Nonparametric tests for treatment effect heterogeneity in observational studies

Maozhu DAI, Weining SHEN*, and Hal S. STERN

Department of Statistics, University of California, Irvine, California, USA

Key words and phrases: Causal inference; observational study; reweighting; subgroup analysis; $U$-statistics.

MSC 2020: 62D20; 62G10.

Abstract: We consider the problem of testing for treatment effect heterogeneity in observational studies and propose a nonparametric test based on multisample $U$-statistics. To account for potential confounders, we use reweighted data where the weights are determined by estimated propensity scores. The proposed method does not require any parametric assumptions on the outcomes and bypasses the need for modelling the treatment effect for each study subgroup. We establish the asymptotic normality for the test statistic and demonstrate its superior numerical performance over several competing approaches via simulation studies. Two real data applications are discussed: an employment programme evaluation study and a mental health study of China’s one-child policy.

1. INTRODUCTION

Treatment effect heterogeneity has attracted a great deal of attention in various research areas, including social sciences (Bitler, Gelbach & Hoynes, 2006; Feller & Holmes, 2009), health care (Ginsburg & Willard, 2009; Kent et al., 2016) and criminology (Pate & Hamilton, 1992; Na, Loughran & Paternoster, 2015). It is now well recognized that “one size does not fit all” in many disease studies because subjects with different characteristics may respond quite differently to the same treatment. To better account for patient heterogeneity while evaluating the treatment effect and providing accurate, personalized treatment recommendation, subgroup analysis (Byar, 1985) has been commonly used to identify subpopulations among subjects and examine the localized treatment effects within subpopulations. In some studies, subjects may be divided into several strata based on baseline characteristics that are expected to be...
associated with treatment effects and recommendations are made based on inference conducted within each stratum. Though commonly used, subgroup analysis involves various problems such as multiple testing and loss of statistical power (Cook, Gebski & Keech, 2004). However, when there is enough treatment effect heterogeneity across strata, subgroup analysis can be beneficial. Therefore, a test identifying the existence of treatment effect heterogeneity can be valuable.

There is an emerging body of literature on developing hypothesis testing approaches for examining treatment effect heterogeneity (e.g., Chang, Lee & Whang, 2015; Ding, Feller & Miratrix, 2016; Hsu, 2017) under different definitions of heterogeneity and different modelling assumptions. In this article, we focus on testing whether the average treatment effects across multiple prespecified subpopulations are identical to each other. The earliest work towards this goal was the likelihood ratio test (LRT) developed by Gail & Simon (1985) under normality assumptions for the stratum-specific treatment effect estimates. Regression methods have also been considered, where the heterogeneity of treatment effects is tested by examining interaction terms between treatment assignment and potential effect modifiers (Krishnan, Sokka & Hannonen, 2003). More recently, several nonparametric approaches have been proposed in the literature. Crump et al. (2008) proposed a test based on sieve estimation for treatment effects. This method was later generalized by Sant’Anna (2020) to test for heterogeneity in duration outcomes under endogenous treatment assignment. More recently, Dai & Stern (2022) proposed a $U$-statistic-based test ($U$-test) that does not require estimating stratum-specific treatment effects. Compared with the LRT and other parametric tests, the nonparametric tests in general require weaker modelling assumptions on the outcome distributions. However, they still either require a specified model for estimating the treatment effects (Crump et al., 2008; Sant’Anna, 2020) or only consider situations where baseline covariates are well balanced within each stratum (Dai & Stern, 2022). Motivated by these observations, we propose a nonparametric test that bypasses the need for estimating treatment effects while still being applicable to observational studies where there exist confounding variables that need to be addressed.

In this article, we focus on testing the equality of the average treatment effects across multiple strata while adjusting for potential confounding variables in observational studies. We propose a new testing procedure based on an adjusted four-sample $U$-statistic that can be viewed as a weighted version of the original $U$-statistic developed by Dai & Stern (2022). Assuming the strata are mutually independent, the main idea is to first construct an adjusted $U$-statistic for comparing the treatment effects between two strata, and then formulate an overall test statistic as a function of those pairwise adjusted $U$-statistics. For each stratum, the weights in the adjusted $U$-statistic are carefully chosen by covariate matching and propensity score estimation (Li, Morgan & Zaslavsky, 2018) such that the baseline covariate distributions for both the treatment and control groups are the same as the marginal distribution for the target population. To derive the asymptotic distribution for the proposed test, we find the main challenge is that our adjusted $U$-statistic no longer belongs to the generalized $U$-statistic family; therefore, classical projection theory is not directly applicable. To solve this problem, we use an idea from Satten, Kong & Datta (2018), a manuscript that studies adjusted two-sample $U$-statistics, to obtain an asymptotic normality result. Based on the derived asymptotic theory, we then conduct several numerical studies to compare the performance of our proposed test with that of the LRT (Gail & Simon, 1985) and the unadjusted $U$-test (Dai & Stern, 2022). Numerical results confirm the excellent operating characteristics for the proposed method even under propensity score model misspecification, and also clearly demonstrate the advantage of our method over the LRT and the unadjusted $U$-test when the data are generated from a non-Gaussian distribution or when the baseline covariates are not well balanced.

The remainder of the article is structured as follows. In Section 2, we provide a review of the $U$-test that assesses treatment effect heterogeneity across strata with balanced baseline
covariates. In Section 3, we introduce our adjusted $U$-test for treatment effect heterogeneity that allows for the existence of confounding variables. In Section 4, we conduct simulation studies to demonstrate the asymptotic validity and efficiency of the adjusted $U$-test, and also explore the impact of model misspecification. In Section 5, we further demonstrate the use of our method through two case studies: an employment programme evaluation study in labour economics, and another study on the evaluation of China’s one-child policy on children’s mental health. We conclude with some remarks in Section 6.

2. REVIEW OF UNADJUSTED $U$-STATISTIC-BASED TEST FOR TREATMENT EFFECT HETEROGENEITY

Dai & Stern (2022) proposed a $U$-statistic-based test ($U$-test) to assess the consistency of average treatment effects in several independent strata, assuming there are no confounding variables. Compared with its parametric counterpart, the likelihood ratio test (LRT) introduced by Gail & Simon (1985), their proposed $U$-test can retain power better when the outcomes deviate far away from a normal distribution. As the method we propose in this article is based on their $U$-test, we start with a review of their method.

Assume there are $S$ strata. Within each stratum $s$ ($s \in \{1, \ldots, S\}$), let $\tau_s$ be the additive treatment effect, $Y^t_i = \{Y^t_{si}, i = 1, \ldots, n^t_s\}$ be the outcomes of subjects in the treatment group, and $Y^c_i = \{Y^c_{si}, i = 1, \ldots, n^c_s\}$ be the outcomes of subjects in the control group. The total sample size across all strata is denoted as $N = \sum_{s=1}^{S} (n^t_s + n^c_s)$. To test for treatment effect heterogeneity across all strata, Dai & Stern (2022) consider the hypotheses

$$H_0 : P\left( Y^t_p - Y^c_p < Y^t_q - Y^c_q \right) + \frac{1}{2} P\left( Y^t_p - Y^c_p = Y^t_q - Y^c_q \right) = \frac{1}{2}$$

for any $1 \leq p < q \leq S$

$$\iff H_a : \text{the equation does not hold for at least one pair of } (p, q).$$

(1)

When $Y^t_s - Y^c_s$ ($s = 1, \ldots, S$) follow a common distribution up to a location shift, or $Y^t_s$ and $Y^c_s$ follow the same distribution up to a stratum-specific location shift within each stratum $s$ ($s = 1, \ldots, S$), the hypotheses in (1) are equivalent to

$$H_0 : \tau_1 = \cdots = \tau_S \iff H_a : \text{there exist } i, j \text{ such that } \tau_i \neq \tau_j,$$

where $\tau_s = \mathbb{E}(Y^t_s) - \mathbb{E}(Y^c_s)$. More discussions about these hypotheses can be found in Section 3.3 of Dai & Stern (2022).

The test statistic is constructed by combining all pairwise $U$-statistics that compare treatment effects in two strata. To compare the treatment effects in the first two strata, a four-sample $U$-statistic is constructed as

$$U^{(1,2)} = \frac{1}{n^t_1 n^c_1 n^t_2 n^c_2} \sum_{i=1}^{n^t_1} \sum_{j=1}^{n^c_1} \sum_{k=1}^{n^t_2} \sum_{l=1}^{n^c_2} \phi^{(1,2)}(i, j, k, l),$$

(3)

where the kernel function $\phi^{(1,2)}(i, j, k, l) = I(Y^t_i - Y^c_i < Y^t_j - Y^c_j) + \frac{1}{2} I(Y^t_i - Y^c_i = Y^t_j - Y^c_j)$. The latter term is used to account for possible ties for discrete distributions. Although Dai & Stern (2022) focus on additive treatment effects, other forms of treatment effects, such as the ratio of outcomes between different treatment groups, can also be incorporated. Dai & Stern (2022) show that under three assumptions: (1) the outcomes $(Y^t_1, Y^t_2, Y^c_1, Y^c_2)$ are mutually independent; (2) there exist positive constants $0 < \lambda_s < 1$ for every $s \in \{1, 2\}$ and $\omega \in \{t, c\}$.

DOI: 10.1002/cjs.11728

The Canadian Journal of Statistics / La revue canadienne de statistique
such that \( n_{ho}^o \rightarrow \lambda_h \) as \( N \rightarrow \infty \); (3) \( 0 < \text{Var}(h_{s,o}^{(1,2)}(Y_{s,o})) < \infty \) for \( s \in \{1, 2\} \) and \( \omega \in \{t, c\} \), we have

\[
\sqrt{N} (U^{(1,2)} - \theta^{(1,2)}) \xrightarrow{D} \mathcal{N}(0, \sigma_{1,2}^2), \quad \text{when } N \rightarrow \infty, \tag{4}
\]

where \( \theta^{(1,2)} = \mathbb{E}(U^{(1,2)}) \), \( h_{s,o}^{(1,2)}(x) = \mathbb{E}[\phi^{(1,2)}(1, 1, 1, 1)|Y_{s,o} = x] - \theta^{(1,2)} \) for \( s \in \{1, 2\} \) and \( \omega \in \{t, c\} \), and \( \sigma_{1,2}^2 = \frac{1}{N} \text{Var}(h_{1,1}^{(1,2)}(Y_{1}^t)) + \frac{1}{N} \text{Var}(h_{1,1}^{(1,2)}(Y_{1}^c)) + \frac{1}{N} \text{Var}(h_{1,2}^{(1,2)}(Y_{2}^c)) + \frac{1}{N} \text{Var}(h_{2,2}^{(1,2)}(Y_{2}^c)) \) is the asymptotic variance of \( \sqrt{N} U^{(1,2)} \). Under the null hypothesis in (1), the expectation of \( \phi^{(1,2)}(i, j, k, l) \) is \( \frac{1}{2} \), thus \( \theta^{(1,2)} \) is also \( \frac{1}{2} \).

