities p and q with respect to a common σ-finite measure μ, the Hellinger distance is defined by $H^2(P,Q) := \frac{1}{2}\int(\sqrt{p}-\sqrt{q})^2 d\mu$, the total variation distance is $\mathrm{TV}(P,Q) := \frac{1}{2}\int |p-q|\, d\mu$ and they satisfy $\mathrm{TV}(P,Q) \le \sqrt{2}\, H(P,Q)$.

We work under the following four assumptions.

Assumption 5. *The estimator $\hat{z}_n$, the integers (q_y, q_x), the selected index sets, and the tie-breaking rule are $\mathcal{A}$-measurable. In addition, there exist deterministic sequences $\bar{q}_y \to \infty$ and $\bar{q}_x \to \infty$ with*

$\bar{q}_y/n_y \to 0$ *and* $\bar{q}_x/n_x \to 0$*, and a sequence* $\eta_n \downarrow 0$*, such that*

$$P\{q_y \le \bar{q}_y \text{ and } q_x \le \bar{q}_x\} \ge 1 - \eta_n \ .$$

Assumption 6. *There exist sequences* $\delta_n \downarrow 0$ *and* $\varepsilon_n \downarrow 0$ *such that* $P\{|\hat{z}_n - z_0| > \delta_n\} \le \varepsilon_n$*.*

Assumption 7. *There exist constants* $c_Y, c_X > 0$ *and* $r_0 > 0$ *such that, for all* $r \in (0, r_0)$ *and* $W \in \{Y, X\}$*,*

$$P\{|Z - z_0| \le r\} \ge c_W\, r \ ,$$

where Z *denotes the covariate of the sample associated with* W*.*

Assumption 8. *There exist constants* $L_Y, L_X < \infty$ *and a neighborhood* $\mathcal{N}$ *of* z_0 *such that, for all* $z \in \mathcal{N}$ *and* $W \in \{Y, X\}$*,*

$$H\big(P_{W,z}, P_{W,z_0}\big) \le L_W\, |z - z_0| \ .$$

The four assumptions play distinct roles. Assumption 5 is new to this section: it disciplines the data-dependent choices by requiring the estimator $\hat{z}_n$, the tuning parameters (q_y, q_x), the selected index sets, and the tie-breaking rule to be $\mathcal{A}$-measurable, and by imposing deterministic envelopes $(\bar{q}_y, \bar{q}_x)$ that cap (q_y, q_x) with probability approaching one. It holds in our setting: for deterministic (q_y, q_x) it is satisfied trivially with $\bar{q}_y = q_y$, $\bar{q}_x = q_x$, and $\eta_n = 0$, while for the data-dependent rule proposed in Section 5.1 we verify there that the requirement holds. Assumption 6 is the same requirement on $\hat{z}_n$ as Assumption 1 in the fixed-q analysis, now stated quantitatively in terms of a rate δ_n because this rate appears in the finite-sample bound of Theorem 3; it holds with $\delta_n = Mn^{-1/2}$ for any $\sqrt{n}$-consistent estimator, the sample median of the pooled Z observations being the leading example. Finally, Assumptions 7 and 8 strengthen Assumptions 2 and 3, respectively: local denseness of Z at z_0 is strengthened to mass proportional to the window length, and sup-norm continuity of the conditional CDFs is strengthened to a local Hellinger–Lipschitz condition. Assumption 8 holds, in particular, whenever the conditional densities of W given $Z = z$ are differentiable in quadratic mean at z_0 with finite Fisher information; see Bugni et al. (2025). Both strengthened conditions accommodate z_0 at the boundary of the support of Z.

Theorem 3. *Let Assumptions 5–8 hold. Then there exist constants* $C < \infty$ *and* $c > 0$ *such that,*

$$\mathrm{TV}\Big(\mathcal{L}\big(q_y, q_x, \hat{S}_n\big),\, \mathcal{L}\big(q_y, q_x, S^\circ\big)\Big) \ \le\ C\left(\frac{\bar{q}_y^{\,3/2}}{n_y} + \frac{\bar{q}_x^{\,3/2}}{n_x}\right) \ +\ C\left(\sqrt{\bar{q}_y} + \sqrt{\bar{q}_x}\right)\delta_n \ +\ \mathrm{Rem}_n \ , \tag{4.5}$$

where $\mathrm{Rem}_n := \varepsilon_n + \eta_n + e^{-c\,\bar{q}_y} + e^{-c\,\bar{q}_x}$*. Consequently,*

$$E_P[\phi(\hat{S}_n)] \ \le\ \alpha + C\left(\frac{\bar{q}_y^{\,3/2}}{n_y} + \frac{\bar{q}_x^{\,3/2}}{n_x}\right) \ +\ C\left(\sqrt{\bar{q}_y} + \sqrt{\bar{q}_x}\right)\delta_n \ +\ \mathrm{Rem}_n \quad \text{for all } P \in \mathbf{P}_0 \ . \tag{4.6}$$

In particular, the test is asymptotically level α *whenever*

$$\bar{q}_y = o(n_y^{2/3}) \ , \quad \bar{q}_x = o(n_x^{2/3}) \ , \quad \sqrt{\bar{q}_y}\,\delta_n \to 0 \ , \quad \sqrt{\bar{q}_x}\,\delta_n \to 0 \ . \tag{4.7}$$

If, in addition, $\min\{q_y, q_x\} \xrightarrow{p} \infty$, *then* $\lim_{n\to\infty} E_{P'}[\phi(\hat{S}_n)] = 1$ *for any* $P' \in \mathbf{P}_1$.

In (4.6), $\phi(\hat{S}_n)$ denotes the test in (3.7) computed with the realized pair (q_y, q_x); that is, the critical value is $c_\alpha(q_y, q)$ with $q = q_y + q_x$. The three terms in the bound (4.5) have distinct interpretations. The first, $C\big(\bar{q}_y^{3/2}/n_y + \bar{q}_x^{3/2}/n_x\big)$, is the cost of approximating the local experiment by the ideal one at a *known* conditioning point, and coincides with the rates for induced order statistics in Bugni et al. (2025) under quadratic mean differentiability. The second, $C\big(\sqrt{\bar{q}_y}+\sqrt{\bar{q}_x}\big)\delta_n$, is the price of *estimating* the conditioning point. The last, Rem_n, collects the low-probability events on which the localization of $\hat{z}_n$ or the envelopes for (q_y, q_x) fail, together with exponential terms that vanish as $\bar{q}_y, \bar{q}_x \to \infty$; under Assumptions 5 and 6 it vanishes automatically, so the growth conditions in (4.7) are what remains to ensure the bound tends to zero.

Remark 8. *When* $\hat{z}_n = z_0$ *and* (q_y, q_x) *are deterministic, the estimation term* $C\big(\sqrt{\bar{q}_y} + \sqrt{\bar{q}_x}\big)\delta_n$ *vanishes* ($\delta_n = 0$) *and the failure probabilities satisfy* $\varepsilon_n = \eta_n = 0$, *so the bound* (4.5) *reduces to*

$$\mathrm{TV}\big(\mathcal{L}(S_n), \mathcal{L}(S)\big) \;\le\; C\left(\frac{q_y^{3/2}}{n_y} + \frac{q_x^{3/2}}{n_x}\right) + e^{-cq_y} + e^{-cq_x}\,, \tag{4.8}$$

so the test is asymptotically level α *provided* $q_y = o(n_y^{2/3})$ *and* $q_x = o(n_x^{2/3})$. *These are the growth conditions we rely on in Section 5.1.* ■

Remark 9. *The proof of Theorem 3 does not rely on the conditional representations for nearest neighbors at a fixed center that underlie the rates in Bugni et al. (2025); those arguments require the center to be nonrandom relative to the sample used to select neighbors. Instead, we use that, conditional on* $\mathcal{A}$, *the selected outcomes are independent with conditional laws* P_{W,Z_i}, *regardless of how* $\hat{z}_n$ *and* (q_y, q_x) *were constructed from the* Z *observations. This approach is well suited under quadratic-mean-differentiability, where the pointwise rate in Assumption 8 is sharp; see Bugni et al. (2025).* ■

Remark 10. *In a framework where* $q \to \infty$, *one could alternatively define a test using a critical value based on the* $1-\alpha$ *quantile of the limiting distribution of the (scaled) KS statistic in* (3.4). *Denote this limiting critical value by* $c_\alpha(\infty)$. *Numerical evaluations show that our scaled critical value is smaller than the limiting critical value:* $\sqrt{q_y q_x/q}\, c_\alpha(q_y, q) \le c_\alpha(\infty)$ *for* $\alpha \in \{0.1, 0.05, 0.01\}$ *and* $(q_y, q_x) \in \{10, 20, \ldots, 500\}^2$, *with strict inequality in most cases. For example, when* $\alpha = 5\%$ *and* $(q_y, q_x) = (70, 70)$—*a value consistent with our simulations—we obtain* $\sqrt{q_y q_x/q}\, c_\alpha(q_y, q) = 1.1832 < 1.2239 = c_\alpha(\infty)$. *The smaller critical value associated with our test translates into an improvement in power of roughly* 4% *under a local-power approximation for location-shift alternatives. We omit the details for brevity, but they are available upon request.* ■

5 Discussion and extensions

5.1 Data-dependent choice of tuning parameters

We now discuss the practical considerations for implementing our test and propose a data-dependent rule for choosing the two tuning parameters q_y and q_x. Our goal is to provide practical guidance based on the data, rather than claiming optimality of any sort. The rule is constructed in two steps that mirror the structure of our asymptotic analysis: the rates of convergence in Section 4.2 determine how fast (q_y, q_x) may grow with (n_y, n_x), and a bias–standard deviation calculation for the estimators of the conditional CDFs, in the spirit of Armstrong and Kolesár (2018) and similar to the approach in Bugni and Canay (2021), determines the constants multiplying these rates. We examine the performance of this rule via Monte Carlo simulations in Section 6 and use it in the empirical application in Section C.3.

We start with the rate, which is pinned down by results that do not involve any tuning constants. By (4.6), the size distortion of the feasible test obeys a finite-sample bound in which the restriction on the tuning parameters is governed by the coupling term $q_y^{3/2}/n_y + q_x^{3/2}/n_x$; for a $\sqrt{n}$-consistent conditioning point, the growth conditions in (4.7) reduce to $q_y = o(n_y^{2/3})$ and $q_x = o(n_x^{2/3})$. This rate is sharp at conditioning points on the boundary of the support of Z by the results in Bugni et al. (2025) (the relevant case in regression discontinuity applications), so no faster bound is available for the class of distributions we allow for. Any rule of the form $q_y \propto n_y^{\gamma}$ with $\gamma < 2/3$ is thus consistent with asymptotic validity, and the only remaining question is which exponent to use within this range. To choose one, we balance two approximation errors that move in opposite directions as q_y grows. Taking q_y larger worsens the coupling between $\hat{S}_n$ and the ideal experiment S°, contributing a size distortion of order $q_y^{3/2}/n_y$. Taking q_y smaller degrades the experiment itself: S° consists of q_y i.i.d. draws from $F_Y(\cdot|z_0)$, so the Gaussian approximation to its rejection probability under local alternatives carries an error of order $q_y^{-1/2}$; see Bugni et al. (2025) for this decomposition in a general setting. The sum $q_y^{3/2}/n_y + q_y^{-1/2}$ is minimized at $q_y \propto n_y^{1/2}$, and we accordingly set $\gamma = 1/2$. This exponent lies strictly inside the range required for validity.

Having fixed the rate, we write the rule as $q_y^* = c_y\, n_y^{1/2}$ and calibrate the constant c_y through the bias and standard deviation of the implicit estimator of $F_Y(\cdot|z_0)$. Assume that the random variable Z is continuous with a density function $f_Z(\cdot)$ satisfying

$$|f_Z(z_1) - f_Z(z_2)| \leq C_Z|z_1 - z_2| \quad \text{for a Lipschitz constant } C_Z < \infty\ ,$$

and any values $z_1, z_2 \in \mathcal{Z}$. In addition, suppose that the conditional CDF of Y satisfies

$$\left|\frac{\partial F_Y(t|z)}{\partial z}\right| \leq C_Y \quad \text{for a constant } C_Y < \infty\ ,$$

and that the conditional CDF of X satisfies the same condition with a constant C_X. It can be

shown that the standardized bias B_{n_y,q_y} associated with the estimator of the conditional CDF $F_Y(\cdot|z_0)$ satisfies

$$|B_{n_y,q_y}| \;\leq\; \frac{q_y^{3/2}}{n_y} \cdot \frac{F_Y(t|z_0)\, C_Z + f_Z(z_0)\, C_Y}{4 f_Z^2(z_0)} \;. \tag{5.1}$$

Here, B_{n_y,q_y} is the bias multiplied by $\sqrt{q_y}$, so it is expressed on the same scale as the sampling fluctuations of the empirical CDFs. The two terms in the numerator capture the two sources of local bias: the curvature of the density of Z near z_0, and the variation of the conditional CDF in z, scaled by the density at the target point. For calibration, we compare this standardized bias with the sampling variability of the two-sample CDF difference that underlies the test. Under the least favorable configuration $F_Y(\cdot|z_0) = F_X(\cdot|z_0) =: F$, $\Delta_{q_y,q}(u)$ in (3.5) has variance $F(1-F)(1/q_y + 1/q_x)$. Under the balanced convention $q_x \approx q_y$, its standard deviation on the same $\sqrt{q}_y$ scale is $\sqrt{2F(1-F)}$. We calibrate the constant by the simplest bias-sd trade-off: requiring the bias to be no larger than one standard deviation, i.e. $|B_{n_y,q_y}| \leq \sqrt{2F(1-F)}$. Solving for q_y yields the benchmark

$$\tilde{q}_y \;=\; n_y^{2/3} \left(2F(1-F)\right)^{1/3} \left(\frac{4 f_Z^2(z_0)}{F\, C_Z + f_Z(z_0)\, C_Y}\right)^{2/3} \;. \tag{5.2}$$

The benchmark $\tilde{q}_y$ grows exactly at the $n_y^{2/3}$ rate. Our rule therefore keeps the constant in (5.2) and combines it with the exponent $\gamma = 1/2$ derived above.

It remains to choose the value of F at which to evaluate (5.2), since the factor $(2F(1-F))^{1/3}$ has no lower bound as t varies over $\mathbf{R}$. We evaluate it at $F = 1/2$, for two reasons. First, the least favorable configuration within the null hypothesis is $F_Y(\cdot|z_0) = F_X(\cdot|z_0)$, as discussed in Section 5.3, so the relevant comparison is between two estimators of a common conditional CDF. Second, under this configuration the two-sample empirical process $\Delta_{q_y,q}(u)$ in (3.5) has variance proportional to $u(1-u)$, which is maximal at $u = 1/2$: the KS supremum concentrates around the median, so the median is where the comparison between bias and sampling variability is relevant for the behavior of the test. Setting $F = 1/2$ in (5.2) and imposing the $n_y^{1/2}$ rate, we obtain

$$q_y^* \;=\; n_y^{1/2} \left(\frac{1}{2}\right)^{1/3} \left(\frac{4 f_Z^2(z_0)}{\frac{1}{2} C_Z + f_Z(z_0)\, C_Y}\right)^{2/3} \;. \tag{5.3}$$

The constants C_Z and C_Y cannot be chosen adaptively: it is impossible to estimate them in a way that delivers uniformly valid inference for testing (2.1) (see, e.g., Armstrong and Kolesár, 2018). Since any data-dependent rule for q_y must take a stand on C_Z and C_Y, we prioritize simplicity and compute them under a bivariate normal working model for (Y, Z), which gives

$$C_Z = \frac{1}{\sigma_Z^2 \sqrt{2\pi e}} \;, \qquad C_Y = \frac{|\rho_Y|}{\sigma_Z \sqrt{1-\rho_Y^2}} \frac{1}{\sqrt{2\pi}} \;, \qquad f_Z(z_0) = \phi_{\mu_Z,\sigma_Z}(z_0) \;,$$

where $\phi_{\mu_Z,\sigma_Z}(\cdot)$ denotes the density of a normal distribution with mean μ_Z and variance σ_Z^2, and ρ_Y is the correlation between Y and Z. The first constant is the maximal slope of the normal

density, attained at $\mu_Z \pm \sigma_Z$; the second is the maximal slope of the conditional CDF in z. We propose choosing q_y and q_x using the following data-dependent rules:

$$q_y^* := n_y^{1/2} \left(\frac{1}{2}\right)^{1/3} \left(\frac{4\,\phi^2_{\hat{\mu}_Z,\hat{\sigma}_Z}(\hat{z}_n)}{\dfrac{1}{2\hat{\sigma}_Z^2\sqrt{2\pi e}} + \phi_{\hat{\mu}_Z,\hat{\sigma}_Z}(\hat{z}_n)\,\dfrac{|\hat{\rho}_Y|}{\hat{\sigma}_Z\sqrt{1-\hat{\rho}_Y^2}}\,\dfrac{1}{\sqrt{2\pi}}} \right)^{2/3} \tag{5.4}$$

and

$$q_x^* := n_x^{1/2} \left(\frac{1}{2}\right)^{1/3} \left(\frac{4\,\phi^2_{\hat{\mu}_Z,\hat{\sigma}_Z}(\hat{z}_n)}{\dfrac{1}{2\hat{\sigma}_Z^2\sqrt{2\pi e}} + \phi_{\hat{\mu}_Z,\hat{\sigma}_Z}(\hat{z}_n)\,\dfrac{|\hat{\rho}_X|}{\hat{\sigma}_Z\sqrt{1-\hat{\rho}_X^2}}\,\dfrac{1}{\sqrt{2\pi}}} \right)^{2/3}, \tag{5.5}$$

where $\hat{\mu}_Z$ is the sample median of Z, $\hat{\sigma}_Z$ is the interquartile range of Z divided by $2\Phi^{-1}(3/4) \approx 1.349$, and $\hat{\rho}_Y := \sin(\pi\hat{\tau}_Y/2)$, with $\hat{\tau}_Y$ denoting Kendall's rank correlation between Y and Z in the first sample; $\hat{\rho}_X$ is defined analogously using the second sample. Each estimator is consistent for the corresponding parameter under the working model — in particular, $\tau = \frac{2}{\pi}\arcsin(\rho)$ for the bivariate normal, so $\sin(\pi\hat{\tau}/2)$ recovers the correlation exactly — while being insensitive to heavy tails and outliers, which is in line with treating normality as a reference only.

Two implementation details are worth making explicit. First, we round q_y^* and q_x^* up to the nearest integer and impose a small floor, so that the rule always returns a well-defined, non-degenerate neighborhood. Second, we compute the rank correlations $\hat{\rho}_Y$ and $\hat{\rho}_X$ from observations lying outside the test neighborhood, a leave-out rule, rather than from the full sample. This is what makes the selected (q_y, q_x) compatible with the measurability requirement of Theorem 3 (Assumption 5), so that the large-q validity result applies to the feasible test based on the data-dependent rule. Both samples share the same conditioning point, and the leave-out construction and the precise measurability statement are described in Appendix C.4.

The constant multiplying $n_y^{1/2}$ in (5.3) is intuitive for two reasons. First, it reflects that a density that varies more steeply with z, or a steeper conditional CDF in z, calls for smaller q_y: in such cases, nearby observations provide a poorer approximation of the quantities at z_0. Since the maximal slopes are determined by C_Z and C_Y, the rule is decreasing in both constants. Second, the rule accounts for low density at z_0. When $f_Z(z_0)$ is small, the q_y nearest observations are likely to be farther from z_0, again requiring smaller q_y.

Remark 11. *The rules in (5.4) and (5.5) are invariant to affine transformations of Z and, because they depend on the outcomes only through Kendall's rank correlation, to strictly increasing transformations of Y and X. The test itself shares both invariance properties, as the nearest-neighbor sets are unchanged by affine transformations of Z and the KS statistic is rank-based. A rule based on the Pearson correlation would instead select different values of (q_y, q_x) for, say, Y and $\log Y$,*

even though the test treats the two problems identically. ■

5.2 Refined critical value for discrete data

While our default critical value in (3.6) is valid as long as the distributions of random variables Y and X have at most finitely many discontinuities, it is possible to construct a smaller *refined* critical value when both variables are discretely distributed with a limited number of support points. Our test with the refined critical value still maintains the asymptotic validity, though it comes at the cost of additional computational complexity.

To motivate the refined critical value, let $\mathbf{Y}$ and $\mathbf{X}$ denote the support of Y and X, respectively. Our proof for the asymptotic validity (specifically, Theorem 5 in the appendix) relies on the following inequalities:

$$P\left\{\sup_{u\in\mathcal{U}} \Delta_{q_y,q}(u) > c_\alpha(q_y,q)\right\} \;\le\; P\left\{\sup_{u\in(0,1)} \Delta_{q_y,q}(u) > c_\alpha(q_y,q)\right\} \;\le\; \alpha.$$

where $\mathcal{U} := \cup_{t\in\mathbf{Y}}\{u = F_X(t|z_0)\}$ is the set of values that $F_X(t|z_0)$ takes as t varies over $\mathbf{Y}$ and $\Delta_{q_y,q}(u)$ is given in (3.5). Once we replace the set $\mathcal{U}$ with the interval $(0,1)$, our default critical value $c_\alpha(q_y,q)$, defined as a quantile of $\sup_{u\in(0,1)}\Delta_{q_y,q}(u)$ in (3.6), ensures the probability is bounded below α. However, replacing the set $\mathcal{U}$ with $(0,1)$ could be unnecessarily conservative when $\mathcal{U}$ contains only a few points—either because $\mathbf{Y}$ contains only a few points or because $F_X(t|z_0)$ takes few distinct values as t varies. In fact, the cardinality of the set $\mathcal{U}$ is determined by the smaller support size of Y and X.

