Optimal Tests of the Composite Null Hypothesis Arising in Mediation Analysis

Caleb H. Miles, Antoine Chambaz *

Abstract

The indirect effect of an exposure on an outcome through an intermediate variable can be identified by a product of two regression coefficients under certain causal and regression modeling assumptions. In this context, the null hypothesis of no indirect effect is a composite null hypothesis, as the null holds if either regression coefficient is zero. A consequence is that traditional hypothesis tests are severely underpowered near the origin (i.e., when both coefficients are small with respect to standard errors). We propose hypothesis tests that (i) preserve level alpha type 1 error, (ii) meaningfully improve power when both true underlying effects are small relative to sample size, and (iii) preserve power when at least one is not. One approach gives a closed-form test that is minimax optimal with respect to local power over the alternative parameter space. Another uses sparse linear programming to produce an approximately optimal test for a Bayes risk criterion. We discuss adaptations for performing large-scale hypothesis testing as well as modifications that yield improved interpretability. We provide an R package that implements our proposed methodology.

Keywords: Bayes risk optimality, Causal inference, Large-scale hypothesis testing, Non-uniform asymptotics, Similar test

*Caleb H. Miles is Assistant Professor, Department of Biostatistics, Columbia University, New York, NY 10032, USA (email: cm3825@cumc.columbia.edu); and Antoine Chambaz is Professor of Applied Mathematics, Université Paris Cité, CNRS, MAP5, F-75006 Paris, France.

1. INTRODUCTION

Mediation analysis is a widely popular discipline that seeks to understand the mechanism by which an exposure affects an outcome by learning about intermediate events that transmit part or all of its effect. For instance, if one is interested in the effect of an exposure A on an outcome Y, one might posit that at least in part, A affects Y by first affecting some intermediate event M, which in turn affects Y. Despite its long history and broad application, with an exception from recent work (van Garderen and van Giersbergen, 2024), existing hypothesis tests of a single mediated effect are overly conservative and underpowered in a certain region of the alternative hypothesis space. In this article, we demonstrate the reason for this suboptimal behavior, and develop new hypothesis tests that deliver optimal power in some decision theoretic senses while preserving type 1 error uniformly over the composite null subspace of the parameter space. We focus on the case in which the indirect effect is identified by a product of two coefficients or functions of coefficients that can be estimated with uniform joint convergence to a bivariate normal distribution with diagonal covariance matrix. In fact, our approach is not limited to the realm of mediation analysis, but can be applied to test for any product of two coefficients that are estimable in the above sense.

Briefly, the inferential problem that arises in mediation analysis is that when the mediated effect is identified by a product of two coefficients, even while the joint distribution of estimators of these coefficients may converge uniformly to a bivariate normal distribution, neither their product nor the minimum of their absolute values will converge uniformly to a Gaussian law. Thus, asymptotic approximation-based univariate test statistics that are a function of either of these summaries of the two coefficient estimates will have poor properties in an important region of the parameter space for any given finite sample.

Previously, MacKinnon et al. (2002) and Barfield et al. (2017) presented simulation

studies comparing hypothesis tests of a mediated effect. Both articles observed that all existing tests are either underpowered in some scenarios or do not preserve nominal type 1 error in all scenarios. Both articles found the so-called joint significance test to be the best overall performing test, while Barfield et al. (2017) found a bootstrap-based test to have comparable performance across some scenarios. Huang (2019) proved that the joint significance test yields a smaller p-value than those of both normality-based tests and normal product-based tests. Barfield et al. (2017) highlighted the problem of large-scale hypothesis testing for mediated effects, which has become a popular inferential task in genomics, where genomic measures such as gene expression or DNA methylation are hypothesized to be potential mediators of the effect of genes or other exposures on various outcomes. In fact, the conservative behavior of mediation hypothesis tests is most evident in large-scale hypothesis testing, since one can compare the distribution of p-values with the uniform distribution. Huang (2019) showed that the joint significance and delta method tests both generate p-values forming a distribution that stochastically dominates the uniform distribution. However, while this behavior is readily observable in large-scale hypothesis testing, the problem is no different for a single hypothesis test; it is merely less obvious.

Recently, Huang (2019), Dai et al. (2022), Liu et al. (2022), and Du et al. (2023) developed large-scale hypothesis testing procedures for mediated effects that account for the conservativeness of standard tests. The latter compared these methods as well as traditional methods in a large simulation study. The method proposed by Huang (2019) is based on a distributional assumption about the two coefficients among all distributions being tested that satisfy the null hypothesis. The methods of Dai et al. (2022), Liu et al. (2022), and Du et al. (2023) involve estimating the proportions of three components of the composite null hypothesis: when both coefficients are zero, when one is zero and the other is not, and vice versa. Dai et al. (2022) and Du et al. (2023) then relaxed the threshold of the joint significance test and delta method-based test, respectively, based on these estimated

proportions. Liu et al. (2022) used the estimated proportions to construct a new test statistic based on p-values attained under each of these components of the null hypothesis. Our proposed methods resolve the conservativeness issue across the entire null hypothesis space even for a single test, hence they require neither assumptions about the distributions of the coefficients nor estimation of proportions of components of the null hypothesis space. The latter is important, because in finite samples, there is no single distribution of a test statistic under the sub-case of the null hypothesis in which one parameter is null and the other is not. The distribution of a test statistic in this sub-case can vary from being well-approximated by the distribution when one parameter is zero and the other is infinite to being well-approximated by the distribution when both parameters are zero. Instead, we can simply extend our method to the large-scale hypothesis test setting while adjusting for multiple testing using standard Bonferroni or Benjamini–Hochberg corrections.

van Garderen and van Giersbergen (2024) is the most closely related work to this article. The authors develop a test for a mediated effect based on defining a rejection region in $\mathbb{R}^2$. In their original version (van Garderen and van Giersbergen, 2020), they proved that a similar test does not exist within a certain class of tests. In this article, we show that a similar test does in fact exist in a slightly expanded class of tests, and that it is the unique such test in this class (in the more recent version, the authors now recognize the existence of this test). They focus on developing an almost-similar test, i.e., they approximate a similar test by optimizing a particular objective function involving a penalty for deviations from the nominal type 1 error level. By contrast, we consider two approaches for optimizing decision theoretic criteria that characterize power while preserving type 1 error. The solution to one of these has a closed form and is an exactly similar test. The other is an approximate Bayes risk optimal test inspired by the test of Rosenblum et al. (2014).

The remainder of the article is organized as follows. In Section 2, we formalize the problem and explain the shortcomings of traditional tests. In Sections 3 and 4, we present the

minimax optimal and Bayes risk optimal tests, respectively. In Section 5, we discuss adaptations for large-scale mediation hypothesis testing. In Section 6, we discuss interpretability challenges with the proposed tests, and introduce modifications leading to improved interpretability. In Section 7, we present results from simulation studies. In Section 8, we apply our methodology to two data sets: one to test whether cognition mediates the effect of cognitive remediation therapy on self-esteem in patients with schizophrenia, and another to test many mediation hypotheses of whether a host of DNA methylation CpG sites mediate the effect of smoking status on lung function. In Section 9, we conclude with a discussion.

2. PRELIMINARIES

We begin with the general problem statement, which we will then connect to mediation analysis. Suppose we have n independent, identically distributed (i.i.d.) observations from a distribution $P_{\boldsymbol{\delta}}$ indexed by the parameter $\boldsymbol{\delta} = (\delta_x, \delta_y)^\top \in \mathbb{R}^2$, and we wish to test the null hypothesis H_0 : "$\delta_x \delta_y = 0$" against its alternative H_1 : "$\delta_x \delta_y \neq 0$". Further, suppose we have a uniformly asymptotically normal and unbiased estimator (Robins and Ritov, 1997) $(\hat{\delta}_x, \hat{\delta}_y)$ for (δ_x, δ_y), i.e.,

$$\sup_{\delta_x, \delta_y} \left| \Pr_{\delta_x, \delta_y} \left[\sqrt{n} \boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1/2} \left\{ (\hat{\delta}_x, \hat{\delta}_y)^\top - (\delta_x, \delta_y)^\top \right\} \leq [t_x, t_y]^\top \right] - \Phi_2([t_x, t_y]^\top) \right| \to 0 \qquad (1)$$

as $n \to \infty$ for all $(t_x, t_y)^\top \in \mathbb{R}^2$, where $\boldsymbol{\Sigma}_{\boldsymbol{n}}$ is a diagonal matrix that is a consistent estimator of the asymptotic covariance matrix of $(\hat{\delta}_x, \hat{\delta}_y)^\top$, and Φ_2 is the cumulative distribution function of the bivariate Gaussian distribution with identity covariance. We wish to test H_0 against H_1 with at most α type 1 error for each (δ_x, δ_y) satisfying $\delta_x \delta_y = 0$ and maximizing power elsewhere. Clearly, H_0 is a very particular type of composite null hypothesis, which in the space spanned by δ_x and δ_y consists of the x and y axes.

The composite null hypothesis H_0 arises naturally in mediation analysis under certain modeling assumptions. Suppose we observe n i.i.d. copies of $(\boldsymbol{C}^\top, A, M, Y)^\top$, where A

is the exposure of interest, Y is the outcome of interest, M is a potential mediator, and $\boldsymbol{C}$ is a vector of baseline covariates that is assumed to control for confounding such that the indirect effect is identified. The *natural indirect effect* (NIE) is the mediated effect of A on Y through M. We formally define the NIE and discuss its identification in the Supplementary Materials. If the linear models with main effect terms given by

$$E(M \mid A = a, \boldsymbol{C} = \boldsymbol{c}) = \beta_0 + \beta_1 a + \boldsymbol{\beta_2}^\top \boldsymbol{c}, \tag{2}$$

$$E(Y \mid A = a, M = m, \boldsymbol{C} = \boldsymbol{c}) = \theta_0 + \theta_1 a + \theta_2 m + \boldsymbol{\theta_3}^\top \boldsymbol{c} \tag{3}$$

are correctly-specified, then the identification formula for the natural indirect effect reduces to $\beta_1\theta_2$. The *product method* estimator is the product of estimates of β_1 and θ_2 obtained by fitting the above regression models. Under models (2) and (3) and standard regularity conditions, the two factors of the product method estimator will jointly satisfy the uniform convergence statement in (1). In fact, (1) will hold for a more general class of models, though there are important limitations to this class. For instance, consider the outcome model with exposure-mediator interaction replacing model (3):

$$E(Y \mid A = a, M = m, \boldsymbol{C} = \boldsymbol{c}) = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta_4}^\top \boldsymbol{c}. \tag{4}$$

Under models (2) and (4), the natural indirect effect comparing treatment levels a' and a'' is identified by $(\theta_2 + \theta_3 a')\beta_1(a' - a'')$, which is naturally estimated by $(\hat{\theta}_2 + \hat{\theta}_3 a')\hat{\beta}_1(a' - a'')$, where $\hat{\theta}_2$, $\hat{\theta}_3$, and $\hat{\beta}_1$ are estimated by fitting the linear regression models (2) and (4). Setting $\delta_x = \theta_2 + \theta_3 a'$, $\delta_y = \beta_1(a' - a'')$, $\hat{\delta}_x = \hat{\theta}_2 + \hat{\theta}_3 a'$, and $\hat{\delta}_y = \hat{\beta}_1(a' - a'')$, the joint estimator will also satisfy (1) under standard regularity conditions.

The problem with traditional tests of the composite null H_0 against its alternative H_1 stems from the fact that while $(\hat{\delta}_x, \hat{\delta}_y)$ converges uniformly in $\boldsymbol{\delta}$, certain scalar functions of these estimators, e.g., the product method estimator $\hat{\delta}_x\hat{\delta}_y$, do not. This can be seen by examining the delta method approach proposed by Sobel (1982) (often referred to as the Sobel test). When $\boldsymbol{\delta} \neq (0, 0)^\top$, the delta method yields $\sqrt{n}(\hat{\delta}_x\hat{\delta}_y - \delta_x\delta_y) \rightsquigarrow \mathcal{N}(0, \delta_y^2\sigma_x^2 + \delta_x^2\sigma_y^2)$,

where σ_x^2 and σ_y^2 are the diagonal elements of the limit in probability of $\boldsymbol{\Sigma}_n$. Thus, when precisely one of δ_x or δ_y is zero, $Z_n^{\text{prod}} := n^{1/2}\hat{\delta}_x\hat{\delta}_y/(\hat{\delta}_y^2 s_x^2 + \hat{\delta}_x^2 s_y^2)^{1/2} \rightsquigarrow \mathcal{N}(0,1)$, where s_x^2 and s_y^2 are consistent estimates of σ_x^2 and σ_y^2, respectively. However, as $\boldsymbol{\delta} \to (0,0)^\top$, this no longer holds, as the denominator also goes to zero. Liu et al. (2022) showed that this instead converges to a centered normal distribution with variance 1/4. Thus, the convergence in distribution of the delta method test statistic Z_n^{prod} is not uniform, and for any given sample size, there will be a region in the parameter space around $(0,0)$ where the standard normal approximation is poor. This is especially problematic, as the region around $(0,0)$ is precisely where we expect the truth to often lie. As $\boldsymbol{\delta}$ approaches $(0,0)$, Z_n^{prod} begins to more closely resemble $N(0,1/4)$, as we show in a Monte Carlo sampling approximation of density functions in Figure S1.a in the Supplementary Materials, where we fix $\delta_y = 0$ and vary δ_x from 0 to 0.3. The delta method uses an approximation based on the standard normal distribution, when in reality, the test statistic's true distribution is much more concentrated about zero. Hypothesis tests based on inverting bootstrap confidence intervals are also known to perform poorly near the origin of the parameter space due to the singularity in the pointwise asymptotic distribution at the origin.

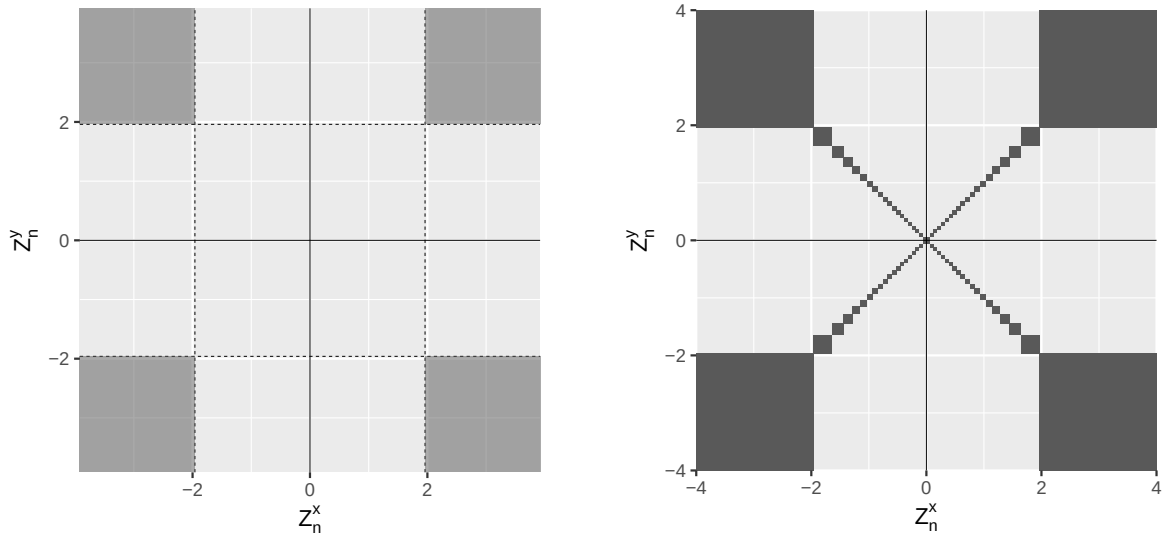
Another popular test of the composite null H_0 against its alternative H_1 is known as the joint significance test (Cohen et al., 2013), which is based on the logic that both δ_x and δ_y must be nonzero for $\delta_x\delta_y$ to be nonzero. The joint significance test with nominal level α amounts to testing the two null hypotheses H_0^x : "$\delta_x = 0$" and H_0^y : "$\delta_y = 0$" against their alternatives H_1^x : "$\delta_x \neq 0$" and H_1^y : "$\delta_y \neq 0$" separately using the Wald statistics $Z_n^x := \sqrt{n}\hat{\delta}_x/s_x$ and $Z_n^y := \sqrt{n}\hat{\delta}_y/s_y$, and rejecting H_0 for H_1 only if H_0^x and H_0^y are both rejected for their alternatives with nominal level α. As noted, this test has been shown to perform better in simulations than other common tests of H_0 against H_1. However, it still suffers from the lack of power near the origin of the parameter space we see in other tests. This is again due to a lack of uniform convergence near the origin. The joint significance

test implicitly relies on the approximation $Z_n^{\min} := |Z_n^x| \wedge |Z_n^y| \sim |\mathcal{N}|(0,1)$, where $|\mathcal{N}|(0,1)$ represents the folded standard normal distribution. This approximation is poor when δ_x and δ_y are both close to zero, as is shown in Figure S1.b in the Supplementary Materials.

To avoid the problems that arise due to non-uniform convergence, we instead construct a rejection region in the parameter space spanned by the vector (δ_x, δ_y) in $\mathbb{R}^2$, rather than that spanned by $\delta_x\delta_y$ in $\mathbb{R}$. In the former, convergence is uniform and there are no discontinuities in the asymptotic distribution. Studying the rejection regions of the traditional tests in $\mathbb{R}^2$ illustrates their conservative behavior from another perspective. Consider a sequence of local noncentrality parameters with respect to a given direction $(\delta_x, \delta_y)^\top$, $(\delta_{x,n}, \delta_{y,n})^\top := (\delta_x/n^{1/2}, \delta_y/n^{1/2})^\top$, and define $\boldsymbol{\delta}^* = (\delta_x^*, \delta_y^*)^\top := \text{plim}_{n\to\infty} n^{1/2}\boldsymbol{\Sigma_n}^{-1/2}(\delta_{x,n}, \delta_{y,n})^\top = \boldsymbol{\Sigma}^{-1/2}(\delta_x, \delta_y)^\top$ where plim is the limit in probability and $\boldsymbol{\Sigma} := \text{plim}_{n\to\infty}\boldsymbol{\Sigma}_n$. Define the bivariate test statistic $(Z_n^x, Z_n^y)^\top := n^{1/2}\boldsymbol{\Sigma_n}^{-1/2}(\hat{\delta}_x, \hat{\delta}_y)^\top$ so that $(Z_n^x, Z_n^y)^\top$ is approximately distributed as $\mathcal{N}\{(\delta_x^*, \delta_y^*)^\top, \boldsymbol{I_2}\}$, and let $(Z_*^x, Z_*^y)^\top \sim \mathcal{N}\{(\delta_x^*, \delta_y^*)^\top, \boldsymbol{I_2}\}$. Based on this asymptotic approximation, the probability of the Wald test rejecting H_0^x for H_1^x at nominal level α is α under all elements of $\{\boldsymbol{\delta}^* : \delta_x^* = 0\}$, and likewise for the Wald test rejecting H_0^y for H_1^y for all $\{\boldsymbol{\delta}^* : \delta_y^* = 0\}$. The rejection region in $\mathbb{R}^2$ corresponding to the joint significance test is the intersection of these two rejection regions, as shown in the plot in Figure 1(a). Because we take the intersection of rejection regions of two size-α tests, the resulting test is necessarily conservative. In fact, when $\boldsymbol{\delta}^* = (0,0)^\top$, the probability of being in the joint significance test rejection region is α^2. By continuity of the rejection probability in $\boldsymbol{\delta}^*$, the rejection probability is likewise well below α in a region around the origin, hence the underpoweredness of the joint significance test. That is, the worst-case type 2 error of the joint significance test is

$$\sup_{\boldsymbol{\delta}^*:\delta_x^*\delta_y^*\neq 0} \Pr_{\boldsymbol{\delta}^*}\{(Z_*^x, Z_*^y) \notin R\} = 1 - \min_{\boldsymbol{\delta}^*:\delta_x^*\delta_y^*=0} \Pr_{\boldsymbol{\delta}^*}\{(Z_*^x, Z_*^y) \in R\} = 1 - \alpha^2.$$

Huang (2019) and van Garderen and van Giersbergen (2024) showed that the delta method-based test's rejection region is contained in that of the joint significance test, and therefore



(a) (b)

Figure 1: (a) The rejection region of the joint significance test is shown in darker gray. Dashed lines indicate the boundaries of the rejection regions of the Wald tests for H_0^x and H_0^y. (b) The rejection region R_{mm} for $\alpha = 0.05$ is shown in dark gray.

the former is uniformly more conservative than the latter. In the following two sections, we define new tests by constructing a rejection region that preserves type 1 error uniformly over the null hypothesis space, but improves power in the alternative hypothesis space, particularly near the origin of the parameter space.