With \( S \) strata, all pairwise \( U \)-statistics \( U^{(p,q)} (1 \leq p < q \leq S) \) can be constructed in the exact same way. Specifically, for every pair of \( (p, q) \), we can define \( U^{(p,q)} \), \( \theta^{(p,q)} \) and \( h_{s,o}^{(p,q)} \) \( (\omega \in \{t, w\}, s \in \{p, q\}) \) similarly with \( U^{(1,2)} \), \( \theta^{(1,2)} \) and \( h_{s,o}^{(1,2)} \) by replacing \( (1, 2) \) with \( (p, q) \). Assuming for each pair of strata \( (p, q) \), the three assumptions before Equation (4) are satisfied, Dai & Stern (2022) show that when \( N \rightarrow \infty \),

\[
\sqrt{N} (U^{(1,2)} - \theta^{(1,2)}, U^{(1,3)} - \theta^{(1,3)}, \ldots, U^{(S-1,S)} - \theta^{(S-1,S)})^T \xrightarrow{D} \mathcal{N}(0, \Sigma), \tag{5}
\]

where \( \Sigma = \frac{1}{2} \Sigma^t + \frac{1}{2} \Sigma^c + \cdots + \frac{1}{2} \Sigma^c + \frac{1}{2} \Sigma^c + \Sigma^S \) and \( \Sigma^o_s \) is the covariance matrix of \( \left( \hat{h}_{s,o}^{(1,2)}(Y_{s,o}), \hat{h}_{s,o}^{(1,3)}(Y_{s,o}), \ldots, \hat{h}_{s,o}^{(S-1,S)}(Y_{s,o}) \right) \) for all \( s \in \{1, \ldots, S\} \) and \( \omega \in \{t, c\} \). Here, \( \hat{h}_{s,o}^{(p,q)}(x) = h_{s,o}^{(p,q)}(x)(I(s = p \text{ or } s = q) \)

To apply this method, \( \Sigma \) is estimated by a weighted average of \( \Sigma^o_s \) \( (s \in \{1, \ldots, S\}, \omega \in \{t, c\} \). For each \( s \) and \( \omega \), \( \Sigma^o_s \) can be estimated by the corresponding sample covariance matrix. As the \( \hat{h} \) terms are unknown, they need to be estimated as well. Though \( h_{s,o}^{(p,q)}(x) = \mathbb{E}[\phi^{(p,q)}(i, j, k, l)|Y_{s,o} = x] - \theta^{(p,q)} \), the constant term \( \theta^{(p,q)} \) can be ignored when calculating the covariance matrices. So Dai & Stern (2022) take the method-of-moment estimator for the expectation term \( \mathbb{E}[\phi^{(p,q)}(i, j, k, l)|Y_{s,o} = x] \) as the estimator of \( h_{s,o}^{(p,q)}(x) \). For instance, the estimator of \( \hat{h}_{1,1}^{(1,2)}(x) = \hat{h}_{1,2}^{(1,2)}(x) = \frac{1}{n_{1,t}^o n_{1,c}^o} \sum_{i=1}^{n_{1,t}^o} \sum_{j=1}^{n_{1,c}^o} \sum_{k=1}^{n_{1,c}^o} \sum_{l=1}^{n_{1,t}^o} \left\{ I(x - Y_{i,c} < Y_{j,t} - Y_{k,c}) + \frac{1}{2} I(x - Y_{i,c} > Y_{j,t} = Y_{k,c}) \right\} \right\}. \]

Similar calculations are repeated for all other \( h \) functions, and then used for computing the sample covariance \( \hat{\Sigma}^o_s \) \( (s \in \{1, \ldots, S\}, \omega \in \{t, c\} \), which leads to the final estimator of \( \Sigma \) as \( \hat{\Sigma} = \frac{1}{2} \hat{\Sigma}^t + \frac{1}{2} \hat{\Sigma}^c + \cdots + \frac{1}{2} \hat{\Sigma}^c + \frac{1}{2} \hat{\Sigma}^c + \frac{1}{2} \hat{\Sigma}^c \).

To test the null hypothesis \( H_0 : \theta = \frac{1}{2} 1_{S(S-1)/2} \), where \( \theta = (\theta^{(1,2)}, \theta^{(1,3)}, \ldots, \theta^{(S-1,S)})^T \), Dai & Stern (2022) focuses on a one-dimensional overall test statistic \( U_h = N \cdot \sum_{1 \leq p < q \leq S} (U^{(p,q)} - \frac{1}{2})^2 \). Though the asymptotic reference distribution of \( U_h \) does not have an analytic form, it can be approximated by simulation, that is, after generating a large number of independent samples \( \{r_1, \ldots, r_L\} \) from \( \mathcal{N}(0, \hat{\Sigma}) \), the empirical distribution of \( \{|r_1|^2, \ldots, |r_L|^2\} \) approximates the asymptotic reference distribution of \( U_h \).

3. ADJUSTED U-TEST OF TREATMENT EFFECT HETEROGENEITY

The test described in Section 2 can only be used in situations where all baseline covariates are well balanced between different treatment groups in each stratum, such as in stratified randomized experiments. In observational studies, directly applying that method may lead to misleading conclusions due to the existence of potential confounding variables. Even in the
3.1. Notation, Assumptions and Setup

We introduce some additional notation here. For each stratum \( s \), where \( s \in \{1, \ldots, S\} \), we use \( X^t_s = \{X^t_{si}, i = 1, \ldots, n^t_s\} \) to denote the collection of baseline covariates for subjects in the treatment group where the first element of each vector \( X^t_{si} \) is 1, corresponding to an intercept term. Similarly \( X^c_s = \{X^c_{si}, i = 1, \ldots, n^c_s\} \) is used to denote the covariates for subjects in the control group. Let \( X_s = X^t_s \cup X^c_s \) be the collection of covariates for all subjects in stratum \( s \), where we assume the first \( n^t_s \) elements are from the treatment group, and the rest are from the control. We use \( T_s = \{T_{si}, i = 1, \ldots, n_s\} \) to denote the indicators of treatment, i.e., the first \( n^t_s \) elements are ones and the rest are zeroes. The within-stratum propensity score, \( P(T_s = 1|X_s) \), is denoted by \( e(X_s) = \{e(X_{si}), i = 1, \ldots, n_s\} \). Similarly, \( e(X^t_s) = \{e(X^t_{si}), i = 1, \ldots, n^t_s\} \) denotes the first \( n^t_s \) elements in \( e(X_s) \) and \( e(X^c_s) = \{e(X^c_{si}), i = 1, \ldots, n^c_s\} \) denotes the rest. Unlike Dai & Stern (2022), we allow the covariate distribution in the treatment group to be different from that in the control group for each stratum. Therefore, our test for treatment effect heterogeneity is based on the comparison between outcome distributions conditioning on the baseline covariates.

We make the following assumptions for the rest of this article.

\(\text{(A1) Stable unit treatment value assumption (SUTVA): The subject outcomes are independent of each other and there is only one version of each treatment option, that is, there is no variation in each treatment, such as dosage.}\)

\(\text{(A2) No unmeasured confounding: For each study subject in stratum } s \text{, the potential outcomes are independent with the treatment assignment conditional on the covariates } X_s.\)

\(\text{(A3) Positivity: } 0 < e(X_s) < 1 \text{ for all } s \in \{1, \ldots, S\}.\)

\(\text{(A4) For each stratum } s \text{ and treatment assignment } \omega \in \{t, c\}, \text{ the outcomes } Y^\omega_{si} \text{ are assumed to be i.i.d. for } i \in \{1, \ldots, n_s^\omega\}.\)

Assumptions (A1)–(A3) are standard assumptions in the literature to make the causal estimands identifiable. Assumption (A4) is a standard sampling assumption.

3.2. Balancing Baseline Covariates Within One Stratum

One way to balance confounding variables is to weight the subjects such that within each stratum all baseline covariates from the two treatment groups have the same distributions. As we assume the strata are mutually independent, here we only focus on how to balance the covariates in one stratum, and the same process can be applied to the others. For simplicity, here we omit the stratum indicator \( s \) in the subscript. In one stratum, for baseline covariate \( X \), let its marginal density function (or probability mass function if \( X \) is discrete) be \( f(x) \), and its conditional density functions (or probability mass functions) in the treatment and control groups be \( f^t(x) \) and \( f^c(x) \), respectively. Our goal is to find weight functions, \( w^t(x) \) and \( w^c(x) \), in the treatment and control groups, respectively, such that \( f^t(x)w^t(x) = f^c(x)w^c(x) \). As discussed in Li, Morgan & Zaslavsky (2018), different choices of weight functions will lead to different target populations of interest. They propose the use of a general function \( h(x) \) to define the population of interest with \( h(x)f(x) \) as its marginal distribution. For example, when \( h(x) = 1 \), the target population has a marginal distribution of \( f(x) \), which corresponds to the distribution of \( X \) in the combined population of treatment and control groups. When \( h(x) = e(x) \) or \( 1 - e(x) \), the target population
refers to the subjects in the treatment or control groups. And when \( h(x) = e(x)(1 - e(x)) \), the
target population is the so-called overlap population (Li, Morgan & Zaslavsky, 2018).

For a given \( h(x) \), the weight functions \( w'(x) \) and \( w^c(x) \) should satisfy

\[
w'(x)f^4(x) \propto w^c(x)f^c(x) \propto f(x)h(x).
\]

Since \( f^4(x) \propto f(x)e(x) \) and \( f^c(x) \propto f(x)(1 - e(x)) \), (6) implies

\[
w'(x) \propto \frac{h(x)}{e(x)}, \quad \text{and} \quad w^c(x) \propto \frac{h(x)}{1 - e(x)}.
\]

When \( h(x) = 1 \), the induced weight functions yield the classical ability weighting (IPW) (Horvitz
& Thompson, 1952).

The aforementioned weighting method can be incorporated in \( U \)-statistics as well. For
example, Satten, Kong & Datta (2018) adopted it to adjust two-sample \( U \)-statistics with the
goal of testing for treatment effect in observational studies. For our study, we
also use this method to adjust the pairwise \( U \)-statistics introduced in Section 2 in order to test
for treatment effect heterogeneity in observational studies. We take \( U^{(1,2)} \) in Equation (3) as an
example, which is the average of several kernel functions. Each kernel function \( \phi^{(1,2)}(i, j, k, l) \)
is constructed by the outcomes of four independent subjects, and each subject needs to be
weighted. As the outcomes are independent, \( \phi^{(1,2)}(i, j, k, l) \) should be weighted by the product
of the weights for the four subjects: \( w'(X_{ij}) \cdot w^c(X_{ij}) \cdot w'(X_{kl}) \cdot w^c(X_{kl}) \).

The choice of the weight functions depends on \( h(x) \), which in principle can be chosen
as any positive function. However, we further require \( h(x) \) to be a constant or a function of
\( e(x) \), and we require it to be differentiable with respect to \( e(x) \). These requirements will later
greatly help with the efficient estimation of the asymptotic reference distribution for the adjusted
\( U \)-statistics without requiring approximation or sampling methods such as bootstrap. In practice,
the choice of \( h(x) \) is determined by the target population of interest. For example, if we are
interested in average treatment effects (ATEs), then \( h(x) \) should be chosen as 1. If we are
interested in the average treatment effect on the treatment group, then we can choose \( h(x) \) as
\( e(x) \). In our simulation study in Section 4 and the application study on only children’s mental
health in Section 5.2, we focus on \( h(x) = 1 \). In the employment programme evaluation study in
Section 5.1, we choose \( h(x) = e(x) \).