To define our refined critical value, let r denote the smaller support size of Y and X,

$$r := \min\{|\mathbf{Y}|, |\mathbf{X}|\}\ ,$$

and let $\mathbf{U}_r$ denote the collection of all ordered r-tuples of distinct points in (0,1),

$$\mathbf{U}_r := \{(u_1,u_2,\ldots,u_r)\in(0,1)^r : u_1 < u_2 < \cdots < u_r\}\ .$$

We denote an arbitrary element of $\mathbf{U}_r$ by $\mathcal{U}_r$. Our refined critical value is defined as

$$c^r_\alpha(q_y,q) := \inf_{x\in\mathbf{R}}\left\{\inf_{\mathcal{U}_r\in\mathbf{U}_r} P\left\{\sup_{u\in\mathcal{U}_r}\Delta_{q_y,q}(u)\le x\right\}\ge 1-\alpha\right\}\ , \tag{5.6}$$

with $\Delta_{q_y,q}(u)$ as in (3.5), and the refined test for the null hypothesis in (2.1) when either Y or X is discretely distributed with a limited number of support points is thus

$$\phi^r(S_n) := I\{T(S_n) > c^r_\alpha(q_y,q)\}\ . \tag{5.7}$$

Section C.1 presents the general version of this test for the case $L > 1$.

The power gains of using $c_\alpha^r(q_y, q)$ over $c_\alpha(q_y, q)$ are most pronounced when r is small. Our numerical analysis shows the largest gains for $r \leq 10$. Thus, this refinement is most effective for discrete data with a limited number of support points, rather than all discrete settings. Moreover, the computational cost of $c_\alpha^r(q_y, q)$ increases with r, and so as r grows, the gains diminish while the cost rises.

We propose to compute $c_\alpha^r(q_y, q)$ numerically, by solving the optimization problem

$$c_\alpha^r(q_y, q) = \min \left\{ x \in [c_{\text{lb}}, c_{\text{ub}}] \cap \mathbf{T} : \inf_{\mathcal{U}_r \in \mathbf{U}_r} P\left\{ \max_{u \in \mathcal{U}_r} \Delta_{q_y,q}(u) \leq x \right\} \geq 1 - \alpha \right\} , \tag{5.8}$$

where $\mathbf{T}$ is the support of $\Delta_{q_y,q}(u)$ in (3.5), $c_{\text{ub}} = c_\alpha(q_y, q)$, and c_{lb} is given by

$$c_{\text{lb}} := \min_{x \in \mathbf{T}} \left\{ P\left\{ \max_{u \in \left\{ \frac{1}{1+r}, \frac{2}{1+r}, \ldots, \frac{r}{1+r} \right\}} \Delta_{q_y,q}(u) \leq x \right\} \geq 1 - \alpha \right\} .$$

That $c_{\text{ub}} = c_\alpha(q_y, q)$ is a valid upper bound is unsurprising given the preceding discussion, while c_{lb} is a valid lower bound because $\left\{ \frac{1}{1+r}, \ldots, \frac{r}{1+r} \right\}$ is a specific element of $\mathbf{U}_r$. In our numerical evaluations we often found $c_\alpha^r(q_y, q) = c_{\text{lb}}$, but not always, so the additional optimization in (5.8) cannot in general be avoided.

Remark 12. *The support* $\mathbf{T}$ *of* $\Delta_{q_y,q}(u)$ *is a discrete subset of* $[-1, 1]$ *that can be enumerated for modest values of* q*: its cardinality is bounded by* $(q_y + 1)(q_x + 1)$*, and it is independent of the realizations of the random variables and of the value* $u \in (0, 1)$*.* ■

The optimization in (5.8) is inexpensive for a given application. The refined critical value depends on the data only through (q_y, q_x) and the support sizes, not on the observed outcome values, so it need not be recomputed as the data change and can be tabulated in advance for the relevant (q_y, q, r, α). The outer search ranges over $[c_{\text{lb}}, c_{\text{ub}}] \cap \mathbf{T}$, which in our experience contains only a handful of points (recall $|\mathbf{T}| \leq (q_y + 1)(q_x + 1)$ by Remark 12); the cost that grows with the support size r comes instead from the inner minimization over r-tuples in $\mathbf{U}_r$. This is why we recommend the refinement only for small r (say $r \leq 10$), where the power gains are largest and the computation is light. The burden therefore becomes noticeable only when $c_\alpha^r(q_y, q)$ must be recomputed across many Monte Carlo replications; for a single empirical application it is negligible. Code that computes $c_\alpha^r(q_y, q)$ is provided in the replication package.

5.3 Properties of our test in the limit experiment

In this section, we study the properties of the test $\phi(\cdot)$ in (3.7) in the limit experiment associated with the asymptotic framework in Section 4.1. By Theorem 1, this test is equivalent to $\phi(S)$, where

$$S = (S_1, \ldots, S_q), \quad S_j \sim F_Y(\cdot|z_0) \text{ for } j \leq q_y \text{ and } S_j \sim F_X(\cdot|z_0) \text{ for } j > q_y . \tag{5.9}$$

In words, in the limit experiment, we observe one random sample of size q_y from the distribution $F_Y(\cdot|z_0)$ and the other independent random sample of size q_x from the distribution $F_X(\cdot|z_0)$. The KS test statistic in (3.4) is a function of S, and the critical value $c_\alpha(q_y, q)$ remains unchanged. We begin our discussion by focusing on the case where S is *continuously distributed*.

The finite-sample properties of the two-sample one-sided KS statistic have been extensively studied in the literature of testing equality of two continuous (unconditional) distributions. Early works established that the test statistic's finite-sample distribution is pivotal under the null and developed algorithms for its computation (Gnedenko and Korolyuk (1951); Korolyuk (1955); Blackman (1956); Hodges (1958); Hájek and Šidák (1967), and Durbin (1973)). Our critical value is obtained from this pivotal finite-sample distribution, despite the different null hypothesis.

This connection to the literature of testing equality of two distributions arises from the observation that the distribution $F_Y(\cdot|z_0) = F_X(\cdot|z_0)$ is the least favorable within the set of null distributions $\mathbf{P}_0$ in (2.3) satisfying stochastic dominance, a point first made by Lehmann (1951) and later reiterated by Hodges (1958); McFadden (1989), and Goldman and Kaplan (2018). We define the subset of continuous distributions in $\mathbf{P}_0$ that satisfy $F_Y(\cdot|z_0) = F_X(\cdot|z_0)$ as $\mathbf{P}_0^*$, and denote a generic element in $\mathbf{P}_0^*$ by P^*. Although these papers studied the distribution of KS statistic under P^*, they did not provide a formal proof that P^* determines the size of the test under the null hypothesis in (2.1). For completeness, we formally state this result in Lemma 1 below, and provide its proof.

Lemma 1. *Let $\mathbf{P}_0^* \subset \mathbf{P}_0$ be the subset of distributions P^* of the random variable S in (5.9), such that for a continuous CDF F, $S_j \sim F$ for all $j = 1, \ldots, q$. Let $\phi(\cdot)$ be the test defined in (3.7). Then, for any $P^* \in \mathbf{P}_0^*$, we have*

$$\sup_{P \in \mathbf{P}_0} E_P[\phi(S)] = E_{P^*}[\phi(S)] = \bar{\alpha} \ ,$$

where $\bar{\alpha} \leq \alpha$ is defined in (4.3).

Lemma 1 shows that when $S \sim P^* \in \mathbf{P}_0^*$, our test is 'exact' in that it achieves the closest possible rejection rate to α, defined as $\bar{\alpha}$ in (4.3). In this case, we have

$$T(S) \overset{d}{=} T(U) \quad \text{where} \quad \{U_j \sim U[0,1] : 1 \leq j \leq q\} \text{ are i.i.d.}$$

This demonstrates that the analytical (finite-sample) critical value $c_\alpha(q_y, q)$ for the one-sided KS test can be accurately approximated by simulating uniform random variables. However, when $S \sim P \in \mathbf{P}_0$ is such that $P\{S_i = S_j : i \neq j\} > 0$, the connection $T(S) \overset{d}{=} T(U)$ need not hold. In this case, $c_\alpha(q_y, q)$ is no longer the finite-sample *analytical* quantile of $T(S)$, but rather a valid upper bound.

When S is continuously distributed, the proposed test $\phi(S)$ is equivalent to a non-randomized permutation test. This connection establishes the validity of permutation tests for testing stochastic

dominance. While Hodges (1958) and McFadden (1989) suggested that a permutation test could be used for this purpose, they did not provide a formal justification. To the best of our knowledge, our proof of this result is novel.

To formally define a permutation test, we introduce the following notation. Let $\mathbf{G}$ denote the set of all permutations $\pi = (\pi(1), \ldots, \pi(q))$ of $\{1, \ldots, q\}$. The permuted values of S are given by

$$S^\pi = (S_{\pi(1)}, \ldots, S_{\pi(q)}) \ .$$

The (non-randomized) permutation test is then defined as follows:

$$\begin{aligned} \phi^{\mathrm{p}}(S) &:= I\left\{T(S) > c_\alpha^{\mathrm{p}}(S)\right\} \\ c_\alpha^{\mathrm{p}}(S) &:= \inf_{x \in \mathbf{R}} \left\{ \frac{1}{|\mathbf{G}|} \sum_{\pi \in \mathbf{G}} I\{T(S^\pi) \leq x\} \geq 1 - \alpha \right\} \ . \end{aligned} \tag{5.10}$$

It follows from standard arguments (see, e.g., Lehmann and Romano (2005, Ch. 15)) that when S is invariant to permutations, i.e., $S \overset{d}{=} S^\pi$, the randomized version of the test $\phi^{\mathrm{p}}(S)$ is exact in finite samples. However, under the null hypothesis in (2.1), we have that $S \overset{d}{\neq} S^\pi$ for some $P \in \mathbf{P}_0$, and so invariance (or the so-called randomization hypothesis) fails. Therefore, the traditional finite-sample arguments for validity no longer apply. Alternative arguments that claim validity of permutation tests when invariance does not hold typically require $q \to \infty$; see Chung and Romano (2013); Canay et al. (2017), and Bugni et al. (2018), among others. In our current setting, where q is fixed, such arguments do not apply.

We contribute to this literature by demonstrating that, when S is continuously distributed, our test is equivalent to a non-randomized permutation test. The formal statement of this result follows.

Lemma 2. *For any random variable $\tilde{S} \in \mathbf{R}^q$ with $P\{\tilde{S}_i \neq \tilde{S}_j : i \neq j\} = 1$, we have*

$$P\{c_\alpha^{\mathrm{p}}(\tilde{S}) = c_\alpha(q_y, q)\} = 1 \ . \tag{5.11}$$

Moreover, (5.11) no longer holds if $\tilde{S}$ is such that $P\{\tilde{S}_i = \tilde{S}_j : i \neq j\} > 0$.

Lemma 2 shows that our data-independent critical value in (3.6) is equivalent to the critical value of a permutation test when the random variable $\tilde{S}$ has no ties. Lemma 2 immediately implies when S_n in (3.3) is continuously distributed, we have

$$\phi^{\mathrm{p}}(S_n) \overset{a.s}{=} \phi(S_n) \ . \tag{5.12}$$

It follows from (5.12) and Theorem 2 that, when S is continuously distributed, a non-randomized permutation test controls the limiting rejection probability under the null hypothesis in (2.1). Importantly, this result holds even though invariance does not hold for all $P \in \mathbf{P}_0$.

Remark 13. *Lemmas 1 and 2 establish the validity of permutation tests for testing stochastic dominance in finite-sample settings. This result illustrates an instance where permutation tests can provide finite-sample valid inference even in settings where the randomization hypothesis does not hold. The only similar result we are aware of is that of Caughey et al. (2023), who consider a design-based framework with the null hypothesis $\tau_i \leq 0$, where τ_i represents a unit-level treatment effect. Like our work, their result is valid under the condition that the random variables are continuously distributed or that a random tie-breaking rule is applied to handle ties.* ■

The results in Lemmas 1 and 2 reveal interesting and novel connections between our test and classical arguments involving finite-sample critical values and permutation tests. However, these results critically depend on the random variable S being continuously distributed and do not extend to cases where S is discretely distributed.

When S is discrete and ties occur with positive probability, i.e., $P\{S_i = S_j\} > 0 : i \neq j$, the finite-sample distribution of the KS test statistic $T(S)$ depends on the number and location of these ties. A natural approach to handle ties is to redefine the test statistic to randomly break them, effectively making the test a randomized one. While this would allow us to establish an analog of Lemma 2 for discrete data, we do not pursue such an extension, as randomized tests are rarely used in practice. Despite our best efforts, we were unable to demonstrate that a permutation test could control size under the null hypothesis of stochastic dominance in (2.1) when S is discrete, without relying on random tie-breaking rules.

6 Simulations

In this section, we evaluate the finite-sample performance of the test in (3.7) for $L = 1$ or the test proposed in Section C.1 for $L > 1$ through a simulation study. We present a variety of data-generating processes to illustrate both the strengths and potential limitations of our test.

We consider seven distinct designs with four cases, (a) to (d), in each design as follows:

- **Case (a)**: the null hypothesis holds with equality and $L = 1$
- **Case (b)**: the null hypothesis holds with strict inequality and $L = 1$
- **Case (c)**: the null hypothesis holds with equality and $L = 2$
- **Case (d)**: the null hypothesis is violated and $L = 1$

The first three designs are based on the following location–scale model:

$$Y = \mu_Y(Z) + \sigma_Y(Z)U \quad \text{and} \quad X = \mu_X(Z) + \sigma_X(Z)V \ , \tag{6.1}$$

where U and V are random variables with specified distributions, and the conditioning variable Z follows a non-negative Beta$(2,2)$ distribution. Under this location–scale model, whenever $\sigma_Y(z_\ell) = \sigma_X(z_\ell)$, the null hypothesis in (2.1) holds as long as

$$\mu_Y(z_\ell) \geq \mu_X(z_\ell) \ . \tag{6.2}$$

Design 1 satisfies (6.2) for all $z \in (0,1)$. Design 2 satisfies (6.2) at $z_\ell = 0.5$, but violates the inequality for $z > 0.5$, which may affect the performance of our test in finite samples due to its reliance on induced order statistics. Design 3 is such that U and V are $U[0,1]$, which guarantees that (2.1) holds even when $\sigma_Y(z_\ell) < \sigma_X(z_\ell)$, provided $\mu_Y(z_\ell) - \mu_X(z_\ell) \geq \sigma_X(z_\ell) - \sigma_Y(z_\ell)$.

Design 4 is a slight variation of the location-scale model to accommodate an RDD:

$$Y = \mu_Y(Z) + \sigma_Y(Z)U \quad \text{if } Z \geq 0, \quad \text{and} \quad X = \mu_X(Z) + \sigma_X(Z)V \quad \text{if } Z < 0 \ . \tag{6.3}$$

Following Shen and Zhang (2016), we take U and V to be independent $N(0,1)$ random variables and $Z \sim 2\text{Beta}(2,2) - 1$. Both mean functions share the common baseline

$$\mu(z) = 0.61 - 0.02z + 0.06z^2 + 0.17z^3 \ , \tag{6.4}$$

with $\mu_Y(z)$ and $\mu_X(z)$ each equal to $\mu(z)$ or a constant shift of it, as specified in Table 4.

Design 5 is such that U and V follow log-normal distributions with the bottom 20% censored. This setup reflects features of wage distributions, which often exhibit a point mass at the minimum wage. Design 6 defines conditional probabilities $P\{X = k|Z\}$ and $P\{Y = k|Z\}$ as

$$P\{X = k|Z\} = \frac{e^{\theta_k^x(\frac{3}{2} - Z)}}{\sum_{j=1}^3 e^{\theta_j^x(\frac{3}{2} - Z)}} \quad \text{and} \quad P\{Y = k|Z\} = \frac{e^{\theta_k^y(\frac{3}{2} - Z)}}{\sum_{j=1}^3 e^{\theta_j^y(\frac{3}{2} - Z)}} \text{ for } k = 1,2,3,$$

where, for $\mu_Y(z) \in [-1,1]$,

$$\theta_1^y = \theta_1^x - \mu_Y(z), \quad \theta_2^y = \theta_2^x + \mu_Y(z), \quad \text{and} \quad \theta_3^y = \theta_3^x \ .$$

The parameter $\theta_k^x = (-0.5, -1.5, -2)$ controls the baseline log-odds of each category for X, while the factor $(3/2 - Z)$ introduces a monotonic dependence on Z. Finally, Design 7 considers

$$X|Z \sim B\Big([25Z], \frac{1}{2}\Big) \quad \text{and} \quad Y|Z \sim B\Big([25Z] + \mu_Y(Z), \frac{1}{2}\Big) \ ,$$

where $B(\cdot,\cdot)$ denotes a binomial distribution and $[x]$ represents the nearest integer to x.

The parameter values for all Designs are reported in Appendix C.4, where we also report the mean values of q_y^* and q_x^* across simulations. The results in this section compute the data dependent rule for q using the leave-out rule described in Appendix C.4. The results based on the alternative rules dicussed in the appendix are quite similar and are not reported here.

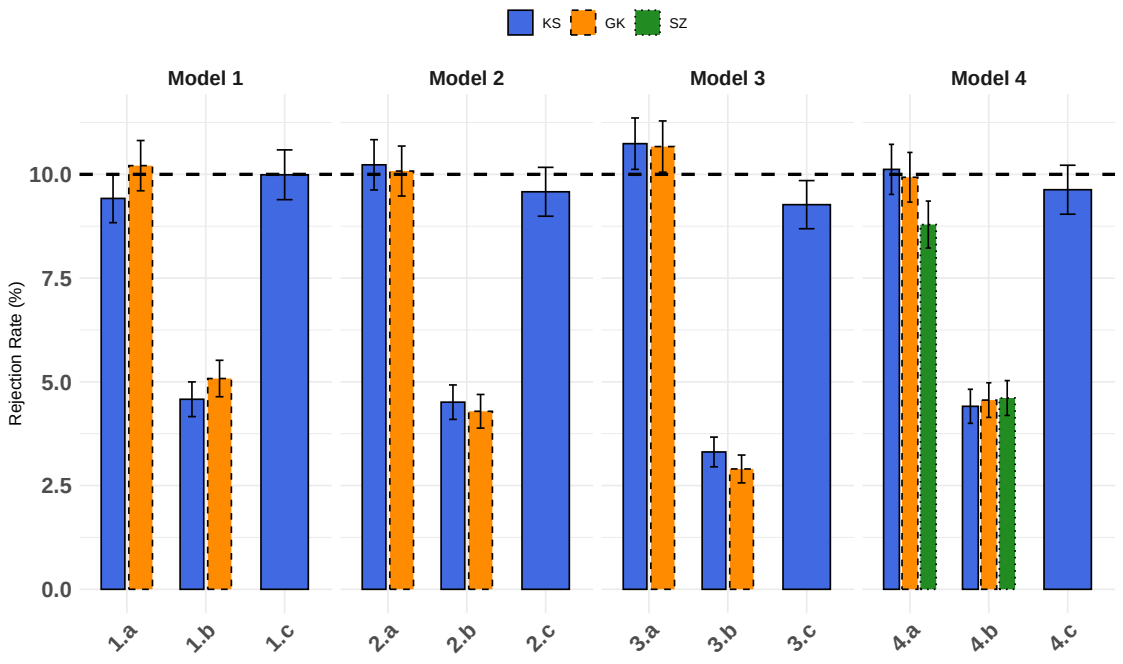

Figure 1: Rejection rates under H_0 for Designs 1-4 in cases (a)-(c): $n = 1,000$, $\alpha = 10\%$, $MC = 10,000$.

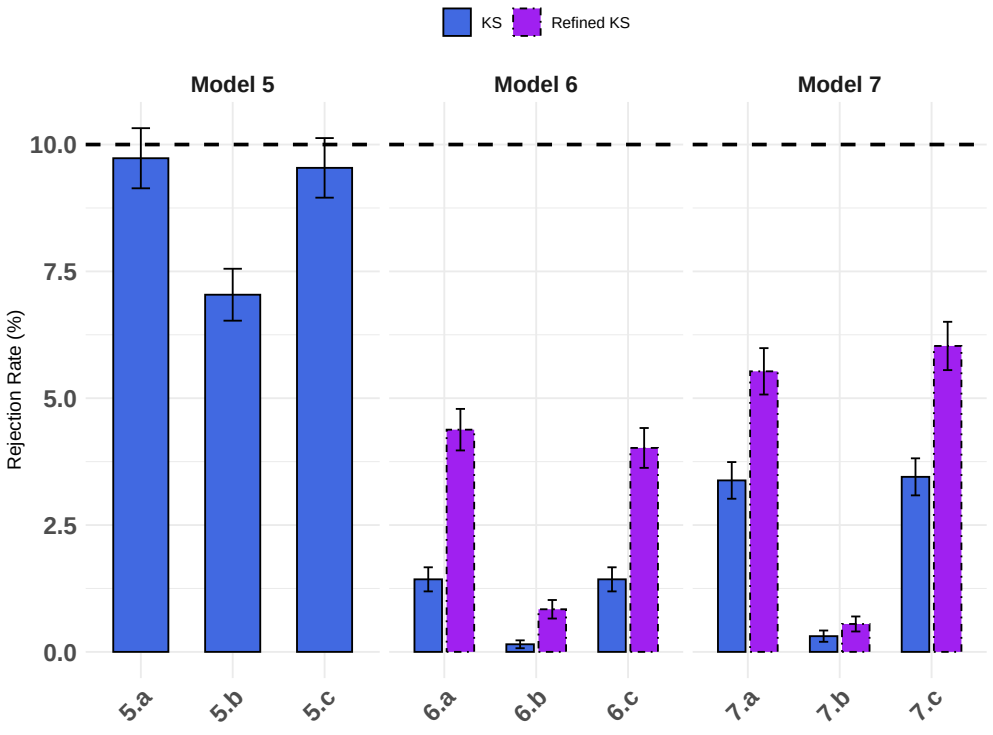

Figure 2: Rejection rates under H_0 for Designs 5-7 in cases (a)-(c): $n = 1,000$, $\alpha = 10\%$, $MC = 10,000$.