3. MINIMAX OPTIMAL TEST

In this section, we seek a test defined by a rejection region $R^* \subset \mathbb{R}^2$ satisfying $\sup_{\boldsymbol{\delta}^*:\delta_x^*\delta_y^*=0} \Pr_{\boldsymbol{\delta}^*}\{(Z_*^x, Z_*^y) \in R^*\} = \alpha$ and the minimax optimality criterion

$$\inf_{\boldsymbol{\delta}^*:\delta_x^*\delta_y^*\neq 0} \Pr_{\boldsymbol{\delta}^*}\{(Z_*^x, Z_*^y) \in R^*\} = \sup_{R\subset\mathbb{R}^2} \inf_{\boldsymbol{\delta}^*:\delta_x^*\delta_y^*\neq 0} \Pr_{\boldsymbol{\delta}^*}\{(Z_*^x, Z_*^y) \in R\}. \tag{5}$$

The rejection region R^* generates a test with type 1 error α achieving the minimax risk of the 0-1 loss function, i.e., yielding the largest worst-case power. By continuity of the power function in $\boldsymbol{\delta}^*$, the minimax optimal power is upper bounded by α. Moreover, if (5) is attainable, a minimax optimal test generated by the rejection region R must satisfy

$$\Pr_{\boldsymbol{\delta}^*}\{(Z_*^x, Z_*^y) \in R\} = \alpha \quad \text{for all } \boldsymbol{\delta}^* \in \mathbb{R}^2 \text{ such that } \delta_x^* \delta_y^* = 0. \tag{6}$$

A test satisfying (6) is known as a *similar test*. Similarity is an important property that is closely tied to uniformly most powerful unbiased tests in the classical hypothesis testing literature. In our case, it is important because a non-similar test must have type 1 error either greater than or less than α somewhere in the null hypothesis space. In the former case, the test will fail to preserve type 1 error. In the latter, the rejection probability will necessarily be less than α in some neighborhood that also contains the alternative hypothesis space. Thus, a non-similar test will necessarily be a biased test and underpowered in a region of the alternative hypothesis space.

3.1. A Minimax Optimal and Similar Test for Unit Fraction α

We now define a rejection region that generates a similar test for any unit fraction α, i.e., any α such that α^{-1} is a positive integer, hence a test that satisfies the minimax optimality criterion (5). Let $a_k = \Phi^{-1}(k\alpha/2)$ for $k = 0, \ldots, 2/\alpha$, where Φ is the standard normal distribution function. Define $R_{mm} := \{\bigcup_{k=1}^{2/\alpha}(a_{k-1}, a_k) \times (a_{k-1}, a_k)\} \cup \{\bigcup_{k=1}^{2/\alpha}(a_{k-1}, a_k) \times (-a_k, -a_{k-1})\}$. The region R_{mm} for $\alpha = 0.05$ is depicted in the plot in Figure 1(b). Clearly, R_{mm} contains the rejection region of the joint significance test, hence it yields a test that is uniformly more powerful.

Theorem 1. *The rejection region R_{mm} satisfies (6) and generates a similar test.*

The essence of the proof of Theorem 1 is that when $\delta_y^* = 0$ (resp. $\delta_x^* = 0$), whatever level Z_*^x (resp. Z_*^y) takes, the conditional probability of Z_*^y (resp. Z_*^x) being in R_{mm} given

Z_*^x (resp. Z_*^y) equals α, since the probability of being in each vertical interval is $\alpha/2$ under the standard normal law.

Corollary 1. *The rejection region R_{mm} satisfies the minimax optimality criterion (5) with $\inf_{\boldsymbol{\delta}^*:\delta_x^*\delta_y^*\neq 0} \Pr_{\boldsymbol{\delta}^*}\{(Z_*^x, Z_*^y) \in R_{mm}\} = \alpha$.*

van Garderen and van Giersbergen (2020) gave a result stating that there is no similar test of H_0 against H_1 within a particular class of tests defined by a rejection region generated by rotating and reflecting a monotonically increasing function from zero to infinity. While Corollary 1 may appear to contradict this result, R_{mm} does not in fact belong to the class considered by van Garderen and van Giersbergen (2020), but rather belongs to a class that relaxes the restriction that the function must be monotonically increasing to allow for it to be monotonically nondecreasing. In fact, the following theorem states that the test generated by R_{mm} is the only similar test in this relaxed class up to a set of measure zero.

Theorem 2. *Let $\mathcal{F}$ be the class of all monotonically nondecreasing functions mapping all x from the nonnegative real numbers to $[0, x]$, and for any $f \in \mathcal{F}$, let R_f be the region generated by taking all possible negations and permutations of the region $\{(x, y) : y \in (f(x), x]\}$, or equivalently reflecting this region about the x, y, $y = x$, and $y = -x$ axes, and rotating it $\pi/2$, π, and $3\pi/2$ radians about the origin. The function $x \mapsto f_{mm}(x) := \sum_{k=1/\alpha}^{2/\alpha} a_{k-1} I(a_{k-1} \leq x < a_k)$, which generates R_{mm}, is the unique function in $\mathcal{F}$ (up to a set of measure zero) that generates a similar test.*

This test has a simple, exact closed form, making it very fast and straightforward to implement. Additionally, it is nonrandom and symmetric with respect to negation and permutations. One less desirable property is that the test will reject for (Z_n^x, Z_n^y) arbitrarily close to $(0, 0)$, which we discuss further in Section 6. However, the following result shows this is necessary in order to attain the best worst-case power.

Theorem 3. *There is no similar test of H_0 against H_1 with a rejection region that is bounded away from $\{\boldsymbol{\delta}^* : \delta_x^* \delta_y^* = 0\}$.*

Thus, one cannot have a hypothesis test of H_0 that is both similar/unbiased and that fails to reject for all values of (Z_n^x, Z_n^y) that are arbitrarily close to the null hypothesis space.

3.2. Dealing With Non-Unit Fraction Values of α

The rejection region R_{mm} is only defined above for unit fractions α. For other values of α, the class of tests generated by $\mathcal{F}$ will not contain a similar test, though nonrandom similar tests may exist outside of this class. Tests at other values of α are typically of interest for two reasons: using multiple testing adjustment procedures and defining p-values. We discuss the latter in the Supplementary Materials. The restriction that α must be a unit fraction is not a limitation for the Bonferroni correction procedure, which involves dividing the familywise error rate α by the number of tests, provided the familywise error rate α itself is a unit fraction. However, it is a limitation for the Benjamini–Hochberg procedure, which involves multiplying the false discovery rate α by a sequence of rational numbers, only some of which are unit fractions. We discuss the latter procedure in Section 5.

We now define an extension of the minimax optimal test to non-unit fraction values of $\alpha \in (0, 1)$. When α is not a unit fraction, then this extension will not yield a minimax optimal test or a similar test, as similar tests do not exist in this class of tests for such values. However, they do preserve type 1 error, and have type 1 error that approximates α under the least-favorable distribution in the null hypothesis very closely as α goes to zero. Let $b_{-1} := 0$ and $b_k := \Phi^{-1}\{1 - (\lfloor \alpha^{-1} \rfloor - k)\alpha/2\}$ for $k = 0, \ldots, \lfloor \alpha^{-1} \rfloor$ so that $0 = b_{-1} \leq b_0 < b_1 < \ldots < b_{\lfloor \alpha^{-1} \rfloor} = +\infty$. We define the rejection region for any $\alpha \in (0, 1)$ to be $R'_{mm} := \{\bigcup_{k=0}^{\lfloor \alpha^{-1} \rfloor}(b_{k-1}, b_k) \times (b_{k-1}, b_k)\} \cup \{\bigcup_{k=0}^{\lfloor \alpha^{-1} \rfloor}(b_{k-1}, b_k) \times (-b_k, -b_{k-1})\} \cup \{\bigcup_{k=0}^{\lfloor \alpha^{-1} \rfloor}(-b_k, -b_{k-1}) \times (b_{k-1}, b_k)\} \cup \{\bigcup_{k=0}^{\lfloor \alpha^{-1} \rfloor}(-b_k, -b_{k-1}) \times (-b_k, -b_{k-1})\}$. When α is a unit fraction, R'_{mm} coincides with R_{mm}, hence R'_{mm} generates the minimax optimal test

discussed earlier. Otherwise, the test generated by R'_{mm} corresponds closely to the minimax optimal test in that if $\delta_y^* = 0$, then given $Z_*^x = z$ with $z \notin [-b_0, b_0]$, the probability of being in the rejection region is exactly α. Precisely:

Theorem 4. *The rejection region R'_{mm} is such that*

$$\begin{aligned} \lfloor \alpha^{-1} \rfloor \alpha^2 + (1 - \lfloor \alpha^{-1} \rfloor \alpha)^2 &= \inf_{\boldsymbol{\delta}^* : \delta_x^* \delta_y^* = 0} \Pr_{\boldsymbol{\delta}^*} \{ (Z_*^x, Z_*^y) \in R'_{mm} \} \\ &\leq \sup_{\boldsymbol{\delta}^* : \delta_x^* \delta_y^* = 0} \Pr_{\boldsymbol{\delta}^*} \{ (Z_*^x, Z_*^y) \in R'_{mm} \} = \alpha. \end{aligned} \tag{7}$$

Thus, R'_{mm} generates a size-α test. For all α between any two consecutive unit fractions, say $\alpha \in ((K+1)^{-1}, K^{-1}]$, the difference between the lower bound in (7) and the nominal type 1 error α is maximized at the midpoint, where it equals $\{4K(K+1)\}^{-1} = O(K^{-2})$. Thus, the difference between the nominal and actual worst-case type 1 error is maximized at $\alpha = 3/4$, where the actual worst-case type 1 error is $5/8$, $1/8$ less than the nominal type 1 error. However, as noted, this maximal difference shrinks at a quadratic rate in $\lfloor \alpha^{-1} \rfloor$, and is at most $1/1680 \approx 0.0006$ for all $\alpha < 0.05$.

3.3. Asymptotic approximation considerations

The results we have given so far have been in terms of the limit distribution, i.e., that of (Z_*^x, Z_*^y). However, if the joint convergence of $(\hat{\delta}_x, \hat{\delta}_y)$ is uniform in $\boldsymbol{\delta}$ as in (1), then the power $\beta_n(\delta_n) := \Pr_{\hat{\delta}_x, \hat{\delta}_y} \{ (Z_n^x, Z_n^y) \in R_{mm} \}$ for the observed data distribution will converge uniformly to $\beta_*(\delta^*) := \Pr_{\boldsymbol{\delta}^*} \{ (Z_*^x, Z_*^y) \in R_{mm} \}$, i.e., that of the limiting distribution. Thus, for any arbitrarily small tolerance $\epsilon > 0$, there is an N such that $\beta_n(\delta_n)$ will be within ϵ of $\beta_*(\delta^*)$ for all $\boldsymbol{\delta}$ and $n \geq N$. In particular, this means violations of type 1 error can be made arbitrarily small uniformly over H_0 given a sufficiently large sample size. The asymptotic approximation can be improved for a given finite sample n by replacing all instances of the normal distribution with a t distribution with $(n-1)$ degrees of freedom.

4. BAYES RISK OPTIMAL TEST

We now consider a Bayes risk optimality criterion. We draw inspiration from and modify the testing procedure of Rosenblum et al. (2014), which gives approximately Bayes risk optimal tests for the distinct purpose of simultaneously testing for treatment effects both in the overall population and in two subpopulations. Their approach was to discretize the test statistic space into a fine grid of cells, then cast the problem as a constrained optimization problem, where the unknown parameters are the rejection probabilities in each cell, resulting in a rejection region that optimizes their particular Bayes risk.

We let $M : \mathbb{R}^2 \times [0,1] \to \{0,1\}$ denote a generic testing function of H_0 against H_1 that characterizes the randomized test consisting of rejecting H_0 for H_1 if and only if $M(Z_*^x, Z_*^y, U) = 1$, where U is a uniform random variable on $[0,1]$. As in Rosenblum et al. (2014), it is purely for computational reasons that we consider a randomized test, as doing so yields a linear programming problem rather than an integer programming problem. However, our solution is almost entirely deterministic, with very few cells having non-degenerate rejection probabilities. Let $L : \{0,1\} \times \mathbb{R}^2 \to \mathbb{R}$ be a bounded loss function, and Λ denote a prior distribution on $\boldsymbol{\delta}^*$. We consider a 0-1 loss function given by

$$L\{M(Z_*^x, Z_*^y, U), \boldsymbol{\delta}^*\} = M(Z_*^x, Z_*^y, U) I(\delta_x^* \delta_y^* = 0) + \{1 - M(Z_*^x, Z_*^y, U)\} I(\delta_x^* \delta_y^* \neq 0),$$

and a bounded quadratic loss function given by

$$L\{M(Z_*^x, Z_*^y, U), \boldsymbol{\delta}^*\} = \{1 - M(Z_*^x, Z_*^y, U)\}\{(\delta_x^* \delta_y^*)^2 \wedge 4\Phi^{-1}(1-\alpha/2)^2\}.$$

We consider $\mathcal{N}\{0, 2\Phi^{-1}(1-\alpha/2)\boldsymbol{I_2}\}$ for the prior Λ. The constrained Bayes optimization problem is the following: For given $\alpha \in (0,1)$, L, and Λ, find the function M minimizing

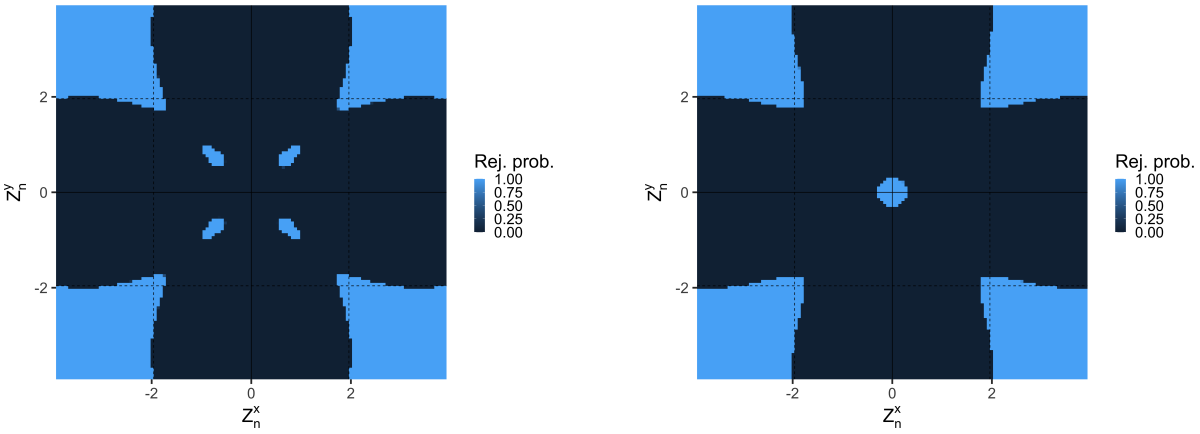
$$\int E_{\boldsymbol{\delta}^*}\left[L\left\{M(Z_*^x, Z_*^y, U), \boldsymbol{\delta}^*\right\}\right] d\Lambda(\boldsymbol{\delta}^*) \tag{8}$$

subject to the type 1 error constraint

$$\Pr_{\boldsymbol{\delta}^*}\{M(Z_*^x, Z_*^y, U) = 1\} \leq \alpha \text{ for all } \boldsymbol{\delta}^* \in \mathbb{R}^2 \text{ such that } \delta_x^* \delta_y^* = 0. \tag{9}$$

To formulate an approximation of this problem, we begin by making the test coincide outside of $B := [-b, b] \times [-b, b]$ with the joint significance test, where we select $b = 2\Phi^{-1}(1 - \alpha/2)$. This is because the event $(Z_*^x, Z_*^y) \notin B$ is only likely under $\boldsymbol{\delta}^*$ far enough away from the origin and very unlikely otherwise, so that the joint significance test will only be negligibly conservative. We then discretize the null hypothesis space into a grid of $8K + 1$ points and B into a $(2K + 1) \times (2K + 1)$ grid consisting of squares $R_{k,k'}$, where K is a suitably large integer. Within B, we define a class of random rejection regions whose realizations are constant within each square. That is, if $(z^x, z^y) \in R_{k,k'}$, then $M(z^x, z^y, u) = M(z^{x'}, z^{y'}, u)$ for all $(z^{x'}, z^{y'}) \in R_{k,k'}$ and all $u \in [0, 1]$. The rejection probabilities corresponding to each square are the unknown arguments defining the rejection region that are to be optimized over. A discretized version of the constrained optimization problem is then defined corresponding to these grids. Collectively, these form a linear optimization problem that approximates the Bayes risk optimization problem. Full details are available in the Supplementary Materials.

The plots in Figure 2 show the solutions to the two linear approximation problems with $\alpha = 0.05$ and $K = 64$. For both tests, the rejection probabilities are nearly all zero or one. If one prefers a nonrandom test, one can obtain a slightly more conservative test by setting all nondegenerate probabilities to zero. Both rejection regions almost contain the entire rejection region of the joint significance test. Moreover, both rejection regions take interesting, somewhat unexpected shapes, with four roughly-symmetric disconnected regions inside of the acceptance region in the case of the 0-1 loss, and a single region inside the acceptance region centered around (0,0) in the case of the quadratic loss. As with the minimax optimal test, it seems that it is important to have part of the rejection region along the diagonals at least somewhat close to the origin, so that smaller alternatives in



(a) (b)

Figure 2: (a) The rejection probabilities of the approximate Bayes risk optimal test with respect to the (a) 0-1 loss function, and (b) bounded quadratic loss function. The rejection region for each is defined to be equivalent to that of the joint significance test outside of the displayed figure, which is outlined by the continuation of the dashed lines. The dashed lines indicate the boundaries of the rejection regions of the Wald tests for H_0^x and H_0^y.

both δ_x^* and δ_y^* have larger rejection probabilities to offset the conservativeness induced by lying in the intersection of two acceptance region bands. As with the minimax optimal test, one might reasonably take issue with rejecting test statistics that are arbitrarily close to the null hypothesis space. We discuss this further in Section 6.

While solving the linear programming problem to produce these tests takes a nontrivial amount of time to run, this need only be done once, after which the object is stored and can be quickly loaded on demand. The approximate Bayes risk optimal tests can be solved for any size $\alpha \in (0, 1)$. However, the linear optimization problem must be solved separately for each value of α, and so cannot be obtained immediately.

As with the minimax optimal test, the approximate optimality of the tests defined in this section is in terms of the limit distribution, i.e., of (Z_*^x, Z_*^y). Assuming uniform convergence

of $(\hat{\delta}_x, \hat{\delta}_y)$ as in (1), the (sample) approximations to the (computational) approximate Bayes risk optimality objective functions and type 1 error constraints can be made arbitrarily small uniformly over the parameter space given a sufficiently large sample size. As with the minimax optimal test, the asymptotic approximation can be improved by replacing all instances of the normal distribution with the t distribution with $(n-1)$ degrees of freedom.