In practice, the propensity scores are unknown, and it is common to use a logistic regression
model between treatment indicators and associated covariates \( X_s \) for their estimation. Formally,
within stratum \( s (s \in \{1, \ldots, S\}) \), we consider the following model with parameter \( \beta_s \),

\[
\log \left( \frac{e(X_{si})}{1 - e(X_{si})} \right) = \beta_s^T X_{si}, \quad i = 1, \ldots, n_s.
\]

Note the model specification here is flexible and can be extended to include quadratic (or
other nonlinear) functions of \( X_s \) and interaction terms as needed. The model does not impose
any assumptions on the response variable, and in practice it is convenient to conduct model
diagnostics for (8), such as those based on Austin (2008). The estimate of \( \beta_s \), denoted by \( \hat{\beta}_s \), can be
obtained by solving the estimating equation of logistic regression, denoted as \( \sum_{j=1}^{n_s} S_{sj}(\hat{\beta}_s) = 0 \).

As the propensity scores are functions of \( \beta_s \), for simplicity, we denote the weights for subjects
in the treatment and control groups by \( w'_{sj}(\beta_s) \ (i = 1, \ldots, n_s' \) and \( w^c_{sj}(\beta_s) \ (i = 1, \ldots, n_s^c \),
respectively for \( s \in \{1, \ldots, S\} \). In practice, these weights can be estimated by their plug-in
estimates.
3.3. Testing Treatment Effect Heterogeneity Between Two Strata

We start with constructing a test statistic that compares the treatment effects between the first two strata. After weighting, the $U$-statistic in (3) becomes

$$U_a^{(1,2)} = \frac{1}{n_1 n_2 n_1^c n_2^c} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_1^c} \sum_{l=1}^{n_2^c} w_i^c w_j^c w_k^c w_l^c \left( \frac{\hat{\beta}_i \hat{\beta}_j \hat{\beta}_k \hat{\beta}_l \phi_{i,j,k,l}}{\hat{\beta}_i \hat{\beta}_j \hat{\beta}_k \hat{\beta}_l} \right),$$

where $\overline{w}_s^c(\hat{\beta}_s) = \frac{1}{n_s^c} \sum_{i=1}^{n_s^c} w_i^c(\hat{\beta}_s)$ for $s \in \{1, \ldots, S\}$ and $\omega \in \{t, c\}$.

Though $U_a^{(1,2)}$ looks like a generalized $U$-statistic (Korolyuk & Borovskich, 2013), unfortunately it is not, because $\hat{\beta}_1$ and $\hat{\beta}_2$ are functions of all outcomes in the corresponding strata. Therefore, the classical projection theorem cannot be directly applied to $U_a^{(1,2)}$. The key observation is that, if we replace $\hat{\beta}_1$ and $\hat{\beta}_2$ by their respective estimands, $\beta_1$ and $\beta_2$, then we obtain a generalized $U$-statistic. Moreover, if $\hat{\beta}_1$ and $\hat{\beta}_2$ are consistent estimates, we would expect the asymptotic properties (e.g., normality) of the generalized $U$-statistics will still hold for our adjusted $U$-statistic. This is indeed the case by the following theorem. The proof is in Appendix A and it is based on the idea in Satten, Kong & Datta (2018) where they derived the asymptotic normality for adjusted two-sample $U$-statistics.

**Theorem 1.** Suppose that $\hat{\beta}_1$ and $\hat{\beta}_2$ are consistent estimators for $\beta_1$ and $\beta_2$, respectively. We make the following assumptions: (1) the outcomes $(Y_1, Y_2, Y_1', Y_2')$ are mutually independent; (2) there exist positive constants $0 < \lambda_s < 1$ for every $s \in \{1, 2\}$ and $\omega \in \{t, c\}$ such that $\frac{\omega_s}{n_s} \to \lambda_s$ as $n_1 + n_2 \to \infty$; (3) $0 < e(X_s) < 1$ for all $s \in \{1, 2\}$ where $e(X_s)$ is defined in Section 3.1; and (4) $0 < \text{Var}\left[\eta_s^{(1,2)}(Y_s')\right] < \infty$ for $s \in \{1, 2\}$ and $\omega \in \{t, c\}$, where the $\eta$ functions are defined in the proof in Appendix A. Then as $n_1 + n_2 \to \infty$, we have

$$\sqrt{(n_1 + n_2)} \left( U_a^{(1,2)} - \theta_a^{(1,2)} \right) \xrightarrow{D} \mathcal{N} \left( 0, \sigma_{1,2}^2 \right),$$

where $\theta_a^{(1,2)} = \lim_{n_1 + n_2 \to \infty} E \left[ U_a^{(1,2)} \right]$ and $\sigma_{1,2}^2 = \lim_{n_1 + n_2 \to \infty} \text{Var} \left[ \eta_1^{c(1,2)}(Y_1') \right] + \frac{n_1 + n_2}{n_1^2} \text{Var} \left[ \eta_2^{c(1,2)}(Y_2') \right] + \frac{n_1 + n_2}{n_2^2} \text{Var} \left[ \eta_2^{c(1,2)}(Y_2') \right]$ is the asymptotic variance of $\sqrt{(n_1 + n_2)} U_a^{(1,2)}$.

Theorem 1 establishes the asymptotic distribution for our proposed adjusted $U$-statistic. Assumptions (1)–(4) are mild and commonly used in the literature. For example, Assumption (2) requires that within each stratum, the proportion of each group is not negligible, which is satisfied in most applications. Assumption (3) requires the propensity score to be bounded away from 0 and 1, which is called probabilistic assignment and is commonly used in the causal inference literature (Imbens & Rubin, 2015).

To estimate the asymptotic variance $\sigma_{1,2}^2$, we first estimate the $\eta$ values, denoted by $\left\{ \hat{\eta}_s^{(1,2)}(Y_s'), i = 1, \ldots, n_s^\omega \right\}$ for $s \in \{1, 2\}$, by replacing all $\beta$s by their consistent estimates and replacing $\bar{h}$ functions by their method-of-moment estimators in the same way as discussed in Section 2. Then we use the sample variance of each set $\left\{ \hat{\eta}_s^{(1,2)}(Y_s'), i = 1, \ldots, n_s^\omega \right\}$.
to estimate $\text{Var}[\eta_s^{\alpha_s(1,2)}(Y_{s}^o)]$, i.e.,
\[
\text{Var}[\eta_s^{\alpha_s(1,2)}(Y_{s}^o)] = \frac{1}{n_s-1} \sum_{i=1}^{n_s} (\eta_s^{\alpha_s(1,2)}(Y_{s}^o) - \bar{\eta}_s^{\alpha_s(1,2)}(Y_{s}^o))^2,
\]
where $\bar{\eta}_s^{\alpha_s(1,2)}(Y_{s}^o)$ is the average of $\{\eta_s^{\alpha_s(1,2)}(Y_{s}^o), i = 1, \ldots, n_s\}$. Then $\sigma_{1,2}^2$ can be consistently estimated by
\[
\hat{\sigma}_{1,2}^2 = \frac{n_1+n_2}{n_1^2} \text{Var}[\eta_1^{\alpha_s(1,2)}(Y_1^o)] + \frac{n_1+n_3}{n_1^2} \text{Var}[\eta_1^{\alpha_s(1,2)}(Y_1^c)] + \frac{n_1+n_2}{n_2^2} \text{Var}[\eta_2^{\alpha_s(1,2)}(Y_2^o)] + \frac{n_1+n_3}{n_2^2} \text{Var}[\eta_2^{\alpha_s(1,2)}(Y_2^c)].
\]

3.4. Testing Treatment Effect Heterogeneity in Multiple Strata

Next we consider testing for treatment effect heterogeneity in multiple strata, $1, 2, \ldots, S$, with $S > 2$, by extending the adjusted $U$-statistic in the previous section. For every pair of strata $p$ and $q$ satisfying $1 \leq p < q \leq S$, we can define an adjusted $U$-statistic $U_{a}^{(p,q)}$ in the same way as $U_{a}^{(1,2)}$. Then it is natural to consider a vector of all pairwise adjusted $U$-statistics $U_a = (U_{a}^{(1,2)}, U_{a}^{(1,3)}, \ldots, U_{a}^{(S-1,S)})^T$. In the next theorem, we derive its joint asymptotic distribution.

**Theorem 2.** Suppose that Assumptions (1)–(4) in Theorem 1 are satisfied for every stratum. Then as the total sample size $N \to \infty$,
\[
\sqrt{N}(U_a - \theta_a) \xrightarrow{D} \mathcal{N}(0, \Sigma_a),
\]
where $\theta_a = \lim_{N \to \infty} E(U_a)$ and $\Sigma_a = \frac{1}{S} \sum_s \Sigma_s + \frac{1}{S} \sum_{s' \neq s} \Sigma_{s'}$ is the asymptotic covariance matrix of $\sqrt{N} U_a$, $\Sigma_s^o$ is the covariance matrix of $\{\tilde{\eta}_s^{\alpha_s(1,2)}, \ldots, \tilde{\eta}_s^{\alpha_s(S-1,S)}\}$ for $s \in \{1, \ldots, S\}$ and $\omega \in \{t, c\}$, where $\tilde{\eta}_s^{\alpha_s(p,q)} = \eta_s^{\alpha_s(p,q)}(Y_s^o)I(s = p \text{ or } s = q)$.

The asymptotic covariance matrix $\Sigma_a$ in Theorem 2 can be conveniently estimated in a similar way as for the univariate case in Theorem 1. That is, we first estimate the $\eta$ terms and $\hat{\eta}$ functions, and then use the sample covariance matrix of estimated $\hat{\eta}_s^{\alpha_s(p,q)}$ to estimate $\Sigma_s^o$ for $s \in \{1, \ldots, S\}$ and $\omega \in \{t, c\}$.

Given the estimated $\Sigma_a$, we can construct a global test statistic by considering a transformation on $U_a$. For instance, under $H_0 : (Y_1^t - Y_1^c)|X_s$ are identically distributed for $s \in \{1, \ldots, S\}$, we have $\theta_a = \frac{1}{2} I$; therefore, a one-dimensional test statistic can be constructed as $T_a = N(U_a - \frac{1}{2}I)^T (U_a - \frac{1}{2}I)$. Though the analytic form of $T_a$’s distribution is not available, we can still approximate it via simulations. This can be done by drawing a large number of samples $\{r_1, \ldots, r_L\}$ from $\mathcal{N}(0, \hat{\Sigma}_a)$, and then using $\{(|r_1|^2, \ldots, |r_L|^2)\}$ as the empirical reference distribution. Other functions of $U_a$ such as $\sqrt{N} \max_{1 \leq p < q \leq S} |U^{(p,q)} - \frac{1}{2}|$, can also be used as the global test statistic, whose reference distribution can be approximated by simulations. In the numerical studies, we focus on using $T_a$, and propose to reject the null hypothesis when $T_a$ is greater than or equal to the 100$(1 - \alpha)$th percentile of $\{(|r_1|^2, \ldots, |r_L|^2)\}$, where $\alpha$ is a predetermined significance level. Note that an alternative test statistic to $T_a$ is the Hotelling’s $T^2$ statistic: $N(U_a - \frac{1}{2}I)^T \hat{\Sigma}_a^{-1} (U_a - \frac{1}{2}I)$, whose limiting distribution is known to be a $\chi^2$ distribution. However, we find that the performance of the alternative test is not as good as that of $T_a$ in our numerical experiments because the determinant of $\hat{\Sigma}_a$ is often very close to 0. Hence, we decide to use $T_a$ for the test. Please see Section 2.1 of the Supplementary Material for more details.