We report results for sample size $n = 1,000$ and nominal level $\alpha = 10\%$, based on $10,000$ Monte Carlo simulations to test the null hypothesis in (2.1). To implement the test $\phi(\cdot)$ in (3.7), denoted by 'KS' in the figures, we select the tuning parameters (q_y, q_x) using the data-dependent rules in (5.4) and (5.5). For Designs 1-3 with $L = 1$, we compare our test with the one proposed in Goldman and Kaplan (2018, GK). For the RDD Design 4 with $L = 1$, we compare our test with the method in Shen and Zhang (2016, SZ).[2] For the mixed and discrete designs, we present results for our test in (3.7) as well as for its refined version in (5.7).

Figure 1 reports rejection probabilities under the null hypothesis for the continuously distributed designs. When the data-generating process satisfies the null in (2.1) with equality (case (a)), the KS test's rejection probabilities closely align with the nominal level and consistently outperform the GK and SZ tests. When the null holds with strict inequality (case (b)), rejection rates fall below α, consistent with our critical value serving as a valid upper bound for the true quantile. Case (c), where $L = 2$, shows behavior similar to case (a), where $L = 1$. Overall, the KS test exhibits excellent size control.

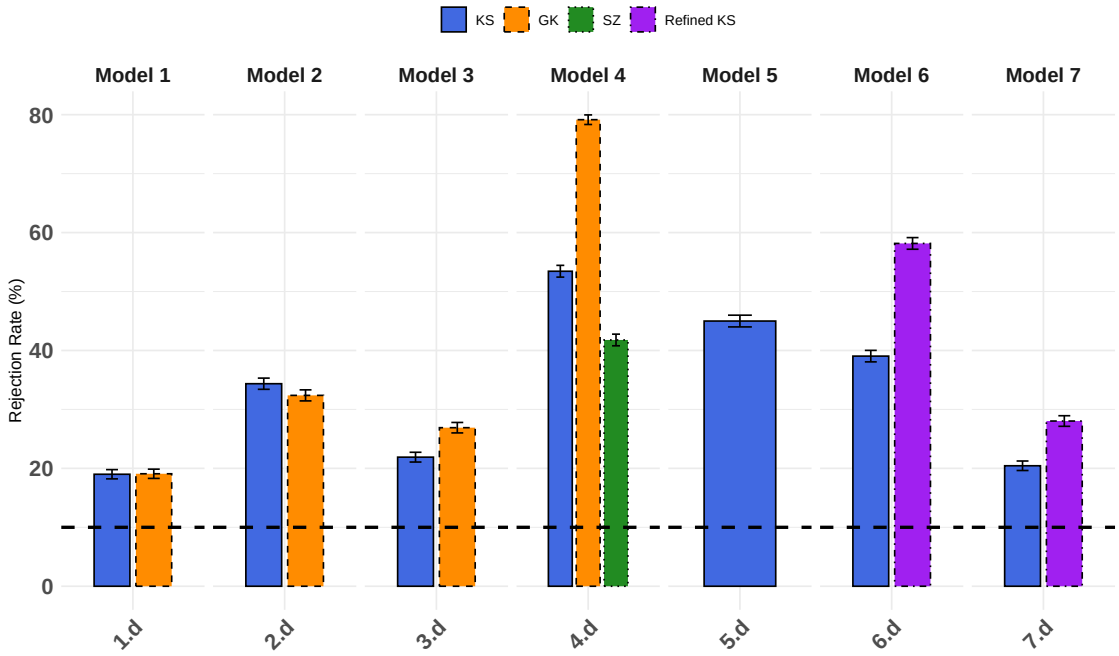

Figure 3: Rejection rates under H_1 for Designs 1-7: $n = 1,000$, $\alpha = 10\%$, $MC = 10,000$.

Figure 2 reports rejection probabilities under the null hypothesis for the mixed and discretely distributed designs. When the data-generating process satisfies the null in (2.1) with equality, the KS test we propose in (3.7) works well when the data is mixed, but it is conservative when the data is discrete with few support points. Designs 6 and 7 illustrate that the refined critical value

[2]SZ is implemented using the authors' rule of thumb, whereas for GK we use our own rule of thumb, as the original paper does not provide guidance on how to select their tuning parameter. Note that neither of these tests is defined when $L > 1$.

$c^{\mathrm{r}}_{\alpha}(q, q_y)$ in (5.6) offers a more accurate approximation of the true quantile of the KS test statistic, though it may still be somewhat conservative.

Figure 3 reports rejection probabilities under the alternative hypothesis for all designs. The power of the KS test is similar to that of GK and could be above, below, or roughly the same. The power of the KS test is much higher than that of SZ for this design. This is worth noting given that the KS test uses about 90 observations (for both q_y and q_x) while the SZ test uses 276 effective observations on each side of the threshold. The power results also demonstrate that the refined critical value $c^{\mathrm{r}}_{\alpha}(q, q_y)$ in (5.6) enhances power in discrete cases.

7 Concluding remarks

This paper introduces a novel test for conditional stochastic dominance (CSD) at target points, offering a flexible, nonparametric approach that avoids kernel smoothing while ensuring computational efficiency. By leveraging induced order statistics, our method constructs empirical CDFs using observations closest to the target conditioning point. We establish asymptotic validity of our test under two frameworks: one in which the number of neighbors is held fixed, where the induced order statistics converge to independent draws from the conditional distributions at the target point; and one in which it grows with the sample size, where we obtain an explicit rate that accommodates both an estimated target point and a data-dependent choice of the number of neighbors. Additionally, we extend our framework to better handle discrete data, proposing a refined critical value that enhances the power of the test with minimal additional information. Monte Carlo simulations align with our theoretical results and suggest that our test performs well in finite samples, making our test readily applicable to empirical research in economics, finance, and public policy.

An important feature of our test is its simplicity. Once the key tuning parameters are computed, the test only requires a standard test statistic with a deterministic critical value, without the need for kernels, local polynomials, bias correction, bandwidth selection or resampling techniques such as the bootstrap. Furthermore, our test admits a clear interpretation in the limit experiment, which allows us to connect it with classical analytical critical values and permutation-based tests. In this sense, our findings contribute to the broader literature on stochastic dominance testing by refining conditional inference methods and establishing new links between permutation-based and rank-based approaches. One open question we did not address in this paper concerns the validity of permutation-based tests for the hypothesis of stochastic dominance when both random variables, Y and X, are discrete. Despite attempts to formalize this result, we were unable to prove or disprove it. Extensive Monte Carlo simulations (not reported here) suggest that the test may be valid, and this is an area we plan to explore further.

Appendix A Proof of the main results

A.1 Proof of Theorem 2

By Theorem 1,

$$S_n \xrightarrow{d} S = (S_1, \cdots, S_q)$$

where the elements of S are independent, and $S_j \sim F_Y(\cdot|z_0)$ for $j = 1, \cdots, q_y$ and $S_j \sim F_X(\cdot|z_0)$ for $j = q_y + 1, \cdots, q$. By the almost-sure representation theorem, we have a sequence of random vectors $\{\tilde{S}_n : 1 \leq i \leq \infty\}$ and a random vector $\tilde{S}$ defined on a common probability space $(\Omega, \mathcal{A}, \tilde{P})$ such that

$$\tilde{S}_n \overset{d}{=} S_n, \ \tilde{S} \overset{d}{=} S, \text{ and } \tilde{S}_n \xrightarrow{a.s.} \tilde{S} \ .$$

Define the vector of pairwise weak-order relations of q, $R : \mathbf{R}^q \mapsto \{0,1\}^{q^2}$ such that

$$R(s) := \big(I\{s_i \leq s_j\} : 1 \leq i, j \leq q\big) \ .$$

Define the event that the rank of the two vectors coincides as follows,

$$F_n = \{R(\tilde{S}_n) = R(\tilde{S})\} \ .$$

To reach the conclusion, it suffices to show that

$$\tilde{P}\{F_n\} \to 1 \ . \tag{A-1}$$

To see this, consider the following argument,

$$E_P[\phi(S_n)] \overset{(1)}{=} E_{\tilde{P}}[\phi(\tilde{S}_n)] \overset{(2)}{=} E_{\tilde{P}}[\phi(\tilde{S})I\{F_n\} + \phi(\tilde{S}_n)I\{F_n^c\}] \overset{(3)}{\to} E_{\tilde{P}}[\phi(\tilde{S})] \overset{(4)}{\leq} \alpha \ ,$$

where (1) holds by $\tilde{S}_n \overset{d}{=} S_n$, (2) because $R(\tilde{S}_n) = R(\tilde{S})$ implies $T(\tilde{S}_n) = T(\tilde{S})$, (3) by (A-1) and $\phi(\cdot) \in \{0,1\}$, and (4) by $P \in \mathbf{P}_0$ and Theorem 5.

We devote the remainder of the proof to establishing (A-1). Define $\mathcal{D} = \mathcal{D}_X(z_0) \cup \mathcal{D}_Y(z_0)$, where $\mathcal{D}_X(z_0)$ and $\mathcal{D}_Y(z_0)$ denote the sets of discontinuity points as specified in Assumption 4. For any $\varepsilon > 0$, let

$$\begin{aligned}
E_1(\varepsilon) &:= \big\{\{|\tilde{S}_i - \tilde{S}_j| > \varepsilon\} \cup \{\tilde{S}_i = \tilde{S}_j \in \mathcal{D}\} : i \neq j = 1, \ldots, q\big\}, \\
E_{n,2}(\varepsilon) &:= \big\{|\tilde{S}_{n,k} - \tilde{S}_k| < \varepsilon/2 : k = 1, \ldots, q\big\}, \\
E_{n,3} &:= \big\{\{\tilde{S}_k \in \mathcal{D}\} \subseteq \{\tilde{S}_k = \tilde{S}_{n,k}\} : k = 1, \ldots, q\big\} \ .
\end{aligned}$$

Observe that

$$E_1(\varepsilon) \cap E_{n,2}(\varepsilon) \cap E_{n,3} \subseteq F_n. \tag{A-2}$$

To establish this, consider the following argument. For any $i, j = 1, \ldots, q$, there are three possible cases: (i) $\tilde{S}_i < \tilde{S}_j$, (ii) $\tilde{S}_i > \tilde{S}_j$, or (iii) $\tilde{S}_i = \tilde{S}_j$. First, consider case (i), where $\tilde{S}_i < \tilde{S}_j$. Under $E_1(\varepsilon)$, this implies $\tilde{S}_i < \tilde{S}_j - \varepsilon$. Under $E_{n,2}(\varepsilon)$, we have $\tilde{S}_{n,i} - \varepsilon/2 < \tilde{S}_i$ and $\tilde{S}_j < \tilde{S}_{n,j} + \varepsilon/2$. Combining these inequalities yields $\tilde{S}_{n,i} < \tilde{S}_{n,j}$, as required. Case (ii) follows identically by reversing the roles of i and j. Finally, consider

case (iii), where $\tilde{S}_i = \tilde{S}_j$. Under $E_1(\varepsilon)$, this implies $\tilde{S}_i = \tilde{S}_j \in \mathcal{D}$. By $E_{n,3}$, it follows that $\tilde{S}_{n,i} = \tilde{S}_{n,j} \in \mathcal{D}$. Since this argument holds for all $i, j = 1, \ldots, q$, we conclude that F_n follows, as desired.

By (A-2), (A-1) follows that there exits $\varepsilon > 0$ such that

$$\tilde{P}\{E_1(\varepsilon) \cap E_{n,2}(\varepsilon) \cap E_{n,3}\} \to 1. \tag{A-3}$$

For arbitrary $\delta > 0$, (A-3) follows from finding $\varepsilon = \varepsilon(\delta) > 0$ and $N(\delta)$ such that $\tilde{P}\{E_1(\varepsilon) \cap E_{n,2}(\varepsilon) \cap E_{n,3}\} \geq 1 - \delta$ for all $n \geq N(\delta)$. Let $\varepsilon_1 = \inf\{\|\tilde{d} - d\|/2 : d < \tilde{d} \in \mathcal{D}\} > 0$. By Lemma 4, $\exists \varepsilon_2 > 0$ such that, for $i \neq j = 1, \ldots, q$,

$$\tilde{P}\{\{|\tilde{S}_i - \tilde{S}_j| < \varepsilon_2\} \cap \{\tilde{S}_i, \tilde{S}_j \in \mathcal{D}^c\}\} < \delta/(9q(q-1)) \ , \tag{A-4}$$

$$\tilde{P}\{\cup_{d \in \mathcal{D}}\{|\tilde{S}_i - d| < \varepsilon_2\} \cap \{\tilde{S}_i \in \mathcal{D}^c\}\} < \delta/(9q(q-1)) \ . \tag{A-5}$$

Finally, set $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\} > 0$ for the remainder of the proof. By elementary arguments, it suffices to show that: (i) $\tilde{P}\{E_1(\varepsilon)^c\} \leq \delta/3$, (ii) $\exists N_2(\delta) \in \mathbf{N}$ s.t. $\tilde{P}\{E_{n,2}(\varepsilon)^c\} \leq \delta/3$ for all $n \geq N_2(\delta)$, and (iii) $\exists N_3(\delta) \in \mathbf{N}$ s.t. $\tilde{P}\{E_{n,3}(\varepsilon)^c\} \leq \delta/3$ for all $n \geq N_3(\delta)$. We divide the rest of the proof into three results.

First, we show that $\tilde{P}\{E_1(\varepsilon)^c\} \leq \delta/3$. Pick $i \neq j$ arbitrarily, and set $B_k := \big(\cup_{d \in \mathcal{D}}\{|\tilde{S}_k - d| < \varepsilon\}\big) \cap \{\tilde{S}_k \in \mathcal{D}^c\}$. On $\{\tilde{S}_i = \tilde{S}_j \in \mathcal{D}\}^c$, if $\tilde{S}_i, \tilde{S}_j \in \mathcal{D}$ they must differ, so $|\tilde{S}_i - \tilde{S}_j| \geq 2\varepsilon_1 \geq \varepsilon$ and the event $\{|\tilde{S}_i - \tilde{S}_j| < \varepsilon\}$ is empty; hence

$$\{|\tilde{S}_i - \tilde{S}_j| < \varepsilon\} \cap \{\tilde{S}_i = \tilde{S}_j \in \mathcal{D}\}^c \subseteq \big(\{|\tilde{S}_i - \tilde{S}_j| < \varepsilon\} \cap \{\tilde{S}_i, \tilde{S}_j \in \mathcal{D}^c\}\big) \cup B_i \cup B_j \ .$$

Each of the three sets has probability at most $\delta/(9q(q-1))$ by (A-4) and (A-5), both of which follow from Lemma 4 and require only independence (not identical distributions) so they apply to cross-sample pairs $i \leq q_y < j$. Summing the resulting bound $\delta/(3q(q-1))$ over $i \neq j$ gives $\tilde{P}\{E_1(\varepsilon)^c\} \leq \delta/3$, as desired.

Second, we show that $\exists N_2(\delta) \in \mathbf{N}$ such that $\tilde{P}\{E_{n,2}(\varepsilon)^c\} \leq \delta/3$ for all $n \geq N_2(\delta)$. To see this, note that

$$\tilde{P}\{E_{n,2}(\varepsilon)^c\} \leq \sum_{k=1}^{q} P\{|\tilde{S}_{n,k} - \tilde{S}_k| > \varepsilon/2\} \ . \tag{A-6}$$

By $\tilde{S}_n \overset{a.s.}{\to} \tilde{S}$, $\exists N_2(\delta)$ such that the right-hand side is less than $\delta/3$, as desired. Finally, we show that $\exists N_3(\delta) \in \mathbf{N}$ such that $\tilde{P}\{E_{n,3}(\varepsilon)^c\} \leq \delta/3$ for all $n \geq N_3(\delta)$. To see this, note that

$$\begin{aligned}\tilde{P}\{E_{n,3}(\varepsilon)^c\} &= \tilde{P}\{\exists k = 1, \ldots, q : \{\{\tilde{S}_k \in \mathcal{D}\} \subseteq \{\tilde{S}_k = \tilde{S}_{n,k}\}\}^c\} \\ &\leq \sum_{k=1}^{q} \tilde{P}\{\{\tilde{S}_k \in \mathcal{D}\} \cap \{\tilde{S}_k \neq \tilde{S}_{n,k}\}\} \ .\end{aligned}$$

By Lemma 3, $\exists N_3(\delta)$ such that the right-hand side is less than $\delta/3$, as desired. This completes the proof of (A-1) and the theorem.

A.2 Proof of Theorem 3

We focus on the case $L = 1$ for simplicity. Let

$$\mathbb{S} := \bigsqcup_{(a,b)\in\mathbf{N}^2} \{(a,b)\} \times \mathbf{R}^{a+b} , \qquad \mathbf{N} := \{1, 2, \dots\} ,$$

be equipped with the sigma-algebra

$$\mathcal{B} := \left\{ B \subseteq \mathbb{S} : B_{a,b} := \{ s \in \mathbf{R}^{a+b} : ((a,b), s) \in B \} \in \mathcal{B}(\mathbf{R}^{a+b}) \text{ for every } (a,b) \in \mathbf{N}^2 \right\} ,$$

where $\mathcal{B}(\mathbf{R}^{a+b})$ denotes the Borel sigma-algebra of $\mathbf{R}^{a+b}$. A function $\psi : \mathbb{S} \to \mathbf{R}$ is $\mathcal{B}$-measurable if and only if the section $s \mapsto \psi((a,b), s)$ is Borel measurable for every $(a,b) \in \mathbf{N}^2$; in particular, the test in (3.7), viewed as the map $((a,b), s) \mapsto I\{T(s) > c_\alpha(a, a+b)\}$, is $\mathcal{B}$-measurable.

For each $(a,b) \in \mathbf{N}^2$ with $a \le n_y$ and $b \le n_x$, let $\hat{M}_y(a)$ denote the index set of the a nearest neighbors of $\hat{z}_n$ in the first sample and $\hat{M}_x(b)$ that of the b nearest neighbors of $\hat{z}_n$ in the second sample, and let $\hat{S}_n^{a,b}$ denote the vector of induced order statistics in (3.3) formed from $\hat{M}_y(a)$ and $\hat{M}_x(b)$; for all remaining $(a,b) \in \mathbf{N}^2$, set $\hat{S}_n^{a,b} := 0 \in \mathbf{R}^{a+b}$, a choice that is immaterial in what follows. The sets $\hat{M}_y(a)$ and $\hat{M}_x(b)$ are nested in a and b and, being functions of the Z observations, $\hat{z}_n$, and the tie-breaking rule alone, they are $\mathcal{A}$-measurable for every fixed (a,b) by Assumption 5; this is the property used in Step 1 below. Similarly, let $\{S^\circ(a,b) : (a,b) \in \mathbf{N}^2\}$ be a family of random vectors, independent of the data, whose components are mutually independent with $Y_i^\circ \sim P_{Y,z_0}$ for $i \le a$ and $X_j^\circ \sim P_{X,z_0}$ for $j \le b$. With these definitions, $\hat{S}_n = \hat{S}_n^{q_y,q_x}$, and $S^\circ = S^\circ(q_y, q_x)$ by (4.4). We also record two conventions. First, $1 \le q_y \le n_y$ and $1 \le q_x \le n_x$ almost surely, as is automatic from the construction of the test; this is used when writing finite sums below. Second, since each $\mathbf{R}^{a+b}$ is a standard Borel space, regular conditional distributions of $\hat{S}_n^{a,b}$ given $\mathcal{A}$ exist, and their total variation distance to any fixed probability measure can be computed as a supremum over a countable field generating $\mathcal{B}(\mathbf{R}^{a+b})$; all conditional total variation and Hellinger quantities appearing below are therefore well defined and $\mathcal{A}$-measurable.

We proceed in four steps: a conditional product representation at fixed (a,b), a bound on the conditional Hellinger distance on a suitable event, the removal of the conditioning, and the implications for size and power.