5. LARGE-SCALE HYPOTHESIS TESTING

As mentioned above, applying the Bonferroni correction to the proposed methods in order to control for familywise error rate is straightforward even without the use of p-values. For J hypotheses, one can simply run any of the proposed tests with level α/J instead of comparing the p-value to a threshold of α/J. Adapting the Benjamini–Hochberg procedure to control false discovery rate requires a bit more explanation if p-values are not available (though the p-value proposed in the Supplementary Materials can be used, which will be conservative). A modified Benjamini–Hochberg procedure works as follows: (i) find the largest value j such that at least j of the J tests reject at level $\alpha j/J$, (ii) reject all hypotheses that the test rejects at level $\alpha j/J$. It can be shown that this procedure controls the false discovery rate at α even for tests that are not monotone in α. A (potentially) conservative version of this procedure can avoid an exhaustive search for the largest j in step (i). This procedure starts with $j = 1$ and increases j until fewer than $\alpha j/J$ hypotheses are rejected for κ consecutive iterations for some pre-specified κ, after which the largest j in this sequence at which at least $\alpha j/J$ hypotheses are rejected is chosen for part (i).

When testing for the presence of multiple mediated effects, one needs to assume that the true mediators do not affect one another, which would result in some mediators being exposure-induced confounders, yielding the NIE non-identifiable. This may not be realistic in many scenarios; however, this problem is beyond the scope of this article.

6. INTERPRETATION AND MODIFICATIONS

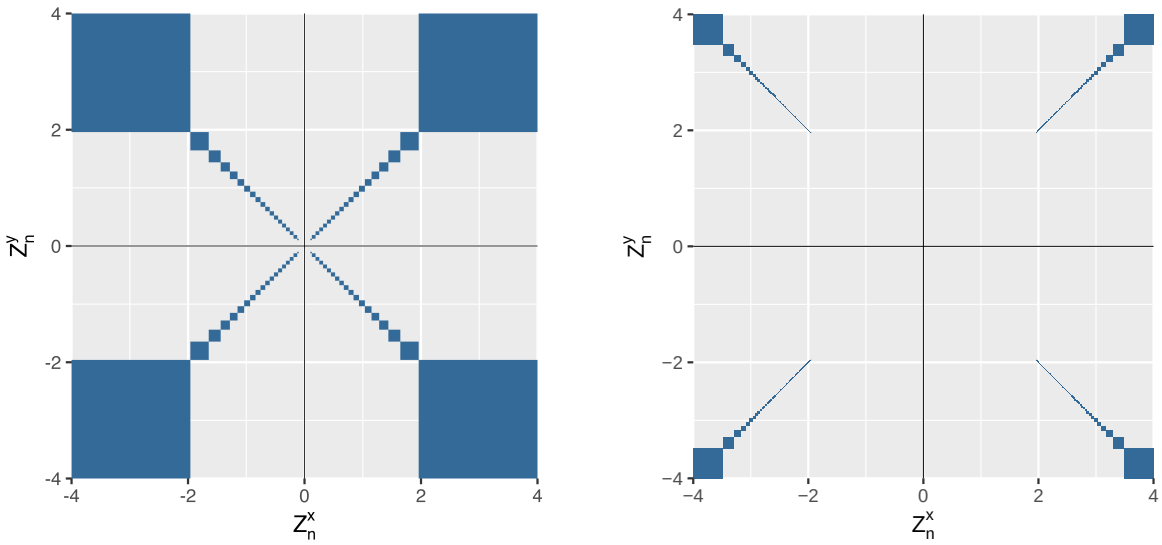
As we have pointed out, the rejection regions for the minimax and Bayes risk optimal tests have some peculiar features that some may regard as undesirable. Tests with such properties are known to arise in multi-parameter hypothesis testing and have received criticism in Perlman and Wu (1999) as being flawed, though the philosophical debate about what criteria should be used to compare among statistical tests is perhaps not entirely settled (Berger, 1999). We take a pluralistic stance and provide a menu of options, among which we hope anyone across this ideological spectrum can find a test that suits their preferences. Furthermore, we discuss differences in interpretation of different tests.

At least two of the criticisms of Perlman and Wu (1999) apply to our tests. First, the minimax optimal test can reject when the point estimate is arbitrarily close to the null. How then is one to interpret a test that rejects when both test statistics are small? One perspective is that one need not rely entirely on the result of a hypothesis test to determine whether an association or an effect is "practically significant", so long as the hypothesis test has correct type 1 error. One could always pre-specify a minimally meaningful effect size or proportion mediated, and never reject if the point estimate does not exceed it.

Second, the minimax optimal test is not monotonic in α, nor is the Bayes risk optimal test guaranteed to be. Thus, in the large-scale hypothesis testing case, it is possible for the Bonferroni- or BH-adjusted minimax optimal test to reject based on a particular test statistic when there are, say, J hypotheses to test, but not when there are $J-1$. While this is true for a fixed test statistic, the expected number of rejected nulls will always be lower with a smaller number of hypotheses, and FWER or FDR are preserved for any number of hypotheses. In the context of large-scale hypothesis testing, one might care less about the interpretation of the rejection of a particular hypothesis, and more about minimizing false positives, as rejected nulls may be viewed as being flagged for further scrutiny in subsequent

analyses. However, these are not the only options. In the Supplementary Materials, we introduce a p-value corresponding to the minimax optimal test, which can itself be used to define a test that rejects at level α whenever the p-value is less than α. Such a test will be monotone in α by construction. A constrained variant of the Bayes risk optimal test can be devised for a given decreasing sequence of α's by setting the rejection value to be no less than the largest rejection probability among all tests with larger α values preceding it in the sequence. Therefore, if one is concerned about coherence, then one might opt to use this p-value based test at the cost of power, but which is still more powerful than traditional tests while preserving FWER or FDR, or alternatively this constrained version of the Bayes risk optimal test.

To circumvent the first issue, we propose modifications of the minimax optimal and Bayes risk optimal tests. For the minimax optimal test, one can simply truncate the rejection region according to any user-specified distance, say d, from the null hypothesis space. If one wishes to preserve monotonicity in α, then one can do the same for the p-value-based test we have just discussed. For a test with rejection region R, the rejection region of the corresponding truncated version will simply be $R \setminus \{(Z_n^x, Z_n^y)^\top : |Z_n^x| \wedge |Z_n^y| \leq d\}$. The rejection region of the truncated minimax optimal test with $d = 0.1$ is shown in panel (a) of Figure 3, where we chose 0.1 for the sake of comparison with van Garderen and van Giersbergen (2024). Such truncated tests will clearly be dominated by their corresponding non-truncated versions in terms of power. One rebuttal to the claim of Perlman and Wu (1999) that the rejection region ought to be bounded away from the null hypothesis space is that this distance is not specified, nor is there a clearly-defined criterion specifying what this distance should be (Berger, 1999). If one believes that this distance should be determined according to the likelihood ratio test, then one would set d to be $\Phi^{-1}(1-\alpha/2)$, in which case the joint significance test (i.e., the likelihood ratio test) would be recovered in the single hypothesis test case. However, if one is conducting large-scale hypothesis



(a) (b)

Figure 3: (a) The rejection region of the truncated minimax optimal test with $d = 0.1$ and $\alpha = 0.05$. (b) The rejection region of the truncated minimax optimal test with $d = \Phi^{-1}(0.975)$ and $\alpha = 0.05/100$, as one might use for a Bonferroni correction when testing 100 mediation hypotheses.

testing and wishes to control for familywise error rate (for instance), and insists on a d of $\Phi^{-1}(1 - \alpha/2)$, then this does generate a novel test with greatly-improved power over the Bonferroni-corrected joint significance test. For example, for a Bonferroni correction for 100 tests at level $\alpha = 0.05$, one can use the truncated minimax optimal test with α level $0.05/100$ and with $d = \Phi^{-1}(1 - \alpha/2)$, which yields the rejection region in panel (b) of Figure 3. That is, one can apply the multiple testing-adjusted minimax optimal test with level 0.0005, and then use the joint significance test with α level 0.05 to screen out effects deemed too small. The resulting test will have 100 times the worst-case power per test of that of the Bonferroni-corrected joint significance test.

For the Bayes risk optimal tests, the same truncation could be easily applied. However,

we also formulate a constrained version of the Bayes risk optimization problem in this article. For $d = 0.1$, the Bayes risk optimal test with 0-1 loss function will not change, since there is no part of the rejection region within 0.1 of the null hypothesis space. On the other hand, the Bayes risk optimal test with quadratic loss will change, since its rejection region contains a region centered on the origin (see the supplementary materials).

Another important challenge has to do with the traditional interpretation of a hypothesis test rejection decision: if the null hypothesis were true, then the probability of observing an event as or more extreme than the observed event is at most α. In the product of coefficients case, such an interpretation can be applied to a partial ordering, where one pair of test statistics is "as or more extreme" than another if both elements in the first pair are at least as large in magnitude as those in the second. van Garderen and van Giersbergen (2022) refer to a test that rejects for all pairs of test statistics that are more extreme than another pair for which it also rejects as "information coherent". Clearly none of the tests we have proposed nor the test of van Garderen and van Giersbergen (2024) satisfy this property, hence these tests lack this traditional interpretation. Instead, they admit a weaker interpretation: if the null hypothesis were true, then the observed event would have been unlikely to occur in the sense that its rejection probability would be at most α. Alternatively, if the joint significance test and one of our proposed tests both reject, then the stronger interpretation can be applied. In a case where they disagree and one is applying one of our tests, then one must be prepared to accept the possibility that one pair of test statistics that are smaller in magnitude may reject while another does not.

van Garderen and van Giersbergen (2022) showed that the joint significance test is the most powerful test in the class of information coherent tests. Thus, there is a fundamental trade-off between power and preserving the former stronger interpretation over the latter weaker interpretation. If one deems the underpoweredness of the joint significance test to be an important problem, then one has no choice but to sacrifice a degree of interpretability.

In proposing more powerful tests, we wish to present options to gain power at the cost of some interpretability for those who would deem it a worthwhile trade-off. One may feel more comfortable with this trade-off in large-scale hypothesis testing scenarios where the focus is on making multiple discoveries while controlling the number of errors. Thus, our proposed methodology is perhaps best suited for this latter setting.

Lastly, we reiterate that the validity of each of our proposed tests is predicated on satisfaction of identification assumptions. As is typical when comparing tests, the most pronounced improvements in power occur when effect sizes are small relative to standard errors. However, analyses in such cases are also qualitatively more vulnerable to identification failures, such as residual confounding. Thus, it is essential to prioritize rigorous study design, comprehensive covariate measurement, and domain-informed confounder selection before concerning oneself with maximizing power.

7. SIMULATION STUDIES

7.1. Single mediation hypothesis testing

To compare the finite sample performances of our proposed tests with those of the others we have discussed, we performed a simulation study in which we sampled from independent t distributions with noncentrality parameters $\boldsymbol{\delta}$. We considered four scenarios: (a) We varied δ_x and δ_y jointly from 0 to 0.4. (b) We fixed δ_y to 0.2 and varied δ_x from 0 to 0.4. (c) Using a normal approximation, we fixed δ_y to 0 and varied δ_x from 0 to 0.4, thereby exploring empirical type 1 error across a range of parameter values within the composite null hypothesis space. (d) We used the same setting as in (c), but using a t-distribution with $(n-1)$ degrees of freedom approximation instead of the normal approximation. We drew 100,000 Monte Carlo samples with $n = 50$ for each value of δ_x, and applied each test of H_0 : "$\delta_x\delta_y = 0$" against H_1 : "$\delta_x\delta_y \neq 0$" to each sample. In particular, we used

three versions of the Bayes risk optimal test: the tests with either the 0-1 loss or quadratic loss function, and the constrained test with quadratic loss function. Likewise, we used three version of the minimax optimal test: the standard version, the test based on the corresponding p-value defined in the Supplementary Materials, and a truncated minimax optimal test. In both the constrained and truncated tests, we used $d = 0.1$ in order to make a fair comparison with the van Garderen and van Giersbergen (2024) (henceforth vGvG) test. The Monte Carlo power estimates are displayed in the plots in Figure 4. In scenario (a), apart from the minimax optimal p-value-based test, all of the minimax optimal and Bayes risk optimal tests and the vGvG test have very close to nominal type 1 error at $\boldsymbol{\delta} = (0, 0)^\top$. All of these tests greatly outperform the delta method and joint significance tests in terms of power for smaller values of $\delta_x = \delta_y$, with the minimax optimal p-value-based test having power somewhere in between, but converging more quickly to the other tests. The delta method test remains very conservative over the entire range of parameter values. The minimax optimal (apart from the p-value based test), Bayes risk optimal, and vGvG tests perform very similarly over the range of parameter values, with the quadratic loss versions of the Bayes risk optimal test trading off some power loss in the 0.1–0.2 range for a slight improvement in power in the 0.3–0.4 range, demonstrating the greater emphasis the quadratic loss places on larger alternatives. The truncated minimax optimal test begins slightly conservative relative to 0.05, but catches up with the most powerful tests quickly. Despite not being similar tests, the Bayes risk optimal and vGvG tests suffer very little in terms of power near the least favorable distribution at $\boldsymbol{\delta} = (0, 0)^\top$.

In scenario (b), the trends are largely the same as in scenario (a). The main difference is that the quadratic loss versions of the Bayes risk optimal test are slightly conservative under the null, albeit less conservative than the minimax optimal p-value-based test. These tests trade off some power loss for smaller values of δ_x ($\lesssim$0.25) for slightly improved power for larger values of δ_x ($\approx$0.35–0.55). Since δ_y is fixed at 0.2, power for all tests appears

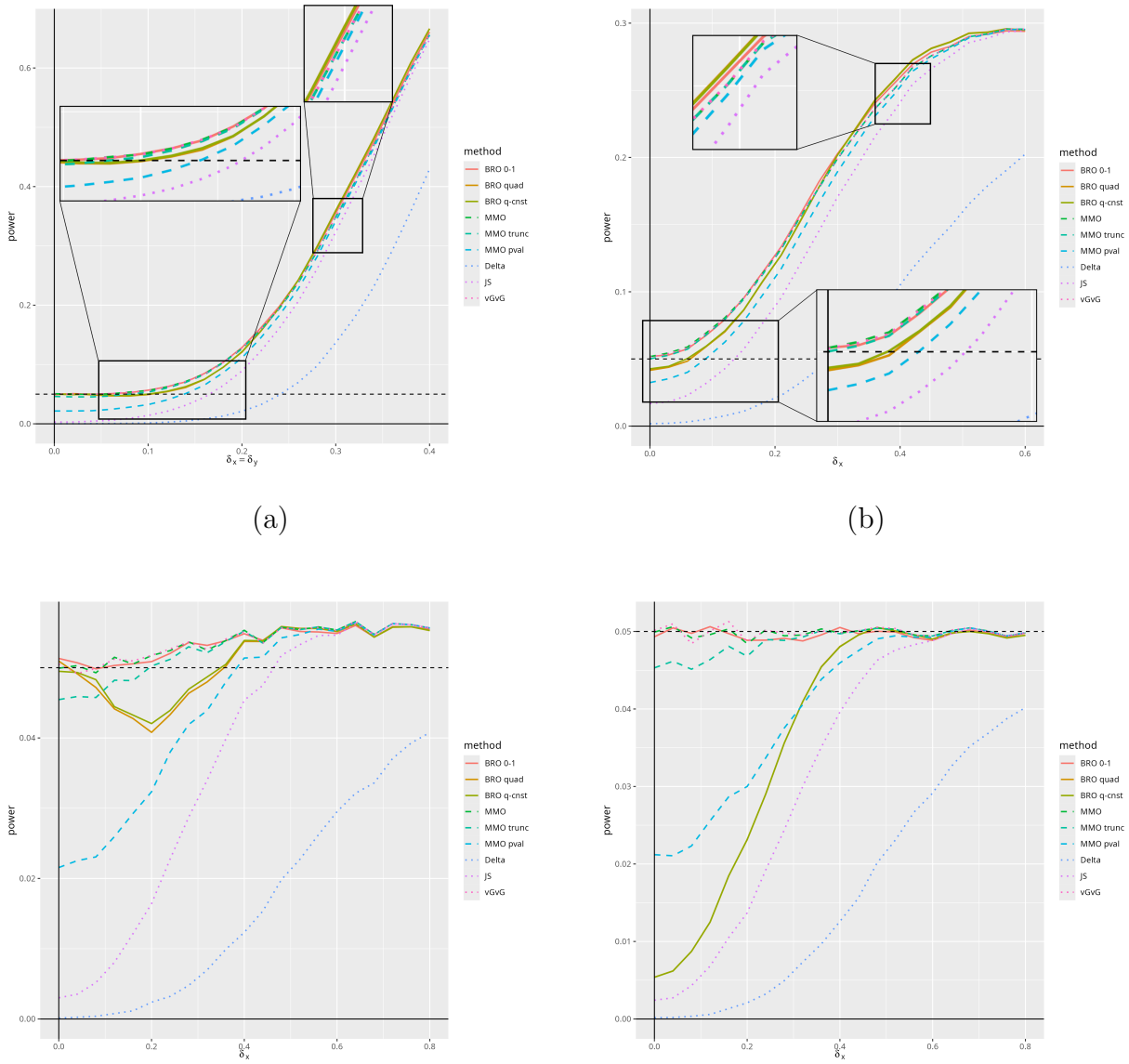


Figure 4: Monte Carlo rejection probabilities of the following: Bayes risk optimal with 0-1 loss (BRO 0-1), Bayes risk optimal with quadratic loss (BRO quad), constrained Bayes risk optimal with quadratic loss (BRO q-cnst), delta method (Delta), joint significance (JS), minimax optimal (MMO), p-value-based test of the minimax optimal test (MMO pval), truncated minimax optimal (MMO trunc), and vGvG. (a) $\delta_x = \delta_y$ varying from 0 to 0.4. (b) δ_y fixed at 0.2, δ_x from 0 to 0.6. (c) δ_y fixed at 0, δ_x from 0 to 0.8 with normality-based tests. (d) δ_y fixed at 0, δ_x from 0 to 0.8 with t-distribution-based tests. In panel d, the BRO quad and BRO q-cnst lines are identical.

to plateau a little under 0.3. In scenarios (c) and (d), we see that all tests approximately preserve type 1 error; however, there is some anti-conservative behavior in scenario (c) due to the normal approximation being a bit too concentrated for a sample size of 50. This is corrected with the use of the t-distribution approximation, which yields type 1 error much closer to 0.05 across the null hypothesis parameter space. In both cases, we see the performances of all tests except the delta method to converge by around $\delta_x = 0.55$.

7.2. Large-scale mediation hypothesis testing

Du et al. (2023) ran a large simulation study comparing a number of methods for performing large-scale hypothesis testing. We reproduce one of their simulation settings and apply our proposed methodology to compare performance with the methods they compared. In particular, we performed 2,000 replications under their setting (a) with $J = 100,000$ mediated effect null hypotheses, $n = 200$ (sample size), $\tau = 0.1$, which controls how dispersed the true parameters δ_x and δ_y are under their respective alternatives, and H_0^x and H_0^y hold for 99.8% of tests, H_0^x and H_a^y hold for 0.1% of tests, and H_a^x and H_0^y hold for 0.1% of tests. See Du et al. (2023) for more details about the simulation set-up. We compare the results of our proposed minimax optimal test with those of the tests compared in Du et al. (2023). The ratios of the actual false positive rates (FPRs) to the true FPRs at different cut-off values are displayed in Table 1 (except for the Sobel/delta method test, which performs worse than the joint significance test). The minimax optimal test yields FPR ratios closer to one and with smaller standard deviations than any other test apart from the joint significance test, which was excessively conservative.