3.5. Trimming Sample

In the causal inference literature, it is common to exclude subjects with estimated propensity scores too close to 0 or 1 (Dehejia & Wahba, 1999; Crump et al., 2009; Imbens & Rubin, 2015).
This trimming procedure has been shown to effectively improve the covariate balance between different treatment groups for several reasons. One is that those subjects whose true propensity scores are equal to 0 or 1 should not be used because there are no counterparts in the alternative group. Another reason is that for those subjects whose estimated propensity scores are very close to 0 or 1, their counterparts will be associated with extremely large weights, which will then lead to a large variance for the estimated treatment effects.

There are two popular trimming rules. One is to set a hard threshold for propensity scores to be included in treatment effect estimates, for example, \([\gamma, 1 - \gamma]\) for \(0 < \gamma < \frac{1}{2}\) as recommended in Crump et al. (2009), and subjects with propensity scores outside this range should be removed. The other trimming rule is that we only use the subjects whose propensity scores are within the overlap region (Dehejia & Wahba, 1999). Specifically, we remove all subjects in the control group whose propensity scores are smaller than the minimum propensity score in the treatment group, and remove all subjects in the treatment group whose propensity scores are larger than the maximum propensity score in the control group. In practice, those two rules can be applied simultaneously. Another option is to use the robust inference procedure proposed for IPW estimators in Ma & Wang (2020). This method is shown to be robust to small weight probabilities and different trimming thresholds.

It is worth mentioning that although the trimming procedure in general improves the treatment effect estimation accuracy, it changes the reference population, hence there is a trade-off. Under this trade-off, people usually still prefer trimming because a reliable estimate for a subpopulation is generally considered more valuable than an estimate for the original population based on extrapolation or with large variance. In the simulation studies, we present both results with and without trimming to demonstrate the effect of trimming. More specifically, when implementing trimming, we first remove subjects outside of the propensity score overlap region, and then re-run the same propensity score model for the remaining subjects to obtain the weights for our adjusted \(U\)-tests. We have conducted several numerical experiments and found that the type I error is better controlled with the new propensity scores. Therefore, we choose to implement this trimming procedure for both case studies in this article.

3.6. Comparison with Satten, Kong & Datta (2018) Test

Satten, Kong & Datta (2018) proposed an adjusted \(U\)-statistic to test for the equality of distributions across multiple groups in the presence of confounding covariates. Their methodology and theory are closely related to ours, but there are differences. We discuss the two main differences here. The first difference is in the hypotheses of interest. We focus on testing for the equality of treatment effects conditional on the baseline covariates across multiple strata, while Satten, Kong & Datta (2018) compare the outcome values from different strata. For example, consider two strata, \([1, 2]\) in our case. We can divide patients into four groups based on strata and treatment assignment \([t, c]\), that is, \(\{(1, t), (1, c), (2, t), (2, c)\}\). The test of Satten, Kong & Datta (2018) checks if outcomes from the four groups, denoted by \(Y^t_1, Y^c_1, Y^t_2\) and \(Y^c_2\), all have the same distribution; while our test is about \(Y^t_1 - Y^c_1\) having the same distribution as \(Y^t_2 - Y^c_2\) conditioning on the covariates. In other words, the null hypothesis in Satten, Kong & Datta (2018) is more stringent than ours and their test cannot be directly applied to our case. The second noteworthy difference concerns the choice of \(h(x)\). Satten, Kong & Datta (2018) only consider \(h(x) = 1\) or \(e(x)\), and we generalize the results to allow \(h(x)\) to be any function of the propensity scores satisfying certain smoothness condition. Our asymptotic results are hence more general in this regard.

4. SIMULATION

We conduct simulation studies to evaluate the empirical performance of the proposed adjusted \(U\)-statistic test and compare it with the likelihood ratio test (LRT), the \(U\)-test developed in Dai &
Stern (2022), and the nonparametric tests in Sant’Anna (2020). Here, we focus on the case where the target population is the union of the treatment and control groups, that is, where \( h(x) = 1 \). We consider the adjusted \( U \)-tests with and without the trimming procedure, and denote them as AUT-T and AUT, respectively.

4.1. Implementation Details

We first discuss the computational implementation of both our proposed \( U \)-tests and the LRT. The \( U \)-test statistic in (11) is a function of \( S(S - 1)/2 \) pairwise adjusted \( U \)-statistics, and the computation of each adjusted \( U \)-statistic can be expensive in simulation studies. Therefore, instead of calculating the complete adjusted \( U \)-statistics, we randomly sample some of the \( \phi \) functions in each of the adjusted \( U \)-statistics. Take \( U_a^{(1,2)} \) in (9) as an example, for each stratum \( s \in \{1, 2\} \) and treatment group \( \omega \in \{t, c\} \), we randomly choose \( M = 1000N \) subjects with replacement, where \( N \) is the total sample size over all strata, denoted by \( \{ (y_{1i}^t, y_{1i}^c, y_{2i}^t, y_{2i}^c), i = 1, \ldots, M \} \). Then we calculate the kernel function \( \phi_i \) based on \( (y_{1i}^t, y_{1i}^c, y_{2i}^t, y_{2i}^c) \) and use the weighted average of \( \{ \phi_i, i = 1, \ldots, M \} \) to approximate \( U_a^{(1,2)} \). Because we also use the weighted kernel functions to estimate \( \hat{H}_s^\omega (Y_m^\omega) \) for \( i \in \{1, \ldots, n_m^\omega\} \), \( s \in \{1, 2\} \) and \( \omega \in \{t, c\} \), which are required to obtain \( \hat{S}_a \), we need to make sure that each subject is sampled at least once. This requirement is usually satisfied given a large sampling size \( M \), and we redo the sampling process on the rare occasion that this requirement is not met. The sampling size \( M = 1000N \) was selected by running a series of different simulation scenarios with three strata and \( N \) ranging from 60 to 3000; this choice of \( M \) ensured the variance of the approximated test statistic \( \frac{L_a}{N} = \frac{1}{\sum_{p<q \leq S} (U(p,q) - \frac{1}{2})^2} \) to be smaller than 0.003. In order to approximate the reference distribution of \( \frac{L_a}{N} \), \( 10^5 \) samples \( \{r_i, i = 1, \ldots, 10^5\} \) are generated independently from the estimated reference distribution \( \mathcal{N}(0, \frac{1}{N} \hat{S}_a) \). Then \( \{\|r_i\|^2, i = 1, \ldots, 10^5\} \) are used to obtain the empirical reference distribution \( \frac{L_a}{N} \). The sample size of \( 10^5 \) is chosen to ensure that the variance of the 95\(^{th}\) percentile of \( \{\|r_i\|^2, i = 1, \ldots, 10^5\} \) is below 0.0001.

Next we give a brief review of the LRT proposed by Gail & Simon (1985) because it is a competitive approach for testing the treatment effect homogeneity. With \( S \) strata, they test the null hypothesis that the average treatment effects \( \tau_s \) (\( s \in \{1, \ldots, S\} \)) are the same across all of the strata versus the alternative that at least two of them are unequal, that is, the hypotheses in (2). Assuming the treatment effect estimates \( \hat{\tau}_s \) (\( s \in \{1, \ldots, S\} \)) follow normal distributions as \( \hat{\tau}_s \sim \mathcal{N}(\tau_s, \sigma^2_s) \), then a test statistic is constructed as

\[
H = \sum_{s=1}^S (\hat{\tau}_s - \hat{\tau})^2 / s^2 \chi^2_{S-1},
\]

where \( \hat{\tau} = \frac{\sum_{s=1}^S \hat{\tau}_s / s^2}{\sum_{s=1}^S 1/s^2} \), and \( s^2 \) is a consistent estimator of \( \sigma^2_s \) for \( s \in \{1, \ldots, S\} \). For an \( \alpha \)-level test, we reject the null hypothesis when \( H \) is greater than or equal to the 100(1 - \( \alpha \))th percentile of \( \chi^2_{S-1} \).

In randomized experiments where we can directly compare the outcomes of different treatment groups to estimate the treatment effect, \( \hat{\tau}_s \) can be the difference of the outcome averages. In observational studies, a method for estimating \( \hat{\tau}_s \) that adjusts for confounding variables should be used. Any methods that can provide a normally distributed \( \hat{\tau}_s \), and a consistent estimator for \( \sigma^2_s \) in stratum \( s \) for \( s \in \{1, \ldots, S\} \) can be used. For instance, when the outcome follows a continuous distribution, a linear regression model between the outcome and the treatment indicator and other confounding variables can be fitted within each stratum. Under the assumption that the outcomes
are independent, the normality assumption for $\hat{\tau}_s$ will be satisfied when the stratum sample size $n_s$ goes to infinity. In this simulation, we fit a linear regression in each stratum $s$ ($s \in \{1, \ldots, S\}$) to obtain $\hat{\tau}_s$ and $\hat{\sigma}^2_s$. We focus on the case that $Y'_s - Y'_c|X_s$ ($s \in \{1, \ldots, S\}$) follow the same distribution up to a location shift. Thus, the hypotheses of the adjusted $U$-tests are equivalent to those of the LRT; hence those two tests are directly comparable.

The nonparametric test of Sant’Anna (2020) is based on directly estimating $h(X) = E[Y|X,T = 1] - E[Y|X,T = 0]$ and testing whether $h(X)$ is equal to a constant. To implement Sant’Anna’s test, we add stratum indicators, i.e., $I(s = 2)$ and $I(s = 3)$, as new covariates (assuming $S = 3$) and then include them in the propensity score model. We implement both types of tests found in Sant’Anna (2020): Cramér–von Mises-type test and Kolmogorov–Smirnov-type test.

4.2. Simulation Design

We consider three strata ($S = 3$), where each stratum has the same sample size, $n_1 = n_2 = n_3 = n$. For each stratum $s$, we generate the data from an outcome model $Y_s = 1 + \beta_{s,1} T_s + Z_s + \epsilon_s$ for $s \in \{1, 2, 3\}$, where the treatment indicator $T_s \sim \text{Bernoulli}(p_s)$, and the residual terms $\epsilon_s$ follow a common distribution $F_{\epsilon}$ across all strata. The probability of being assigned to the treatment group $p_s$ is also a function of the confounding variable $Z_s$, for which we assume $\text{logit}(p_s) = \gamma_s Z_s$. In the following simulations, we fix $Z_1 \sim \mathcal{N}(0, 1)$, $Z_2 \sim \mathcal{N}(0, 1)$, and $Z_3 \sim \text{Unif}(-0.5, 0.5)$, and choose $\gamma_1 = 1$, $\gamma_2 = -1$, and $\gamma_3 = 1$, such that the confounding variables either follow different distributions or satisfy different relationships with the treatment assignment among the three strata. Also, the treatment effects are set to $\beta_{1,1} = 1$, $\beta_{2,1} = 1 + \Delta$, $\beta_{3,1} = 1 + 2\Delta$, where the constant $\Delta$ is treated as the effect size. Note that when $\Delta = 0$, there still exists a treatment effect within each stratum although there is no treatment effect heterogeneity, in other words, the null hypothesis is true. For all of the simulation scenarios, we fix the significance level at 0.05, and repeat the data generating mechanism $L = 2000$ times to obtain the empirical rejection rates.

In addition to the simulation design described above, we also consider several other designs with unequal sample size and different error distributions across the three strata. The simulation designs and results are very similar to those in Dai & Stern (2022), so we choose not to present them in this article.