Step 1: Fix $(a,b) \in \mathbf{N}^2$ with $a \le n_y$ and $b \le n_x$. By Assumption 5 and the preliminaries, conditionally on $\mathcal{A}$, the quantities $\hat{z}_n$, the index sets $(\hat{M}_y(a), \hat{M}_x(b))$, and the ordering of the selected indices are deterministic. Since observations are independently drawn within each sample, and the two samples are mutually independent, and each outcome is independent of all remaining Z observations conditional on its own, it follows that

$$\mathcal{L}\big(\hat{S}_n^{a,b} \,\big|\, \mathcal{A}\big) \;=\; \bigotimes_{i\in\hat{M}_y(a)} P_{Y,Z_i} \;\otimes\; \bigotimes_{j\in\hat{M}_x(b)} P_{X,Z_j} \;, \tag{A-7}$$

which repeats the conditional representation used in the proof of Theorem 1, the only difference being that the selection is governed by $\hat{z}_n$; the $\mathcal{A}$-measurability imposed in Assumption 5 is what guarantees that this difference is immaterial. Since the family $\{S^\circ(a,b)\}$ is independent of the data,

$$\mathcal{L}\big(S^\circ(a,b) \,\big|\, \mathcal{A}\big) \;=\; \big(P_{Y,z_0}\big)^a \otimes \big(P_{X,z_0}\big)^b \;. \tag{A-8}$$

Next, recall that the Hellinger affinity $1 - H^2$ is multiplicative across independent components, so for product

measures with components indexed by ℓ,

$$H^2\Big(\bigotimes_\ell P_\ell, \bigotimes_\ell Q_\ell\Big) = 1 - \prod_\ell \big(1 - H^2(P_\ell, Q_\ell)\big) \overset{(1)}{\le} \sum_\ell H^2(P_\ell, Q_\ell) \ ,$$

where (1) follows by induction from $1-(1-u)(1-v) \le u+v$ for $u, v \in [0,1]$. Applying this to (A-7) and (A-8) yields, for every (a,b) with $a \le n_y$ and $b \le n_x$,

$$H^2\Big(\mathcal{L}\big(\hat{S}_n^{a,b}|\mathcal{A}\big),\ \big(P_{Y,z_0}\big)^a \otimes \big(P_{X,z_0}\big)^b\Big) \ \le \sum_{i \in \hat{M}_y(a)} H^2\big(P_{Y,Z_i}, P_{Y,z_0}\big) + \sum_{j \in \hat{M}_x(b)} H^2\big(P_{X,Z_j}, P_{X,z_0}\big) \ . \tag{A-9}$$

Step 2: Define the radii

$$r_y := \frac{2}{c_Y} \cdot \frac{\bar{q}_y}{n_y} \ , \qquad r_x := \frac{2}{c_X} \cdot \frac{\bar{q}_x}{n_x} \ , \tag{A-10}$$

and the counts

$$N_y := \sum_{i=1}^{n_y} I\{|Z_i - z_0| \le r_y\} \ , \qquad N_x := \sum_{j=1}^{n_x} I\{|Z_j - z_0| \le r_x\} \ .$$

Since $\bar{q}_y/n_y \to 0$, we have $r_y < r_0$ for all n large enough with $r_0 > 0$ as in Assumption 7. Thus, for all n large enough, Assumption 7 yields $p_y := P\{|Z_i - z_0| \le r_y\} \ge c_Y r_y = 2\bar{q}_y/n_y$ and hence $E[N_y] = n_y p_y \ge 2\bar{q}_y$. As N_y is a sum of n_y i.i.d. indicators, the multiplicative Chernoff bound (see, e.g., Dubhashi and Panconesi, 2009, Theorem 1.1) gives, for all n large enough,

$$P\{N_y < \bar{q}_y\} \le P\{N_y \le E[N_y]/2\} \le \exp\big(-E[N_y]/8\big) \le e^{-\bar{q}_y/4} \ . \tag{A-11}$$

An analogous argument yields $P\{N_x < \bar{q}_x\} \le e^{-\bar{q}_x/4}$ for all n large enough. Define the event

$$\mathcal{G}_n := \big\{|\hat{z}_n - z_0| \le \delta_n\big\} \cap \big\{q_y \le \bar{q}_y,\ q_x \le \bar{q}_x\big\} \cap \big\{N_y \ge \bar{q}_y,\ N_x \ge \bar{q}_x\big\} \ , \tag{A-12}$$

and note that $\mathcal{G}_n \in \mathcal{A}$ and, by Assumptions 5–6 and (A-11),

$$P\{\mathcal{G}_n^c\} \le \varepsilon_n + \eta_n + e^{-\bar{q}_y/4} + e^{-\bar{q}_x/4} \ . \tag{A-13}$$

On $\mathcal{G}_n$, at least $\bar{q}_y$ observations of the first sample lie within r_y of z_0, and hence within $r_y + \delta_n$ of $\hat{z}_n$; consequently, for every $a \le \bar{q}_y$, each of the a nearest neighbors of $\hat{z}_n$ satisfies $|Z_i - \hat{z}_n| \le r_y + \delta_n$, and therefore

$$|Z_i - z_0| \le |Z_i - \hat{z}_n| + |\hat{z}_n - z_0| \le r_y + 2\delta_n \ , \qquad i \in \hat{M}_y(a) \ , \quad a \le \bar{q}_y \ . \tag{A-14}$$

An analogous argument yields

$$|Z_j - z_0| \le r_x + 2\delta_n \ , \qquad j \in \hat{M}_x(b) \ , \quad b \le \bar{q}_x \ . \tag{A-15}$$

For all n large enough, $r_y + 2\delta_n$ and $r_x + 2\delta_n$ lie inside the neighborhood $\mathcal{N}$ of Assumption 8, so, on $\mathcal{G}_n$ and for every (a,b) with $a \le \bar{q}_y$ and $b \le \bar{q}_x$, we obtain

$$H^2\Big(\mathcal{L}\big(\hat{S}_n^{a,b}|\mathcal{A}\big),\ \big(P_{Y,z_0}\big)^a \otimes \big(P_{X,z_0}\big)^b\Big) \ \overset{(1)}{\le} \ L_Y^2\, a\, (r_y + 2\delta_n)^2 + L_X^2\, b\, (r_x + 2\delta_n)^2$$

$$\overset{(2)}{\leq} C_1 \left\{ a \left(\frac{\bar{q}_y}{n_y} + \delta_n \right)^2 + b \left(\frac{\bar{q}_x}{n_x} + \delta_n \right)^2 \right\} , \tag{A-16}$$

where C_1 depends only on (L_Y, L_X, c_Y, c_X), (1) holds by (A-9), (A-14), (A-15), and Assumption 8, and (2) by (A-10). By increasing the constant C_1 if necessary, (A-16) can be extended to all n.

Step 3: Since the events $\{q_y = a, q_x = b\}$ partition the sample space (by the convention $1 \leq q_y \leq n_y$, $1 \leq q_x \leq n_x$) and $\hat{S}_n = \hat{S}_n^{a,b}$ on $\{q_y = a, q_x = b\}$, for every $B \in \mathcal{B}$,

$$I\{((q_y, q_x), \hat{S}_n) \in B\} = \sum_{a \leq n_y} \sum_{b \leq n_x} I\{q_y = a, q_x = b\}\, I\{\hat{S}_n^{a,b} \in B_{a,b}\} ,$$

and the same identity holds with $S^\circ(a,b)$ in place of $\hat{S}_n^{a,b}$. Therefore,

$$P\{((q_y, q_x), \hat{S}_n) \in B \mid \mathcal{A}\} = \sum_{a \leq n_y} \sum_{b \leq n_x} I\{q_y = a, q_x = b\}\, P\{\hat{S}_n^{a,b} \in B_{a,b} \mid \mathcal{A}\} , \tag{A-17}$$

$$P\{((q_y, q_x), S^\circ) \in B \mid \mathcal{A}\} = \sum_{a \leq n_y} \sum_{b \leq n_x} I\{q_y = a, q_x = b\} \left(P_{Y,z_0}\right)^a \otimes \left(P_{X,z_0}\right)^b (B_{a,b}) , \tag{A-18}$$

where (A-17) and (A-18) use that $I\{q_y = a, q_x = b\}$ is $\mathcal{A}$-measurable and that the sums have finitely many terms, and the latter also uses (A-8). Letting

$$\Delta_{a,b}(B_{a,b}) := P\big\{\hat{S}_n^{a,b} \in B_{a,b} \mid \mathcal{A}\big\} - \left(P_{Y,z_0}\right)^a \otimes \left(P_{X,z_0}\right)^b (B_{a,b}) ,$$

we obtain

$$\begin{aligned}
\mathrm{TV}\Big(\mathcal{L}\big(q_y, q_x, \hat{S}_n\big), \mathcal{L}\big(q_y, q_x, S^\circ\big)\Big) &\overset{(1)}{=} \sup_{B \in \mathcal{B}} \Big| E\Big[P\big\{((q_y, q_x), \hat{S}_n) \in B \mid \mathcal{A}\big\} - P\big\{((q_y, q_x), S^\circ) \in B \mid \mathcal{A}\big\} \Big] \Big| \\
&\overset{(2)}{=} \sup_{B \in \mathcal{B}} \Big| E\Big[\sum_{a \leq n_y} \sum_{b \leq n_x} I\{q_y = a, q_x = b\}\, \Delta_{a,b}(B_{a,b}) \Big] \Big| \\
&\overset{(3)}{\leq} E\left[\sup_{B \in \mathcal{B}} \Big| \sum_{a \leq n_y} \sum_{b \leq n_x} I\{q_y = a, q_x = b\}\, \Delta_{a,b}(B_{a,b}) \Big| \right] \\
&\overset{(4)}{=} E\left[\sum_{a \leq n_y} \sum_{b \leq n_x} I\{q_y = a, q_x = b\}\, \mathrm{TV}\Big(\mathcal{L}\big(\hat{S}_n^{a,b} \mid \mathcal{A}\big), \left(P_{Y,z_0}\right)^a \otimes \left(P_{X,z_0}\right)^b \Big) \right] ,
\end{aligned} \tag{A-19}$$

where (1) holds by the definition of total variation on $(\mathbb{S}, \mathcal{B})$ and the law of iterated expectations, (2) by (A-17) and (A-18), (3) by the triangle inequality and moving the supremum inside the expectation, and (4) because, for each realization, exactly one indicator is nonzero, so the inner supremum equals the total variation distance between the corresponding pair of laws on $\mathbf{R}^{a+b}$, attained by optimizing $B_{a,b}$ at the realized $(a,b) = (q_y, q_x)$.

Next, split the expectation in (A-19) over $\mathcal{G}_n$ and its complement:

$$\begin{aligned}
&E\left[\sum_{a \leq n_y} \sum_{b \leq n_x} I\{q_y = a, q_x = b\}\, \mathrm{TV}\Big(\mathcal{L}\big(\hat{S}_n^{a,b} \mid \mathcal{A}\big), \left(P_{Y,z_0}\right)^a \otimes \left(P_{X,z_0}\right)^b \Big) \right] \\
&\qquad \overset{(5)}{\leq} E\left[I\{\mathcal{G}_n\} \sum_{a \leq \bar{q}_y} \sum_{b \leq \bar{q}_x} I\{q_y = a, q_x = b\}\, \sqrt{2}\, H\Big(\mathcal{L}\big(\hat{S}_n^{a,b} \mid \mathcal{A}\big), \left(P_{Y,z_0}\right)^a \otimes \left(P_{X,z_0}\right)^b \Big) \right] + P\{\mathcal{G}_n^c\}
\end{aligned}$$

$$\overset{(6)}{\leq} \sqrt{2C_1}\left\{\sqrt{\bar{q}_y}\left(\frac{\bar{q}_y}{n_y}+\delta_n\right)+\sqrt{\bar{q}_x}\left(\frac{\bar{q}_x}{n_x}+\delta_n\right)\right\}+P\{\mathcal{G}_n^c\}$$
$$\overset{(7)}{\leq} C\left(\frac{\bar{q}_y^{3/2}}{n_y}+\frac{\bar{q}_x^{3/2}}{n_x}+\sqrt{\bar{q}_y}\,\delta_n+\sqrt{\bar{q}_x}\,\delta_n\right)+\varepsilon_n+\eta_n+e^{-\bar{q}_y/4}+e^{-\bar{q}_x/4}\,, \tag{A-20}$$

where (5) holds because on $\mathcal{G}_n$ the realized pair satisfies $q_y \leq \bar{q}_y$ and $q_x \leq \bar{q}_x$, TV $\leq \sqrt{2}\,H$, and on $\mathcal{G}_n^c$ the sum is bounded by one, since it contains a single nonzero total variation term, (6) by (A-16) applied at the (unique) realized pair $(a,b) = (q_y, q_x)$, which satisfies $a \leq \bar{q}_y$ and $b \leq \bar{q}_x$ on $\mathcal{G}_n$, together with $\sqrt{u+v} \leq \sqrt{u}+\sqrt{v}$ and $a \leq \bar{q}_y$, $b \leq \bar{q}_x$, and (7) by (A-13) and collecting constants. This establishes (4.5) with $c = 1/4$.

Step 4: The test ϕ in (3.7), computed with the realized pair (q_y, q_x), is the indicator of a set in $\mathcal{B}$. Hence, directly from the definition of total variation on $(\mathbb{S}, \mathcal{B})$,

$$\left|E_P[\phi(\hat{S}_n)] - E_P[\phi(S^\circ)]\right| \leq \mathrm{TV}\Big(\mathcal{L}\big(q_y, q_x, \hat{S}_n\big), \mathcal{L}\big(q_y, q_x, S^\circ\big)\Big)\,, \tag{A-21}$$

which parallels the argument in (A-19). Conditional on (q_y, q_x), the vector S° consists of q_y i.i.d. draws from P_{Y,z_0} and q_x i.i.d. draws from P_{X,z_0}, mutually independent, and the critical value is $c_\alpha(q_y, q)$. Therefore, for any $P \in \mathbf{P}_0$, Theorem 5 applied at the realized pair (q_y, q_x) yields

$$E_P\big[\phi(S^\circ)\,\big|\,q_y, q_x\big] \leq \alpha \tag{A-22}$$

Taking expectations in (A-22) and combining with (A-21) and (A-20) proves (4.6), and the asymptotic level claim follows under (4.7) since every term on the right-hand side of (A-20) then vanishes as $n \to \infty$.

For the power claim, fix $P' \in \mathbf{P}_1$, and let t^*, ε, and κ_α be as in Lemma 7. Since the family $\{S^\circ(a,b)\}$ is independent of the data and (q_y, q_x) is $\mathcal{A}$-measurable,

$$E_{P'}\big[\phi(S^\circ)\,\big|\,q_y, q_x\big] = h(q_y, q_x)\,, \qquad \text{where} \qquad h(a,b) := E_{P'}\big[\phi(S^\circ(a,b))\big]\,.$$

Fix any $m \in \mathbf{N}$ with $2\kappa_\alpha/\sqrt{m} < \varepsilon/3$. By Lemma 7(ii), $h(a,b) \geq 1 - 2e^{-2m\varepsilon^2/9}$ whenever $\min\{a,b\} \geq m$, and therefore

$$E_{P'}[\phi(S^\circ)] \;\geq\; E\big[h(q_y,q_x)\,I\{\min\{q_y,q_x\} \geq m\}\big] \;\geq\; \big(1-2e^{-2m\varepsilon^2/9}\big)\,P\{\min\{q_y,q_x\} \geq m\}\,.$$

Since $\min\{q_y, q_x\} \xrightarrow{p} \infty$ by hypothesis, $P\{\min\{q_y,q_x\} \geq m\} \to 1$ for every fixed m, so $\liminf_{n\to\infty} E_{P'}[\phi(S^\circ)] \geq 1-2e^{-2m\varepsilon^2/9}$. Letting $m \to \infty$ yields $E_{P'}[\phi(S^\circ)] \to 1$. Combining with (A-21) and (A-20), whose right-hand side vanishes under (4.7), gives $E_{P'}[\phi(\hat{S}_n)] \to 1$, as desired.

A.3 Proof of Lemma 1

Note that

$$E_{P^*}[\phi(S)] \overset{(1)}{\leq} \sup_{P\in\mathbf{P}_0} E_P[\phi(S)] \overset{(2)}{\leq} \bar{\alpha},$$

where (1) holds by $P^* \in \mathbf{P}_0$ and (2) by Theorem 5. To complete the proof, it suffices to show that $E_{P^*}[\phi(S)] = \bar{\alpha}$. To this end, consider the following argument:

$$
\begin{aligned}
& E_{P^*}[\phi(S)] \\
& \overset{(1)}{=} P^* \left\{ \sup_{t \in \mathbf{R}} \left(\frac{1}{q_y} \sum_{j=1}^{q_y} I\{S_j \leq t\} - \frac{1}{q_x} \sum_{j=q_y+1}^{q} I\{S_j \leq t\} \right) > c_\alpha(q_y, q) \right\} \\
& \overset{(2)}{=} P^* \left\{ \sup_{t \in \mathbf{R}} \left(\frac{1}{q_y} \sum_{j=1}^{q_y} I\{U_j \leq F(t)\} - \frac{1}{q_x} \sum_{j=q_y+1}^{q} \{U_j \leq F(t)\} \right) > c_\alpha(q_y, q) \right\} \\
& \overset{(3)}{=} P^* \left\{ \sup_{u \in (0,1)} \left(\frac{1}{q_y} \sum_{j=1}^{q_y} I\{U_j \leq u\} - \frac{1}{q_x} \sum_{j=q_y+1}^{q} I\{U_j \leq u\} \right) > c_\alpha(q_y, q) \right\} \\
& \overset{(4)}{=} \bar{\alpha} \ ,
\end{aligned}
\tag{A-23}
$$

where (1) holds by (3.4), (2) holds by Pollard (2002, Eq. 36) and the same arguments used in the proof of Theorem 5, (3) follows from the continuity of F guaranteeing that

$$
\sup_{t \in \mathbf{R}} \Delta_{q_y,q}\left(F(t)\right) = \sup_{u \in (0,1)} \Delta_{q_y,q}(u)
$$

for $\Delta_{q_y,q}(u) := \frac{1}{q_y} \sum_{j=1}^{q_y} I\{U_j \leq u\} - \frac{1}{q-q_y} \sum_{j=q_y+1}^{q} I\{U_j \leq u\}$ as defined in (3.5), and (4) by definition of $\bar{\alpha}$ in (4.3).

A.4 Proof of Lemma 2

Let $Q := \{Q_1, \ldots, Q_q\} = \{1, 2, \ldots, q\}$ and denote by $Q^\pi := \{Q_{\pi(1)}, Q_{\pi(2)} \ldots, Q_{\pi(q)}\}$ the permutation $\pi = (\pi(1), \pi(2), ..., \pi\,(q))$ of Q. Let $R(s)$ denote the rank of s, which maps s to a permutation of Q. Since the KS statistic $T(\cdot)$ in (3.4) is a rank statistic, it follows that for any s

$$
T(s) \ = \ T^*(R(s)) \ , \tag{A-24}
$$

where T^* is a known function; see Hajek et al. (1999, page 99). That is, the KS test statistic depends on S only through $R(S)$. Define

$$
c_\alpha^{\mathrm{p}} \ := \ \inf_{x \in \mathbf{R}} \left\{ \frac{1}{|\mathbf{G}|} \sum_{\pi \in \mathbf{G}} I\left\{T(Q^\pi) \leq x\right\} \geq 1 - \alpha \right\} \ . \tag{A-25}
$$

We divide the rest of the argument into four steps.

Step 1. For any $s \in \mathbf{R}^q$ with $s_i \neq s_j$ for $i \neq j$, and $c_\alpha^{\mathrm{p}}(s)$ as in (5.10),

$$
c_\alpha^{\mathrm{p}}(s) \ = \ c_\alpha^{\mathrm{p}} \ .
$$

To establish this, consider the following derivation,

$$\begin{aligned} c_\alpha^{\mathrm{p}}(s) &= \inf_{x\in\mathbf{R}}\left\{\frac{1}{|\mathbf{G}|}\sum_{\pi\in\mathbf{G}} I\left\{T(s^\pi)\le x\right\}\ge 1-\alpha\right\} \\ &\overset{(1)}{=} \inf_{x\in\mathbf{R}}\left\{\frac{1}{|\mathbf{G}|}\sum_{\pi\in\mathbf{G}} I\left\{T^*(R(s^\pi))\le x\right\}\ge 1-\alpha\right\} \\ &\overset{(2)}{=} \inf_{x\in\mathbf{R}}\left\{\frac{1}{|\mathbf{G}|}\sum_{\pi\in\mathbf{G}} I\left\{T^*((Q^{\bar\pi})^\pi)\le x\right\}\ge 1-\alpha\right\} \\ &\overset{(3)}{=} \inf_{x\in\mathbf{R}}\left\{\frac{1}{|\mathbf{G}|}\sum_{\pi\in\mathbf{G}} I\left\{T^*(Q^\pi)\le x\right\}\ge 1-\alpha\right\} \\ &\overset{(4)}{=} c_\alpha^{\mathrm{p}}\ . \end{aligned} \tag{A-26}$$

Here, (1) follows by (A-24), (2) follows since $s_i \neq s_j$ for $i \neq j$ implies that $R(s^\pi) = (R(s))^\pi$ and $R(s) = Q^{\bar\pi(s)}$ for some $\bar\pi(s) \in \mathbf{G}$, (3) by $\mathbf{G} = \tilde{\mathbf{G}} := \{\pi \circ \bar\pi(s) : \pi \in \mathbf{G}\}$, which follows from the fact that $\mathbf{G}$ is a group, and (4) by (A-25).

Step 2. $c_\alpha^{\mathrm{p}} = c_\alpha(q_y, q)$, where $c_\alpha(q_y, q)$ is defined in (3.6).

Let $\{U_i : 1 \le i \le q\}$ be i.i.d. with $U_i \sim U(0,1)$ and let $\hat\pi$ be a uniformly chosen permutation from $\mathbf{G}$, independent of U. Note that $c_\alpha(q_y, q)$ is the $(1-\alpha)$-quantile of $T(U)$ and, by (A-25), c_α^{p} is the $(1-\alpha)$-quantile of the CDF $\frac{1}{|\mathbf{G}|}\sum_{\pi\in\mathbf{G}} I\{T^*(Q^\pi)\le x\}$. The desired result then follows from noting that $\frac{1}{|\mathbf{G}|}\sum_{\pi\in\mathbf{G}} I\{T^*(Q^\pi)\le x\}$ is the CDF of $T(U)$, as we show next.