Table 1: Mean (standard deviation) of the ratio of the actual FPR to the nominal FPR using the joint significance, JT-comp (Huang, 2019), HDMT (Dai et al., 2022), Sobel-comp (Du et al., 2023), DACT (Liu et al., 2022), and minimax optimal tests.

Cutoff	Joint sig.	JT-comp	HDMT	Sobel-comp	DACT	MM-opt
1e−3	0.00 (0.01)	1.11 (0.10)	1.02 (0.16)	0.89 (0.34)	1.47 (1.81)	1.00 (0.10)
1e−4	0.00 (0.01)	1.48 (0.37)	1.02 (0.46)	0.81 (0.66)	2.90 (4.19)	1.00 (0.32)
1e−5	0.00 (0.04)	3.02 (1.70)	1.17 (1.30)	1.04 (1.37)	6.54 (10.61)	0.98 (0.99)
1e−6	0.01 (0.32)	9.99 (10.09)	1.81 (4.46)	1.81 (4.63)	16.50 (31.56)	1.01 (3.21)
5e−7	0.00 (0.00)	15.19 (17.36)	2.11 (6.62)	2.32 (7.17)	22.65 (47.46)	0.91 (4.31)

8. DATA EXAMPLES

8.1. DCTRS Data Analysis

We applied our methodology to data from the Database of Cognitive Training and Remediation Studies (DCTRS), which consists of data from several randomized trials testing the efficacy of cognitive remediation therapy (CRT) in patients with schizophrenia. CRT targets patients' cognitive outcomes with the long-term goal of cognitive gains translating into improvements in functioning and quality of life. We used data from three trials described in Wykes et al. (1999), Wykes et al. (2007a), and Wykes et al. (2007b) consisting of 128 patients with complete treatment, mediator, and outcome data. In each study, patients were randomized to a CRT arm or a control arm. There are a number of issues (e.g., potential unmeasured confounding, variable follow-up time, and the fact that the data come from studies with different protocols) that we do not entirely account for in this analysis, and so our results should be viewed as purely illustrative of the proposed methodology.

Our exposure, A, was an indicator of whether the patient was assigned to the treatment arm; our outcome, Y, was the patient's Rosenberg self-esteem scale score at follow-up, and the mediator, M, was the patient's Wechsler Adult Intelligence Scale (WAIS) working memory digit span test at their end of treatment, which measures both working memory and attention. The mean self-esteem scale score was 33.6 with sample range from 17 to 50. Since the studies were all randomized trials, we only need to assume that variables included in $\boldsymbol{C}$ and A are sufficient to control for confounding of the effect of M on Y. Here, we let $\boldsymbol{C}$ consist of the patients' sex, race, education, which study they participated in, and their self-esteem scale and working memory digit span test scores measured at baseline. We provide more detailed discussion about the variable choices and modeling assumptions as well as unadjusted and more comprehensively adjusted analyses in the Supplementary Materials. However, we remind the reader that the assumption of no unmeasured mediator-outcome confounding is inherently untestable, and covariate selection should ideally be based primarily on causal understanding and domain expertise. We prioritize the results in Table 2, as these are based on the most relevant subject-matter based covariate selection rather than data-driven selection, as the latter may also introduce post-selection inference problems. All tests agree in the additional analyses based on the more comprehensive covariate adjustment presented in the Supplementary Materials. The relationship between covariate adjustment and agreement of the various tests is complex. Adjusting for more covariates can influence the underlying test statistics in either direction, and doing so does not consistently make the various tests more or less likely to agree. Previously, Wykes et al. (2007b) found cognitive remediation to improve working memory digit span test scores; Wykes et al. (1999) found cognitive remediation to improve Rosenberg self-esteem scale scores. We are testing whether there is an effect of cognitive remediation on patients' self-esteem that is mediated by its effect on working memory and attention.

We fit models (2) and (4) using OLS, and estimated the NIE by the corresponding

estimator given in Section 2. We then applied the three versions of the minimax optimal test and the three versions of the Bayes risk optimal test from Section 7, and compared these with the joint significance, delta method, and vGvG tests, all at level $\alpha = 0.05$, using the t-distribution approximation. The estimated total effect is 0.70, which is interpreted as CRT modestly improving the self-esteem scale score by an average of 0.70 (0.10 standard deviations). The estimated natural indirect effect is 0.29, which is interpreted as CRT modestly improving self-esteem by an average of 0.29 (0.04 standard deviations) through its effect on working memory and attention as measured by the digit span test. The estimated proportion mediated is 41.0%. The test statistic is plotted on the rejection regions of the minimax and Bayes risk optimal tests in Figures S4–S8 of the Supplementary Materials.

Table 2: Hypothesis test results for the NIE

Hypothesis test	Reject	p-value
Bayes risk optimal, 0-1 loss	No	Undefined
Bayes risk optimal, quadratic loss	No	Undefined
Bayes risk optimal, quadratic loss, constrained	No	Undefined
Minimax optimal	Yes	0.039
Minimax optimal, truncated	Yes	0.039
Minimax optimal, p-value	Yes	0.039
Delta method	No	0.256
Joint significance	No	0.115
van Garderen and van Giersbergen	Yes	Undefined

Hypothesis test results are presented in Table 2, where we observe the improved power of the different versions of the minimax optimal test to reject the null hypothesis that the NIE is zero in favor of the alternative that it is nonzero. The interpretation for any of

these tests is that if the true NIE were in fact zero, then the event we observed would have been unlikely to have occurred in the sense that the test would have rejected with at most 5% probability. Since these tests disagree with the joint significance test, they lack the stronger interpretation that the probability of observing an event as or more extreme would be at most 5% if the true NIE were zero. Having established that what we observed is "unlikely" in the former sense, one may proceed by examining whether the estimated effect size is meaningful. An effect size of 0.04 standard deviations of the self-esteem scale score may not be deemed relevant for understanding the mechanism by which CRT affects self-esteem. However, had the estimate been larger, one might conclude that there was a meaningful indirect effect through working memory, and that such an estimate would have been unlikely to have been observed if there were no true underlying indirect effect. Alternatively, one might consider a proportion mediated exceeding 10%, say, to be meaningful, and would then conclude that the indirect effect through working memory indeed explains a meaningful amount of the total effect of cognitive remediation. However, as the total effect is not statistically significant in this case, the proportion mediated may be a less compelling criterion for evaluating practical significance. In this case, neither of the traditional tests nor any of the Bayes risk optimal tests reject, but the vGvG test does reject, agreeing with the minimax optimal tests. While we have conducted nine tests for purposes of illustration, in practice one should decide *a priori* which test to use based on how they wish to navigate the power–interpretability trade-off to avoid data dredging.

8.2. Normative Aging Study Analysis

Liu et al. (2022) performed large-scale hypothesis testing on the Normative Aging Study (NAS) to test the mediated effect of smoking status on forced expiratory flow at 25%–75% of the Forced Expiratory Vital capacity, a measure of lung function, through 484,613 DNA methylation CpG cites among men in Eastern Massachusetts. We apply our proposed

methodology to this data and compare results. See Liu et al. (2022) for details regarding this data. Based on the minimax optimal test with a Bonferroni correction controlling FWER at 0.05, we detect significant mediated effects through sites cg03636183, cg05575921, cg06126421, and cg21566642. The conservative version of the Benjamini–Hochberg correction controlling FDR at 0.05 additionally detects a mediated effect through the site cg05951221. The Bonferroni procedure ran locally on a MacBook Pro with 2.3 GHz Intel Core i5 in 7.3 seconds; the Benjamini–Hochberg procedure in 18.9 seconds. Liu et al. (2022) note that each of these sites has previously been found to be related to smoking status and/or lung cancer risk. The FDR-adjusted joint significance test also identified all of these, while the FWER-adjusted joint significance test identified all except cg05951221. The FDR-adjusted DACT method (Liu et al., 2022) found all of these to be significant as well as 14 others, the latter of which do not appear to have been found to be linked with either smoking or lung cancer. In our simulation study the DACT was found to have an inflated proportion of false positives—over 22 times that of the nominal proportion when using the appropriate cutoff—whereas our minimax optimal test never identified false positives at a rate any higher than 1.01 times the nominal proportion. These considerations cast doubt on whether the additional 14 CpG sites identified by the DACT method but not by ours are true positives. Thus, the FDR-adjusted minimax optimal test identified more CpG sites than the standard joint significance or delta method tests while also preserving approximate FDR control as shown theoretically and in the simulation study.

9. DISCUSSION

We have proposed two novel classes of tests for the composite null hypothesis H_0 : "$\delta_x\delta_y = 0$" against H_1 : "$\delta_x\delta_y \neq 0$", which arises commonly in the context of mediation analysis. This is a challenging inferential task due to the non-uniform asymptotics of univariate test

statistics of the product of coefficients. We have constructed these tests to be both optimal in a decision theoretic sense, and to preserve type 1 error uniformly over the null hypothesis space—exactly so on both counts in the case of the minimax optimal test, and approximately so in the case of the Bayes risk optimal test. We have described procedures for carrying out large-scale hypothesis testing of many product-of-coefficient hypotheses, controlling for both familywise error rate and false discovery rate. We have also considered some shortcomings of these tests in terms of interpretability, and have developed modifications to remedy some of these. Each of these comes at some cost of power, and we have illustrated a fundamental trade-off between the objectives of power and interpretability. Lastly, we have provided an R package to implement the proposed tests.

It is natural to ask which of the proposed tests should be used in practice. This is highly subjective, and depends on the criteria by which one judges hypothesis tests, as well as how much value one places on power in different parts of the alternative hypothesis space. We summarize the trade-offs with respect to power, scalability for large-scale hypothesis testing, and interpretability of the various tests discussed in this article in Table 3. We have introduced tests that improve on the power of traditional tests, but not all of which have rejection regions bounded away from the null hypothesis space or that are monotonic in α. If one is only concerned with maximizing power subject to the constraint of uniformly preserving type 1 error, then these tests will be desirable. If one is concerned about the possibility of rejecting with small values of the test statistic, then one may employ a truncated or constrained version of these tests. If one finds the nonmonotocity property inadmissible, then one can use the test based on the p-value corresponding to the minimax optimal test. If one is concerned about both, then one can use a truncated version of the p-value based test. The minimax optimal test has the advantage of being a closed form exact solution to its corresponding optimization problem as well as preserving type 1 error exactly in the limit. The minimax optimal test also can be generated almost instantaneously for

Table 3: Summary of properties of various tests

Hypothesis test	Power	Scalability to mult. testing	Monotonicity in α	Information coherence	Bounded away from null
Bayes risk optimal:					
– 0-1 loss	highest	no	can be	no	yes
– quad. loss	highest	no	can be	no	no
– quad. loss, constr.	higher	no	can be	no	yes
Minimax optimal:					
– standard	highest	yes	no	no	no
– truncated	higher	yes	no	no	yes
– p-value	moderate	yes	yes	no	no
Delta method	lowest	yes	yes	yes	yes
Joint significance	lower	yes	yes	yes	yes
vGvG	highest	no	yes	no	yes

all unit fraction values of α, and the generalized test can be generated for other values of α, whereas for the Bayes risk optimal test, the optimization step to generate the rejection region must be performed each time a test for a new value of α is desired. It is easy to obtain p-values for the minimax optimal test. Lastly, the minimax optimal test is a strictly deterministic test, whereas the Bayes risk optimal test contains a few cells in the rejection region where one rejects with non-degenerate probability. If a non-random test is desired, the latter can be converted to a slightly more conservative non-random test.

We close with some practical recommendations for implementation of our proposed methodology. We reiterate that which test to use is subjective and comes down to how the investigator thinks about interpretability and weighs the properties summarized in Table

3. One will often wish to avoid reporting significant findings that are not "practically relevant", in which case we would recommend pre-specifying one's criterion of "practical relevance". This might be in terms of the proportion mediated, say if the objective is to quantify how much of a significant total effect is transmitted through the candidate mediator, or it could be in terms of the indirect effect estimate itself if this is considered to be the more relevant measure to threshold. Alternatively, if one wishes to take into account uncertainty, then one might prefer to threshold based on the test statistics, which would correspond to using the truncated minimax optimal or constrained Bayes risk optimal test.

SOFTWARE

An R package implementing the minimax optimal test is available at https://github.com/achambaz/mediation.test.

ACKNOWLEDGMENTS

This publication was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant Number KL2TR001874. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

DATA AVAILABILITY STATEMENT

We do not have permission from the NIMH to share the DCTRS data, but interested parties can apply to the NIMH Data Archive for access. Summary statistics used for the NAS data analysis are available in the supplementary materials of Liu et al. (2022).

CONFLICT OF INTEREST

The authors report there are no competing interests to declare.

References

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. In *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pages 357–363.

Barfield, R., Shen, J., Just, A. C., Vokonas, P. S., Schwartz, J., Baccarelli, A. A., VanderWeele, T. J., and Lin, X. (2017). Testing for the indirect effect under the null for genome-wide mediation analyses. *Genetic Epidemiology*, 41(8):824–833.

Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Berger, R. (1999). Comment on Perlman and Wu, "The Emperor's New Tests" (with rejoinder by authors). *Statistical Science*, 14(4):370–381.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2013). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Routledge.

Dai, J. Y., Stanford, J. L., and LeBlanc, M. (2022). A multiple-testing procedure for

high-dimensional mediation hypotheses. *Journal of the American Statistical Association*, 117(537):198–213.

Ding, P. and Vanderweele, T. J. (2016). Sharp sensitivity bounds for mediation under unmeasured mediator-outcome confounding. *Biometrika*, 103(2):483–490.

Du, J., Zhou, X., Clark-Boucher, D., Hao, W., Liu, Y., Smith, J. A., and Mukherjee, B. (2023). Methods for large-scale single mediator hypothesis testing: Possible choices and comparisons. *Genetic Epidemiology*, 47(2):167–184.

Huang, Y.-T. (2019). Genome-wide analyses of sparse mediation effects under composite null hypotheses. *The Annals of Applied Statistics*, 13(1):60–84.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309.

Kang, H., Lee, Y., Cai, T. T., and Small, D. S. (2020). Two robust tools for inference about causal effects with invalid instruments. *Biometrics*.

Liu, Z., Shen, J., Barfield, R., Schwartz, J., Baccarelli, A. A., and Lin, X. (2022). Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies. *Journal of the American Statistical Association*, 117(537):67–81.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1):83.

Miles, C. H., Kanki, P., Meloni, S., and Tchetgen Tchetgen, E. J. (2017a). On partial identification of the natural indirect effect. *Journal of Causal Inference*, 5(2).

Miles, C. H., Shpitser, I., Kanki, P., Meloni, S., and Tchetgen Tchetgen, E. J. (2017b). Quantifying an adherence path-specific effect of antiretroviral therapy in the Nigeria PEPFAR program. *Journal of the American Statistical Association*, 112(520):1443–1452.

Miles, C. H., Shpitser, I., Kanki, P., Meloni, S., and Tchetgen Tchetgen, E. J. (2020). On semiparametric estimation of a path-specific effect in the presence of mediator-outcome confounding. *Biometrika*, 107(1):159–172.

Perlman, M. D. and Wu, L. (1999). The emperor's new tests. *Statistical Science*, 14(4):355–369.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319.

Rosenblum, M., Liu, H., and Yen, E.-H. (2014). Optimal tests of treatment effects for the overall population and two subpopulations in randomized trials, using sparse linear programming. *Journal of the American Statistical Association*, 109(507):1216–1228.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, 13:290–312.

van Garderen, K. J. and van Giersbergen, N. (2020). Almost similar tests for mediation effects and other hypotheses with singularities. *arXiv preprint arXiv:2012.11342*.

van Garderen, K. J. and van Giersbergen, N. (2022). On the optimality of the LR test for mediation. *Symmetry*, 14(1):178.

van Garderen, K. J. and van Giersbergen, N. P. (2024). A nearly similar powerful test for mediation. *Review of Economics and Statistics*, pages 1–26.

VanderWeele, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

VanderWeele, T. J. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, 2:457–468.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20(7):557–585.

Wykes, T., Newton, E., Landau, S., Rice, C., Thompson, N., and Frangou, S. (2007a). Cognitive remediation therapy (CRT) for young early onset patients with schizophrenia: An exploratory randomized controlled trial. *Schizophrenia Research*, 94(1-3):221–230.

Wykes, T., Reeder, C., Corner, J., Williams, C., and Everitt, B. (1999). The effects of neurocognitive remediation on executive processing in patients with schizophrenia. *Schizophrenia Bulletin*, 25(2):291–307.

Wykes, T., Reeder, C., Landau, S., Everitt, B., Knapp, M., Patel, A., and Romeo, R. (2007b). Cognitive remediation therapy in schizophrenia: Randomised controlled trial. *The British Journal of Psychiatry*, 190(5):421–427.

Supplemental Materials: Optimal Tests of the Composite Null Hypothesis Arising in Mediation Analysis

A. Mediation analysis details

The composite null hypothesis H_0 arises naturally in mediation analysis under certain modeling assumptions. In order to define the indirect effect of interest, we first introduce notation. First, suppose we observe n i.i.d. copies of $(\boldsymbol{C}^\top, A, M, Y)^\top$, where A is the exposure of interest, Y is the outcome of interest, M is a potential mediator that is temporally intermediate to A and Y, and $\boldsymbol{C}$ is a vector of baseline covariates that we will assume throughout to be sufficient to control for various sorts of confounding needed for the indirect effect to be identified.

We now introduce counterfactuals in order to define the natural indirect effect. Let $Y(a, m)$ be the counterfactual outcome that we would have observed (possibly contrary to fact) had A been set to the level a and M been set to the level m. Similarly, let $M(a)$ be the counterfactual mediator value we would have observed (possibly contrary to fact) had A been set to a. Lastly, define the nested counterfactual $Y\{a', M(a'')\}$ to be the counterfactual outcome we would have observed had A been set to a' and M been set to the counterfactual value it would have taken had A instead been set to a''. The *natural direct effect* (NDE) and *natural indirect effect* (NIE) on the difference scale are then defined to be

$$\begin{aligned}
\text{NDE}(a', a'') &:= E\left[Y\{a', M(a'')\}\right] - E\left[Y\{a'', M(a'')\}\right] \\
\text{NIE}(a', a'') &:= E\left[Y\{a', M(a')\}\right] - E\left[Y\{a', M(a'')\}\right].
\end{aligned}$$

These additively decompose the total effect,

$$\text{TE}(a', a'') := E[Y\{a', M(a')\}] - E[Y\{a'', M(a'')\}] = E\{Y(a')\} - E\{Y(a'')\},$$

where $Y(a)$ is the counterfactual outcome we would have observed (possibly contrary to fact) had A been set to a. Our focus will be on the natural indirect effect.

The natural indirect effect is nonparametrically identified under certain causal assumptions, viz., consistency, positivity, and a number of no unobserved confounding assumptions, which we will not review here for the sake of brevity and because our focus is on statistical inference rather than identification. These assumptions are well documented in the causal inference literature (see VanderWeele (2015) for an overview). The identification formula for the natural indirect effect (also known as the mediation formula) is

$$\int_{\mathbf{c}} \int_{m} E(Y \mid M, A = a', \boldsymbol{C}) \\ \times \left\{ f_{M|A,\boldsymbol{C}}(m \mid A = a', \boldsymbol{C}) - f_{M|A,\boldsymbol{C}}(m \mid A = a'', \boldsymbol{C}) \right\} d\mu(m) f_{\boldsymbol{C}}(c) d\mu(\boldsymbol{c}),$$

where $\mu(m)$ and $\mu(\boldsymbol{c})$ are dominating measures with respect to the distributions of M and $\boldsymbol{C}$.