4.3. Simulation Results

We first check the type I error of our proposed adjusted $U$-test with and without trimming (AUT-T and AUT) when $\Delta = 0$, $n = 200$, $F_{\epsilon} = \mathcal{N}(0, 1)$, and compare them with the unadjusted $U$-test reviewed in Section 2. Based on 2000 Monte-Carlo replications, the type I error rates for the AUT-T and AUT are 0.051 and 0.058: both are very close to the nominal level of 0.05, whereas the unadjusted $U$-test has a (false) rejection rate of 1.000. The invalidity of the unadjusted $U$-test is not surprising, because the unweighted outcome distributions are quite different between treatment and control groups in each stratum, as shown in Figure 1. This finding clearly demonstrates the need for confounder adjustment when testing for treatment effect heterogeneity. In Figure 2, we plot the empirical $P$-values with the expected uniformly distributed $P$-values for both the AUT-T and AUT methods, and the two nonparametric tests by Sant’Anna (2020). We find that the empirical distribution for the $P$-values is very close to the uniform distribution under the null hypothesis, which confirms both the validity of the asymptotic null distribution derived in Theorem 2 and the accuracy of random sampling when calculating the test statistics. Compared with AUT, the results for AUT-T are less perfect due to the fact that the population has changed after trimming the propensity score. To demonstrate the effect of trimming, we present the average number of removed subjects for each stratum in Table 1, and find that the effect of trimming is minor because less than 8% of the subjects are removed from each stratum. For the tests in Sant’Anna (2020), we did not perform trimming.
FIGURE 1: Density plots for the unadjusted outcomes in the treatment and control groups.

FIGURE 2: Empirical and expected $P$-values for our proposed $U$-tests (AUT and AUT-T) and the tests proposed by Sant’Anna (2020) (Cramer-von Mises and Kolmogorov-Smirnov) under the null hypothesis.
Next we investigate the power for the proposed adjusted $U$-tests under different values for the sample size $n$, effect size $\Delta$ and error distributions $F_e$. We also use the results from the regression-based LRT and the nonparametric tests by Sant’Anna (2020) as a benchmark for power comparison. We choose four distributions for $F_e$: $\mathcal{N}(0, 1)$, Unif($-2, 2$), $t_4$, and $0.5\mathcal{N}(-5,1) + 0.5\mathcal{N}(5,1)$. For each of them, we consider four effect sizes (including 0) and then present the empirical rejection rates for the five tests in Figure 3. We first note that under all four scenarios, the type I error rates are very close to the nominal level of 0.05. There is a minor discrepancy for the trimmed $U$-test, especially when the sample size is small. This is to be expected because trimming changes the reference population, although the number of trimmed subjects (see Figure 4) is quite small (between 2% and 15%). Therefore, it is fair to compare the power of those three tests given that their type I errors are at the same level.

When effect size $\Delta > 0$, we first notice that the power increases quickly to its maximum of one as either the sample size $n$ or the effect size $\Delta$ increases for the adjusted $U$-tests (AUT-T and AUT) and the LRT. The two nonparametric tests by Sant’Anna (2020) (marked by lines with triangles and circles) clearly have lower power under all four scenarios. By comparing the power between the two adjusted $U$-tests (AUT-T and AUT), we find that the overall AUT-T has a higher power, although the advantage is not significant. This is to be expected because we only remove a small percentage of subjects by trimming. We then compare the power of the AUT and LRT, and find that LRT is more powerful than AUT if the error distribution $F_e$ is normal or has lighter tails than normal distribution (e.g., uniform distribution). On the other hand, our proposed AUT is more powerful than LRT when $F_e$ has heavy tails (e.g., $t_4$) or large deviations from a normal distribution (e.g., a bimodal distribution such as $0.5\mathcal{N}(-5, 1) + 0.5\mathcal{N}(5, 1)$). Those findings confirm that the LRT is still the most powerful test under the normality assumption. However, our proposed method will gain efficiency in testing against the null hypothesis as the true error distribution starts to move away from a normal distribution, with a more significant improvement in power over LRT when the error distribution becomes bimodal.

### 4.4. Sensitivity Analysis

Because our proposed adjusted $U$-test is based on a propensity score model, in this section, we conduct a sensitivity analysis to evaluate the performance of our method under misspecification of the propensity score model. Despite recent advances in propensity score model diagnosis (Imbens & Rubin, 2015; Vegetabile, Gillen & Stern, 2020) by measuring the degree of covariance balance from the weighted samples in the treatment and control groups, measuring covariate balance still remains challenging, especially when the number of covariates is large. Therefore, it remains important to explore the sensitivity of the proposed test. We consider several different null cases where there is no treatment effect heterogeneity, and explore the sensitivity of the adjusted $U$-tests with and without trimming by checking the distributions of empirical $P$-values when the propensity score models are misspecified.

For data generation, we consider three strata ($S = 3$), each with a sample size of 200, and a confounding variable $Z_s$ ($s = 1, 2, 3$) in each stratum satisfying $Z_1 \sim \mathcal{N}(0, 0.5^2)$.
Figure 3: Power analysis: empirical rejection rates for five tests under various error distributions, sample sizes and effect sizes, based on 2000 Monte-Carlo replications.
The sample size for each stratum is 200. The confounder $Z_s$ with $T_s$ with $Z_s \sim \text{Bernoulli}(p_s)$ and $\epsilon_s \sim F_e$ for $s \in \{1, 2, 3\}$. Note that there is no treatment effect heterogeneity in this scenario, that is, the null hypothesis is true. Furthermore, we set $\beta_{1,2} = \gamma_{1,2} = 2$, $\beta_{2,2} = \gamma_{2,2} = -2$, $\beta_{3,2} = \gamma_{3,2} = 2$, $\gamma_{1,0} = -0.5$, $\gamma_{2,0} = 0.5$, and $\gamma_{3,0} = -1/6$ to make the coefficients for $Z$ and $Z^2$ the same in both outcome and propensity score models in every stratum. Values of $\gamma_{s,0}$ $(s = 1, 2, 3)$ are chosen to avoid propensity scores being too close to 0 or 1. Here we explore the extent to which the empirical distributions of $P$-values for the adjusted $U$-tests deviate from the expected uniform distribution when the propensity model is fitted without the quadratic term. As with earlier simulations, we consider four choices for the error distribution $F_e$ as $\mathcal{N}(0, 1)$, $\text{Unif}(-2, 2)$, $t_4$, and $0.5\mathcal{N}(-2, 1) + 0.5\mathcal{N}(2, 1)$.

Figure 5 shows the relationship between the empirical $P$-values versus the expected uniform $P$-values for the adjusted $U$-tests with and without trimming under each of the four error distributions. AUT-T is always more robust to model misspecification than AUT. This finding suggests subject trimming based on propensity scores may improve model robustness. We also show the average number of trimmed subjects for each stratum in Table 2 and find that proportion to be reasonably small ($< 5\%$).

A possible explanation for the advantage of trimming in this simulation scenario is that after removing subjects with extreme propensity scores, a linear function can better approximate the relationship between the log-odds of the propensity scores and the confounders for the remaining subjects. To examine this conjecture and assess whether trimming helps more generally, we consider a different scenario where the $R^2$ of the linear regression $\text{logit}(p_s) \sim Z_s$ drops after trimming subjects. For data generation, again we consider three strata, and the sample size for each stratum is 200. The confounder $Z_s$ within each stratum satisfies $Z_1 \sim \mathcal{N}(0, 1)$, $Z_2 \sim \mathcal{N}(0, 1)$, and $Z_3 \sim \text{Unif}(-3, 3)$. The outcome and propensity score models are $Y_s = T_s + a_s W + \epsilon_s$ and $\text{logit}(p_s) = a_s W$, where $W = (-1.875 + Z_s)I(Z_s \leq -1.5) + (1.875 + Z_s)I(Z_s \geq 1.5) + Z_s^2 I(-1.5 < Z_s < 1.5)$, $T_s \sim \text{Bernoulli}(p_s)$ and $\epsilon_s \sim F_e$ for $s \in \{1, 2, 3\}$. We set $a_1 = a_3 = 1$ and $a_2 = -1$ to make the confounders either follow different distributions or have different relationships with outcomes and treatment assignments across the three strata. For the misspecified propensity score model, we fit logistic regressions regressing $T_s$ only on $Z_s$ for $s \in \{1, 2, 3\}$. We consider four choices for $F_e$, $\mathcal{N}(0, 1)$, $\text{Unif}(-2, 2)$, $t_4$, and $0.5\mathcal{N}(-2, 1) + 0.5\mathcal{N}(2, 1)$.

Figure 4: Power analysis: average number of trimmed subjects for four error distributions based on 2000 Monte-Carlo replications.
Figure 5: Empirical $P$-values of misspecified AUT and AUT-T versus expected $P$-values.

Table 2: Sensitivity analysis: average number of trimmed subjects (out of 200) for each stratum by trimmed $U$-test based on 2000 Monte-Carlo replications.

|                | Stratum 1          | Stratum 2          | Stratum 3          |
|----------------|--------------------|--------------------|--------------------|
|                | Treatment | Control | Treatment | Control | Treatment | Control |
| $\mathcal{N}(0,1)$ | 9.00      | 0.21     | 0.23      | 9.12     | 2.21      | 1.18    |
| $U(-2, 2)$     | 8.95      | 0.23     | 0.20      | 8.95     | 2.16      | 1.17    |
| $t_4$          | 8.85      | 0.22     | 0.23      | 9.33     | 2.18      | 1.20    |
| $0.5\mathcal{N}(-5, 1) + 0.5\mathcal{N}(5, 1)$ | 9.03      | 0.19     | 0.21      | 8.99     | 2.20      | 1.19    |
Table 3: Sensitivity analysis: average $R^2$ of linear regression logit($p_s$) $\sim Z_s$ for untrimmed and trimmed samples within each stratum based on 2000 Monte-Carlo replications.

| Stratum 1 | Stratum 2 | Stratum 3 |
|-----------|-----------|-----------|
| Untrimmed | Trimmed   | Untrimmed | Trimmed   | Untrimmed | Trimmed   |
| 0.88      | 0.79      | 0.88      | 0.79      | 0.97      | 0.88      |

Table 3 shows the average $R^2$ of linear regression logit($p_s$) $\sim Z_s$ for the original and trimmed samples within each stratum over 2000 Monte-Carlo replications. The values are the same for the four scenarios. It clearly shows that trimming decreases $R^2$ in all strata. Figure 6 shows the relationships between the empirical $P$-values of the AUT and AUT-T with misspecified propensity score models and the expected uniform $P$-values when the tests are valid. It indicates that AUT is sensitive to misspecified propensity scores, but trimming again effectively leverages the effect in all scenarios. Though there is no proof that trimming will always improve performance, the two different scenarios considered here suggest it is generally a good idea.

We have conducted another sensitivity analysis by changing $h(x)$ from 1 to $e(x)$ (while keeping other settings the same) and checking the validity and power of our proposed test. The results are summarized in Section 2.2 of the Supplementary Material. The same conclusions are made between $h(x) = 1$ and $h(x) = e(x)$, which confirms the desired robustness property of the proposed tests.