Let $E := \{U_i \neq U_j \text{ for } i \neq j\}$. For any $x \in \mathbf{R}$, our desired result follows from this derivation:

$$\begin{aligned} P\{T(U)\le x\} &\overset{(1)}{=} P\{T(U^{\hat\pi})\le x\} \\ &\overset{(2)}{=} P\left\{\left\{T(U^{\hat\pi})\le x\right\}\cap E\right\} \\ &\overset{(3)}{=} \int_{s\in E}\frac{1}{|\mathbf{G}|}\sum_{\pi\in\mathbf{G}} I\{T(u^\pi)\le x\}\, dP_U^*(u) \\ &\overset{(4)}{=} \int_{u\in E}\frac{1}{|\mathbf{G}|}\sum_{\pi\in\mathbf{G}} I\{T^*(Q^\pi)\le x\}\, dP_U^*(u) \\ &\overset{(5)}{=} \frac{1}{|\mathbf{G}|}\sum_{\pi\in\mathbf{G}} I\{T^*(Q^\pi)\le x\}\ . \end{aligned}$$

Here, (1) holds by $U \overset{d}{=} U^{\hat\pi}$, (2) and (5) by $P\{E\} = 1$, (3) by $\hat\pi \perp U$ and $\hat\pi$ uniformly chosen in $\mathbf{G}$, and (4) by repeating the arguments used to derive (A-26).

Step 3. By Step 1, $\{S_i \neq S_j \text{ for } i \neq j\} \subseteq \{c_\alpha^{\mathrm{p}}(S) = c_\alpha^{\mathrm{p}}\}$, and so $P\{c_\alpha^{\mathrm{p}}(S) = c_\alpha^{\mathrm{p}}\} = 1$ holds by our assumption. By Step 2, $c_\alpha^{\mathrm{p}} = c_\alpha(q_y, q)$. The desired result follows from combining these points.

Step 4. To show the last statement, consider $S = \{1 : 1 \le j \le q\}$. Then, $T(S^\pi) = T(S) = 0$ for all $\pi \in \mathbf{G}$, and so $c_\alpha^{\mathrm{p}}(S) = 0$. On the other hand, Step 2 implies $c_\alpha^{\mathrm{p}} = c_\alpha(q_y, q)$, which are positive for typical values of (q_y, q, α). For example, $q_y = 1$, $q = 2$, and $\alpha = 0.1$ yield $c_\alpha^{\mathrm{p}} = c_\alpha(q_y, q) = 1$.

Appendix B Supporting technical results

For any $\varepsilon > 0$, we use $o_\varepsilon(1)$ to denote an expression that converges to zero as $\varepsilon \to 0$. Analogously, for any $n \in \mathbf{N}$, we use $o_n(1)$ to denote an expression that converges to zero as $n \to \infty$.

B.1 Auxiliary theorems

Theorem 4. *Let Assumptions 1, 2, and 3 hold with $\mathcal{Z} = \{z_1, z_2, \ldots, z_L\}$, and let $\hat{S}_n^\ell$ be defined as in (C-1) based on $\hat{z}_{n,\ell}$ instead of z_ℓ, with distance ties broken by any $\mathcal{A}$-measurable rule. Then,*

$$((\hat{S}_n^1)', \ldots, (\hat{S}_n^L)') \xrightarrow{d} ((S^1)', \ldots, (S^L)') \ , \tag{B-1}$$

where, for any $(s_1^\ell, \ldots, s_{q_y^\ell}^\ell, s_{q_y^\ell+1}^\ell, \ldots, s_{q^\ell}^\ell) \in \mathbf{R}^{q^\ell}$ for each $\ell = 1, \ldots, L$ with $q^\ell = q_y^\ell + q_x^\ell$ fixed, $((S^1)', \ldots, (S^L)')$ has the following distribution:

$$P\left\{\bigcap_{\ell=1}^{L}\bigcap_{i=1}^{q^\ell}\{S_i^\ell \le s_i^\ell\}\right\} \ = \ \prod_{\ell=1}^{L}\prod_{i=1}^{q_y^\ell} F_Y(s_i^\ell|z_\ell)\prod_{j=1}^{q_x^\ell} F_X(s_{j+q_y^\ell}^\ell|z_\ell) \ . \tag{B-2}$$

Proof. For each $\ell = 1, \ldots, L$, let $\hat{g}_\ell(Z) := |Z - \hat{z}_{n,\ell}|$, let M_y^ℓ denote the subset of the indices $i = 1, \ldots, n_y$ corresponding to the q_y^ℓ first $\hat{g}_\ell$-order statistics $(\hat{Z}_{n_y,(1)}^\ell, \ldots, \hat{Z}_{n_y,(q_y^\ell)}^\ell)$, and let M_x^ℓ denote the subset of the indices $j = 1, \ldots, n_x$ corresponding to the q_x^ℓ first $\hat{g}_\ell$-order statistics $(\hat{Z}_{n_x,(1)}^\ell, \ldots, \hat{Z}_{n_x,(q_x^\ell)}^\ell)$. Since each $\hat{z}_{n,\ell}$ is $\mathcal{A}$-measurable by Assumption 1 and distance ties are broken by an $\mathcal{A}$-measurable rule, the sets M_y^ℓ, M_x^ℓ and the orderings within them are $\mathcal{A}$-measurable. Let E_n denote the following event:

$$E_n \ = E_{n_y,y} \cap E_{n_x,x} \ , \text{ where } \ E_{n_y,y} := \bigcap_{1\le\ell<\ell'\le L}\left\{M_y^\ell \cap M_y^{\ell'} = \emptyset\right\} \text{ and } E_{n_x,x} := \bigcap_{1\le\ell<\ell'\le L}\left\{M_x^\ell \cap M_x^{\ell'} = \emptyset\right\}.$$

In words, E_n means that the subsets of the data used in each of the L tests are pairwise disjoint within each sample. Note that $E_n \in \mathcal{A}$. We begin by showing that

$$P\{E_n\} \to 1 \ . \tag{B-3}$$

Set $\bar{\varepsilon} := \frac{1}{4}\min\{|z_\ell - z_{\ell'}| : \ell \ne \ell', \ \ell, \ell' = 1, \ldots, L\} > 0$ and fix $\varepsilon \in (0, \bar{\varepsilon})$ arbitrarily. Set

$$G_n := \{\max|\hat{z}_{n,\ell} - z_\ell| \le \varepsilon/2 : \ell = 1, \ldots, L\} \ .$$

For each $\ell = 1, \ldots, L$, let $B_{\ell,y} := \sum_{i=1}^{n_y} I\{|Z_i - z_\ell| \le \varepsilon/2\}$ and $\hat{B}_{\ell,y} := \sum_{i=1}^{n_y} I\{|Z_i - \hat{z}_{n,\ell}| \le \varepsilon\}$, and define $B_{\ell,x}$ and $\hat{B}_{\ell,x}$ analogously for the second sample. Note that

$$G_n \bigcap \bigcap_{\ell=1}^{L}\{B_{\ell,y} \ge q_y^\ell\} \overset{(1)}{\subseteq} \bigcap_{\ell=1}^{L}\bigcap_{i=1}^{q_y^\ell}\{|\hat{Z}_{n_y,(i)}^\ell - z_\ell| \le 3\varepsilon/2\} \overset{(2)}{\subseteq} E_{n_y,y} \ , \tag{B-4}$$

where (1) holds by the fact that, for every $\ell = 1, \ldots, L$, G_n and $\{B_{\ell,y} \ge q_y^\ell\}$ imply that $|\hat{z}_{n,\ell} - z_\ell| \le \varepsilon/2$ and $|\hat{Z}_{n_y,(i)}^\ell - \hat{z}_{n,\ell}| \le \varepsilon$ for $i = 1, \ldots, q_y^\ell$, and (2) by the fact that for $\ell \ne \ell'$ we have $|z_\ell - z_{\ell'}| \ge 4\bar{\varepsilon} > 3\varepsilon$, so no

sample observation can lie within $3\varepsilon/2$ of both z_ℓ and $z_{\ell'}$, and therefore $M_y^\ell \cap M_y^{\ell'} = \emptyset$ for all $\ell \neq \ell'$, which implies $E_{n_y,y}$.

The same argument applied to the second sample yields the analogue of (B-4) with $(B_{\ell,x}, q_x^\ell, \hat{Z}^\ell_{n_x,(j)}, E_{n_x,x})$ in place of $(B_{\ell,y}, q_y^\ell, \hat{Z}^\ell_{n_y,(i)}, E_{n_y,y})$. From here, we have that (B-3) follows if we show that

$$P\{G_n^c\} \to 0 \ , \tag{B-5}$$

$$P\{B_{\ell,y} < q_y^\ell\} \to 0 \ \text{ and } \ P\{B_{\ell,x} < q_x^\ell\} \to 0 \ \text{ for all } \ell = 1, \ldots, L \ , \tag{B-6}$$

since then $P\{E_n^c\} \leq P\{G_n^c\} + \sum_{\ell=1}^L \left(P\{B_{\ell,y} < q_y^\ell\} + P\{B_{\ell,x} < q_x^\ell\}\right) \to 0$. Note that this union bound does not rely on the independence of the two samples: G_n may involve the Z observations from both samples, as is the case when $\hat{z}_{n,\ell}$ is computed from the pooled data.

Note that (B-5) holds by Assumption 1 and L being finite. For (B-6), we only prove the first convergence, as the second is analogous. Since the count $B_{\ell,y}$ is centered at the non-random point z_ℓ, $B_{\ell,y} \sim Bi(n_y, P\{|Z_i - z_\ell| \leq \varepsilon/2\})$, where $P\{|Z_i - z_\ell| \leq \varepsilon/2\} > 0$ by Assumption 2. By this and $\max_{\ell=1,\ldots,L} q^\ell$ being bounded, $\exists N(\varepsilon)$ s.t. for all $n_y \geq N(\varepsilon)$,

$$0 < \frac{1}{2} P\{|Z_i - z_\ell| \leq \varepsilon/2\} \leq P\{|Z_i - z_\ell| \leq \varepsilon/2\} - \max_{\ell=1,\ldots,L} \frac{q_y^\ell}{n_y} \ . \tag{B-7}$$

Then, for all $n_y \geq N(\varepsilon)$, we have

$$\begin{aligned} P\{B_{\ell,y} < q_y^\ell\} &= P\left\{B_{\ell,y}/n_y - P\{|Z_i - z_\ell| \leq \varepsilon/2\} < q_y^\ell/n_y - P\{|Z_i - z_\ell| \leq \varepsilon/2\}\right\} \\ &\overset{(1)}{\leq} P\left\{|B_{\ell,y}/n_y - P\{|Z_i - z_\ell| \leq \varepsilon/2\}| > \tfrac{1}{2} P\{|Z_i - z_\ell| \leq \varepsilon/2\}\right\} \overset{(2)}{\to} 0 \ , \end{aligned}$$

as desired, where (1) holds by (B-7) and (2) holds by the LLN as $n_y \to \infty$.

We are now ready to prove (B-1). For any $(s_1^\ell, \ldots, s_{q_y^\ell}^\ell, s_{q_y^\ell+1}^\ell, \ldots, s_{q^\ell}^\ell) \in \mathbf{R}^{q^\ell}$ for each $\ell = 1, \ldots, L$ with $q^\ell = q_y^\ell + q_x^\ell$, we have

$$\begin{aligned} & P\left\{\bigcap_{\ell=1}^L \bigcap_{i=1}^{q^\ell} \left\{\hat{S}_{n,i}^\ell \leq s_i^\ell\right\}\right\} \\ & \overset{(1)}{=} E\left[P\left\{\bigcap_{\ell=1}^L \left\{\bigcap_{i=1}^{q_y^\ell} \left\{\hat{Y}_{n,[i]}^\ell \leq s_i^\ell\right\} \bigcap \bigcap_{j=1}^{q_x^\ell} \left\{\hat{X}_{n,[j]}^\ell \leq s_{j+q_y^\ell}^\ell\right\}\right\} \middle| \mathcal{A}\right\}\right] \\ & \overset{(2)}{=} E\left[P\left\{\bigcap_{\ell=1}^L \left\{\bigcap_{i=1}^{q_y^\ell} \left\{\hat{Y}_{n,[i]}^\ell \leq s_i^\ell\right\} \bigcap \bigcap_{j=1}^{q_x^\ell} \left\{\hat{X}_{n,[j]}^\ell \leq s_{j+q_y^\ell}^\ell\right\}\right\} \middle| \mathcal{A}\right\} I\{E_n\}\right] + o_n(1) \\ & \overset{(3)}{=} E\left[\prod_{\ell=1}^L P\left\{\left\{\bigcap_{i=1}^{q_y^\ell} \left\{\hat{Y}_{n,[i]}^\ell \leq s_i^\ell\right\} \bigcap \bigcap_{j=1}^{q_x^\ell} \left\{\hat{X}_{n,[j]}^\ell \leq s_{j+q_y^\ell}^\ell\right\}\right\} \middle| \mathcal{A}\right\} I\{E_n\}\right] + o_n(1) \\ & \overset{(4)}{=} E\left[\prod_{\ell=1}^L P\left\{\left\{\bigcap_{i=1}^{q_y^\ell} \left\{\hat{Y}_{n,[i]}^\ell \leq s_i^\ell\right\} \bigcap \bigcap_{j=1}^{q_x^\ell} \left\{\hat{X}_{n,[j]}^\ell \leq s_{j+q_y^\ell}^\ell\right\}\right\} \middle| \mathcal{A}\right\}\right] + o_n(1) \end{aligned}$$

$$\overset{(5)}{=} E\left[\prod_{\ell=1}^{L}\left\{\prod_{i=1}^{q_y^\ell} F_Y(s_i^\ell|\hat{Z}^\ell_{n_y,(i)})\prod_{j=1}^{q_x^\ell} F_X(s^\ell_{j+q_y^\ell}|\hat{Z}^\ell_{n_x,(j)})\right\}\right] + o_n(1)\ , \tag{B-8}$$

where (1) holds by the LIE, (2) and (4) by (B-3) and the conditional probability being bounded by one, (3) by Assumption 1 and the fact that, conditional on $E_n \in \mathcal{A}$, the ℓ-th bracket term is a function of $\{Y_i : i \in M_y^\ell\}$ and $\{X_j : j \in M_x^\ell\}$ only, which are pairwise disjoint observations across $\ell = 1, \ldots, L$, and (5) by Assumption 1 and the $\mathcal{A}$-measurable tie-breaking rule, the indices of the selected observations and their ordering by $\hat{g}_\ell$ are $\mathcal{A}$-measurable, so conditional on $\mathcal{A}$ the vector $(\hat{Y}^\ell_{n,[1]}, \ldots, \hat{Y}^\ell_{n,[q_y^\ell]}, \hat{X}^\ell_{n,[1]}, \ldots, \hat{X}^\ell_{n,[q_x^\ell]})$ has independent components with conditional conditional CDFs $F_Y(\cdot|\hat{Z}^\ell_{n_y,(i)})$ and $F_X(\cdot|\hat{Z}^\ell_{n_x,(j)})$.

Next, we show that

$$\hat{Z}^\ell_{n_y,(i)} \xrightarrow{p} z_\ell \quad \text{for all} \quad i = 1, \ldots, q_y^\ell \text{ and } \quad \ell = 1, \ldots, L\ , \tag{B-9}$$

$$\hat{Z}^\ell_{n_x,(j)} \xrightarrow{p} z_\ell \quad \text{for all} \quad j = 1, \ldots, q_x^\ell \text{ and } \quad \ell = 1, \ldots, L\ . \tag{B-10}$$

We only show (B-9), as (B-10) can be shown analogously. Fix $\varepsilon \in (0, \bar{\varepsilon})$ arbitrarily, with G_n and $B_{\ell,y}$ defined as before. By (B-4)-(B-6), we conclude that

$$P\Big\{\bigcap_{\ell=1}^{L}\bigcap_{i=1}^{q_y^\ell}\{|\hat{Z}^\ell_{n_y,(i)} - z_\ell| \le 3\varepsilon/2\}\Big\} \to 1\ .$$

Since the choice of $\varepsilon \in (0, \bar{\varepsilon})$ was arbitrary, (B-9) follows.

By (B-9), (B-10), and Assumption 3, $F_Y(s_i^\ell|\hat{Z}^\ell_{n_y,(i)}) \xrightarrow{p} F_Y(s_i^\ell|z_\ell)$ and $F_X(s^\ell_{j+q_y^\ell}|\hat{Z}^\ell_{n_x,(j)}) \xrightarrow{p} F_X(s^\ell_{j+q_y^\ell}|z_\ell)$ for every i, j, ℓ, and every $(s_1^\ell, \ldots, s_{q^\ell}^\ell)$. By this and the fact that the integrand in (B-8) is bounded by one and converges in probability to a constant, it converges in L^1 and so

$$\lim_{n\to\infty} P\left\{\bigcap_{\ell=1}^{L}\bigcap_{i=1}^{q^\ell}\left\{\hat{S}^\ell_{n,i} \le s_i^\ell\right\}\right\} \;=\; \prod_{\ell=1}^{L}\left\{\prod_{i=1}^{q_y^\ell} F_Y(s_i^\ell|z_\ell)\prod_{j=1}^{q_x^\ell} F_X(s^\ell_{j+q_y^\ell}|z_\ell)\right\}\ . \tag{B-11}$$

By definition of convergence in distribution, (B-11) implies (B-1), as desired. ■

Theorem 5. *Let S be the random variable in Theorem 1. Then, for any $P \in \mathbf{P}_0$ and $\alpha \in (0,1)$, we obtain*

$$E_P[\phi(S)] \le \bar{\alpha} \le \alpha\ ,$$

where

$$\bar{\alpha} := P\left\{\sup_{u\in(0,1)}\left(\frac{1}{q_y}\sum_{j=1}^{q_y} I\{U_j \le u\} - \frac{1}{q_x}\sum_{j=q_y+1}^{q} I\{U_j \le u\}\right) > c_\alpha(q_y, q)\right\}\ . \tag{B-12}$$

Moreover, the first inequality becomes an equality under P such that $F_Y(t|z_0) = F_X(t|z_0)$ for all $t \in \mathbf{R}$ and these are continuous functions of $t \in \mathbf{R}$.

Proof. Recall that $S = (S_1, \ldots, S_{q_y}, S_{q_y+1}, \ldots, S_q)$ are independent, and such that $S_j \sim F_Y(t|z_0)$ for all $j = 1, \ldots, q_y$ and $S_{q_y+j} \sim F_X(t|z_0)$ for all $j = 1, \ldots, q_x$. Denote by $Q_Y(\cdot|z_0)$ and $Q_X(\cdot|z_0)$ the quantile

functions of $F_Y(t|z_0)$ and $F_X(t|z_0)$. Consider the following argument:

$$
\begin{aligned}
&E_P[\phi(S)] \\
&= P\left\{\max_{k\in\{1,\ldots,q_y\}}\left(\frac{1}{q_y}\sum_{j=1}^{q_y} I\{S_j \le S_k\} - \frac{1}{q_x}\sum_{j=q_y+1}^{q} I\{S_j \le S_k\}\right) > c_\alpha(q_y,q)\right\} \\
&\overset{(1)}{=} P\left\{\sup_{t\in\mathbf{Y}}\left(\frac{1}{q_y}\sum_{j=1}^{q_y} I\{Q_Y(U_j|z_0) \le t\} - \frac{1}{q_x}\sum_{j=q_y+1}^{q} I\{Q_X(U_j|z_0) \le t\}\right) > c_\alpha(q_y,q)\right\} \\
&\overset{(2)}{=} P\left\{\sup_{t\in\mathbf{Y}}\left(\frac{1}{q_y}\sum_{j=1}^{q_y} I\{U_j \le F_Y(t|z_0)\} - \frac{1}{q_x}\sum_{j=q_y+1}^{q} I\{U_j \le F_X(t|z_0)\}\right) > c_\alpha(q_y,q)\right\} \\
&\overset{(3)}{\le} P\left\{\sup_{t\in\mathbf{Y}}\left(\frac{1}{q_y}\sum_{j=1}^{q_y} I\{U_j \le F_X(t|z_0)\} - \frac{1}{q_x}\sum_{j=q_y+1}^{q} I\{U_j \le F_X(t|z_0)\}\right) > c_\alpha(q_y,q)\right\} \\
&\overset{(4)}{=} P\left\{\sup_{u\in\mathcal{U}}\left(\frac{1}{q_y}\sum_{j=1}^{q_y} I\{U_j \le u\} - \frac{1}{q_x}\sum_{j=q_y+1}^{q} I\{U_j \le u\}\right) > c_\alpha(q_y,q)\right\} \\
&\overset{(5)}{\le} P\left\{\sup_{u\in(0,1)}\left(\frac{1}{q_y}\sum_{j=1}^{q_y} I\{U_j \le u\} - \frac{1}{q_x}\sum_{j=q_y+1}^{q} I\{U_j \le u\}\right) > c_\alpha(q_y,q)\right\} \\
&\overset{(6)}{=} \bar{\alpha}\ ,
\end{aligned}
$$

where (1) follows from the quantile transformation and the fact that replacing $\max_{k\in\{1,\ldots,q_y\}}$ with $\sup_{t\in\mathbf{Y}}$ does not affect the magnitude of the test statistic, (2) follows from Pollard (2002, Eq. 36), (3) from $P \in \mathbf{P}_0$, (4) from a simple change of variables and $\mathcal{U} := \cup_{t\in\mathbf{Y}}\{u = F_X(t|z_0)\}$, (5) from $\mathcal{U} \subseteq [0,1]$, and (6) from the definition of $c_\alpha(q_y,q)$ and $\bar{\alpha}$, and the fact that $\{U_j : j = 1,\ldots,q\}$ are i.i.d. distributed as $U(0,1)$.