If the linear models with main effect terms given by

$$E(M \mid A = a, \boldsymbol{C} = \boldsymbol{c}) = \beta_0 + \beta_1 a + \boldsymbol{\beta_2}^\top \boldsymbol{c} \tag{S1}$$

$$E(Y \mid A = a, M = m, \boldsymbol{C} = \boldsymbol{c}) = \theta_0 + \theta_1 a + \theta_2 m + \boldsymbol{\theta_3}^\top \boldsymbol{c}, \tag{S2}$$

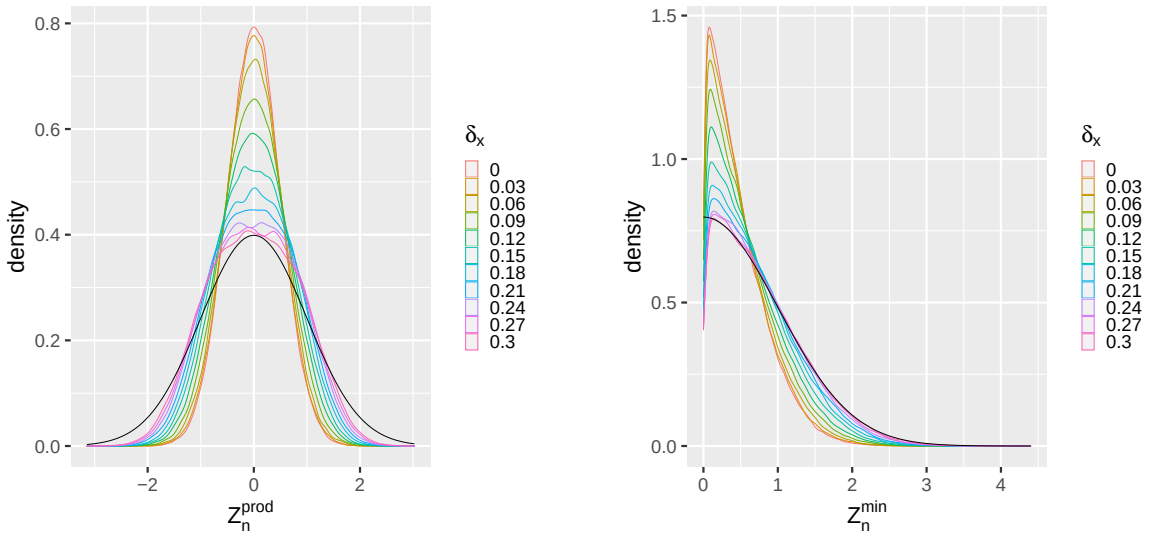
are correctly-specified, then the identification formula for the natural indirect effect reduces to $\beta_1\theta_2$. The *product method* estimator estimates β_1 and θ_2 by fitting the above regression models and takes the product of these coefficient estimates. In fact, following the path analysis literature of Wright (1921), Baron and Kenny (1986) originally defined the indirect effect to be $\beta_1\theta_2$, where $\boldsymbol{C}$ is empty in the above model, rather than in terms of counterfactuals, and proposed the product method estimator. This is an extremely popular estimator

of indirect effects. VanderWeele and Vansteelandt (2009) demonstrated the consistency of the product method for the natural indirect effect under the above linear models. Under models (S1) and (S2) and standard regularity conditions, the two factors of the product method estimator will satisfy the uniform joint convergence statement in (1). In fact, (1) will hold for a more general class of models, though there are important limitations to this class. For instance, consider the outcome model with exposure-mediator interaction replacing model (S2):

$$E(Y \mid A = a, M = m, \boldsymbol{C} = \boldsymbol{c}) = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta_4}^\top \boldsymbol{c}. \tag{S3}$$

Under models (S1) and (S3), the natural indirect effect is identified by $\text{NIE}(a', a'') = (\theta_2 + \theta_3 a')\beta_1(a' - a'')$, hence the estimator $(\hat{\theta}_2 + \hat{\theta}_3 a')\hat{\beta}_1(a' - a'')$, where $\hat{\theta}_2$, $\hat{\theta}_3$, and $\hat{\beta}_1$ are estimated by fitting the linear regression models (S1) and (S3), will also satisfy (1) under standard regularity conditions, with $\delta_x = \theta_2 + \theta_3 a'$ and $\delta_y = \beta_1(a' - a'')$ and with the corresponding plug-in estimators for $\hat{\delta}_x$ and $\hat{\delta}_y$. This is a key extension, as modern causal mediation analysis' ability to account for the presence of exposure-mediator interactions is one of its important advantages over traditional mediation methods such as path analysis and the Baron and Kenny (1986) approach. Unfortunately, the functional form of $E(Y \mid A = a, M = m, \boldsymbol{C} = \boldsymbol{c})$ and $E(M \mid A = a, \boldsymbol{C} = \boldsymbol{c})$ does not yield NIE estimators that factorize to satisfy (1) in general, as the identification formula can result in sums or integrals of products of coefficients. Handling models that yield such identification formulas will be an important direction in which to extend the work done in the present article. Extending the theory in the present article to more general settings than (1) is also especially important because mediated effects through multiple mediators can also be identified by sums of products of coefficients.

B. Figure demonstrating the conservativeness of the asymptotic approximation to the true distributions of the delta method and joint significance tests



(a) (b)

Figure S1: (a) Density plots estimated by Monte Carlo sampling of Z_n^{prod} under $\delta_y = 0$ and varying δ_x with $n = 100$. The standard normal density curve is displayed in black. (b) Density plots estimated by Monte Carlo sampling of $Z_n^{\min}$ under $\delta_y = 0$ and varying δ_x with $n = 100$. The folded standard normal density curve is displayed in black.

C. Plot of the power function surface of the minimax optimal test

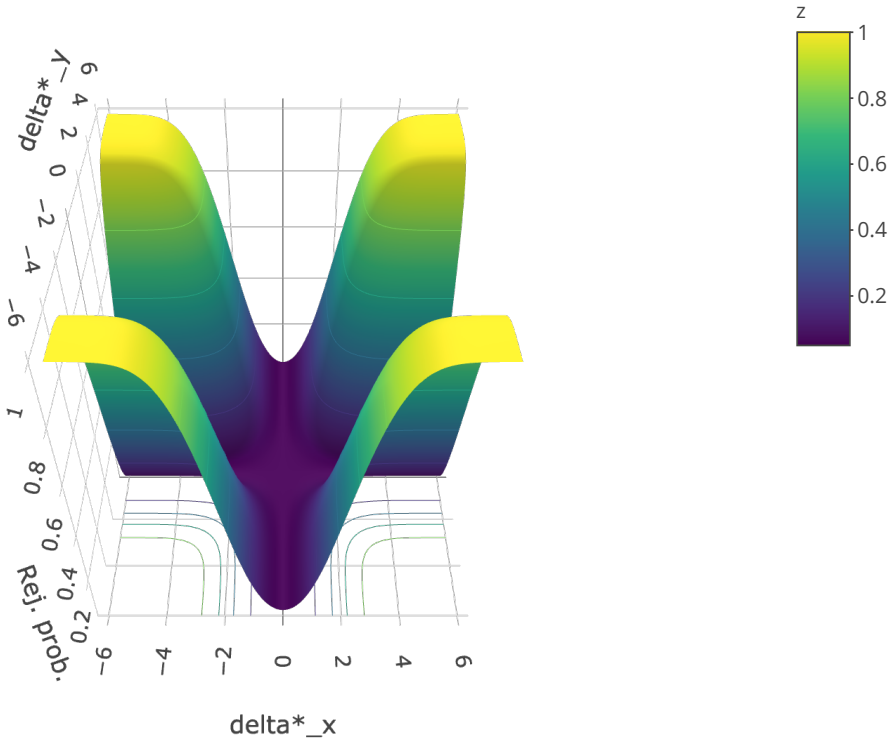


Figure S2: The power function surface for the minimax optimal test over (δ_x^*, δ_y^*). The power function takes the value 0.05 everywhere on the δ_x^* and δ_y^* axes, and is strictly greater than 0.05 everywhere else, while going to one as δ_x^* and δ_y^* both go to infinity.

D. Details of the Bayes risk optimization problem formulation

We denote the full null hypothesis space by $G := \{(\delta_x^*, \delta_y^*) : \delta_x^* \delta_y^* = 0\}$ and a corresponding fine grid within the null hypothesis space by $G' := \{(-2b, 0)^\top, (-2b + b/K, 0)^\top, \ldots, (2b - b/K, 0)^\top, (2b, 0)^\top, (0, -2b)^\top, (0, -2b + b/K)^\top, \ldots, (0, 2b - b/K)^\top, (0, 2b)^\top\}$ for a suitably large integer K. This grid extends beyond B, because for distributions centered outside of, but near B, there will still be a high probability of being in B, and we need to control the type 1 error for such distributions. We then consider the $(2K + 1) \times (2K + 1)$ grid on B, $\{-b, -b(K-1)/K, \ldots, b(K-1)/K, b\}^2$, and denote the squares induced by this grid $R_{k,k'} := [b(k-1)/K, bk/K) \times [b(k'-1)/K, bk'/K)$ for each k and k' in $-K+1, -K+2, \ldots, K-1, K$. Let $\mathcal{R}$ denote the collection of squares $R_{k,k'}$ for all k and k' in $-K+1, -K+2, \ldots, K-1, K$. Within B, we then define a class $\mathcal{M}_\mathcal{R}$ of random rejection regions whose realizations are constant within each square. That is, if $(z^x, z^y) \in R_{k,k'}$, $M(z^x, z^y, u) = M(z^{x'}, z^{y'}, u)$ for all $(z^{x'}, z^{y'}) \in R_{k,k'}$ and all $u \in [0, 1]$. Denote the rejection probability corresponding to each square $r \in \mathcal{R}$ by m_r. These are the unknown arguments defining the rejection region that are to be optimized over.

Under a random test defined by the rejection probabilities m_r on the squares in $\mathcal{R}$, the Bayes risk in (8) decomposes as the sum

$$\begin{aligned} &\int E_{\delta_x^*, \delta_y^*} \left(L \left[I \left\{ (Z_*^x, Z_*^y) \in R_{\bar{B}} \right\}, \delta_x^*, \delta_y^* \right] I \left\{ (Z_*^x, Z_*^y) \in \bar{B} \right\} \right) d\Lambda(\delta_x^*, \delta_y^*) \\ + &\int E_{\delta_x^*, \delta_y^*} \left[L \left\{ M(Z_*^x, Z_*^y, U), \delta_x^*, \delta_y^* \right\} I \left\{ (Z_*^x, Z_*^y) \in B \right\} \right] d\Lambda(\delta_x^*, \delta_y^*), \end{aligned}$$

of which only the second term depends on $(m_r)_{r \in \mathcal{R}}$, being equal to

$$\int \sum_{r \in \mathcal{R}} E_{\delta_x^*, \delta_y^*} \left[L \left\{ M(Z_*^x, Z_*^y, U), \delta_x^*, \delta_y^* \right\} I \left\{ (Z_*^x, Z_*^y) \in r \right\} \right] d\Lambda \left(\delta_x^*, \delta_y^* \right)$$

$$= \int \sum_{r \in \mathcal{R}} E_{\delta_x^*, \delta_y^*} \left[L \left\{ M(Z_*^x, Z_*^y, U), \delta_x^*, \delta_y^* \right\} \mid (Z_*^x, Z_*^y) \in r \right] \Pr_{\delta_x^*, \delta_y^*} \left\{ (Z_*^x, Z_*^y) \in r \right\} d\Lambda \left(\delta_x^*, \delta_y^* \right)$$

$$= \sum_{r \in \mathcal{R}} (1 - m_r) \int L \left(0, \delta_x^*, \delta_y^* \right) \Pr_{\delta_x^*, \delta_y^*} \left\{ (Z_*^x, Z_*^y) \in r \right\} d\Lambda \left(\delta_x^*, \delta_y^* \right).$$

From the above, the objective function is affine in the unknown variables $(m_r)_{r \in \mathcal{R}}$.

We also define a discretized approximation of the type 1 error constraint (9). For each $(\delta_x^*, \delta_y^*) \in G'$,

$$\begin{aligned} \alpha &\geq \Pr_{\delta_x^*, \delta_y^*} \left\{ M(Z_*^x, Z_*^y, U) = 1, (Z_*^x, Z_*^y) \in \bar{B} \right\} + \Pr_{\delta_x^*, \delta_y^*} \left\{ M(Z_*^x, Z_*^y, U) = 1, (Z_*^x, Z_*^y) \in B \right\} \\ &= \Pr_{\delta_x^*, \delta_y^*} \left\{ M(Z_*^x, Z_*^y, U) = 1, (Z_*^x, Z_*^y) \in \bar{B} \right\} + \sum_{r \in \mathcal{R}} m_r \Pr_{\delta_x^*, \delta_y^*} \left\{ (Z_*^x, Z_*^y) \in r \right\}, \end{aligned} \tag{S4}$$

which is also affine in $(m_r)_{r \in \mathcal{R}}$. Lastly, we need the probability constraints $0 \leq m_r \leq 1$ for all $r \in \mathcal{R}$. These inequalities along with (S4) define a linear optimization problem that approximates the Bayes risk optimization problem in $(m_r)_{r \in \mathcal{R}}$, where all other terms are known.

The linear optimization problem is high-dimensional, with the dimension of the unknowns being $4K^2$, and the number of constraints being $8K^2 + 8K + 1$. However, this is also a highly sparse optimization problem, which allows for efficient computation despite the dimensionality. The type 1 error constraints consist of only $8K + 1$ inequalities, and the matrix characterizing the remaining probability constraints contains only one nonzero element per row.

E. Minimax optimal test p-values

A common definition of a p-value for a simple null hypothesis is the probability of observing the same or more extreme values of the test statistic under the null hypothesis. For a composite null hypothesis, a p-value may be defined similarly, but for the least-favorable

distribution in the null hypothesis space. This definition is only useful when rejection regions for each α can be defined as level sets of some function of α, such that "more extreme" than a certain value can be taken to mean the set of values that the test always rejects whenever it rejects that particular value. The tests we have defined in this article admit no such representation. For instance, if $(Z_n^x, Z_n^y) = (\Phi^{-1}(4/5), \Phi^{-1}(5/7))$, the minimax optimal test will reject H_0 for H_1 for $\alpha = 1/3$ but not for $\alpha = 1/2$.

Alternatively, we adopt the more general definition of a p-value as a statistic whose law stochastically dominates that of the uniform distribution with support $[0, 1]$ for all distributions in the null hypothesis space. We can define a tentative p-value to be $\hat{p} := \int_0^1 I\{(Z_*^x, Z_*^y) \notin R_{mm}(\alpha)\}d\alpha$, where $R_{mm}(\alpha)$ is the rejection region for a level α extended minimax optimal test. For the minimax optimal test, we use the extension of the minimax optimal test to non-unit fraction values of α described in Section 3.2. The form of this p-value is derived in Section H. We conjecture that this will result in a valid p-value in the sense that it will dominate the uniform distribution over [0,1] uniformly over the null hypothesis space. The plot in Figure S3 shows the empirical cumulative distribution function of a Monte Carlo sample of this p-value for the minimax optimal test under the least-favorable distribution where $(\delta_x^*, \delta_y^*) = (0, 0)$. We can see that this cumulative distribution function does indeed dominate that of the uniform[0,1] distribution, and yet it is dominated by that of the p-value corresponding to the joint-significance test. Having defined a p-value corresponding to the extended minimax optimal test, one may use these p-values to perform the Benjamini–Hochberg correction to many such independent tests to control the false discovery rate. When generating QQ-plots or volcano plots, no calibrations need to be made to the p-values. A major distinction of our approach to large-scale mediation hypothesis testing from existing approaches is that we make no assumptions about the distribution of the parameters under the null hypothesis and do not attempt to estimate such a distribution. As such, our p-values need to be valid under all parameter

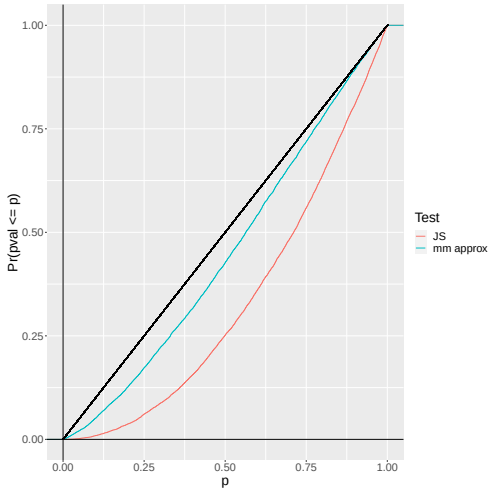


Figure S3: Empirical cumulative distribution function plots of the p-values corresponding to the extended minimax optimal test and joint significance test based on 10,000 Monte Carlo samples from the bivariate normal distribution with $(\delta_x^*, \delta_y^*) = (0, 0)$. The cumulative distribution function of the uniform[0,1] distribution is shown in black.

values under the composite null simultaneously, and we cannot recalibrate to get an exactly uniform distribution.

Since the extended minimax optimal test is conservative for non-unit fraction values of α, the p-value will also be conservative. Thus, if one wishes to employ the standard minimax optimal test at a particular level of α, one would reject based on the test defined by the rejection region $R_{mm}(\alpha)$ rather than comparing the p-value to α, as these may not agree and the hypothesis test will be more powerful. However, as we discuss in the following section, one might prefer to instead define a test based on this p-value.

F. DCTRS data analysis details

F.1. Table of baseline covariates

Table S1: Distribution of baseline covariates across studies

	Study 4155 (N=29)	**Study 8134 (N=32)**	**Study 9212 (N=67)**
Age (months)			
Mean (SD)	466 (117)	215 (28.8)	426 (106)
Median [Min, Max]	444 [228, 768]	216 [168, 276]	408 [204, 732]
Sex			
Female	5 (17.2%)	11 (34.4%)	16 (23.9%)
Male	24 (82.8%)	21 (65.6%)	51 (76.1%)
Race			
Asian	3 (10.3%)	2 (6.3%)	4 (6.0%)
Black or African American	2 (6.9%)	8 (25.0%)	12 (17.9%)
Unknown or not reported	0 (0%)	4 (12.5%)	2 (3.0%)
White	24 (82.8%)	18 (56.3%)	49 (73.1%)
Education (years)			
Mean (SD)	12.3 (1.97)	11.2 (1.47)	11.4 (2.24)
Median [Min, Max]	12.0 [9.00, 17.0]	11.0 [8.00, 16.0]	11.0 [2.00, 17.0]
WAIS working memory digit span test			
Mean (SD)	12.1 (4.55)	11.7 (3.29)	14.8 (3.95)
Median [Min, Max]	12.0 [3.00, 23.0]	11.0 [7.00, 18.0]	15.0 [6.0, 23.0]
Missing	2 (6.9%)	2 (6.3%)	0 (0%)
Rosenberg self-esteem scale	48		
Mean (SD)	31.5 (7.61)	33.4 (6.63)	32.0 (7.65)
Median [Min, Max]	30.5 [12.0, 50.0]	34.0 [19.0, 48.0]	34.0 [10.0, 48.0]
Missing	1 (3.4%)	0 (0%)	0 (0%)

F.2. Variable selection, causal assumptions, and statistical models

Although many additional measures of cognition and functioning at baseline are available in the data, we only included baseline measures of the WAIS working memory digit span test and the Rosenberg self-esteem scale for three reasons. The first is that since we are only concerned with controlling for confounding of the effect of working memory at the end of treatment on self-esteem at the end of follow-up, their corresponding baseline measures would seem to be the most relevant in terms of capturing baseline common causes from the cognition and functioning/quality-of-life domains. The second reason is that the sample size is not very large—certainly not large enough to include all baseline cognition and functioning measures—and so we favored a more parsimonious model that adjusted for the most relevant baseline measures of cognition and functioning/quality-of-life. The third is that many of these measures are not observed in all three studies. As is always the case with mediation analysis (even in the context of randomized experiments), our analyses are subject to bias due to residual confounding. Given that the purpose of this analysis is to illustrate our methodology, a full sensitivity analysis is beyond the scope of this article, though methods for such sensitivity analyses are available (Imai et al., 2010; Ding and Vanderweele, 2016), and will be important to incorporate into our testing procedures in future work. However, we do explore the impact of including additional covariates in $\boldsymbol{C}$ in Section F.5.