5. CASE STUDY

5.1. Comparing Effects of an Employment Program on People with Different Ages

We apply the proposed method to an employment programme evaluation study in labour economics, which evaluates the effect of the National Support Work (NSW) Demonstration on trainee earnings. The NSW was conducted in the mid-1970s with the goal of helping disadvantaged workers gain working experience. More details about this programme can be found in LaLonde (1986) and Dehejia & Wahba (1999). In this programme, applicants were randomly assigned to the treatment and control groups; and the treatment effect can be easily assessed by directly comparing the outcomes between those two groups. In order to evaluate whether observational studies can replicate results from randomized experiments, LaLonde (1986) compared the treated subjects in the experiment to two nonexperimental comparison groups, namely, the Panel Study of Income Dynamics (PSID-1) and Current Population Survey-Social Security Administration File (CPS-1), as well as several subsets of them. The collected pretreatment covariates include age, education, marital status, indicator of “no degree”, race indicators and earnings in 1974 (RE74) and 1975 (RE75). The outcome of interest is earnings in 1978.

We focus on the data set constructed by Dehejia & Wahba (1999), which is a subset of the original data set in LaLonde (1986) that includes data collected from male participants who have earnings information in 1974. The data are available at https://users.nber.org/~rdehejia/data/.nswdata2.html. It has been shown by Dehejia & Wahba (1999) that there is a positive treatment effect. Our goal here is to investigate whether there is treatment effect heterogeneity across different age groups for the treated subjects. Two strata are created based on the median age (25 years old) of the treatment group, that is, Stratum 1 for subjects with age $\leq 25$ and Stratum 2 for age $>25$. Figure 7 shows the outcome distributions of the treated subjects in the two strata, and it is clear that both distributions are highly right-skewed, which suggests that nonparametric $U$-tests should be preferred to the LRT.
We compare the NSW treatment group to the NSW control group and CPS-1 separately. The first three columns of Table 4 show the summary statistics of baseline covariates in both strata for the three groups. To compare the NSW treatment group with its control, we notice that the baseline covariates between groups are similarly distributed, so the unadjusted $U$-test can be applied to assess the treatment effect heterogeneity between the two strata. We obtain an estimated unadjusted $U$-statistic of 0.554 with a $P$-value of 0.181, which suggests that the treatment effect in the younger group (Stratum 1) is smaller than that in the elder group (Stratum 2), although this difference is not statistically significant (note that a $U$-statistic value of 0.5 indicates that there is no heterogeneity between those two strata, and a value larger than 0.5 indicates that Stratum 1 has a smaller treatment effect than that of Stratum 2).

We then study the comparison between the NSW treatment group and CPS-1 group. We apply the proposed adjusted $U$-test with trimming. In both strata, we use logistic regressions to estimate propensity scores. For Stratum 1, we use the following covariates: age, age$^2$, age$^3$, education, education$^2$, $I$(married), $I$(no degree), $I$(black), $I$(Hispanic), $RE74$, $RE75$, $I(RE74 = 0)$, $I(RE75 = 0)$, $RE74 \times I$(married), and $RE74 \times I$(no degree). In Stratum 2, we...
consider age, age\(^2\), age\(^3\), education, education\(^2\), \(I\) (married) + \(I\) (no degree), \(I\) (black), \(I\) (Hispanic), RE74, RE75, \(I\) (RE74 = 0), \(I\) (RE75 = 0) and education \(\times\) RE74. Most of those covariates are also included in the study of Dehejia & Wahba (1999). Subjects are weighted according to (7) with \(h(x) = e(x)\). We present summary statistics for the baseline covariates after trimming and weighting as in the fourth column of Table 4. The weighted distributions of baseline covariates in CPS-1 are very similar to the NSW treatment group. Due to the large sample size, we randomly sample \(M = 1000N\) (\(N = 4022\) is the total sample size) weighted kernel functions to approximate the adjusted \(U\)-statistics as illustrated in Section 4.1. The estimated adjusted \(U\)-statistic comparing the treatment effects in the two strata is 0.541 with a \(P\)-value of 0.508, which leads to the same conclusion as the randomized data comparison (NSW treatment versus its control). Meanwhile, if an unadjusted \(U\)-test is applied to conduct the same comparison, then the estimated \(U\)-statistic would be 0.426 with a \(P\)-value of 0.004, which will lead to the opposite conclusion. This finding confirms the benefit of our proposed methodology and also highlights the necessity of appropriately adjusting for covariate balance between groups when testing for treatment heterogeneity effect.

We also apply our method without trimming. The weighted distributions of baseline covariates without trimming are presented in the last column of Table 4. Table 4 shows that without trimming, the covariates are also balanced quite well in both strata. The adjusted \(U\)-statistic without trimming is 0.574 with a \(P\)-value of 0.445. This can be compared with the \(U\)-statistic of 0.541 and a \(P\)-value of 0.508 with trimming. The conclusion regarding the treatment effect remains the same.

5.2. Assessing Heterogeneity of the Effect of Being an Only Child on Mental Health

From 1979 to 2015, China’s one-child policy was implemented to slow the rapid growth of the nation’s population. Though the policy has led to economic benefits for China, it has been criticized for introducing a series of social problems such as forced abortions, female infanticide and a heavy burden of elderly support (Hesketh & Zhu, 1997). Apart from these problems, the psychological well-being of the massive number of only children resulting from the policy has been a great concern because it has been widely recognized that siblings have a large
Table 4: Sample means (standard deviations) of baseline characteristics for NSW and CPS-1 data in two age strata.

| Stratum 1 | NSW treated | NSW control | CPS-1 | Weighted trimmed CPS-1 | Weighted CPS-1 without trimming |
|-----------|-------------|-------------|-------|------------------------|-------------------------------|
| Sample size | 106         | 161         | 4676  | 2169                   | 4676                          |
| Age       | 21.09 (2.76) | 20.75 (2.75) | 20.82 (2.82) | 20.97 (2.51) | 20.99 (2.51) |
| Education | 10.29 (1.77) | 9.93 (1.43)  | 11.91 (2.14) | 10.2 (1.54)       | 10.19 (1.53)                     |
| Black     | 0.82 (0.39)  | 0.8 (0.4)    | 0.08 (0.28)  | 0.85 (0.36)       | 0.85 (0.36)                     |
| Hispanic  | 0.08 (0.26)  | 0.13 (0.33)  | 0.07 (0.26)  | 0.06 (0.24)       | 0.06 (0.24)                     |
| Married   | 0.11 (0.32)  | 0.09 (0.28)  | 0.36 (0.48)  | 0.1 (0.3)         | 0.1 (0.3)                      |
| No degree | 0.72 (0.45)  | 0.89 (0.3)   | 0.34 (0.47)  | 0.78 (0.41)       | 0.78 (0.41)                     |
| RE74      | 2129.02 (5489.7) | 2195.81 (6240.8) | 7044.39 (7156.6) | 1845.71 (4032.9) | 1817.98 (4012.65)              |
| RE75      | 1215.97 (2409.7) | 1125.32 (6240.8) | 7665.79 (7156.6) | 1068.04 (4032.9) | 1049.83 (4012.65)              |
| Stratum 2 |             |             |       |                        |                               |
| Sample size | 79         | 99          | 11316 | 1668                   | 11316                          |
| Age       | 32.15 (6.24) | 32.05 (6.24) | 38.35 (8.9)  | 32.25 (5.97)       | 32.15 (5.92)                   |
| Education | 10.42 (2.28) | 10.35 (1.84) | 12.07 (3.12) | 10.47 (2.1)        | 10.46 (2.08)                   |
| Black     | 0.87 (0.33)  | 0.87 (0.33)  | 0.07 (0.26)  | 0.89 (0.32)        | 0.89 (0.32)                     |
| Hispanic  | 0.04 (0.2)   | 0.07 (0.26)  | 0.07 (0.26)  | 0.03 (0.17)        | 0.03 (0.17)                     |
| Married   | 0.29 (0.46)  | 0.26 (0.44)  | 0.86 (0.35)  | 0.24 (0.42)        | 0.23 (0.42)                     |
| No degree | 0.7 (0.46)   | 0.74 (0.44)  | 0.28 (0.45)  | 0.67 (0.47)        | 0.68 (0.47)                     |
| RE74      | 2050.7 (4957.2) | 1962.64 (4611.6) | 16897.94 (8936.6) | 1993.3 (4772.0) | 1924.58 (4723.73)             |
| RE75      | 1956.17 (4204.0) | 1497.18 (3178.3) | 16123.93 (8876.9) | 1909.62 (4093.3) | 1849.18 (4057.04)             |

impact on children’s social behaviour and mental health (Dunn, 1988; McHale, Updegraff & Whiteman, 2012). Only children in China are generally perceived to be more self-centred and less trustworthy. However, the difference between only and non-only children may vary with geographic area and gender for two reasons. First, parents living in urban and rural areas differ in many aspects including education level, family income and lifestyle. Second, a preference for male children was prevalent at that time, especially in rural areas. For these reasons, the literature assessing the effects of being an only child is typically carried out in different strata that are determined by the type of region (urban/rural) and gender (male/female). For example, Wu (2014) found that only children have worse mental health than children with siblings on average in China, but this negative effect mainly came from rural males, whereas Zeng, Li & Ding (2020) found that the negative effects were more significant in urban areas. It is therefore of interest to apply the adjusted \( U \)-test to study whether there is significant treatment effect
Table 5: Unweighted and weighted sample means (standard deviations) of baseline characteristics and responses in treatment and control groups.

| Unweighted Trimmed and weighted |
|--------------------------------|
| Only children | Children with siblings | Only children | Children with siblings |
|----------------|------------------------|----------------|------------------------|
| Sample size    | 971                    | 3216           | 968                    | 3216                   |
| Baseline covariates |                   |                  |                        |                        |
| Maternal education (year) | 7.95 (4.28) | 4.19 (4.29) | 4.44 (4.64) | 4.98 (4.49) |
| Paternal education (year) | 8.72 (3.98) | 6.41 (4.36) | 6.21 (4.56) | 6.88 (4.36) |
| Age (year)      | 24.99 (3.38)         | 25.19 (3.51) | 25.38 (3.65) | 25.17 (3.50) |
| Han ethnicity   | 0.96 (0.20)          | 0.89 (0.32)   | 0.88 (0.32)  | 0.91 (0.30)  |
| Family annual income (Chinese Yuan) | 56,957.5 (58,152.7) | 37,403.1 (44,133.1) | 41,324.5 (51,470.2) | 42,793.1 (54,362.3) |
| Parental age at birth (year) | 26.83 (3.81) | 27.66 (5.11) | 27.92 (5.68) | 27.45 (4.99) |
| Maternal age at birth (year) | 25.09 (3.44) | 25.7 (4.54) | 25.9 (4.67) | 25.55 (4.44) |
| Divorce         | 0.03 (0.17)          | 0.01 (0.10)   | 0.01 (0.10)  | 0.01 (0.10)  |
| Urban area      | 0.78 (0.41)          | 0.39 (0.49)   | 0.43 (0.50)  | 0.48 (0.50)  |
| Male            | 0.59 (0.49)          | 0.47 (0.50)   | 0.49 (0.50)  | 0.50 (0.50)  |
| Outcomes        |                       |                  |                        |                        |
| Confidence      | 3.96 (0.92)          | 4.02 (0.95)    | 3.95 (0.95)  | 4.02 (0.94)  |
| Anxiety         | 4.62 (0.67)          | 4.60 (0.69)    | 4.63 (0.69)  | 4.61 (0.68)  |
| Desperation     | 4.68 (0.62)          | 4.72 (0.61)    | 4.69 (0.62)  | 4.73 (0.61)  |

*P(only child) is modelled by a logistic regression with all baseline covariates, (maternal age at birth)$^3$, (paternal age at birth)$^3$, and (family income)$^3$.