To conclude the proof, it suffices to show that: (i) inequality (3) holds as an equality under the condition $F_Y(t \mid z_0) = F_X(t \mid z_0)$ for all $t \in \mathbf{R}$, and (ii) inequality (5) holds as an equality when these functions are continuous in t. The first claim is immediate. For the second, it follows from the fact that the continuity of the CDFs implies $\mathcal{U} = (0,1)$. By elementary properties of CDFs, $\lim_{t\to-\infty} F_X(t \mid z_0) = 0$ and $\lim_{t\to\infty} F_X(t \mid z_0) = 1$. By the intermediate value theorem, for any $u \in (0,1)$, there exists $t \in \mathbf{R}$ such that $u = F_X(t \mid z_0)$, implying $u \in \mathcal{U}$, as desired. ■

B.2 Auxiliary lemmas

Lemma 3. *Suppose that $S_n \sim P$ with P satisfying Assumptions 2, 3, and 4 hold. Consider a sequence of random vectors $\{\tilde{S}_n : 1 \le n < \infty\}$ and a random vector $\tilde{S}$ defined on a common probability space $(\Omega, \mathcal{A}, \tilde{P})$ such that*

$$\tilde{S}_n \overset{d}{=} S_n, \quad \tilde{S} \overset{d}{=} S, \quad \text{and} \quad \tilde{S}_n \overset{a.s.}{\to} \tilde{S}\ . \tag{B-13}$$

Then, for any $j \in \{1,\ldots,q\}$,

$$\tilde{P}\{\tilde{S}_{n,j} \ne \tilde{S}_j,\ \tilde{S}_j \in \mathcal{D}\} = o_n(1)\ .$$

Proof. We focus on an arbitrary $j \in \{1,\ldots,q_y\}$. The argument for $j \in \{q_y+1,\ldots,q\}$ follows analogously

by replacing Y with X. By Lemma 6, it suffices to show that

$$\sup_{t\in\mathbb{R}} \left|F_{\tilde{S}_{n,j}}(t) - F_{\tilde{S}_j}(t)\right| = o_n(1) \ , \tag{B-14}$$

where $F_{\tilde{S}_{n,j}}$ and $F_{\tilde{S}_j}$ denote the distribution functions of $\tilde{S}_{n,j}$ and $\tilde{S}_j$ in $(\Omega, \mathcal{A}, \tilde{P})$.

By the proof in Theorem 4 (with $L = 1$ and $\mathcal{Z} = \{z_0\}$), we have $F_{S_j}(t) = F_Y(t|z_0)$ and $F_{S_{n,j}}(t) = E[F_Y(t|Z_{n_y,(j)})]$ where $Z_{n_y,(j)} \xrightarrow{p} z_0$. By these and (B-13), (B-14) follows from

$$\sup_{t\in\mathbf{R}} |E[F_Y(t|Z_{n_y,(j)})] - F_Y(t|z_0)| = o_n(1) \ . \tag{B-15}$$

For any $\varepsilon > 0$,

$$\begin{aligned} E[F_Y(t|Z_{n_y,(j)})] &= \int_{|z-z_0|\le\varepsilon} F_Y(t|z) dP_{Z_{n_y,(j)}}(z) + \int_{|z-z_0|>\varepsilon} F_Y(t|z) dP_{Z_{n_y,(j)}}(z) \\ &\le \int_{|z-z_0|\le\varepsilon} F_Y(t|z) dP_{Z_{n_y,(j)}}(z) + P(|Z_{n_y,(j)} - z_0| > \varepsilon) \\ &\overset{(1)}{=} \int_{|z-z_0|\le\varepsilon} F_Y(t|z) dP_{Z_{n_y,(j)}}(z) + o_n(1) \ , \end{aligned}$$

where (1) holds by $Z_{n_y,(j)} \xrightarrow{p} z_0$. Then,

$$\sup_{t\in\mathbf{R}} |E[F_Y(t|Z_{n_y,(j)})] - F_Y(t|z_0)| \le \sup_{t\in\mathbf{R}} \sup_{|z-z_0|\le\varepsilon} |F_Y(t|z) - F_Y(t|z_0)| + o_n(1) \ .$$

Fix $\delta > 0$ arbitrarily. By Assumption 3, $\exists \varepsilon > 0$ such that $\sup_{t\in\mathbf{R}} \sup_{|z-z_0|\le\varepsilon} |F_Y(t|z) - F_Y(t|z_0)| < \delta/2$. For all large enough n_y, the right-hand side is bounded above by δ. Since the choice of δ was arbitrary, (B-15) follows. ■

Lemma 4. *Let V_1 and V_2 be independent random variables that are discontinuous at a finite set of points $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. Then, for any $\delta > 0$, $\exists \varepsilon > 0$ small enough s.t.*

$$\begin{aligned} P\{\cup_{d\in\mathcal{D}_1}\{|V_1 - d| < \varepsilon\} \cap \{V_1 \in \mathcal{D}_1^c\}\} < \delta \ , \\ P\{\{|V_1 - V_2| < \varepsilon\} \cap \{V_1 \in \mathcal{D}_1^c\} \cap \{V_2 \in \mathcal{D}_2^c\}\} < \delta \ . \end{aligned}$$

Proof. Set $\bar{\varepsilon} := \{\min |d - \tilde{d}| : d, \tilde{d} \in \mathcal{D}_1 \cap d \ne \tilde{d}\} > 0$. For any $\varepsilon \in (0, \bar{\varepsilon}/2)$,

$$\begin{aligned} &P\{\cup_{d\in\mathcal{D}_1}\{|V_1 - d| < \varepsilon\} \cap \{V_1 \in \mathcal{D}_1^c\}\} \\ &\le \sum_{d\in\mathcal{D}_1} P\{\{|V_1 - d| < \varepsilon\} \cap \{V_1 \in \mathcal{D}_1^c\}\} \\ &\le_{(1)} \sum_{d\in\mathcal{D}_1} P\{V_1 \in (d-\varepsilon, d) \cup (d, d+\varepsilon)\} \overset{(2)}{=} o_\varepsilon(1) \ , \end{aligned}$$

where (1) holds by $\varepsilon \in (0, \bar{\varepsilon}/2)$ and (2) by the fact that $(d-\varepsilon, d) \cap (d, d+\varepsilon)$ has no discontinuities in the CDF. The right-hand side is less than δ by making ε arbitrarily small, as desired. Also, for any $\varepsilon \in (0, \bar{\varepsilon}/2)$,

$$P\{\{|V_1 - V_2| < \varepsilon\} \cap \{V_1 \in \mathcal{D}_1^c\} \cap \{V_2 \in \mathcal{D}_2^c\}\}$$

$$\overset{(1)}{=} \int_{\mathcal{D}_2^c} P\{\{|V_1 - v_2| < \varepsilon\} \cap \{V_1 \in \mathcal{D}_1^c\}\} dP_{V_2}(v_2) \overset{(2)}{=} o_\varepsilon(1) \ ,$$

where (1) holds by $V_1 \perp V_2$, and (2) holds by Lemma 5 and the dominated convergence theorem. The right-hand side is less than δ by making ε arbitrarily small, as desired. ■

Lemma 5. *Consider a random variable V whose CDF is discontinuous at a finite set of points $\mathcal{D}$. Then, for any $v \in \mathbf{R}$,*

$$P\{\{|V - v| < \varepsilon\} \cap \{V \in \mathcal{D}^c\}\} = o_\varepsilon(1) \ .$$

Proof. Set $\bar{\varepsilon} := \{\min |d - \tilde{d}| : d, \tilde{d} \in \mathcal{D}_1 \cap d \neq \tilde{d}\} > 0$. Fix $\varepsilon < \bar{\varepsilon}/2$. There are two possibilities: $v \in \mathcal{D}$ or $v \notin \mathcal{D}$. First, consider $v \in \mathcal{D}$. In this case,

$$P\{|V - v| < \varepsilon\} \cap \{V \in \mathcal{D}^c\} \leq P\{V \in (v - \varepsilon, v) \cap (v, v + \varepsilon)\} \overset{(1)}{=} o_\varepsilon(1) \ ,$$

where (1) holds because $(d - \varepsilon, d) \cap (d, d + \varepsilon)$ has no mass points. Second, consider $v \notin \mathcal{D}$. Then,

$$P\{\{|V - v| < \varepsilon\} \cap \{V \in \mathcal{D}^c\}\} \leq F_V(v + \varepsilon) - F_V(v - \varepsilon) \overset{(2)}{=} o_\varepsilon(1) \ ,$$

where (1) by the fact that $(v - \varepsilon, v + \varepsilon]$ has no mass points, so F_V is continuous on that interval. ■

Lemma 6. *Let $\{V_n : n \geq 1\}$ be a sequence of random variables that satisfy $V_n \xrightarrow{p} V$, where V is a random variable whose CDF is discontinuous at a finite set of points $\mathcal{D}$. Furthermore, assume $\sup_{t \in \mathbf{R}} |F_{V_n}(t) - F_V(t)| \to 0$. Then,*

$$P\{\{V \in \mathcal{D}\} \cap \{V_n \neq V\}\} \to 0 \ .$$

Proof. Fix $\delta > 0$ arbitrarily. It suffices to find $N = N(\delta)$ such that $P\{V \in \mathcal{D}, V_n \neq V\} < \delta$ for all $n > N$.

Set $\bar{\varepsilon} := \{\min |d - \tilde{d}| : d, \tilde{d} \in \mathcal{D} \cap d \neq \tilde{d}\} > 0$. For any $\varepsilon \in (0, \bar{\varepsilon}/2)$, consider the following argument.

$$\begin{aligned}
&P\{\{V \in \mathcal{D}\} \cap \{V_n \neq V\}\} \\
&= \sum_{d \in \mathcal{D}} P\{\{V = d\} \cap \{V_n \neq d\}\} \\
&\overset{(1)}{=} \sum_{d \in \mathcal{D}} P\{\{V = d\} \cap \{V_n \neq d\} \cap \{|V_n - d| < \varepsilon\}\} + o_n(1) \\
&\leq \sum_{d \in \mathcal{D}} P\{V_n \in (d - \varepsilon, d) \cap (d, d + \varepsilon)\} + o_n(1) \\
&\overset{(2)}{=} \sum_{d \in \mathcal{D}} P\{V \in (d - \varepsilon, d) \cap (d, d + \varepsilon)\} + o_n(1) \\
&\overset{(3)}{=} o_\varepsilon(1) + o_n(1) \ ,
\end{aligned}$$

where (1) holds by $V_n \xrightarrow{p} V$, (2) by $\sup_{t \in \mathbf{R}} |F_{V_n}(t) - F_V(t)| = o_n(1)$, and (3) by the fact that $(d - \varepsilon, d) \cap (d, d + \varepsilon)$ has no discontinuities in the CDF. For all large enough n and small enough ε, the right-hand side is bounded by δ, as desired. ■

Lemma 7. *Let $\alpha \in (0, 1)$ and $P' \in \mathbf{P}_1 := \mathbf{P} \backslash \mathbf{P}_0$ be given, and let $t^* \in \mathbf{R}$ and $\varepsilon := F_Y(t^*|z_0) - F_X(t^*|z_0) > 0$, which exist by the definition of $\mathbf{P}_1$. Let $\kappa_\alpha := \sqrt{\log(2/\alpha)/2}$. Then:*

(i) $c_\alpha(a, a+b) \le \kappa_\alpha\big(a^{-1/2} + b^{-1/2}\big)$ *for all* $(a,b) \in \mathbf{N}^2$*;*

(ii) for every $(a,b) \in \mathbf{N}^2$ *such that* $\kappa_\alpha(a^{-1/2} + b^{-1/2}) < \varepsilon/3$*,*

$$E_{P'}\big[\phi(S^\circ(a,b))\big] \;\ge\; 1 - e^{-2a\varepsilon^2/9} - e^{-2b\varepsilon^2/9} \;.$$

In particular, $E_{P'}[\phi(S^\circ(a_n,b_n))] \to 1$ *along any sequences* (a_n,b_n) *with* $\min\{a_n,b_n\} \to \infty$*.*

Proof. We first prove (i). By (3.6), $c_\alpha(a, a+b)$ is the $1-\alpha$ quantile of $\sup_{u\in(0,1)} \Delta_{a,a+b}(u)$, where $\Delta_{a,a+b}(u) = \hat{G}_a(u) - \hat{H}_b(u)$ and $\hat{G}_a$ and $\hat{H}_b$ denote the empirical CDFs of two independent i.i.d. samples from the uniform distribution on $(0,1)$, of sizes a and b; see (3.5). For every $u \in (0,1)$,

$$\Delta_{a,a+b}(u) = \big(\hat{G}_a(u) - u\big) + \big(u - \hat{H}_b(u)\big) \le \sup_{u\in(0,1)} \big(\hat{G}_a(u) - u\big) + \sup_{u\in(0,1)} \big(u - \hat{H}_b(u)\big) \;.$$

By the one-sided Dvoretzky–Kiefer–Wolfowitz inequality with the tight constant (Massart, 1990, Theorem 1), $P\{\sup_u(\hat{G}_a(u) - u) > \xi\} \le e^{-2a\xi^2}$ for every $\xi > 0$, and the same bound holds for $\sup_u(u - \hat{H}_b(u))$ by symmetry. Setting $\xi_a := \kappa_\alpha a^{-1/2}$ and $\xi_b := \kappa_\alpha b^{-1/2}$, so that $e^{-2a\xi_a^2} = e^{-2b\xi_b^2} = \alpha/2$, the union bound gives

$$P\Big\{ \sup_{u\in(0,1)} \Delta_{a,a+b}(u) > \xi_a + \xi_b \Big\} \le \alpha \;,$$

and therefore $c_\alpha(a, a+b) \le \xi_a + \xi_b = \kappa_\alpha(a^{-1/2} + b^{-1/2})$, as desired.

We now prove (ii). Let $\hat{F}_Y^\circ(t) := a^{-1}\sum_{i\le a} I\{Y_i^\circ \le t\}$ and $\hat{F}_X^\circ(t) := b^{-1}\sum_{j\le b} I\{X_j^\circ \le t\}$ denote the empirical CDFs of the two blocks of $S^\circ(a,b)$, and recall from (3.4) that the test statistic satisfies $T(S^\circ(a,b)) \ge \hat{F}_Y^\circ(t^*) - \hat{F}_X^\circ(t^*)$. Since $a\hat{F}_Y^\circ(t^*)$ is a Binomial random variable with success probability $F_Y(t^*|z_0)$, Hoeffding's inequality yields

$$P'\Big\{\hat{F}_Y^\circ(t^*) \le F_Y(t^*|z_0) - \varepsilon/3\Big\} \le e^{-2a\varepsilon^2/9} \;, \qquad P'\Big\{\hat{F}_X^\circ(t^*) \ge F_X(t^*|z_0) + \varepsilon/3\Big\} \le e^{-2b\varepsilon^2/9} \;.$$

On the intersection of the complements of these two events,

$$T(S^\circ(a,b)) \;\ge\; \hat{F}_Y^\circ(t^*) - \hat{F}_X^\circ(t^*) \;\ge\; \big(F_Y(t^*|z_0) - \varepsilon/3\big) - \big(F_X(t^*|z_0) + \varepsilon/3\big) \;=\; \varepsilon/3 \;>\; c_\alpha(a, a+b) \;,$$

where the last inequality holds by part (i) and the hypothesis $\kappa_\alpha(a^{-1/2}+b^{-1/2}) < \varepsilon/3$; hence the test rejects. Part (ii) then follows from the union bound. The final claim follows from (i) and (ii), since $\kappa_\alpha(a_n^{-1/2} + b_n^{-1/2}) \to 0$ and both exponential terms vanish as $\min\{a_n,b_n\} \to \infty$. ■

Appendix C Extensions and Supplementary Analyses

C.1 Multiple target points

Throughout the paper, we have focused on testing the null hypothesis (2.1) with $L = 1$ to emphasize the main ideas and maintain clarity. This section describes how our method extends when $L > 1$. In this setting, we address the intersection nature of the null hypothesis by using a maximum test.

To formally define the test, we need to update our notation to explicitly reflect dependence on both ℓ and α. Consider the point $z_\ell \in \mathcal{Z}$ for some $\ell \in \{1, \dots, L\}$. Let $g_\ell(Z) := |Z - z_\ell|$ and, for any two values $z, z' \in \mathcal{Z}$, define the ordering $\leq_\ell$ as follows:

$$z \leq_\ell z' \quad \text{if and only if} \quad g_\ell(z) \leq g_\ell(z') \ .$$

For each ℓ, this leads to g-order statistics $Z_{\ell,(i)}$ in each sample, as well as induced order statistics for Y and X, that we denote by

$$Y^\ell_{n_y,[1]}, Y^\ell_{n_y,[2]}, \dots, Y^\ell_{n_y,[q_y]} \ \text{ and } \ X^\ell_{n_x,[1]}, X^\ell_{n_x,[2]}, \dots, X^\ell_{n_x,[q_x]} \ .$$

As with $L = 1$, any ties in the g-ordering can be resolved arbitrarily. Finally, if we let $n := n_y + n_x$ and $q := q_y + q_x$, the effective (pooled) sample is given by

$$S^\ell_n = (S^\ell_{n,1}, \dots, S^\ell_{n,q}) := (Y^\ell_{n_y,[1]}, \dots, Y^\ell_{n_y,[q_y]}, X^\ell_{n_x,[1]}, \dots, X^\ell_{n_x,[q_x]}) \ . \tag{C-1}$$

Our test is entirely based on S^ℓ_n, with the default test statistic being

$$T(S^\ell_n) = \sup_{t \in \mathbf{R}} \left(\hat{F}^\ell_{n,Y}(t) - \hat{F}^\ell_{n,X}(t) \right) = \max_{k \in \{1,\dots,q_y\}} \left(\hat{F}^\ell_{n,Y}(S^\ell_{n,k}) - \hat{F}^\ell_{n,X}(S^\ell_{n,k}) \right) \ ,$$

where the empirical CDFs are,

$$\hat{F}^\ell_{n,Y}(t) := \frac{1}{q_y} \sum_{j=1}^{q_y} I\{S^\ell_{n,j} \leq t\} \quad \text{and} \quad \hat{F}^\ell_{n,X}(t) := \frac{1}{q_x} \sum_{j=1}^{q_x} I\{S^\ell_{n,q_y+j} \leq t\} \ .$$

In order to define our test below in (C-2), we first introduce $\phi_\ell(\alpha)$, where

$$\phi_\ell(S^\ell_n, \alpha) := I\{T(S^\ell_n) > c_\alpha(q_y, q)\} \ .$$

The test $\phi_\ell(S^\ell_n, \alpha)$ depends on the point z_ℓ, or equivalently the index ℓ, in two ways. First, the random variable S^ℓ_n depends on z_ℓ through the induced order statistics, meaning the test statistic $T(S^\ell_n)$ is influenced by the choice of the target point z_ℓ. Second, the critical value $c_\alpha(q_y, q)$ is a function of (q_y, q), which, through the data-dependent rules introduced in Section 5.1, implicitly depends on z_ℓ. The dependence on α is directly evident from the definition of the critical value $c_\alpha(q_y, q)$.

To test the null hypothesis in (2.1) when $L > 1$, we propose the following test:

$$\phi(S^1_n, S^2_n, \dots, S^L_n) := \max_{\ell \in \{1,\dots,L\}} \phi_\ell \left(S^\ell_n, 1 - (1-\alpha)^{1/L} \right) \ . \tag{C-2}$$

In other words, the test $\phi(S^1_n, S^2_n, \dots, S^L_n)$ is the maximum of the L individual tests, each computed at a target point z_ℓ. However, each of these individual tests is performed at a level of $1 - (1-\alpha)^{1/L}$, rather than at the nominal level α. Since $\mathcal{Z}$ consists of a finite number of points, the asymptotic validity of the test $\phi(\alpha)$ follows from the validity of each individual test ϕ_ℓ and the fact that they are asymptotically independent. We formalize this result in Theorem 4.

C.2 Results for other test statistics

Our paper proposes the test in (3.7) based on the one-sided KS test statistic in (3.4) and a critical value in (3.6) constructed by simulating i.i.d. uniform random variables applied to this statistic. In this section, we investigate the properties of analogous tests that replace the KS statistic with other commonly used statistics in the stochastic dominance literature, such as the Cramér–von Mises (CvM) and Anderson–Darling (AD) statistics.

The main takeaway of this section is that the validity of our approach does not extend to tests that replace the KS statistic with the CvM or AD statistics. Specifically, we provide examples of data-generating processes that satisfy the null hypothesis in (2.1) and the test overrejects. Our examples involve discretely distributed data, which is allowed by our assumptions. That said, the CvM-based version of our test remains valid under the assumption of continuously distributed data, as we show later.

For concreteness, we now define the CvM and AD analogs of our test for the null hypothesis in (2.1). Given the pooled sample S_n in (3.3), the one-sided CvM test statistic is given by

$$T_{\text{CvM}}(S_n) = \int \left(\hat{F}_{n,Y}(s) - \hat{F}_{n,X}(s)\right)^+ d\hat{F}_{n,S}(s) = \frac{1}{q}\sum_{j=1}^{q} \left(\hat{F}_{n,Y}(S_{n,j}) - \hat{F}_{n,X}(S_{n,j})\right)^+ , \tag{C-3}$$

where $\hat{F}_{n,S}(s)$ denotes the empirical CDF of S_n and $x^+ = \max\{x, 0\}$[2]. In turn, the one-sided AD test statistic is given by

$$T_{\text{AD}}(S_n) = \int \frac{\left(\hat{F}_{n,Y}(s) - \hat{F}_{n,X}(s)\right)^+}{\hat{F}_{n,S}(s)(1-\hat{F}_{n,S}(s))} d\hat{F}_{n,S}(s) = \frac{1}{q}\sum_{j=1}^{q} \frac{\left(\hat{F}_{n,Y}(S_{n,j}) - \hat{F}_{n,X}(S_{n,j})\right)^+}{\hat{F}_{n,S}(S_{n,j})\left(1-\hat{F}_{n,S}(S_{n,j})\right)} . \tag{C-4}$$

These statistics align with their standard textbook definitions, as discussed in Hajek et al. (1999, p. 101). For both of these statistics, we can define the analog test as in our paper. For the CvM statistic, the corresponding test is

$$\phi_{\text{CvM}}(S) = I\{T_{\text{CvM}}(S) > c_{\alpha,\text{CvM}}(q_y, q)\}. \tag{C-5}$$

where $c_{\alpha,\text{CvM}}(q_y, q) = \inf_{x\in\mathbf{R}}\{P\{T_{\text{CvM}}(U) \leq x\} \geq 1-\alpha\}$ and $U = (U_1, \cdots, U_q)$ is a vector of i.i.d. $U(0,1)$. The testing procedure based on the AD statistic is defined analogously.