Another identification assumption for the NIE is that no common cause of M and Y is affected by A, which, as is always the case in mediation analysis, may be violated in our study. However, this issue is also beyond the scope of this article. To address exposure-induced confounding, one could alternatively estimate partial identification bounds (Miles et al., 2017a), or focus on a different target estimand, such as the path-specific effect not

through the exposure-induced confounder (Avin et al., 2005; Miles et al., 2017b, 2020). Both approaches require observation of all exposure-induced confounders, say $\boldsymbol{L}$. The latter additionally requires the effect of A on $\boldsymbol{L}$ to itself have no residual confounding (which is guaranteed when A is randomized), and no unobserved confounding of the effect of $\boldsymbol{L}$ on M and Y. For a more detailed discussion on the assumptions needed to identify this path-specific effect, see Avin et al. (2005) and Miles et al. (2017b, 2020).

The DCTRS analysis involved pooling data from three different trials. Formally, this is justified if when the study ID indicators are included in $\boldsymbol{C}$, the causal assumptions (consistency, positivity, and no unobserved confounding assumptions) needed for identification of the NIE are satisfied, and if the statistical models for $E(Y \mid A, M, \boldsymbol{C})$ and $E(M \mid A, \boldsymbol{C})$ are correctly specified. Model S3, which only contains main effect terms in study ID indicators, amounts to an assumption that the effect of the exposure on the mediator is homogeneous across studies. Likewise, model S1 amounts to an assumption that the effects of the exposure and the mediator on the outcome are homogeneous across studies.

F.3. Plots of the bivariate test statistic from the DCTRS data analysis on the rejection regions corresponding to various tests

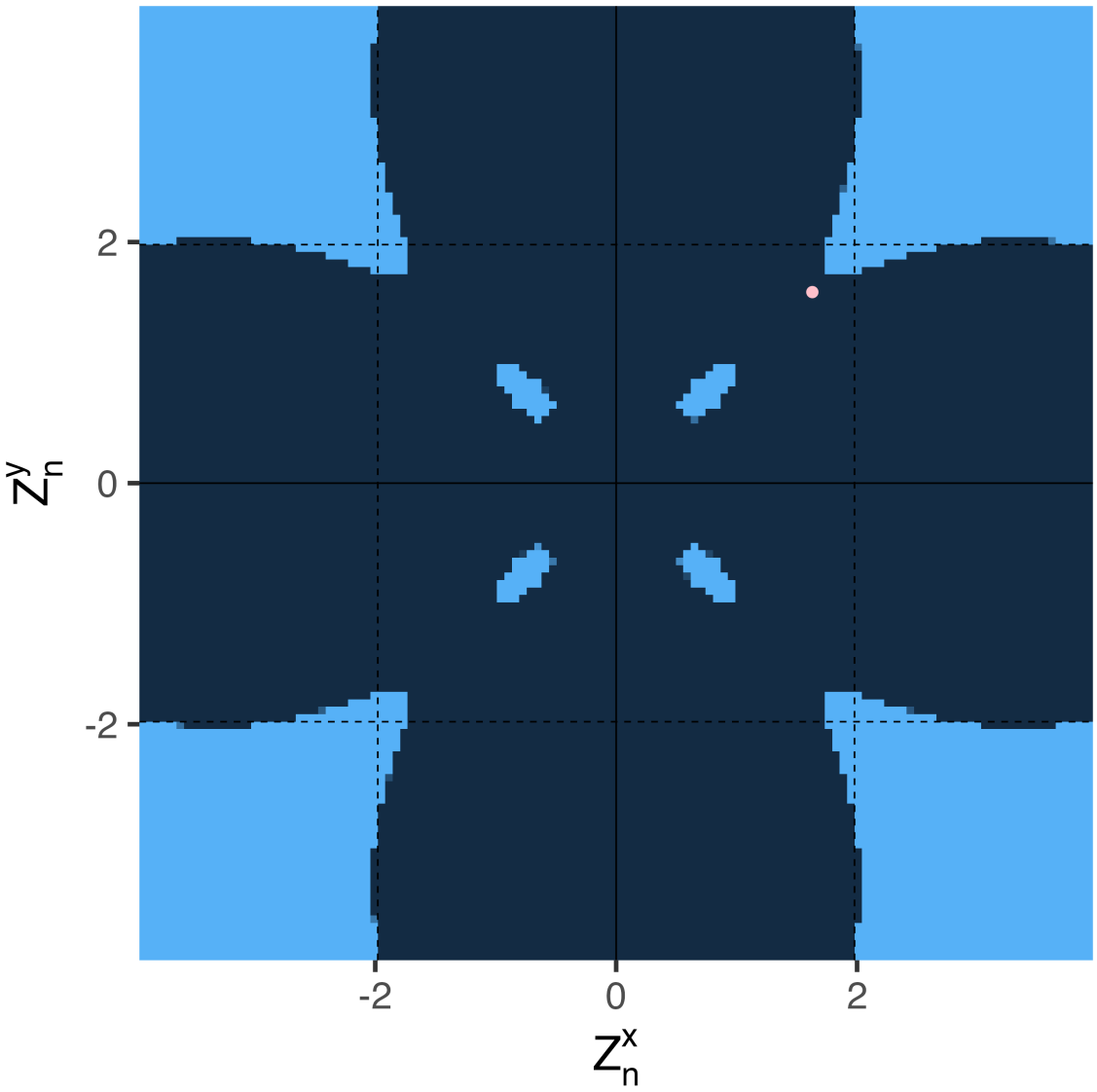


Figure S4: Plot of the test statistic from the DCTRS data analysis on the rejection region of the Bayes risk optimal test with 0-1 loss.

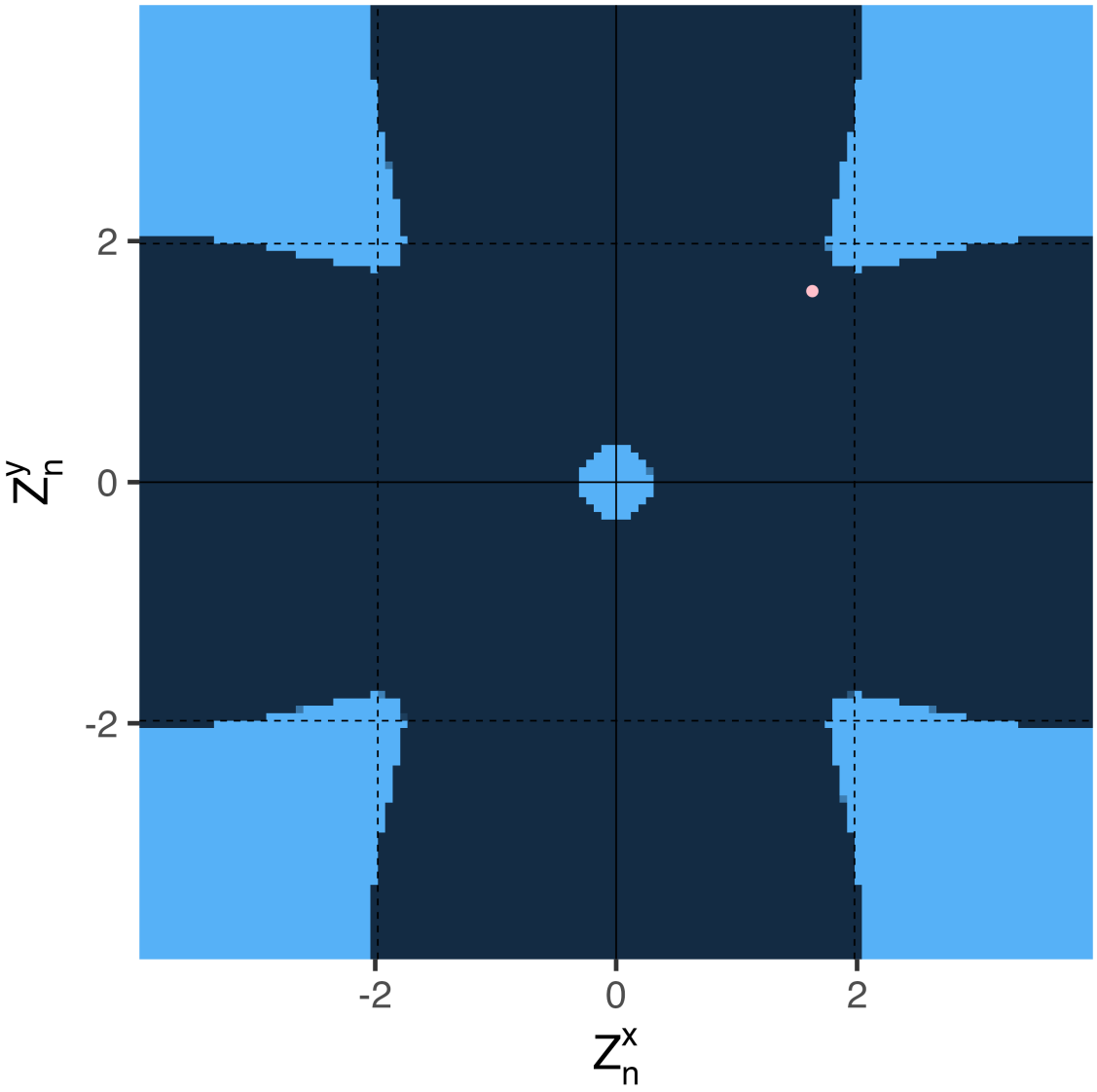


Figure S5: Plot of the test statistic from the DCTRS data analysis on the rejection region of the Bayes risk optimal test with quadratic loss.

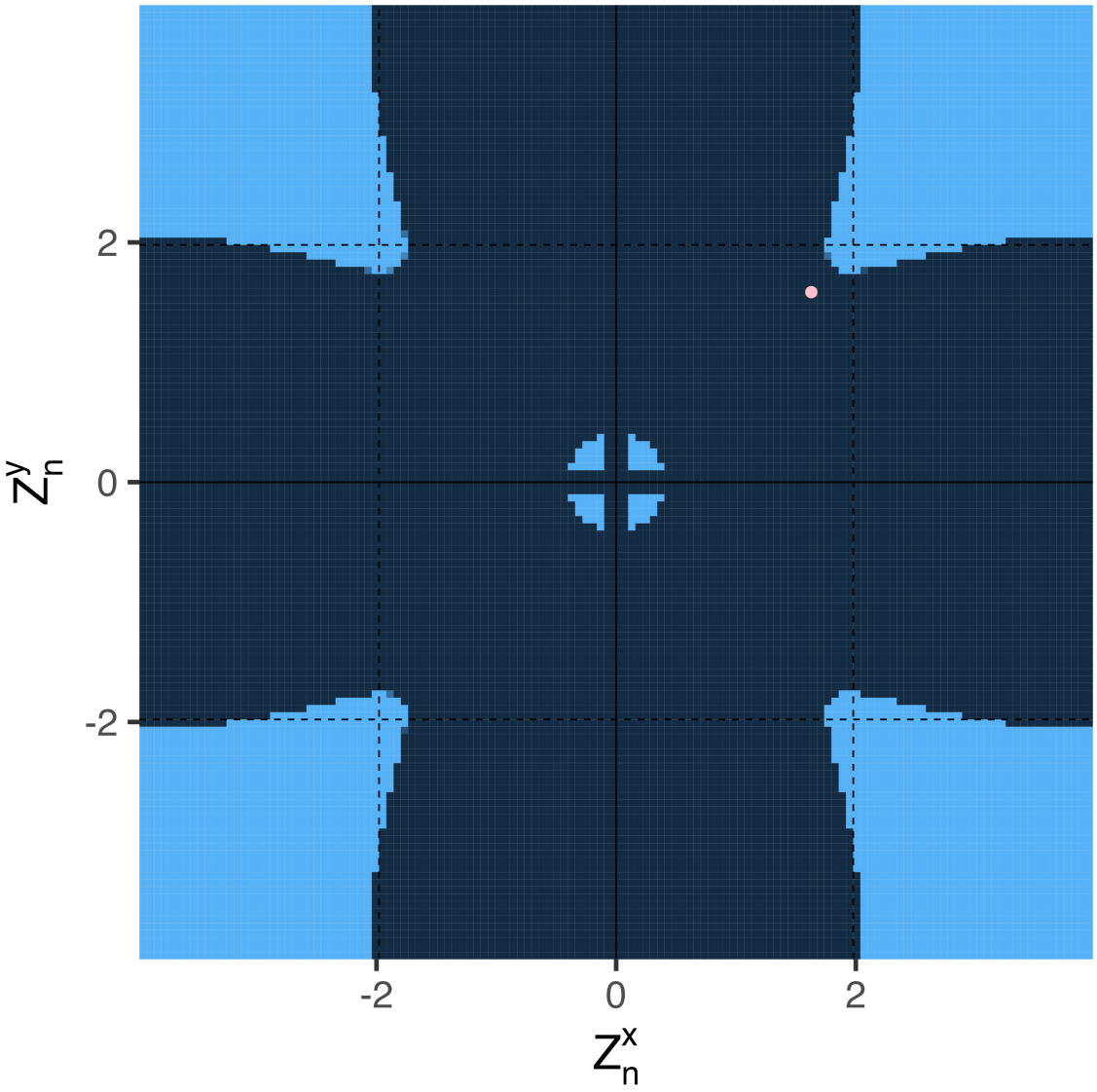


Figure S6: Plot of the test statistic from the DCTRS data analysis on the rejection region of the constrained Bayes risk optimal test with quadratic loss and $d = 0.1$.

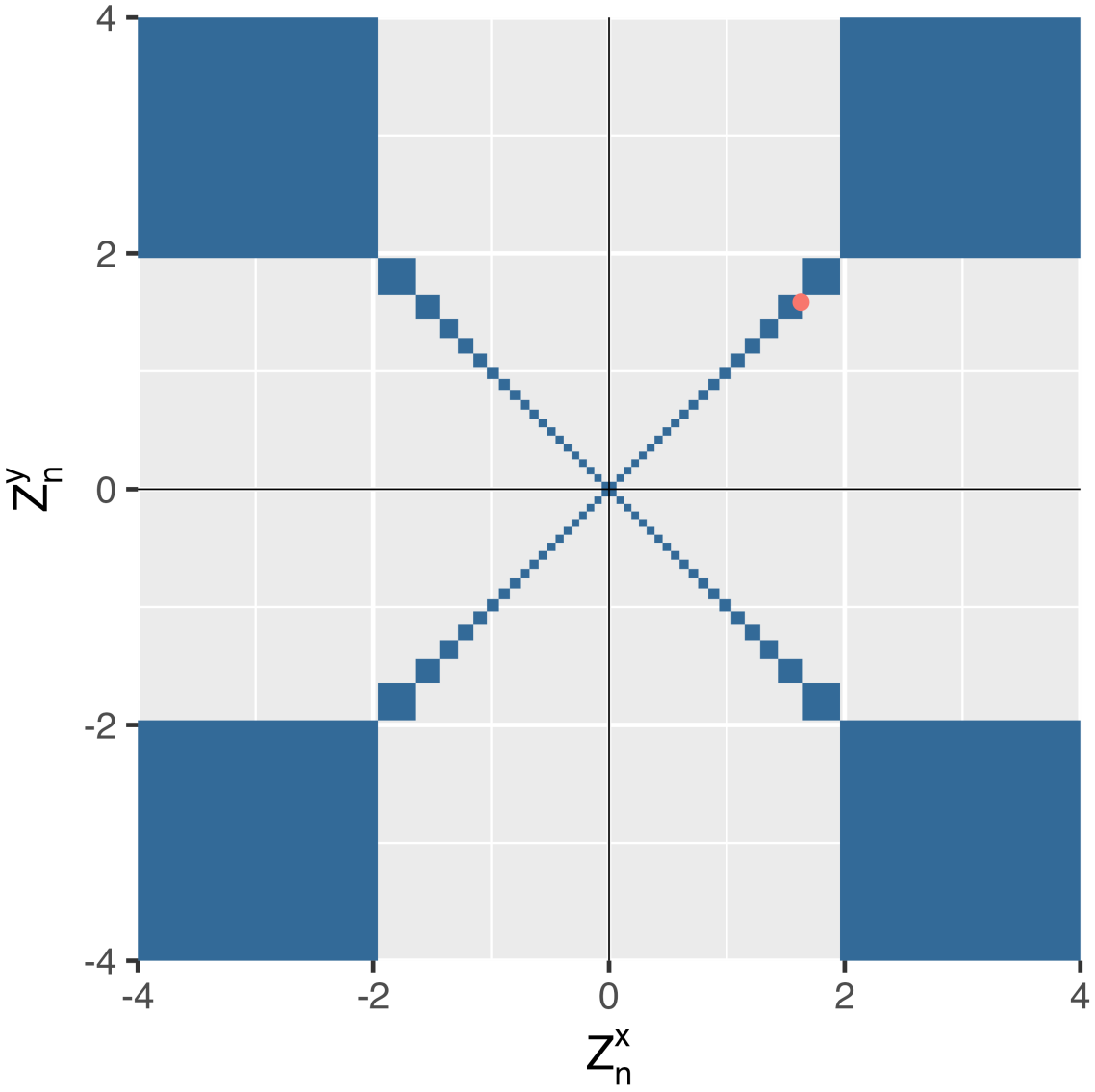


Figure S7: Plot of the test statistic from the DCTRS data analysis on the rejection region of the minimax optimal test.

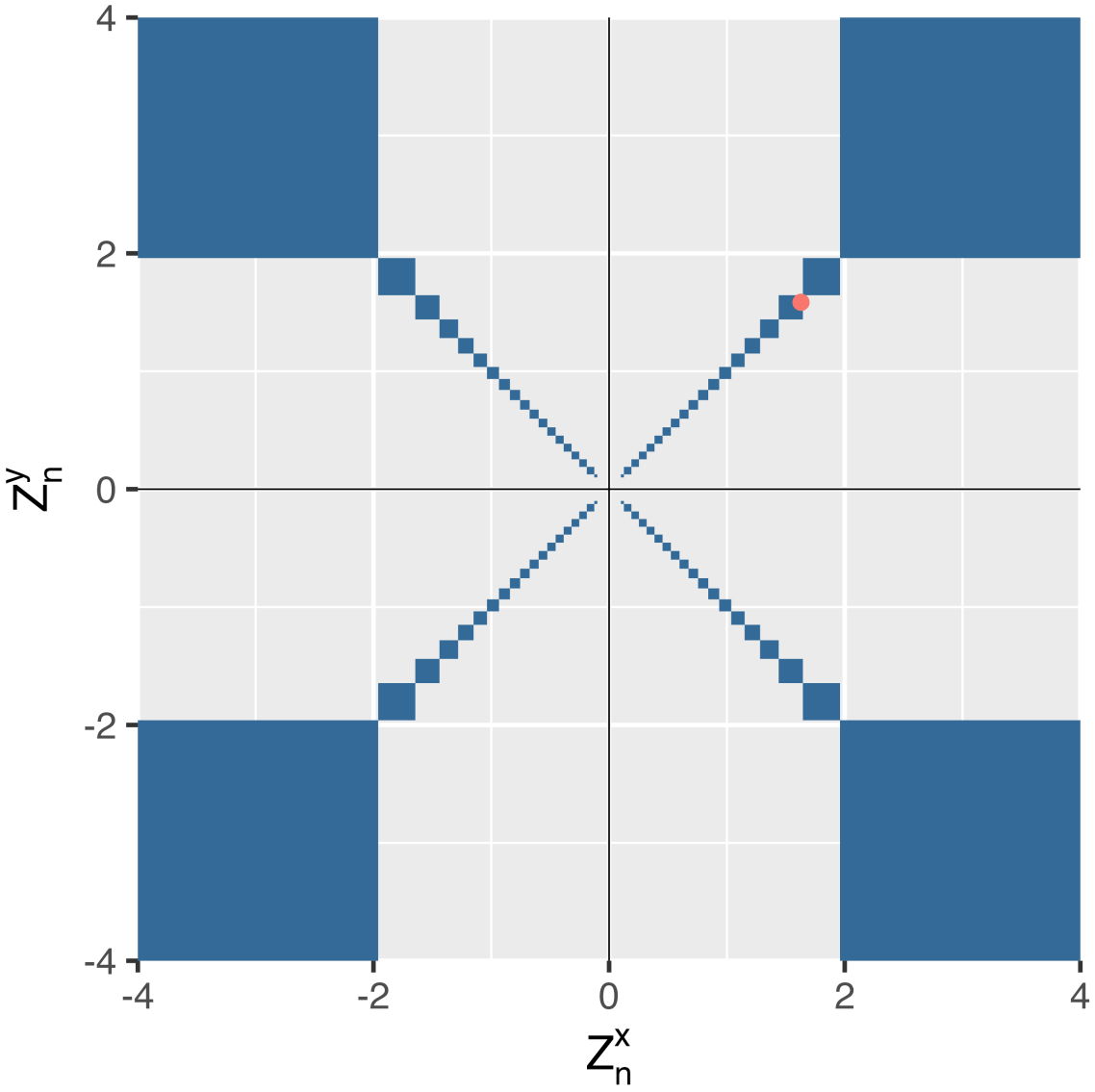


Figure S8: Plot of the test statistic from the DCTRS data analysis on the rejection region of the truncated minimax optimal test with $d = 0.1$.