The data we use were obtained from the Chinese Family Panel Studies (CFPS) (Xie and Hu, 2014), which is a longitudinal survey aimed at documenting changes in various aspects of Chinese society. The baseline survey was conducted in 2010. It covers 25 provinces, municipalities and autonomous regions that represent 95% of the Chinese population. The data set we focus on is a subset of the CFPS baseline sample constructed by Zeng, Li & Ding (2020). It consists of children born after 1979, aged between 20 and 31. The data set is available at https://rss.onlinelibrary.wiley.com/pb-assets/hub-assets/rss/Datasets/RSSA%20183.4/A1595Zeng-1600084584507.zip. For families with more than one child, only the oldest child is included in the data set. Baseline covariates include age, ethnicity (Han or not), parents’ education level (in years), family income in 2010, parents’ marital status (divorced or not), parents’ ages when the child was born, region type (urban/rural) and sex. The responses include three self-rated psychological measures: confidence, anxiety and desperation. All measures take integer values from 1 to 5, with a higher value indicating better mental health. We treat the only children as the treatment group and the other children with siblings as the control group. We also test for treatment heterogeneity among the four subpopulations: urban males, urban females, rural males and rural females.
remove subjects with obviously erroneous information such as a parent’s age below 14 at the time of the child’s birth or any response measure outside the range of the scale. Three children with family annual incomes higher than two million Chinese Yuan are removed because these incomes are dramatically larger than the rest of the sample. The final data set has 4187 subjects, with 971 in the treatment group (only children). The distributions of baseline covariates and outcomes are summarized in the left-hand side of Table 5. We find that parents with only one child have both a higher average education level and family income. Among only children, there are larger proportions of male or urban subjects compared with children with siblings. With respect to the three psychological responses, the summary statistics are very similar between the two treatment groups. Figure 8 shows the distributions of the three responses within each treatment group. Apart from the fact that every outcome is similarly distributed in the treatment and control groups, they are all heavily left-skewed.

We first apply the weighted version of the Mann–Whitney test introduced by Satten, Kong & Datta (2018) to assess the overall average treatment effects with respect to the three outcomes. We standardize all baseline covariates and then fit a logistic regression to estimate propensity scores. After trimming subjects whose estimated propensity scores are outside of the overlap region, we fit the same logistic regression again with the remaining subjects and use the newly estimated propensity scores for weighting. The weights are based on formulas in (7). Because we focus on estimating the average treatment effects, we use $h(x) = 1$. Summary statistics for the baseline and response variables after trimming and weighting are presented in the right-hand side of Table 5. There is a clear improvement in covariate balance, though the summary statistics of

![Figure 8: Distributions of confidence, anxiety and desperation measures in the treatment and control groups.](image)
Table 6: Unweighted and weighted sample means (standard deviations) of baseline characteristics and responses in treatment and control groups of four strata.

| Urban males<sup>a</sup> | Unweighted | Trimmed and weighted |
|--------------------------|------------|----------------------|
|                          | Only children | Children with siblings | Only children | Children with siblings |
| Sample size              | 430         | 580                  | 423           | 558                    |
| Baseline covariates      |             |                      |               |                        |
| Maternal education (year)| 8.66 (4.05) | 5.15 (4.37)          | 6.58 (4.70)   | 6.63 (4.43)            |
| Paternal education (year)| 9.44 (3.76) | 7.29 (4.37)          | 8.05 (4.14)   | 8.20 (4.25)            |
| Age (year)               | 25.38 (3.33) | 25.66 (3.53)        | 25.62 (3.46)  | 25.71 (3.50)           |
| Han ethnicity            | 0.98 (0.14) | 0.93 (0.24)          | 0.96 (0.20)   | 0.96 (0.20)            |
| Family annual income (Chinese Yuan) | 59,449.4 (60,477.7) | 45,266.9 (45,263.7) | 51,219.8 (56,528.3) | 54,606.3 (64,632.9) |
| Paternal age at birth (year) | 26.94 (3.48) | 27.90 (4.90)          | 27.24 (4.35)  | 27.36 (4.52)           |
| Maternal age at birth (year) | 25.16 (3.23) | 26.29 (4.32)          | 25.61 (4.00)  | 25.69 (3.76)           |
| Divorce                  | 0.03 (0.17) | 0.01 (0.10)          | 0.02 (0.14)   | 0.02 (0.14)            |
| Outcomes                 |             |                      |               |                        |
| Confidence               | 4.00 (0.92) | 3.96 (0.98)          | 3.98 (0.94)   | 3.97 (0.93)            |
| Anxiety                  | 4.61 (0.7)  | 4.64 (0.62)          | 4.63 (0.67)   | 4.65 (0.58)            |
| Desperation              | 4.67 (0.62) | 4.74 (0.58)          | 4.67 (0.61)   | 4.75 (0.56)            |

| Urban females<sup>b</sup> | Unweighted | Trimmed and weighted |
|---------------------------|------------|----------------------|
|                          | Only children | Children with siblings | Only children | Children with siblings |
| Sample size              | 331         | 690                  | 330           | 634                    |
| Baseline covariates      |             |                      |               |                        |
| Maternal education (year)| 9.05 (3.74) | 5.83 (4.28)          | 7.03 (4.41)   | 7.2 (4.21)             |
| Paternal education (year)| 9.54 (3.39) | 7.43 (4.13)          | 8.35 (3.74)   | 8.42 (3.91)            |
| Age (year)               | 25.01 (3.37) | 25.69 (3.55)        | 25.43 (3.47)  | 25.4 (3.46)            |
| Han ethnicity            | 0.95 (0.22) | 0.93 (0.24)          | 0.93 (0.26)   | 0.93 (0.24)            |
| Family annual income (Chinese Yuan) | 64,914.3 (59,182.4) | 50,261.4 (61,045.9) | 55,548.9 (51,572.5) | 55,414.0 (53,315.4) |
| Paternal age at birth (year) | 27.07 (3.41) | 27.86 (4.99)          | 27.03 (3.74)  | 27.21 (3.78)           |
| Maternal age at birth (year) | 25.47 (3.04) | 25.91 (4.19)          | 25.33 (3.34)  | 25.54 (3.44)           |
| Divorce                  | 0.03 (0.17) | 0.01 (0.10)          | 0.02 (0.14)   | 0.02 (0.14)            |
| Outcomes                 |             |                      |               |                        |
| Confidence               | 3.89 (0.87) | 3.94 (0.92)          | 3.87 (0.87)   | 3.99 (0.91)            |
| Anxiety                  | 4.67 (0.57) | 4.62 (0.67)          | 4.69 (0.55)   | 4.61 (0.67)            |
| Desperation              | 4.68 (0.60) | 4.73 (0.59)          | 4.69 (0.57)   | 4.73 (0.60)            |
|                          | Unweighted Only children | Unweighted Children with siblings | Trimmed and weighted Only children | Trimmed and weighted Children with siblings |
|--------------------------|--------------------------|----------------------------------|-----------------------------------|--------------------------------------------|
| **Rural males**          |                          |                                  |                                   |                                            |
| **Sample size**          | 146                      | 942                              | 146                               | 927                                        |
| **Baseline covariates**  |                          |                                  |                                   |                                            |
| Maternal education (year)| 4.92 (4.00)              | 2.92 (3.96)                      | 3.54 (3.82)                       | 3.24 (4.10)                                |
| Paternal education (year)| 5.94 (3.90)              | 5.7 (4.30)                       | 5.79 (4.00)                       | 5.73 (4.28)                                |
| Age (year)               | 24.13 (3.37)             | 24.97 (3.44)                     | 24.85 (3.56)                      | 24.83 (3.42)                               |
| Han ethnicity            | 0.92 (0.28)              | 0.87 (0.35)                      | 0.91 (0.28)                       | 0.88 (0.32)                                |
| Family annual income (Chinese Yuan) | 37.539.9 (39,878.4) | 31,437.3 (38,706.5) | 34,498.6 (32,817.1) | 31,936.8 (31,243.0) |
| Paternal age at birth (year) | 25.66 (4.59)        | 27.66 (5.28)                     | 27.02 (5.20)                      | 27.27 (5.11)                               |
| Maternal age at birth (year) | 24.08 (4.13)        | 25.7 (4.79)                      | 24.93 (4.43)                      | 25.39 (4.68)                               |
| Divorce                  | 0.01 (0.10)              | 0.01 (0.10)                      | 0.01 (0.10)                       | 0.01 (0.10)                                |
| **Outcomes**             |                          |                                  |                                   |                                            |
| Confidence               | 4.07 (0.96)              | 4.11 (0.93)                      | 4.06 (0.92)                       | 4.12 (0.94)                                |
| Anxiety                  | 4.57 (0.75)              | 4.57 (0.73)                      | 4.59 (0.77)                       | 4.58 (0.73)                                |
| Desperation              | 4.73 (0.59)              | 4.72 (0.63)                      | 4.71 (0.59)                       | 4.72 (0.63)                                |
| **Rural females**        |                          |                                  |                                   |                                            |
| **Sample size**          | 64                       | 1004                             | 62                                | 950                                        |
| **Baseline covariates**  |                          |                                  |                                   |                                            |
| Maternal education (year)| 4.48 (4.11)              | 3.68 (4.05)                      | 3.55 (3.99)                       | 3.76 (4.07)                                |
| Paternal education (year)| 6.05 (4.41)              | 5.88 (4.34)                      | 5.83 (4.46)                       | 5.95 (4.29)                                |
| Age (year)               | 24.23 (3.27)             | 24.77 (3.48)                     | 25.18 (3.63)                      | 24.9 (3.47)                                |
| Han ethnicity            | 0.92 (0.26)              | 0.87 (0.33)                      | 0.91 (0.28)                       | 0.92 (0.26)                                |
| Family annual income (Chinese Yuan) | 43,360.5 (59,802.2) | 29,620.9 (29,073.6) | 28,334.0 (25,023.5) | 30,437.5 (29,195.6) |
| Paternal age at birth (year) | 27.47 (5.22)        | 27.40 (5.14)                     | 27.71 (5.40)                      | 27.39 (5.15)                               |
| Maternal age at birth (year) | 24.97 (4.41)        | 25.23 (4.60)                     | 25.35 (4.63)                      | 25.14 (4.54)                               |
| Divorce                  | 0.03 (0.17)              | 0.01 (0.10)                      | 0.01 (0.10)                       | 0.01 (0.10)                                |
| **Outcomes**             |                          |                                  |                                   |                                            |
| Confidence               | 3.86 (0.95)              | 4.03 (0.95)                      | 3.77 (0.97)                       | 4.04 (0.95)                                |
| Anxiety                  | 4.56 (0.68)              | 4.6 (0.69)                       | 4.55 (0.65)                       | 4.60 (0.71)                                |
| Desperation              | 4.61 (0.68)              | 4.71 (0.62)                      | 4.64 (0.64)                       | 4.71 (0.62)                                |

*a*P(only child) is modelled by a logistic regression with all baseline covariates, (maternal age at birth)$^2$, and (family income)$^2$.

*b*P(only child) is modelled by a logistic regression with all baseline covariates, (maternal age at birth)$^2$, (paternal age at birth)$^2$, and (family income)$^2$.