We are now ready to argue that the CvM and AD analogs of our test are invalid under our assumptions. Since the CvM and AD statistics are both rank statistics, we can repeat arguments in Theorem 2 to show that, for any $P \in \mathbf{P}_0$,

$$\lim_{n\to\infty} E_P[\phi_{\text{CvM}}(S_n)] = E_{\tilde{P}}[\phi_{\text{CvM}}(\tilde{S})] \quad \text{and} \quad \lim_{n\to\infty} E_P[\phi_{\text{AD}}(S_n)] = E_{\tilde{P}}[\phi_{\text{AD}}(\tilde{S})].$$

To prove that these tests are invalid, it then suffices to find any $P \in \mathbf{P}_0$ such that $E_{\tilde{P}}[\phi_{\text{CvM}}(\tilde{S})] > \alpha$ and $E_{\tilde{P}}[\phi_{\text{AD}}(\tilde{S})] > \alpha$. To this end, we consider $q_y = 2$, $q_x = 1$, $\{Y|Z = z_0\}$ and $\{X|Z = z_0\}$ distributed according to $Bernoulli(0.5)$. For $\alpha = 5\%$, we get $E_{\tilde{P}}[\phi_{\text{CvM}}(\tilde{S})] \approx 12.5\%$ and $E_{\tilde{P}}[\phi_{\text{AD}}(\tilde{S})] \approx 12.5\%$, as desired.

A notable feature of the examples above is that the data are discrete. This naturally raises the question of whether the validity of these tests can be recovered in settings with continuously distributed data. While we were unable to establish this result for the AD test, we were able to do so for the CvM test.

Theorem 6. *Let* $\mathbf{P}$ *the space of distributions that satisfy Assumptions 2, 3, and* $Y|Z=z_0$ *and* $X|Z=z_0$ *are continuously distributed, and let* $\mathbf{P}_0$ *be the subset of* $\mathbf{P}$ *such that (2.1) holds. Let* $\alpha \in (0,1)$ *be given,* T_{CvM} *be the CvM test statistic in (C-3), and* $\phi_{\text{CvM}}(\cdot)$ *be the test in (C-5). Then,*

$$\limsup_{n\to\infty} E_P[\phi_{\text{CvM}}(S_n)] \leq \alpha \tag{C-6}$$

whenever $P \in \mathbf{P}_0$.

Proof. This result follows from repeating arguments that prove Theorem 2. The only difference in the proof is that Theorem 5 is replaced by Theorem 7. ■

The previous theorem relies on the following auxiliary result.

Theorem 7. *Let* S *be the random variable in Theorem 1 and let* $\mathbf{P}_0$ *be as in Theorem 6. Then, for any* $P \in \mathbf{P}_0$ *and* $\alpha \in (0,1)$*, we obtain*

$$E_P[\phi_{\text{CvM}}(S)] \leq \bar{\alpha}_{\text{CvM}} \leq \alpha \ ,$$

where

$$\bar{\alpha}_{\text{CvM}} = P\left\{T_{\text{CvM}}(U) > c_{\alpha,\text{CvM}}(q_y, q)\right\}.$$

Moreover, the first inequality becomes an equality under P *such that* $F_Y(t|z_0) = F_X(t|z_0)$ *for all* $t \in \mathbf{R}$.

Proof. For any $t \in \mathbf{R}$, consider the CDF

$$F_M(t) = \frac{q_Y}{q} F_{Y|Z=z_0}(t) + \frac{q_X}{q} F_{X|Z=z_0}(t) \ ,$$

and let Q_M denote its corresponding quantile function. Let Q_Y and Q_X are the quantiles functions associated with $F_Y(\cdot|z_0)$ and $F_X(\cdot|z_0)$, respectively.

Since $Y|Z=z_0$ and $X|Z=z_0$ are continuously distributed, F_M is continuous. In turn, this implies that Q_M is strictly increasing. By $F_{Y|Z=z_0}(t) \leq F_{X|Z=z_0}(t)$ for all $t \in \mathbf{R}$, we also have

$$Q_Y(u) \geq Q_M(u) \geq Q_X(u) \quad \forall u \in (0,1) \ . \tag{C-7}$$

For brevity notation, we denote the critical value as $c_\alpha \equiv c_{\alpha,\text{CvM}}(q_y, q)$. Then,

$$\begin{aligned}
&E_P[\phi_{\text{CvM}}(S)] \\
&= P\left\{\frac{1}{q}\sum_{u=1}^{q}\left(\frac{1}{q_y}\sum_{i=1}^{q_y} I\left\{S_i \leq S_u\right\} - \frac{1}{q_x}\sum_{j=q_y+1}^{q} I\left\{S_j \leq S_u\right\}\right)^+ > c_\alpha\right\} \\
&\overset{(1)}{=} P\left\{\left[\begin{array}{l} \frac{1}{q}\sum_{u=1}^{q_y}\left(\frac{1}{q_y}\sum_{i=1}^{q_y} I\left\{Q_Y\left(U_i\right) \leq Q_Y\left(U_u\right)\right\} - \frac{1}{q_x}\sum_{j=q_y+1}^{q} I\left\{Q_X\left(U_j\right) \leq Q_Y\left(U_u\right)\right\}\right)^+ + \\ \frac{1}{q}\sum_{u=q_y+1}^{q}\left(\frac{1}{q_y}\sum_{i=1}^{q_y} I\left\{Q_Y\left(U_i\right) \leq Q_X\left(U_u\right)\right\} - \frac{1}{q_x}\sum_{j=q_y+1}^{q} I\left\{Q_X\left(U_j\right) \leq Q_X\left(U_u\right)\right\}\right)^+ \end{array}\right] > c_\alpha\right\}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(2)}{\leq} P\left\{\left[\begin{array}{l}
\frac{1}{q}\sum_{u=1}^{q_y}\left(\frac{1}{q_y}\sum_{i=1}^{q_y} I\left\{Q_M\left(U_i\right)\leq Q_M\left(U_u\right)\right\}-\frac{1}{q_x}\sum_{j=q_y+1}^{q} I\left\{Q_M\left(U_j\right)\leq Q_M\left(U_u\right)\right\}\right)^{+}+\\
\frac{1}{q}\sum_{u=q_y+1}^{q}\left(\frac{1}{q_y}\sum_{i=1}^{q_y} I\left\{Q_M\left(U_i\right)\leq Q_M\left(U_u\right)\right\}-\frac{1}{q_x}\sum_{j=q_y+1}^{q} I\left\{Q_M\left(U_j\right)\leq Q_M\left(U_u\right)\right\}\right)^{+}
\end{array}\right]>c_\alpha\right\}\\
&\overset{(3)}{=} P\left\{\left[\begin{array}{l}
\frac{1}{q}\sum_{u=1}^{q_y}\left(\frac{1}{q_y}\sum_{i=1}^{q_y} I\left\{U_i\leq U_u\right\}-\frac{1}{q_x}\sum_{j=q_y+1}^{q_x} I\left\{U_j\leq U_u\right\}\right)^{+}+\\
\frac{1}{q}\sum_{u=q_y+1}^{q}\left(\frac{1}{q_y}\sum_{i=1}^{q_y} I\left\{U_i\leq U_u\right\}-\frac{1}{q_x}\sum_{j=q_y+1}^{q} I\left\{U_j\leq U_u\right\}\right)^{+}
\end{array}\right]>c_\alpha\right\}\\
&\overset{(4)}{=} P\left\{T_{\mathrm{CvM}}(U)>c_\alpha\right\}\overset{(5)}{=}\bar{\alpha}_{\mathrm{CvM}}\overset{(6)}{\leq}\alpha\ ,
\end{aligned}
\tag{C-8}
$$

where (1) holds by the quantile transformation with $(U_1,\cdots,U_q)$ i.i.d. $U(0,1)$, (2) by

$$
\{Q_Y(U_a)\leq Q_X(U_b)\}\subseteq\{Q_M(U_a)\leq Q_M(U_b)\}\subseteq\{Q_X(U_a)\leq Q_Y(U_b)\} \tag{C-9}
$$

for all $a,b=1,\ldots,q$, (3) by the fact that Q_M is strictly increasing, and (4), (5), and (6) by the definitions of T_{CvM} and $\bar{\alpha}_{\mathrm{CvM}}$. We now show that (C-9) holds. For any $a,b=1,\ldots,q$, suppose that $Q_Y(U_a)\leq Q_X(U_b)$. By this and (C-7), we have that $Q_M(U_a)\leq Q_Y(U_a)\leq Q_X(U_b)\leq Q_M(U_b)$, which implies that $Q_M(U_a)\leq Q_M(U_b)$, as desired by the first inclusion. In turn, by this and (C-7), we have that $Q_X(U_a)\leq Q_M(U_a)\leq Q_M(U_b)\leq Q_Y(U_b)$, as desired by the second inclusion.

Since $P\in\mathbf{P}_0$ was arbitrary, (C-8) implies that $\sup_{P\in\mathbf{P}_0}E_P[\phi_{\mathrm{CvM}}(S)]\leq\bar{\alpha}$. Furthermore, under $F_Y(\cdot|z_0)=F_X(\cdot|z_0)$, we have that $Q_Y=Q_X=Q_M$, and so the inequality (2) in (C-8) holds with equality, as desired. ■

C.3 Empirical application

In this section, we consider two empirical applications. The first is an RDD in which the target point is known. The second is a two-independent-sample setting in which the target point is the median of the conditioning variable and must be estimated from the data. In both applications, the outcome variable contains ties. These applications therefore illustrate the use of our procedure in settings where the outcome distributions need not be atomless and, in the second application, where the target conditioning point is estimated from data.

C.3.1 The effect of retirement on household consumption

We apply our methodology to assess the distributional effect of retirement on household expenditures. Our empirical analysis follows Battistin et al. (2009) and Shen and Zhang (2016) (SZ, henceforth). As in these studies, we use data from the 1993–2004 Italian Survey on Household Income and Wealth (SHIW), a repeated cross-section survey that covers roughly 8,000 households across more than 3,000 municipalities; see [dataset] Shen and Zhang (2016). Because the outcomes are survey-reported, they exhibit ties, although the largest empirical point mass does not exceed 1.3%

We use the SHIW data to evaluate whether retirement shifts the distribution of household expenditures downward. As SZ explains, this hypothesis can be examined by framing the problem as an RDD, comparing

the distributions of household expenditures at ages just before and just after retirement. To formulate our problem, let Z denote the age of the household head, z_0 the retirement age, Y household expenditure before retirement, and X household expenditure after retirement. Under continuity of the conditional distributions in the retirement age of the household head, the test of a negative distributional effect of retirement can be formulated as the following CSD testing problem:

$$H_0 : F_Y(t|z_0) \leq F_X(t|z_0) \text{ for all } t \in \mathbf{R}, \qquad H_1 : F_Y(t|z_0) > F_X(t|z_0) \text{ for some } t \in \mathbf{R} \,. \tag{C-10}$$

We note that H_0 in (C-10) is a special case of (2.1) with $\mathcal{Z} = \{z_0\}$. Following the logic of RDDs, conditioning on $Z = z_0$ is key to controlling for confounding factors.

sample (n)	expenditure type	p-value (KS)	q (KS)	p-value (SZ)	q (SZ)	c_L (KS)	c_U (KS)
all households	nondurable	22.64	479	13.69	8,739	0	52.30
(26,914)	food	86.05	466	71.74	8,739	0	42.10
wife already eligible	nondurable	99.55	185	66.72	1,822	0	∞
(4,798)	food	85.96	188	99.69	1,414	0	25.70
wife ineligible	nondurable	16.14	232	0.15*	1,460	0	∞
(7,305)	food	13.95	228	3.94*	1,460	0.75	3.60
single men	nondurable	43.23	296	65.82	2,884	0	16.30
(9,001)	food	41.33	292	31.19	2,537	0	18.60

Table 1: Empirical results for the SHIW data. The table reports p-values and effective sample sizes for our KS test and the SZ test. All p-values are reported in percent. The last two columns report the sensitivity analysis for the KS test, where the baseline effective sample sizes (q_y, q_x) are rescaled as (cq_y, cq_x). The values c_L and c_U are the lower and upper breakpoints at which the 10% test decision changes relative to the baseline choice $c = 1$. Values of 0 and ∞ indicate no change below or above $c = 1$, respectively, over the range considered. The symbol * denotes rejection at the 10% level.

Table 1 reports the results of implementing the test in (C-10) using both our methodology (KS) and the procedure proposed by SZ. Following SZ, we consider expenditures on nondurables and food, and we report results for four samples: the full sample, the subsample of households with wives eligible for retirement, the subsample of households with wives not yet eligible for retirement, and the subsample of households consisting of single men. The table presents the p-values for both tests and their corresponding effective sample sizes. While our test (KS) does not reject H_0 across all subsamples and expenditures, SZ rejects H_0 for both nondurable and food expenditures (at the 1% and 5% levels, respectively) for the subsample of households in which the wife is ineligible for retirement. Consistent with our simulation evidence, the SZ test uses substantially larger effective sample sizes than our KS test.

Furthermore, the last two columns of Table 1 report a sensitivity analysis with respect to the choice of effective sample sizes. Given the baseline choice (q_y, q_x) computed by our rule of thumb, we rerun the test using (cq_y, cq_x), where $c > 0$ is a common scaling factor that varies over an evenly spaced grid with step size 0.05. The lower breakpoint c_L is the first grid point below 1 at which the 10% test decision changes, while the upper breakpoint c_U is defined analogously above 1. Except for food expenditure in the wife ineligible subsample, the interval $[c_L, c_U]$ is wide, with $c_L = 0$ and c_U ranging from 16.30 to ∞, indicating that the empirical conclusions are robust to the choice of effective sample sizes. For food expenditure in the wife ineligible subsample, the interval is narrower, $[0.75, 3.60]$, implying that the rejection at the 10% level is somewhat more sensitive to the choice of effective sample sizes, although the conclusion remains unchanged over a substantial range around the baseline choice.

C.3.2 The effect of class size on student achievement

We next illustrate our methodology using data from the Tennessee Student/Teacher Achievement Ratio (STAR) experiment; see [dataset] Achilles et al. (2008). Following Qu and Yoon (2015), we study the distributional effect of class size on student achievement conditional on teacher experience. The target conditioning point is chosen as the median teacher experience and estimated by its sample analogue. Our analysis focuses on conditional first-order stochastic dominance of the outcome distributions at this target value, rather than the conditional quantile treatment effects considered by Qu and Yoon (2015).

Let Y and X denote the student achievement scores of students assigned to small and regular classes, respectively, and let Z denote teacher experience, measured in years. We consider three achievement outcomes: reading scores, mathematics scores, and their sum. All three outcomes are integer-valued, so ties arise naturally in the recorded outcomes. Table 3 reports the largest empirical point masses in both the full samples and the effective samples used by our test. Following Qu and Yoon (2015), we estimate the median teacher experience by its sample analogue. We consider both dominance directions:

$$\begin{aligned} &H_0^+ : F_Y(t \mid z_0) \le F_X(t \mid z_0) \text{ for all } t \in \mathbf{R}, \quad H_1^+ : F_Y(t \mid z_0) > F_X(t \mid z_0) \text{ for some } t \in \mathbf{R}, \\ &H_0^- : F_X(t \mid z_0) \le F_Y(t \mid z_0) \text{ for all } t \in \mathbf{R}, \quad H_1^- : F_X(t \mid z_0) > F_Y(t \mid z_0) \text{ for some } t \in \mathbf{R}. \end{aligned} \tag{C-11}$$

Here, H_0^+ corresponds to the hypothesis that small classes first-order stochastically dominate regular classes, while H_0^- corresponds to the reverse dominance hypothesis.

null	outcome	q_{small}	q_{regular}	n_{small}	n_{regular}	p-value	c_L	c_U
H_0^+	sum	94	98	301	338	95.5	0	∞
H_0^-						0.06*	0	∞
H_0^+	reading	93	96	302	338	98.9	0	∞
H_0^-						0.05*	0.6	∞
H_0^+	math	95	101	301	339	96.9	0	∞
H_0^-						0.75*	0	∞

Table 2: Empirical results for the Project STAR data. The table reports p-values for the two dominance hypotheses in (C-11). All p-values are reported as percentages. We consider three outcomes: reading, math, and their sum. The effective sample sizes $(q_{\text{small}}, q_{\text{regular}})$ are rescaled as $(cq_{\text{small}}, cq_{\text{regular}})$ in the sensitivity analysis. The values c_L and c_U are the lower and upper breakpoints at which the 10% test decision changes relative to the baseline choice $c = 1$. Values of 0 and ∞ indicate no change below or above $c = 1$, respectively, over the range considered. The symbol * denotes rejection at the 10% level.

Table 2 reports the results for both dominance directions and all three outcomes. For the sum of reading and math scores, we do not reject the hypothesis that the conditional distribution of student achievement in small classes first-order stochastically dominates that in regular classes; the corresponding p-value is 95.5%. By contrast, the reverse dominance hypothesis is rejected at the 10% level, with a p-value of 0.06%. The same qualitative conclusion holds for reading and math scores separately: H_0^+ is not rejected, whereas H_0^- is rejected at the 1% level for both outcomes. These results provide evidence that, conditional on median teacher experience, the distribution of student achievement under small classes first-order stochastically dominates that under regular classes. The last two columns of Table 2 report the same sensitivity analysis as in the previous application. For five of the six outcome–hypothesis combinations, the breakpoints satisfy $c_L = 0$ and $c_U = \infty$. For the reverse dominance hypothesis for reading, $c_L = 0.6$ and $c_U = \infty$. Thus, the conclusions are unchanged over a wide range of effective sample sizes around the baseline choice.

Outcome	Class	Sample	q	n	Score	Count	Largest point mass (%)
reading	small	full	–	1,739	434	48	2.8
reading	regular	full	–	2,006	419	57	2.8
reading	small	effective	93	–	434	6	6.5
reading	regular	effective	96	–	398 and 413	5 each	5.2
math	small	full	–	1,762	489	87	4.9
math	regular	full	–	2,032	478	100	4.9
math	small	effective	95	–	500	7	7.4
math	regular	effective	101	–	449	10	9.9

Table 3: Largest empirical point masses in the full and effective samples for the reading and mathematics scores. Score denotes the score attaining the maximum observed frequency, and Count denotes the corresponding frequency.

C.4 Details on the Simulations and the data dependent rule for q

The parameter values used in the simulations of Section 6 are reported in the following tables. The parameters for Designs 1-4 are summarized in Table 4, those for Designs 5-7 in Table 5, and the parameter values for case (d)—power results—in Table 6. Table 7 reports the mean values of q_y^* and q_x^* across simulations.

C.4.1 Details on the implementation of the data dependent rule for q

Here we also discuss some numerical details on the implementation of the data dependent rule for q introduced in (5.4) and (5.5). We do this for a generic sample $\{(W_i, Z_i) : 1 \le i \le n\}$, suppressing the sample-specific Y and X subscripts. Let $\hat{\mu}_Z$ be the sample median, $\hat{\sigma}_Z = \mathrm{IQR}(Z)/(2\Phi^{-1}(3/4))$, $\hat{f}_0 = \phi_{\hat{\mu}_Z,\hat{\sigma}_Z}(\hat{z}_n)$, and

$$\hat{C}_Z = \frac{1}{\hat{\sigma}_Z^2\sqrt{2\pi e}}, \qquad \hat{C}_W(\rho) = \frac{|\rho|}{\hat{\sigma}_Z\sqrt{1-\rho^2}}\frac{1}{\sqrt{2\pi}} \ .$$

For $|\rho| < 1$, define

$$\hat{c}_n(\rho) := 2^{-1/3}\left(\frac{4\hat{f}_0^2}{\hat{C}_Z/2 + \hat{f}_0\hat{C}_W(\rho)}\right)^{2/3}, \qquad q_n(\rho) := \max\left\{q_{\min}, \lceil n^{1/2}\hat{c}_n(\rho)\rceil\right\} \ . \tag{C-12}$$

Our codes uses $q_{\min} = 10$ and $\hat{\rho} = \mathrm{sgn}(\tilde{\rho})\min\{|\tilde{\rho}|, 0.99\}$, where $\tilde{\rho} = \sin(\pi\hat{\tau}/2)$ and $\hat{\tau}$ is Kendall's rank correlation. Thus $\hat{\rho} \in [-0.99, 0.99]$. The implementation rounds upward and imposes the lower bound $q_{\min}$.

The rule has a deterministic $O(n^{1/2})$ envelope. Since $\hat{C}_W(\rho) \ge 0$, $\hat{f}_0 \le (\hat{\sigma}_Z\sqrt{2\pi})^{-1}$, and $\hat{C}_Z = (\hat{\sigma}_Z^2\sqrt{2\pi e})^{-1}$,

$$\hat{c}_n(\rho) \le 2^{-1/3}\left(\frac{8\hat{f}_0^2}{\hat{C}_Z}\right)^{2/3} \le \left(\frac{16e}{\pi}\right)^{1/3} =: \kappa \approx 2.4012 \ .$$

Consequently,

$$q_n(\rho) \le \bar{q}_n := \max\{q_{\min}, \lceil \kappa n^{1/2}\rceil\} \ . \tag{C-13}$$

Hence $\bar{q}_n = o(n^{2/3})$ and the envelope in Assumption 5 holds with $\eta_n = 0$. If $\delta_n \asymp n^{-1/2}$, then $\sqrt{\bar{q}_n}\delta_n = O(n^{-1/4})$, so the growth conditions in (4.7) hold.