F.4. Analysis results by study

Table S2: Hypothesis test results for the NIE in Study 4155

Hypothesis test	Reject	p-value
Bayes risk optimal, 0-1 loss	No	Undefined
Bayes risk optimal, quadratic loss	No	Undefined
Bayes risk optimal, quadratic loss, constrained	No	Undefined
Minimax optimal	No	0.164
Minimax optimal, truncated	No	0.164
Minimax optimal, p-value	No	0.164
Delta method	No	0.223
Joint significance	No	0.164
van Garderen and van Giersbergen	No	Undefined

Table S3: Hypothesis test results for the NIE in Study 8134

Hypothesis test	Reject	p-value
Bayes risk optimal, 0-1 loss	No	Undefined
Bayes risk optimal, quadratic loss	No	Undefined
Bayes risk optimal, quadratic loss, constrained	Yes	Undefined
Minimax optimal	No	0.303
Minimax optimal, truncated	No	0.303
Minimax optimal, p-value	No	0.303
Delta method	No	0.880
Joint significance	No	0.859
van Garderen and van Giersbergen	No	Undefined

Table S4: Hypothesis test results for the NIE in Study 9212

Hypothesis test	Reject	p-value
Bayes risk optimal, 0-1 loss	No	Undefined
Bayes risk optimal, quadratic loss	No	Undefined
Bayes risk optimal, quadratic loss, constrained	No	Undefined
Minimax optimal	No	0.343
Minimax optimal, truncated	No	0.343
Minimax optimal, p-value	No	0.343
Delta method	No	0.794
Joint significance	No	0.749
van Garderen and van Giersbergen	No	Undefined

F.5. DCTRS analysis with additional baseline covariates

To assess the sensitivity to the choice of adjustment variables, we also conducted the analysis with additional covariates. In particular, we fit the linear models for the outcome and mediator including all variables available at baseline (after a round of imputation using random forest to fill in missing baseline covariates) using LASSO. We then augmented the adjustment set used in the initial analysis with the union of covariates selected in both LASSO fits, as motivated by the post-double-selection method of Belloni et al. (2014) for estimating average treatment effects. In our case, we only need to control for mediator-outcome confounding since the treatment is randomized, hence the determination of which variables to add to the adjustment set by models for the mediator and the outcome. When using "lambda.min" to tune the LASSO in glmnet (i.e., minimizing cross-validated error

MSE), the selected variables consisted of the following baseline measurements: study ID, marital status, Rosenberg self-esteem scale (self-confirmation factor and total scores), Social Behaviour Scale (SBS) items: (attention-seeking behavior, coherence of conversation, concentration, depression, hostility/friendliness, personal appearance and hygiene, laughing and talking to self, socially unacceptable manners or habits, other behaviors that impede progress, panic attacks and phobias, slowness, verbal fluency test using the letters F, A, and S (FAS): total number of correct responses, trailmaking test part A (TMTA) (paper & pencil): number of errors, trailmaking test part B (TMTB) (paper & pencil): number of errors, Wechsler Adult Intelligence Scale (WAIS) (working memory digit span test raw score, picture completion raw and scaled scores, vocabulary raw score), Wisconsin Card Sorting Test (WCST): categories achieved, Positive and Negative Syndrome Scale (PANSS) items: P1 Delusions, P4 Excitement, P6 Suspiciousness, N5 Difficulty in Abstract Thinking, G2 Anxiety, G3 Guilt Feelings, G9 Unusual Thought Content, G12 Lack of Judgment and Insight, G16 Active Social Avoidance.

Augmenting our adjustment set with these variables yielded the results in Table S5. Evidently, the results are indeed sensitive to which variables we adjust for. Rejection by more of the tests in this case could be a result of instability due to the number of covariates being adjusted for being large relative to the sample size. On the other hand, the fact that some of the tests failed to reject in the previous analysis could be a result of residual confounding bias. We cannot be sure of which is the case. However, it is certainly true that the validity of our proposed tests hinge on the validity of the identification assumptions.

To demonstrate that adjusting for more covariates does not necessarily make tests more likely to agree, we also present results from an unadjusted analysis, i.e., where we do not adjust for any baseline covariates. Results from the unadjusted analysis in Table S6 show that all tests agree, whereas adjusting for the covariate set used for the analysis in the main article yields test results that disagree. Thus, we see empirically across Tables 2, S5, and

Table S5: Hypothesis test results for the NIE when using the augmented set of adjustment variables

Hypothesis test	Reject	p-value
Bayes risk optimal, 0-1 loss	Yes	Undefined
Bayes risk optimal, quadratic loss	Yes	Undefined
Bayes risk optimal, quadratic loss, constrained	Yes	Undefined
Minimax optimal	Yes	0.034
Minimax optimal, truncated	Yes	0.034
Minimax optimal, p-value	Yes	0.034
Delta method	No	0.127
Joint significance	Yes	0.040
van Garderen and van Giersbergen	Yes	Undefined

Table S6: Hypothesis test results for the NIE when not adjusting for any baseline covariates

Hypothesis test	Reject	p-value
Bayes risk optimal, 0-1 loss	No	Undefined
Bayes risk optimal, quadratic loss	No	Undefined
Bayes risk optimal, quadratic loss, constrained	No	Undefined
Minimax optimal	No	0.498
Minimax optimal, truncated	No	0.498
Minimax optimal, p-value	No	0.498
Delta method	No	0.523
Joint significance	No	0.498
van Garderen and van Giersbergen	No	Undefined

S6 that adjusting for more covariates does not generally mean that tests are more or less likely to agree.

G. Tests of products of more than two coefficients

We now consider the more general hypothesis testing setting of H_0^ℓ : "$\prod_{j=1}^{\ell} \delta_j = 0$" against H_1^ℓ : "$\prod_{j=1}^{\ell} \delta_j \neq 0$" when there exists an asymptotically normal estimator $\hat{\boldsymbol{\delta}} := (\hat{\delta}_1, \ldots, \hat{\delta}_\ell)^\top$ of $\boldsymbol{\delta} := (\delta_1, \ldots, \delta_\ell)^\top$ such that the convergence $\sqrt{n}\boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1/2}\left(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\right) \rightsquigarrow \mathcal{N}\left(\mathbf{0}_{\boldsymbol{\ell}}, \boldsymbol{I}_{\boldsymbol{\ell}}\right)$ is uniform in $\boldsymbol{\delta} \in \mathbb{R}^\ell$, where $\boldsymbol{\Sigma}_{\boldsymbol{n}}$ is again a consistent estimator of the asymptotic covariance matrix, $\mathbf{0}_{\boldsymbol{\ell}}$ is the ℓ-vector of zeros, and $\boldsymbol{I}_{\boldsymbol{\ell}}$ is the $\ell \times \ell$ identity matrix. This setting arises in linear structural equation modeling when one wishes to test the effect of an exposure along a chain of intermediate variables. Unfortunately, such effects are seldom identifiable in the causal mediation framework; however, this setting can arise in other applications as well. For instance, given multiple candidate instrumental variables, Kang et al. (2020) characterize the null hypothesis of no treatment effect and one valid instrument that is uncorrelated with the other candidates to be the product of multiple coefficients equaling zero.

The over-conservativeness and underpoweredness problems faced by the traditional tests of products of coefficients are in fact exacerbated in higher dimensions. For instance, the generalization of the joint significance test to ℓ rejects when the ℓ Wald tests corresponding to "$\delta_j = 0$" against "$\delta_j \neq 0$" reject the nulls for their corresponding alternatives for all $j = 1, \ldots, \ell$. Thus, under $\boldsymbol{\delta} = \mathbf{0}_{\boldsymbol{\ell}}$, the rejection probability of the joint significance test will be α^ℓ. For $\alpha = 0.05$ and $\ell = 3$, for example, this will be 0.000125.

The minimax optimal test lacks an obvious unique natural extension to the setting with more than two coefficients. Nevertheless, for unit fraction α, we can prove the existence of deterministic similar tests that are symmetric with respect to negations. For $\ell = 3$, these tests can be constructed using Latin squares of order α^{-1}. If the Latin square is totally

symmetric, then the test will also be symmetric with respect to permutations. We only provide the result for $\ell = 3$; however, we conjecture that these can be readily generalized to higher dimensions using Latin hypercubes, provided they exist for the choice of ℓ and α.

Given a Latin square of order α^{-1}, $\boldsymbol{A}$, let $1, \ldots, \alpha^{-1}$ be the set of symbols populating $\boldsymbol{A}$, and let $A_{i,j}$ denote the symbol in the i-th row and j-th column of $\boldsymbol{A}$. Define $c_k := \Phi^{-1}\{(1+k\alpha)/2\}$ for all $k = 0, \ldots, \alpha^{-1}$. We define the rejection region $R^\dagger \in \mathbb{R}^3$ corresponding to $\boldsymbol{A}$ to be the following union of open hyperrectangles in the nonnegative orthant of $\mathbb{R}^3$:

$$R^\dagger := \bigcup_{i=1}^{\alpha^{-1}} \bigcup_{j=1}^{\alpha^{-1}} (c_{i-1}, c_i) \times (c_{j-1}, c_j) \times (c_{A_{i,j}-1}, c_{A_{i,j}}). \tag{S5}$$

The test corresponding to $R^\dagger$ rejects H_0^ℓ for H_1^ℓ if $(|Z_*^1|, |Z_*^2|, |Z_*^3|)^\top \in R^\dagger$.

Theorem S1. *For any unit fraction $\alpha \in (0,1)$ and Latin square of order α^{-1}, $\boldsymbol{A}$, the corresponding test defined by the rejection region $R^\dagger$ in (S5) is a similar test of H_0^3 against H_1^3 that is symmetric with respect to negations. If $\boldsymbol{A}$ is totally symmetric, then the test is also symmetric with respect to permutations.*

For the test to also be consistent, $\boldsymbol{A}$ must satisfy $A_{\alpha^{-1},\alpha^{-1}} = \alpha^{-1}$. This is because under the alternative hypothesis, $\hat{\delta} \to (\infty, \infty, \infty)^\top$ as $n \to \infty$, hence the power will only converge to one if the hyperrectangle $(\Phi^{-1}(1-\alpha/2), \infty) \times (\Phi^{-1}(1-\alpha/2), \infty) \times (\Phi^{-1}(1-\alpha/2), \infty)$ is included in the rejection region. For any given $\boldsymbol{A}$, there exists at least one isotopy satisfying this property, i.e., one can permute the rows, columns, and symbols of $\boldsymbol{A}$ to produce another Latin square with this property. The arrangement of the other items in the Latin square are relatively less important, and likely result in varying amounts of power in different parts of the parameter space. It is not clear whether one Latin square yields a test that uniformly dominates all others in terms of power.

The Bayes risk optimal test can in theory also be extended to tests of products of ℓ coefficients. The grid in $\mathbb{R}^2$ needs to be expanded to $\mathbb{R}^\ell$, and the grid on the null hypothesis

space must be expanded to each of the $\mathbb{R}^{\ell-1}$ hyperplanes composing the new composite null hypothesis space. Further, the prior on the coefficients must be expanded to an ℓ-dimensional distribution. However, the computational burden will clearly escalate rapidly with ℓ, and may not be feasible even for dimensions of four or possibly even three without access to tremendous computing power and/or memory.

H. Proofs

Proof of Theorem 1. When $\delta_y^* = 0$, for any $\delta_x^* \in \mathbb{R}$,

$$\begin{aligned}
&\Pr_{\delta_x^*,\delta_y^*}\{(Z_*^x, Z_*^y) \in R_{mm}\} \\
=&E_{\delta_x^*}\left[\Pr_{\delta_y^*}\{(Z_*^x, Z_*^y) \in R_{mm} \mid Z_*^x\}\right] \\
=&E_{\delta_x^*}\left[\sum_{k=1}^{2/\alpha} I\{Z_*^x \in (a_{k-1}, a_k)\} \Pr_{\delta_y^*}\{Z_*^y \in (a_{k-1}, a_k) \cup (-a_k, -a_{k-1}) \mid Z_*^x\}\right] \\
=&\alpha E_{\delta_x^*}\left[\sum_{k=1}^{2/\alpha} I\{Z_*^x \in (a_{k-1}, a_k)\}\right] \\
=&\alpha.
\end{aligned}$$

The same holds for any $\delta_y^* \in \mathbb{R}$ when $\delta_x^* = 0$ by symmetry. Thus, (6) holds with R_{mm} substituted for R, and the test generated by the rejection region R_{mm} is a similar test (because the null hypothesis space is its own boundary). □

Proof of Theorem 2. For $\delta_y^* = 0$, we wish to find an R such that for all δ_x^*, we will have

$$\alpha = \Pr_{(\delta_x^*,0)}\{(Z_*^x, Z_*^y) \in R^*\} = E_{\delta_x^*}\left[\Pr_{\delta_y^*=0}\{(Z_*^x, Z_*^y) \in R^* \mid Z_*^x\}\right],$$

which implies that $E_{\delta_x^*}\left[\Pr_{\delta_y^*=0}\{(Z_*^x, Z_*^y) \in R^* \mid Z_*^x\} - \alpha\right] = 0$ for all δ_x^*. Since $Z_*^x \sim \mathcal{N}(\delta_x^*, 1)$, which is a full-rank exponential family, Z_*^x is a complete statistic, and

$$\Pr_{\delta_y^*=0}\{(Z_*^x, Z_*^y) \in R^* \mid Z_*^x\} = \alpha \text{ almost everywhere.} \tag{S6}$$

Lemma 1. *Suppose for any* $k \in \{0, 1, \ldots, 1/\alpha - 1\}$, $f(x) \geq \Phi^{-1}(\frac{1+k\alpha}{2})$ *for all* $x > \Phi^{-1}(\frac{1+k\alpha}{2})$. *Then (i)* $f(x) = \Phi^{-1}(\frac{1+k\alpha}{2})$ *for all* $x \in (\Phi^{-1}(\frac{1+k\alpha}{2}), \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\})$, *and (ii)* $f(x) \geq \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\}$ *for all* $x > \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\}$.

Proof of Lemma 1. (i) Suppose $f(x_0) = y_0 > \Phi^{-1}(\frac{1+k\alpha}{2})$ for some

$$x_0 \in \left(\Phi^{-1}\left(\frac{1+k\alpha}{2}\right), \Phi^{-1}\left\{\frac{1+(k+1)\alpha}{2}\right\}\right),$$

and let f^{-1} be the generalized function inverse $f^{-1}(y) := \inf\{x : f(x) \geq y\}$. Then $\Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\} > x_0 \geq \inf\{x : f(x) \geq y_0\} = f^{-1}(y_0)$ by the nondecreasing monotonicity of f^{-1}. For all $x' \in [\Phi^{-1}(\frac{1+k\alpha}{2}), y_0]$,

$$\begin{aligned}
\Pr_{\delta_y^*=0}\{(Z_*^x, Z_*^y) \in R^* \mid Z_*^x = x'\} &= 2\left[\Phi\left\{f^{-1}(x')\right\} - \Phi\left\{f(x')\right\}\right] \\
&\leq 2\left[\Phi\left\{f^{-1}(y_0)\right\} - \Phi\left\{\Phi^{-1}\left(\frac{1+k\alpha}{2}\right)\right\}\right] \\
&< 2\left(\Phi\left[\Phi^{-1}\left\{\frac{1+(k+1)\alpha}{2}\right\}\right] - \frac{1+k\alpha}{2}\right) \\
&= 2\left\{\frac{1+(k+1)\alpha}{2} - \frac{1+k\alpha}{2}\right\} \\
&= \alpha.
\end{aligned}$$

Since $[\Phi^{-1}(\frac{1+k\alpha}{2}), y_0]$ is a set of positive measure if $y_0 > \Phi^{-1}(\frac{1+k\alpha}{2})$, this contradicts (S6). Thus, $f(x) = \Phi^{-1}(\frac{1+k\alpha}{2})$ for all $x \in (\Phi^{-1}(\frac{1+k\alpha}{2}), \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\})$.

(ii) By (S6), for almost every $x \in (\Phi^{-1}(\frac{1+k\alpha}{2}), \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\})$,

$$\begin{aligned}
\alpha &= \Pr_{\delta_y^*=0}\{(Z_*^x, Z_*^y) \in R^* \mid Z_*^x = x\} \\
&= 2\left[\Phi\left\{f^{-1}(x)\right\} - \Phi\left\{f(x)\right\}\right] \\
&= 2\left[\Phi\left\{f^{-1}(x)\right\} - \frac{1+k\alpha}{2}\right],
\end{aligned}$$

hence $f^{-1}(x) = \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\}$ almost everywhere in $(\Phi^{-1}(\frac{1+k\alpha}{2}), \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\})$. By nondecreasing monotonicity of f^{-1}, $f^{-1}(x) = \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\}$ for all $x \in (\Phi^{-1}(\frac{1+k\alpha}{2}), \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\})$.