*c*P(only child) is modelled by a logistic regression with all baseline covariates, (maternal age at birth)$^2$, (family income)$^2$, and divorce $\times$ family income.

*d*P(only child) is modelled by a logistic regression with all baseline covariates, (maternal age at birth)$^2$, (family income)$^2$, and Han $\times$ age.
Table 7: Adjusted Mann–Whitney test statistics (95% CI) for different populations with respect to different response measures.

| Population      | Confidence       | Anxiety         | Desperation     |
|-----------------|------------------|-----------------|-----------------|
| All             | 0.523 (0.486, 0.559) | 0.489 (0.464, 0.515) | 0.519 (0.495, 0.542) |
| Urban males     | 0.497 (0.457, 0.538)  | 0.500 (0.465, 0.535)  | **0.537 (0.505, 0.569)** |
| Urban females   | **0.542 (0.503, 0.582)**  | 0.480 (0.446, 0.514)  | 0.526 (0.494, 0.559)  |
| Rural males     | 0.523 (0.471, 0.575)  | 0.492 (0.447, 0.537)  | 0.507 (0.466, 0.549)  |
| Rural females   | **0.581 (0.511, 0.651)**  | 0.536 (0.472, 0.599)  | 0.540 (0.481, 0.600)  |

The responses do not change much. The adjusted $U$-statistics and corresponding 95% confidence intervals are given in the first row of Table 7. Here, the expectation of the adjusted $U$-statistic is the probability that the outcome in the treatment group is smaller than that in the control group. Thus, a value larger than 0.5 indicates a negative treatment effect, in other words, worse outcomes for only children. The $U$-statistics show that only children are less confident, less anxious and more desperate than children with siblings, with the 95% confidence intervals showing that none of these findings are statistically significant.

We then split the data into four strata based on gender and region type. The sample sizes and distributions of the baseline and response variables in the treatment and control groups are summarized in the left-hand side of Table 6. It shows that the baseline characteristics vary among strata. For instance, urban parents have higher education levels and incomes than rural parents. The proportion of males is higher among only children than children with siblings, especially in rural areas. With respect to the response variables, there are no obvious differences among these subgroups. Adjusted Mann–Whitney tests are implemented in each stratum separately based on the same weighting procedure described above. The baseline covariates are clearly better balanced in all strata. The adjusted Mann–Whitney test statistics and corresponding 95% confidence intervals are listed in the second to fifth rows of Table 7. Most tests show insignificant results except for testing desperation among urban males, and confidence among urban females and rural females. All these significant results suggest that only children’s mental health is worse than children with siblings. Even these should be interpreted with caution, given the large number of tests being carried out. The findings here are related to but not exactly the same as those reported in Zeng, Li & Ding (2020), which found significantly negative treatment effects among both urban female and male strata for almost all responses (except for anxiety of urban females). It is worth noting that the statistically significant findings in both papers are close to the boundary of statistical insignificance; for example, the confidence intervals of our significance tests and the credible intervals in Zeng, Li & Ding (2020) are very close to including the null value, 0.5, in the intervals.

Interpretation of the results here is challenging due to the number of strata and outcomes. A further challenge is that the results in Table 7 suggest similar results across strata in each column but with some attaining significance (at a level of 0.05) and others not. It is natural to ask whether these are significant differences across strata (see, e.g., Gelman & Stern (2006)). The question can be addressed by assessing treatment effect heterogeneity among the four strata. We implement our proposed adjusted $U$-test and calculate the test statistic by randomly selecting $M = 1000N$ ($N = 4030$) kernel terms with replacement as described in Section 4.1. The obtained $P$-values are respectively 0.142, 0.411 and 0.738, for the response variables confidence, anxiety and desperation, which indicates that there is no significant treatment effect heterogeneity among the four subpopulations for each of the three outcomes. Pairwise tests among the four strata to examine treatment effect heterogeneity regarding the three response variables are also conducted.
and the $P$-values of the 18 tests are almost uniformly distributed, which further demonstrates that there does not appear to be treatment effect heterogeneity across gender and region types. We have also analyzed this dataset without trimming. The results are presented in Section 2.3 of the Supplementary Material.

6. DISCUSSION

In this article, we propose a new nonparametric $U$-test for heterogeneity of treatment effects in observational studies. Our method extends the $U$-test in Dai & Stern (2022) for randomized experiments to observational studies by adjusting for the confounding variables using propensity score modelling. Our approach is adaptive to various choices of target population, as long as the general function $h(x)$ used to define the target population is a constant or a differentiable function of the propensity score. Many target populations of interest in practice satisfy this requirement, including subjects in treatment and control groups combined, treated subjects and subjects under control.

Compared with its parametric counterpart, the LRT, the proposed adjusted $U$-test inherits the advantages of nonparametric tests: it requires weaker modelling assumptions about the distribution of the outcome and provides a significant improvement in power for non-normally distributed data. Several simulation scenarios suggest that subject trimming based on propensity scores may improve the robustness of the adjusted $U$-test to model misspecification. There is no analytic proof of this latter result; more exploration needs to be done.

Several future working directions remain open. Firstly, we assume that for our method, all confounding variables are observed, which is untestable and may be subject to violation in practice. It will be of interest to conduct a sensitivity analysis to address this issue. Secondly, we assume there are no missing values of the confounding variables. Multiple imputation (Schafer, 1997) can be used to resolve the issue if the values are missing at random. If they are missing not at random, it will be of interest to extend our work based on ideas from Yang, Wang & Ding (2019). Thirdly, the calculation of $U$-statistics is based on a random sampling procedure over all pairwise comparisons between strata for our method. Developing a more efficient sampling method for faster $U$-statistic computation will be an interesting future working direction. Fourthly, it will be of interest to extend our test statistic for high-dimensional covariates based on the results in He et al. (2021). Lastly, in addition to covariate balancing for observational studies, another potential advantage of weighting is to improve the efficiency of test statistics. For example, Li, Morgan & Zaslavsky (2018) show that under homoscedasticity their proposed weighted average treatment effect estimator has the smallest asymptotic variance when $h(x) \propto e(x)(1 - e(x))$. For our adjusted $U$-statistic, the asymptotic variance formula is more complicated and we do not yet have an analogous result. It will be of interest to identify a class of $h(x)$ functions that lead to variance reduction.

ACKNOWLEDGEMENT

The authors thank the editor Fang Yao, an associate editor and two referees for their constructive comments that significantly helped improve the early version of this article.

REFERENCES

Austin, P. C. (2008). Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidemiology and Drug Safety*, 17(12), 1202–1217.

Bitler, M. P., Gelbach, J. B., & Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, 96(4), 988–1012.

Byar, D. P. (1985). Assessing apparent treatment—Covariate interactions in randomized clinical trials. *Statistics in Medicine*, 4(3), 255–263.
Chang, M., Lee, S., & Whang, Y.-J. (2015). Nonparametric tests of conditional treatment effects with an application to single-sex schooling on academic achievements. *The Econometrics Journal*, 18(3), 307–346.

Cook, D. I., Gebski, V. J., & Keech, A. C. (2004). Subgroup analysis in clinical trials. *Medical Journal of Australia*, 180(6), 289.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3), 389–405.

Crump, R. K., Joseph Hotz, V., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1), 187–199.

Dai, M. & Stern, H. S. (2022). A U-statistic-based test of treatment effect heterogeneity. *Journal of Nonparametric Statistics*, 34(1), 141–163.

Dehejia, R. H. & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448), 1053–1062.

Ding, P., Feller, A., & Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3), 655–671.

Dunn, J. (1988). Sibling influences on childhood development. *Journal of Child Psychology and Psychiatry*, 29(2), 119–127.

Feller, A. & Holmes, C. C. (2009). *Beyond Toplines: Heterogeneous Treatment Effects in Randomized Experiments*. Unpublished manuscript, Oxford University, Oxford.

Gail, M. & Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41(2), 361–372.

Gelman, A. & Stern, H. (2006). The difference between “Significant” and “not significant” is not itself statistically significant. *The American Statistician*, 60(4), 328–331.

Ginsburg, G. S. & Willard, H. F. (2009). Genomic and personalized medicine: Foundations and applications. *Translational Research*, 154(6), 277–287.

He, Y., Gongjun, X., Chong, W., & Pan, W. (2021). Asymptotically independent u-statistics in high-dimensional testing. *Annals of Statistics*, 49(1), 154–181.

Hesketh, T. & Zhu, W. X. (1997). The one child family policy: The good, the bad, and the ugly; health in China, part 3. *British Medical Journal*, 314, 1685–1692.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685.

Hsu, Y.-C. (2017). Consistent tests for conditional treatment effects. *The Econometrics Journal*, 20(1), 1–22.

Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge.

Kent, D. M., Nelson, J., Dahabreh, I. J., Rothwell, P. M., Altman, D. G., & Hayward, R. A. (2016). Risk and treatment effect heterogeneity: Re-analysis of individual participant data from 32 large clinical trials. *International Journal of Epidemiology*, 45(6), 2075–2088.

Korolyuk, V. S. & Borovskich, Y. V. (2013). *Theory of U-statistics*, Vol. 273, Springer Science & Business Media, New York.

Krishnan, E., Sokka, T., & Hannonen, P. (2003). Smoking—Gender interaction and risk for rheumatoid arthritis. *Arthritis Research & Therapy*, 5(3), R158.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76, 604–620.

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521), 390–400.

Ma, X. & Wang, J. (2020). Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115(532), 1851–1860.

McHale, S. M., Updegraff, K. A., & Whiteman, S. D. (2012). Sibling relationships and influences in childhood and adolescence. *Journal of Marriage and Family*, 74(5), 913–930.

Na, C., Loughran, T. A., & Paternoster, R. (2015). On the importance of treatment effect heterogeneity in experimentally-evaluated criminal justice interventions. *Journal of Quantitative Criminology*, 31(2), 289–310.

Pate, A. M. & Hamilton, E. E. (1992). Formal and informal deterrents to domestic violence: The Dade county spouse assault experiment. *American Sociological Review*, 57, 691–697.
Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

Sant’Anna, P. H. C. (2020). Nonparametric tests for treatment effect heterogeneity with duration outcomes. *Journal of Business & Economic Statistics*, 39, 1–17.

Satten, G. A., Kong, M., & Datta, S. (2018). Multisample adjusted U-statistics that account for confounding covariates. *Statistics in Medicine*, 37(23), 3357–3372.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, CRC Press, Boca Raton.

Vegetabile, B. G., Gillen, D. L., & Stern, H. S. (2020). Optimally balanced Gaussian process propensity scores for estimating treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(1), 355–377.

Wu, L. (2014). *Are only children worse off on subjective well-being? evidence from China’s one-child policy*. Master’s thesis, Hong Kong University of Science and Technology.

Xie, Y., Brand, J. E., & Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological Methodology*, 42(1), 314–347.

Xie, Y. & Hu, J. (2014). An introduction to the China family panel studies (CFPS). *Chinese Sociological Review*, 47(1), 3–29.

Yang, S., Wang, L., & Ding, P. (2019). Causal inference with confounders missing not at random. *Biometrika*, 106(4), 875–888.

Zeng, S., Li, F., & Ding, P. (2020). Is being an only child harmful to psychological health? Evidence from an instrumental variable analysis of China’s one-child policy. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4), 1615–1635.

---

*Received 17 September 2021
Accepted 2 February 2022*