The rank correlation $\hat{\rho}$ in (5.4)–(5.5) depends on the outcomes, so estimating it from the full sample is not $\mathcal{A}$-measurable and the resulting rule does not satisfy Assumption 5. As anticipated in Section 5.1, we

therefore estimate $\hat{\rho}$ from the observations that cannot enter the test neighborhood. Let the pilot size be $m_0 = \bar{q}_n$, with $\bar{q}_n$ as in (C-13), and let M_0 contain the m_0 observations closest to $\hat{z}_n$, breaking distance ties by an $\mathcal{A}$-measurable rule. We estimate Kendall's correlation from $\{(W_i, Z_i) : i \notin M_0\}$, transform and clip it to obtain $\hat{\rho}_{\text{lo}}$, and set $q_{\text{lo}} = q_n(\hat{\rho}_{\text{lo}})$. Since $q_{\text{lo}} \leq \bar{q}_n = m_0$, the selected neighborhood is contained in M_0, so the outcomes that determine the test observations and those used to compute $\hat{\rho}_{\text{lo}}$ are disjoint.

This construction meets the requirement of Theorem 3 under a minor extension. Define

$$\mathcal{G}_n = \mathcal{A} \vee \sigma\{W_i : i \notin M_0\} \ ,$$

taking the join of the corresponding σ-fields for the two samples. While q_{lo} is strictly speaking not $\mathcal{A}$-measurable, it is indeed $\mathcal{G}_n$-measurable. The selected indices are also $\mathcal{G}_n$-measurable, while the selected outcomes are not included in $\mathcal{G}_n$ and, conditional on $\mathcal{G}_n$, remain mutually independent with conditional laws P_{W,Z_i}. The conditional product representation (A-7) therefore continues to hold with $\mathcal{G}_n$ in place of $\mathcal{A}$, and the proof of Theorem 3 extends by conditioning on $\mathcal{G}_n$. Because $m_0 = O(n^{1/2})$, the leave-out step discards only a vanishing fraction of the sample, so $\hat{\rho}_{\text{lo}}$ and the full-sample estimate have the same probability limit under standard regularity conditions. In finite samples the rules based on the leave-out and full-sample estimates are numerically very close; we compute both and report them side by side in the robustness checks below.

A simpler $\mathcal{A}$-measurable alternative sets $\rho = 0$: since $\hat{C}_W(\rho) \geq \hat{C}_W(0) = 0$ gives $q_n(\rho) \leq q_n(0)$, this dispenses with the leave-out step and yields the largest tuning parameter in the family we consider, at the cost of ignoring the estimated dependence. We report both this variant and the full-sample rule in the robustness checks below.

Design	$\mu_Y(z)$	$\mu_X(z)$	$\sigma_Y(Z)$	$\sigma_X(Z)$	U, V	z_ℓ
1a	z	z	z^2	z^2	$N(0,1)$	0.5
1b	$1.05z$	z	z^2	z^2	$N(0,1)$	0.5
1c	z	z	z^2	z^2	$N(0,1)$	0.25, 0.75
2a	z	$z^2 + 0.25$	z^2	z^2	$N(0,1)$	0.5
2b	$1.05z$	$0.5z + 0.25$	z^2	z^2	$N(0,1)$	0.5
2c	z	$z - (z - 0.25)(z - 0.75)$	z^2	z^2	$N(0,1)$	0.25, 0.75
3a	z	z	z^2	z^2	$U[0,1]$	0.5
3b	$z + 0.1z^2$	z	$0.95z^2$	z^2	$U[0,1]$	0.5
3c	z	z	z^2	z^2	$U[0,1]$	0.25, 0.75
4a	$\mu(z)$	$\mu(z)$	1	1	$N(0,1)$	0
4b	$\mu(z) + 0.1$	$\mu(z)$	1	1	$N(0,1)$	0
4c	$\mu(z)$	$\mu(z)$	1	1	$N(0,1)$	-0.5, 0.5

Table 4: Parameter values for the continuously distributed designs. Designs 1 to 3 are based on the location-scale model in (6.1), while Designs 4 is based on the RDD model in (6.3).

C.4.2 Robustness checks

In this appendix, we assess the sensitivity of our procedure to three choices: the sample size, the nominal level, and the rule used to select the tuning parameters (q_y, q_x). We consider Design 4 (continuous, RDD) and Design 7 (discrete, binomial) from Section 6, together with $\alpha \in \{0.05, 0.10\}$ and three q-selection rules: the data-dependent rule proposed in the paper (*leave-out*); a version of the same rule that estimates ρ without leaving observations out (*full*); and a variant that ignores the estimated dependence altogether by imposing

Design	$\mu_Y(z)$	$\mu_X(z)$	$\sigma_Y(Z)$	$\sigma_X(Z)$	U, V	z_ℓ
5a	0	0	z^2	z^2	C-logN	0.5
5b	0	0	$1.05z^2$	z^2	C-logN	0.5
5c	0	0	z^2	z^2	C-logN	0.25, 0.75
6a	0	–	–	–	discrete	0.5
6b	1	–	–	–	discrete	0.5
6c	0	–	–	–	discrete	0.25, 0.75
7a	0	–	–	–	Binomial	0.5
7b	1	–	–	–	Binomial	0.5
7c	0	–	–	–	Binomial	0.25, 0.75

Table 5: Parameter values for the discrete and mixed designs. Designs 5 is based on the location-scale model in (6.1), while Designs 6 and 7 are discrete designs as described in the main text.

Design	$\mu_Y(z)$	$\mu_X(z)$	$\sigma_Y(Z)$	$\sigma_X(Z)$	U, V	z_ℓ
1d	$0.95z$	z	z^2	z^2	$N(0,1)$	0.5
2d	z	$0.6z + 0.25$	z^2	z^2	$N(0,1)$	0.5
3d	z	z	$0.90z^2$	z^2	$U[0,1]$	0.5
4d	$\mu(z)$	$\mu(z)$	$0.5 + z^2$	1	$N(0,1)$	0
5d	0	0	$0.90z^2$	z^2	C-logN	0.5
6d	-1/2	–	–	–	discrete	0.5
7d	-1	–	–	–	Binomial	0.5

Table 6: Parameter values for all designs under the alternative hypothesis H_1.

$\rho = 0$ (*rho0*). We further consider two additional sample sizes ($n = 500$ and $n = 2{,}000$, alongside the paper's $n = 1{,}000$) and two unbalanced sample sizes, $(n_y, n_x) = (500, 1{,}500)$ and $(1{,}500, 500)$.

Table 8 reports rejection rates under H_0 and Table 9 under H_1, with $MC = 10{,}000$ replications per cell. Each cell reports the rejection rate (%) under the three q-rules, separated by slashes, in the order full/leave-out/rho0.

The results confirm the conclusions in the main text. For Design 4, size is controlled closely at both nominal levels and across all sample sizes, and the choice of q-rule is essentially immaterial: the dependence between the outcome and the running variable is weak in this design, so $\hat{\rho}$ is close to zero regardless of how it is estimated, and the three rules select nearly identical values of q. For Design 7, rejection rates under H_0 remain below the nominal level throughout – the test never over-rejects in this grid – and the default rule is conservative, as in the main text; the refined critical value brings rejection rates closer to the nominal level. Imposing $\rho = 0$ increases q substantially, which improves power. Overall, the procedure's size and power properties are robust to the sample size, the nominal level, and reasonable variation in how the tuning parameter is selected.

C.4.3 Finite-sample effects of estimating the conditioning point

We investigate the finite-sample effects of estimating the conditioning point using two designs based on Designs 1 and 2 of Section 6. In both, $Z \sim \text{Beta}(2, 2)$, the two samples are independent with $n_y = n_x = n$, the conditioning point is the median of Z, $z_0 = 1/2$, and $\hat{z}_n$ is the sample median of the pooled Z observations. The tuning parameters (q_y, q_x) are chosen by the rule in Section 5.1, evaluated at the same point used to select the nearest neighbors.

As before, Design (D1a) is such that $F_Y(\cdot|z) = F_X(\cdot|z)$ at every z and the null hypothesis in (2.1)

Design	1			2			3			4		
	a	b	c	a	b	c	a	b	c	a	b	c
q_y^*	46.34	45.53	29.28	46.32	45.53	29.27	21.25	20.09	15.17	52.00	51.97	40.22
q_x^*	46.33	46.34	29.28	48.36	56.79	29.18	21.25	21.26	15.17	52.35	52.35	40.22
Design	**5**			**6**			**7**					
	a	b	c	a	b	c	a	b	c			
q_y^*	39.18	39.18	25.94	66.88	74.26	37.72	33.10	33.85	22.60	–	–	–
q_x^*	39.16	39.20	25.92	66.92	66.90	37.74	33.11	33.11	22.59	–	–	–

Table 7: Mean values of q_y^* and q_x^* across simulations: $n = 1,000$, $\alpha = 10\%$, $MC = 10,000$.

(n_y, n_x)	D4 α=.05	D4 α=.10	D7 α=.05 def	D7 α=.05 ref	D7 α=.10 def	D7 α=.10 ref
(500,500)	5.1/5.1/4.8	9.7/9.7/7.4	1.6/1.6/1.3	2.6/2.9/1.4	3.5/3.7/2.5	4.8/4.8/6.6
(1000,1000)	4.8/4.8/3.4	9.6/9.4/7.5	1.4/1.5/1.1	2.3/2.4/2.6	3.5/3.4/2.6	4.8/4.7/3.9
(2000,2000)	4.9/5.1/4.9	9.8/9.6/9.3	1.6/1.5/1.5	2.1/2.1/1.5	3.2/3.1/3.1	4.5/4.6/3.2
(500,1500)	5.2/5.2/5.3	10.4/10.3/10.4	1.6/1.4/1.5	2.4/2.3/2.6	3.6/3.8/3.3	5.8/5.9/5.9
(1500,500)	4.5/4.5/4.6	9.4/9.6/9.6	1.6/1.6/1.4	2.4/2.5/2.8	3.3/3.5/3.5	5.7/5.8/5.9

Table 8: Rejection rates (%) under H_0 for Designs 4 (RDD) and 7 (binomial), $\alpha \in \{.05, .10\}$, default and refined critical values (D7 only), MC = 10,000. Each cell reports `full/leaveout/rho0`, the rejection rate under the three q-selection rules.

(n_y, n_x)	D4 α=.05	D4 α=.10	D7 α=.05 def	D7 α=.05 ref	D7 α=.10 def	D7 α=.10 ref
(500,500)	36.0/35.9/35.7	53.3/53.0/46.2	9.1/8.4/16.7	13.3/12.9/16.9	16.1/15.8/23.6	20.1/19.3/31.1
(1000,1000)	52.6/52.3/46.3	69.1/69.1/64.5	12.2/12.3/19.8	15.3/15.0/32.6	21.5/21.1/32.6	26.9/26.6/40.0
(2000,2000)	72.4/72.2/73.9	85.3/85.3/85.2	15.8/15.5/34.4	20.4/20.1/34.7	25.4/25.1/47.0	32.9/32.3/47.3
(500,1500)	35.6/35.7/36.5	53.0/53.2/54.5	10.8/10.7/21.8	15.0/14.7/29.0	19.2/19.4/33.4	26.3/26.0/44.1
(1500,500)	36.4/36.3/37.7	52.7/52.5/53.4	10.9/11.0/22.0	14.7/14.9/30.5	19.2/19.1/33.7	26.2/26.5/44.5

Table 9: Rejection rates (%) under H_1 for Designs 4 (RDD) and 7 (binomial), $\alpha \in \{.05, .10\}$, default and refined critical values (D7 only), MC = 10,000. Each cell reports `full/leaveout/rho0`, the rejection rate under the three q-selection rules.

holds with equality at every conditioning point. In Design (D1b) the null hypothesis holds in its interior. In Design (D1d) the null hypothesis is violated at every z. The second design (D2a) sets $\mu_y(z) = z$ and $\mu_x(z) = z^2 + 1/4$, so that $\mu_x(z) - \mu_y(z) = (z - 1/2)^2$ with equal conditional variances. The null hypothesis then holds, with equality, only at $z_0 = 1/2$, and is violated at every other conditioning point. This is the most demanding configuration for an estimated conditioning point: any deviation of $\hat{z}_n$ from z_0 centers the neighborhoods at points where the null hypothesis fails. Table 10 reports the rejection rates at $\alpha = 10\%$ using the known conditioning point z_0 versus the estimated $\hat{z}$ (pooled sample median of Z), for $n = 500, 1{,}000, 2{,}000$. D1a/D1b/D2a report size, D1d reports power. Mean $|\hat{z} - z_0|$ measures how well the estimated center concentrates around the true one as n grows.

References

Abadie, A. (2002): "Bootstrap tests for distributional treatment effects in instrumental variable models," *Journal of the American statistical Association*, 97, 284–292.

	$n = 500$				$n = 1{,}000$				$n = 2{,}000$			
	D1a	D1b	D1d	D2a	D1a	D1b	D1d	D2a	D1a	D1b	D1d	D2a
KS (z_0)	9.5	5.1	17.9	10.3	9.9	4.6	19.1	9.6	10.2	3.5	21.1	10.2
KS ($\hat{z}$)	9.4	5.3	18.1	10.4	9.6	4.6	19.4	9.7	10.0	3.7	21.0	10.8
Mean $\lvert\hat{z} - z_0\rvert$	0.009	0.008	0.008	0.008	0.006	0.006	0.006	0.006	0.004	0.004	0.004	0.004

Table 10: Rejection rates (%) at $\alpha = 10\%$ using the known conditioning point z_0 versus the estimated $\hat{z}$, for $n = 500, 1{,}000, 2{,}000$. D1a/D1b/D2a report size, D1d reports power.

Anderson, G. (1996): "Nonparametric tests of stochastic dominance in income distributions," *Econometrica: Journal of the Econometric Society*, 1183–1193.

Andrews, D. W. and X. Shi (2017): "Inference based on many conditional moment inequalities," *Journal of Econometrics*, 196, 275–287.

Armstrong, T. B. and M. Kolesár (2018): "Optimal inference in a class of regression models," *Econometrica*, 86, 655–683.

Barcellos, S. H., L. S. Carvalho, and P. Turley (2023): "Distributional effects of education on health," *Journal of Human Resources*, 58, 1273–1306.

Barrett, G. F. and S. G. Donald (2003): "Consistent tests for stochastic dominance," *Econometrica*, 71, 71–104.

Battistin, E., A. Brugiavini, E. Rettore, and G. Weber (2009): "The retirement consumption puzzle: evidence from a regression discontinuity approach," *American Economic Review*, 99, 2209–2226.

Bhattacharya, P. (1974): "Convergence of sample paths of normalized sums of induced order statistics," *The Annals of Statistics*, 1034–1039.

Blackman, J. (1956): "An extension of the Kolmogorov distribution," *The Annals of Mathematical Statistics*, 513–520.

Bugni, F. A. and I. A. Canay (2021): "Testing Continuity of a Density via g-order statistics in the Regression Discontinuity Design," *Journal of Econometrics*, 221, 138–159.

Bugni, F. A., I. A. Canay, and D. Kim (2025): "On the Rate of Convergence of Induced Ordered Statistics and their Applications," Tech. rep., Northwestern University.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018): "Inference under Covariate Adaptive Randomization," *Journal of the American Statistical Association*, 113, 1784–1796.

Canay, I. A. and V. Kamat (2018): "Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design," *The Review of Economic Studies*, 85, 1577–1608.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017): "Randomization Tests under an Approximate Symmetry Assumption," *Econometrica*, 85, 1013–1030.

Caughey, D., A. Dafoe, X. Li, and L. Miratrix (2023): "Randomisation inference beyond the sharp null: bounded null hypotheses and quantiles of individual treatment effects," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85, 1471–1491.

Chang, M., S. Lee, and Y.-J. Whang (2015): "Nonparametric tests of conditional treatment effects with an application to single-sex schooling on academic achievements," *The Econometrics Journal*, 18, 307–346.

Chung, E. and J. P. Romano (2013): "Exact and asymptotically robust permutation tests," *The Annals of Statistics*, 41, 484–507.

[dataset] Achilles, C. M., H. P. Bain, F. Bellott, J. Boyd-Zaharias, J. Finn, J. Folger, J. Johnston, and E. Word (2008): "Tennessee's Student Teacher Achievement Ratio (STAR) Project," .

[dataset] Shen, S. and X. Zhang (2016): "Replication data for: Distributional Tests for Regression Discontinuity: Theory and Empirical Examples," Harvard Dataverse, version 1.1.

David, H. and J. Galambos (1974): "The asymptotic theory of concomitants of order statistics," *Journal of Applied Probability*, 762–770.

Davidson, R. and J.-Y. Duclos (2000): "Statistical inference for stochastic dominance and for the measurement of poverty and inequality," *Econometrica*, 68, 1435–1464.

——— (2013): "Testing for Restricted Stochastic Dominance," *Econometric Reviews*, 32, 84–125.

Delgado, M. A. and J. C. Escanciano (2013): "Conditional stochastic dominance testing," *Journal of Business & Economic Statistics*, 31, 16–28.

Donald, S. G., Y.-C. Hsu, and G. F. Barrett (2012): "Incorporating covariates in the measurement of welfare and inequality: methods and applications," *The Econometrics Journal*, 15, C1–C30.

Dubhashi, D. P. and A. Panconesi (2009): *Concentration of Measure for the Analysis of Randomized Algorithms*, Cambridge University Press.

Durbin, J. (1973): *Distribution theory for tests based on the sample distribution function*, SIAM.

Gnedenko, B. V. and V. S. Korolyuk (1951): "On the maximum divergence of two empirical distributions (in Russian)," *DAN*, 80, 525.

Goldman, M. and D. M. Kaplan (2018): "Comparing distributions by multiple testing across quantiles or CDF values," *Journal of Econometrics*, 206, 143–166.

Gonzalo, J. and J. Olmo (2014): "Conditional stochastic dominance tests in dynamic settings," *International Economic Review*, 55, 819–838.

Hajek, J., Z. Sidak, and P. K. Sen (1999): *Theory of rank tests*, Academic press, 2nd edition ed.

Hodges, J. (1958): "The significance probability of the Smirnov two-sample test," *Arkiv för matematik*, 3, 469–486.

Hájek, J. and Z. Šidák (1967): *Theory of Rank Tests*, New York: Academic Press.

Kaufmann, E. and R.-D. Reiss (1992): "On conditional distributions of nearest neighbors," *Journal of Multivariate Analysis*, 42, 67–76.

Korolyuk, V. S. (1955): "On the discrepancy of empiric distributions for the case of two independent samples," *Izv. Akad. Nauk SSSR Ser. Mat.*, 19, 81–96.

Lehmann, E. L. (1951): "Consistency and unbiasedness of certain nonparametric tests," *The annals of mathematical statistics*, 165–179.

Lehmann, E. L. and J. P. Romano (2005): *Testing Statistical Hypotheses*, Springer, New York, 3rd ed.

Li, Y. and N. Sunder (2024): "Distributional effects of education on mental health," *Labour Economics*, 88, 102528.

Linton, O., E. Maasoumi, and Y.-J. Whang (2005): "Consistent testing for stochastic dominance under general sampling schemes," *The Review of Economic Studies*, 72, 735–765.

Linton, O., M. H. Seo, and Y.-J. Whang (2023): "Testing stochastic dominance with many conditioning variables," *Journal of Econometrics*, 235, 507–527.

Linton, O., K. Song, and Y.-J. Whang (2010): "An improved bootstrap test of stochastic dominance," *Journal of Econometrics*, 154, 186–202.

Linton, O., Y.-J. Whang, and Y.-M. Yen (2016): "A Nonparametric Test of a Strong Leverage Hypothesis," *Journal of Econometrics*, 194, 153–186.

Lok, T. M. and R. V. Tabri (2021): "An Improved Bootstrap Test for Restricted Stochastic Dominance," *Journal of Econometrics*, 224, 371–393.

Massart, P. (1990): "The Tight Constant in the Dvoretzky–Kiefer–Wolfowitz Inequality," *The Annals of Probability*, 18, 1269–1283.

McFadden, D. (1989): "Testing for stochastic dominance," in *Studies in the economics of uncertainty: In honor of Josef Hadar*, Springer, 113–134.

Pollard, D. (2002): *A User's Guide to Measure Theoretic Probability*, Cambrigde University Press, New York.

Qu, Z. and J. Yoon (2015): "Nonparametric estimation and inference on conditional quantile processes," *Journal of Econometrics*, 185, 1–19.

——— (2019): "Uniform inference on quantile effects under sharp regression discontinuity designs," *Journal of Business & Economic Statistics*, 37, 625–647.

Reiss, R.-D. (1989): *Approximate distributions of order statistics: with applications to nonparametric statistics*, Springer-Verlag, New York.

Shen, S. (2019): "Estimation and inference of distributional partial effects: Theory and application," *Journal of Business & Economic Statistics*, 37, 54–66.

Shen, S. and X. Zhang (2016): "Distributional Tests for Regression Discontinuity: Theory and Empirical Examples," *Review of Economics and Statistics*.

Whang, Y.-J. (2019): "Econometric analysis of stochastic dominance," *Cambridge Books*.

Wu, Q. and D. M. Kaplan (2025): "Multiple testing of stochastic monotonicity," Tech. rep., Department of Economics, University of Missouri.