Now suppose $x_1 > \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\}$ and $f(x_1) < \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\}$. Then for $x'' \in (f(x_1), \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\})$, $f^{-1}(x'') \geq x_1 > \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\}$, which contradicts $f^{-1}(x'') = \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\}$ since $x'' \in (\Phi^{-1}(\frac{1+k\alpha}{2}), \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\})$. Thus, $f(x) \geq \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\}$ for all $x > \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\}$. □

By the constraints on $\mathcal{F}$, $f(x) \geq 0 = \Phi^{-1}(1/2)$ for all $x > 0 = \Phi^{-1}(1/2)$. Applying Lemma 1, for each $k \in \{0, 1, \ldots, 2/\alpha - 1\}$, we have $f(x) = \Phi^{-1}(\frac{1+k\alpha}{2})$ for all $x \in (\Phi^{-1}(\frac{1+k\alpha}{2}), \Phi^{-1}\{\frac{1+(k+1)\alpha}{2}\})$ by induction. That is, $f(x) = a_{k-1}$ for all $x \in (a_{k-1}, a_k)$. Additionally, $f(0) = 0$ must hold. The only set of points on which we have not defined f is

$$\left\{a_{1/\alpha+1}, a_{1/\alpha+2}, \ldots, a_{2/\alpha}\right\} = \left\{\Phi^{-1}\left(\frac{1+\alpha}{2}\right), \Phi^{-1}\left(\frac{1+2\alpha}{2}\right), \ldots, \Phi^{-1}(1)\right\},$$

which is a set of measure zero. Thus, any function f satisfying (S6) and hence generating a similar test must be equal to f_{mm} everywhere except for the above set of measure zero. □

Proof of Theorem 3. Suppose $R^\dagger \subset \mathbb{R}^2$ is a rejection region bounded away from $\{(Z_*^x, Z_*^y) : Z_*^x Z_*^y = 0\}$. Then there is some $\varepsilon > 0$ for which $R^\dagger \cap \{(Z_*^x, Z_*^y)^\top : -\varepsilon < Z_*^x < \varepsilon\} = \emptyset$. For all $Z_*^x \in (-\varepsilon, \varepsilon)$, a set of positive measure, $\Pr\left\{(Z_*^x, Z_*^y)^\top \in R^\dagger \mid Z_*^x\right\} = 0$, hence (S6) does not hold for $R^\dagger$. Thus, $\Pr\left\{(Z_*^x, Z_*^y)^\top \in R^\dagger\right\} = 0.05$ cannot hold for all (δ_x^*, δ_y^*) in H_0, and $R^\dagger$ does not generate a similar test. □

Proof of Theorem 4. It won't come as a surprise that the proof is very similar to that of Theorem 1. For brevity, introduce $B_k := (b_{k-1}, b_k) \cup (-b_k, -b_{k-1})$ for every $k = 1, \ldots, \lfloor \alpha^{-1} \rfloor$. By symmetry, we can focus on the case where $\delta_y^* = 0$ and δ_x^* is fixed arbitrarily in $\mathbb{R}$. Note

that $\Pr_{\delta_y^*=0}(Z_*^y \in B_k \mid Z_*^x) = \alpha$ for every $k = 1, \ldots, \lfloor \alpha^{-1} \rfloor$. Therefore, it holds that

$$
\begin{aligned}
&\Pr_{\delta_x^*, \delta_y^*=0}\{(Z_*^x, Z_*^y) \in R'_{mm}\} \\
&= E_{\delta_x^*}\left[\Pr_{\delta_y^*=0}\left\{(Z_*^x, Z_*^y) \in R'_{mm} \mid Z_*^x\right\}\right] \\
&= E_{\delta_x^*}\Bigg[\sum_{k=1}^{\lfloor \alpha^{-1} \rfloor} I\left\{Z_*^x \in (b_{k-1}, b_k)\right\} \Pr_{\delta_y^*=0}\left(Z_*^y \in B_k \mid Z_*^x\right) \\
&\qquad + \sum_{k=1}^{\lfloor \alpha^{-1} \rfloor} I\left\{Z_*^x \in (-b_k, -b_{k-1})\right\} \Pr_{\delta_y^*=0}\left(Z_*^y \in B_k \mid Z_*^x\right) \\
&\qquad + I\left\{Z_*^x \in (-b_0, b_0)\right\} \Pr_{\delta_y^*=0}\left\{Z_*^y \in (-b_0, b_0) \mid Z_*^x\right\}\Bigg] \\
&= \alpha \Pr_{\delta_x^*}\left\{Z_*^x \notin (-b_0, b_0)\right\} + \Pr_{\delta_x^*}\left\{Z_*^x \in (-b_0, b_0)\right\} \Pr_{\delta_y^*=0}\left\{Z_*^y \in (-b_0, b_0)\right\} \\
&= \alpha \Pr_{\delta_x^*}\left\{Z_*^x \notin (-b_0, b_0)\right\} + \Pr_{\delta_x^*}\left\{Z_*^x \in (-b_0, b_0)\right\} \left(1 - \lfloor \alpha^{-1} \rfloor \alpha\right) \\
&= \alpha + \Pr_{\delta_x^*}\left\{Z_*^x \in (-b_0, b_0)\right\} \left(1 - \lfloor \alpha^{-1} \rfloor \alpha - \alpha\right).
\end{aligned}
$$

Because $(1 - \lfloor \alpha^{-1} \rfloor \alpha - \alpha) \leq 0$, the right-hand side expression is smaller than α, which is also the limit as $\delta_x^* \to +\infty$. Moreover, the right-hand side expression is minimized at $\delta_x^* = 0$, where it equals $\lfloor \alpha^{-1} \rfloor \alpha^2 + (1 - \lfloor \alpha^{-1} \rfloor \alpha)^2$. This completes the proof. □

Derivation of the p-value formula. Set $z = (z_1, z_2) \in \mathbb{R}^2 \setminus \{t \in \mathbb{R}^2 : |t_1| = |t_2|\}$ and define $\bar{z} = \max(|z_1|, |z_2|)$, $\underline{z} = \min(|z_1|, |z_2|)$. For reasons that will become apparent later, suppose moreover that $\Phi(\bar{z}) \notin \{(2 + \ell)/[2(\ell + 1)] : \ell \in \mathbb{N}\}$. If z is the realization of a random variable Z drawn in $\mathbb{R}^2$ from a law dominated by the Lebesgue measure, then z meets all the conditions almost surely.

Set $\ell \geq 1$ and $\alpha \in](\ell + 1)^{-1}, \ell]$, so that $\lfloor \alpha^{-1} \rfloor = \ell$. It holds that $z \in R(\alpha)$ if and only if (iff)

1. $\alpha\ell \leq 2(1 - \Phi(\bar{z}))$, or

2. $\alpha \geq 2(1-\Phi(\underline{z}))$, or

3. $\exists k \in \{1,\ldots,\ell-1\}$ such that

$$\frac{2(1-\Phi(\underline{z}))}{k+1} \leq \alpha \leq \frac{2(1-\Phi(\bar{z}))}{k}. \tag{S7}$$

Note that the three conditions are exclusive. Moreover, there exists at most one k such that (S7) holds true. Therefore,

$$\begin{aligned}
&\int_0^1 \mathbf{1}\{z \in R(\alpha)\} d\alpha \\
&= \sum_{\ell \geq 1} \int_{(\ell+1)^{-1}}^{\ell^{-1}} \mathbf{1}\left\{\alpha\ell \leq 2(1-\Phi(\bar{z}))\right\} d\alpha && \text{(S8)} \\
&\quad + \sum_{\ell \geq 1} \int_{(\ell+1)^{-1}}^{\ell^{-1}} \mathbf{1}\left\{\alpha \geq 2(1-\Phi(\underline{z}))\right\} d\alpha && \text{(S9)} \\
&\quad + \sum_{\ell \geq 1} \int_{(\ell+1)^{-1}}^{\ell^{-1}} \sum_{k=1}^{\ell-1} \mathbf{1}\left\{\frac{2(1-\Phi(\underline{z}))}{k+1} \leq \alpha \leq \frac{2(1-\Phi(\bar{z}))}{k}\right\} d\alpha. && \text{(S10)}
\end{aligned}$$

Next, we study the summands (S8), (S9), (S10) in turn.

Introduce for convenience notation for the jth harmonic number $S_1(j) = \sum_{i=1}^{j} i^{-1}$ for any $j \geq 1$. Set also $S_1(0) = 0$.

Summand (S8). Define

$$L = \left\lfloor \frac{2(1-\Phi(\bar{z}))}{2\Phi(\bar{z})-1} \right\rfloor < \frac{2(1-\Phi(\bar{z}))}{2\Phi(\bar{z})-1}. \tag{S11}$$

The inequality is entailed by the constraints imposed on z. Moreover, $L \geq 1$ iff $\Phi(\bar{z}) \leq \frac{3}{4}$. Now, observe that, for any integer $\ell \geq 1$,

$$\frac{1}{\ell+1} \leq \frac{2(1-\Phi(\bar{z}))}{\ell} \leq \frac{1}{\ell} \text{ iff } \ell \leq L.$$

It follows that

$$\begin{aligned}
\text{(S8)} &= \mathbf{1}\{L \geq 1\} \sum_{\ell=1}^{L} \int_{(\ell+1)^{-1}}^{\ell^{-1}} \mathbf{1}\left\{\alpha\ell \leq 2(1-\Phi(\bar{z}))\right\} d\alpha \\
&= \mathbf{1}\{L \geq 1\} \sum_{\ell=1}^{L} \left(\frac{2(1-\Phi(\bar{z}))}{\ell} - \frac{1}{\ell+1} \right) \\
&= \mathbf{1}\{L \geq 1\} \left[2(1-\Phi(\bar{z}))S_1(L) - S_1(L+1) + 1 \right] \\
&= \mathbf{1}\{L \geq 1\} \left[1 - (2\Phi(\bar{z})-1)S_1(L) - (L+1)^{-1} \right] \\
&= 1 - (2\Phi(\bar{z})-1)S_1(L) - (L+1)^{-1}. \qquad \text{(S12)}
\end{aligned}$$

Summand (S9)**.** It is obvious that

$$\text{(S9)} = 1 - 2(1-\Phi(\underline{z})) = 2\Phi(\underline{z}) - 1. \qquad \text{(S13)}$$

Summand (S10)**.** We first note that, if $k \geq 1$ is an integer, then

$$\frac{2(1-\Phi(\underline{z}))}{k+1} \leq \frac{2(1-\Phi(\bar{z}))}{k} \text{ iff } k \leq K = \left\lfloor \frac{1-\Phi(\bar{z})}{\Phi(\bar{z}) - \Phi(\underline{z})} \right\rfloor.$$

Moreover, $K \geq L$ and $K \geq 1$ iff $2\Phi(\bar{z}) \leq 1 + \Phi(\underline{z})$. Consequently,

$$\begin{aligned}
\text{(S10)} &= \mathbf{1}\{K \geq 1\} \sum_{\ell \geq 1} \int_{(\ell+1)^{-1}}^{\ell^{-1}} \sum_{k=1}^{K} \mathbf{1}\{k \leq \ell - 1\} \mathbf{1}\left\{ \frac{2(1-\Phi(\underline{z}))}{k+1} \leq \alpha \leq \frac{2(1-\Phi(\bar{z}))}{k} \right\} d\alpha \\
&= \mathbf{1}\{K \geq 1\} \sum_{k=1}^{K} \sum_{\ell \geq k+1} \int_{(\ell+1)^{-1}}^{\ell^{-1}} \mathbf{1}\left\{ \frac{2(1-\Phi(\underline{z}))}{k+1} \leq \alpha \leq \frac{2(1-\Phi(\bar{z}))}{k} \right\} d\alpha \\
&= \mathbf{1}\{K \geq 1\} \sum_{k=1}^{K} \int_{0}^{(k+1)^{-1}} \mathbf{1}\left\{ \frac{2(1-\Phi(\underline{z}))}{k+1} \leq \alpha \leq \frac{2(1-\Phi(\bar{z}))}{k} \right\} d\alpha. \qquad \text{(S14)}
\end{aligned}$$

Now, note that, for any $k \geq 1$,

$$\frac{2(1-\Phi(\bar{z}))}{k} \leq \frac{1}{k+1} \quad \text{iff} \quad k \geq \frac{2(1-\Phi(\bar{z}))}{2\Phi(\bar{z})-1} \quad \text{iff} \quad k \geq L+1,$$

the second equivalence following from the inequality in (S11). In light of (S14), it thus holds that

$$
\begin{aligned}
(\text{S10}) &= \mathbf{1}\{K \geq 1 > L = 0\} \sum_{k=L+1}^{K} \left(\frac{2(1-\Phi(\bar{z}))}{k} - \frac{2(1-\Phi(\underline{z}))}{k+1} \right) \\
&\quad + \mathbf{1}\{L \geq 1\} \sum_{k=1}^{L} \left(\frac{1}{k+1} - \frac{2(1-\Phi(\underline{z}))}{k+1} \right) \\
&\quad + \mathbf{1}\{K > L \geq 1\} \sum_{k=L+1}^{K} \left(\frac{2(1-\Phi(\bar{z}))}{k} - \frac{2(1-\Phi(\underline{z}))}{k+1} \right) \\
&= \mathbf{1}\{L \geq 1\}[(2\Phi(\underline{z}) - 1)(S_1(L+1) - 1)] \\
&\quad + [\mathbf{1}\{K \geq 1 > L = 0\} + \mathbf{1}\{K > L \geq 1\}] \sum_{k=L+1}^{K} \left(\frac{2(1-\Phi(\bar{z}))}{k} - \frac{2(1-\Phi(\underline{z}))}{k+1} \right) \\
&= (2\Phi(\underline{z}) - 1)(S_1(L+1) - 1) \\
&\quad + \mathbf{1}\{K > L\} \sum_{k=L+1}^{K} \left(\frac{2(1-\Phi(\bar{z}))}{k} - \frac{2(1-\Phi(\underline{z}))}{k+1} \right) \\
&= (2\Phi(\underline{z}) - 1)(S_1(L+1) - 1) \\
&\quad + \mathbf{1}\{K > L\} \{2(1-\Phi(\bar{z}))[S_1(K) - S_1(L)] - 2(1-\Phi(\underline{z}))[S_1(K+1) - S_1(L+1)]\} \\
&= (2\Phi(\underline{z}) - 1)(S_1(L+1) - 1) \\
&\quad - 2\mathbf{1}\{K > L\}\{(\Phi(\bar{z}) - \Phi(\underline{z}))[S_1(K) - S_1(L)]\} \\
&\quad + 2\mathbf{1}\{K > L\}\{(1-\Phi(\underline{z}))[(L+1)^{-1} - (K+1)^{-1}]\}. \qquad (\text{S15})
\end{aligned}
$$

In view of (S12), (S13), (S15), we obtain

$$
\begin{aligned}
\int_0^1 \mathbf{1}\{z \in R(\alpha)\} d\alpha &= 2\Phi(\underline{z}) - 1 \\
&\quad + 1 - (2\Phi(\bar{z}) - 1)S_1(L) - (L+1)^{-1} \\
&\quad + (2\Phi(\underline{z}) - 1)(S_1(L+1) - 1) \\
&\quad - 2\mathbf{1}\{K > L\}\{(\Phi(\bar{z}) - \Phi(\underline{z}))[S_1(K) - S_1(L)]\}
\end{aligned}
$$

$$
\begin{aligned}
&+ 2\mathbf{1}\{K > L\}\{(1-\Phi(\underline{z}))[(L+1)^{-1} - (K+1)^{-1}]\} \\
&= 1 - 2(1-\Phi(\underline{z}))(L+1)^{-1} - 2(\Phi(\bar{z}) - \Phi(\underline{z}))S_1(L) \\
&\quad - 2\mathbf{1}\{K > L\}\{(\Phi(\bar{z}) - \Phi(\underline{z}))[S_1(K) - S_1(L)]\} \\
&\quad + 2\mathbf{1}\{K > L\}\{(1-\Phi(\underline{z}))[(L+1)^{-1} - (K+1)^{-1}]\} \\
&= 1 - 2(1-\Phi(\underline{z}))(K \vee L + 1)^{-1} - 2(\Phi(\bar{z}) - \Phi(\underline{z}))S_1(K \vee L) \\
&= 1 - 2(1-\Phi(\underline{z}))(K+1)^{-1} - 2(\Phi(\bar{z}) - \Phi(\underline{z}))S_1(K).
\end{aligned}
$$

We conclude that

$$
\int_0^1 \mathbf{1}\{z \notin R(\alpha)\} d\alpha = 2\{1-\Phi(\underline{z})\}(K+1)^{-1} + 2\{\Phi(\bar{z}) - \Phi(\underline{z})\}S_1(K). \tag{S16}
$$

□

Proof of Theorem S1. Let $|Z_*| := (|Z_*^1|, |Z_*^2|, |Z_*^3|)^\top$. When $\delta_3^* = 0$, for any $\delta_1^*, \delta_2^* \in \mathbb{R}$,

$$
\begin{aligned}
&\Pr_{\delta^*}\{|Z_*| \in R^\dagger\} \\
&= E_{\delta_1^*, \delta_2^*}\left[\Pr_{\delta_3^*}\left\{|Z_*| \in R^\dagger \mid Z_*^1, Z_*^2\right\}\right] \\
&= E_{\delta_1^*, \delta_2^*}\left[\sum_{i=1}^{\alpha^{-1}}\sum_{j=1}^{\alpha^{-1}} I\left\{(|Z_*^1|, |Z_*^2|) \in (c_{i-1}, c_i) \times (c_{j-1}, c_j)\right\} \Pr_{\delta_3^*}\left\{|Z_*^3| \in (c_{A_{i,j}-1}, c_{A_{i,j}}) \mid Z_*^1, Z_*^2\right\}\right] \\
&= \alpha E_{\delta_1^*, \delta_2^*}\left[\sum_{i=1}^{\alpha^{-1}}\sum_{j=1}^{\alpha^{-1}} I\left\{(|Z_*^1|, |Z_*^2|) \in (c_{i-1}, c_i) \times (c_{j-1}, c_j)\right\}\right] \\
&= \alpha.
\end{aligned}
$$

We may alternatively represent A in its orthogonal array representation, which is an array with rows corresponding to each entry of A consisting of the triple (row, column, symbol). Any permutation of the coordinates of these triples yields an orthogonal array corresponding to another Latin square known as a conjugate. Let A^{132} be the conjugate

of A generated by exchanging the second and third columns of the orthogonal array representation of A. The region $R^\dagger$ can alternatively be expressed as

$$\bigcup_{i=1}^{\alpha^{-1}} \bigcup_{k=1}^{\alpha^{-1}} (c_{i-1}, c_i) \times (c_{A^{132}_{i,k}-1}, c_{A^{132}_{i,k}}) \times (c_{k-1}, c_k).$$

When $\delta_2^* = 0$, for any $\delta_1^*, \delta_3^* \in \mathbb{R}$, we have the analogous result:

$$\begin{aligned}
&\Pr_{\delta^*}\{|Z_*| \in R^\dagger\} \\
&= E_{\delta_1^*,\delta_3^*}\left[\Pr_{\delta_2^*}\left\{|Z_*| \in R^\dagger \mid Z_*^1, Z_*^3\right\}\right] \\
&= E_{\delta_1^*,\delta_3^*}\left[\sum_{i=1}^{2/\alpha}\sum_{k=1}^{2/\alpha} I\left\{(|Z_*^1|, |Z_*^3|) \in (c_{i-1}, c_i) \times (c_{k-1}, c_k)\right\} \Pr_{\delta_2^*}\left\{|Z_*^2| \in (c_{A^{132}_{i,k}-1}, c_{A^{132}_{i,k}}) \mid Z_*^1, Z_*^3\right\}\right] \\
&= \alpha E_{\delta_1^*,\delta_3^*}\left[\sum_{i=1}^{2/\alpha}\sum_{k=1}^{2/\alpha} I\left\{(|Z_*^1|, |Z_*^3|) \in (c_{i-1}, c_i) \times (c_{k-1}, c_k)\right\}\right] \\
&= \alpha.
\end{aligned}$$

Obviously, the same holds for any $\delta_2^*, \delta_3^* \in \mathbb{R}$ when $\delta_1^* = 0$ by symmetry. Thus, $\Pr_{\delta^*}\{|Z_*| \in R^\dagger\} = \alpha$ for all $(\delta_1^*, \delta_2^*, \delta_3^*) \in \{(\delta_1^*, \delta_2^*, \delta_3^*) : \delta_1^*\delta_2^*\delta_3^* = 0\}$, which is equal to the boundary of the null hypothesis space, so $R^\dagger$ generates a similar test.

Clearly, the test is symmetric with respect to negations by construction. A Latin square is totally symmetric if it is equal to all of its conjugates. Thus, if A is totally symmetric, then $R^\dagger = R^\sigma$ for all permutations σ of (1,2,3), and the test corresponding to a totally symmetric Latin square is also symmetric with respect to permutations. □