Confounder detection in linear mediation models: Performance of kernel-based tests of independence

Wolfgang Wiedermann¹ · Xintong Li¹

Published online: 19 March 2019
© The Psychonomic Society, Inc. 2019

Abstract

It is well-known that the identification of direct and indirect effects in mediation analysis requires strong unconfoundedness assumptions. Even when the predictor is under experimental control, unconfoundedness assumptions must be imposed on the mediator–outcome relation in order to guarantee valid indirect-effect identification. Researchers are therefore advised to test for unconfoundedness when estimating mediation effects. Significance tests to evaluate unconfoundedness usually rely on an instrumental variable (IV)—that is, a variable that is nonindependent of the explanatory variable and, at the same time, independent of all exogenous factors that affect the outcome when the explanatory variable is held constant. Because IVs may be hard to come by, the present study shows that confounders of the mediator–outcome relation can be detected without making use of IVs when variables are nonnormal. We show that kernel-based tests of independence are able to detect confounding under nonnormality. Results from a simulation study are presented that suggest that these tests perform well in terms of Type I error protection and statistical power, independent of the distribution or measurement level of the confounder. A real-world data example from the Job Search Intervention Study (JOBS II) illustrates how the presented approach can be used to minimize the risk of obtaining biased indirect-effect estimates. The data requirements and role of unconfoundedness tests as diagnostic tools are discussed. A Monte Carlo–based power analysis tool for sample size planning is also provided.

Keywords Mediation analysis · Confounder · Exogeneity · Hilbert–Schmidt independence criterion · Nonnormality

Mediation analysis is widely considered a promising method for providing an account of the mechanism through which an intervention has an effect on the targeted outcome (e.g., Gottfredson et al., 2015; MacKinnon, 2008). The total intervention effect is decomposed into a *direct* effect of the treatment and an *indirect* effect component that describes the effect of the treatment on the outcome via one or more mediating variables. In general, the indirect effect serves as a tentative explanation as to why the treatment effect occurs, whereas the direct effect can be interpreted as a general measure of all effects not explained by the mediator. The analysis of mediation processes has been largely dominated by Baron and Kenny's (1986) stepwise

approach, and the majority of research on mediation analysis has focused on the robustness and power of tests of mediation (see, e.g., Fritz & MacKinnon, 2007; Hayes & Scharkow, 2013; Judd & Kenny, 2010; MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; Shrout & Bolger, 2002). However, the recent reconceptualization of total, direct, and indirect effects that uses the counterfactual framework of causation (Imai, Keele, & Tingley, 2010; Imai, Keele, & Yamamoto, 2010; Keele, 2015; Pearl, 2001, 2012; VanderWeele, 2015) has provided a new framework for understanding the exact conditions under which mediation effects can be endowed with a causal interpretation. Due to this reconceptualization, it is now well understood that even when the intervention (the predictor) is under experimental control, unconfoundedness assumptions (similar to those made for purely observational data) must be imposed on the mediator–outcome relation. This observation is not entirely new, because similar statements can already be found in Judd and Kenny's (1981) original exposition.

Blockage or enhancement designs have been proposed (see, e.g., Imai, Keele, Tingley, & Yamamoto, 2011; Pirlott & MacKinnon, 2016) that enable researchers to

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13428-019-01230-4>) contains supplementary material, which is available to authorized users.

✉ Wolfgang Wiedermann
wiedermannw@missouri.edu

¹ Department of Educational, School, and Counseling Psychology,
University of Missouri, Columbia, MO, USA

experimentally control the mediator in addition to the predictor. However, these designs also require strong assumptions. For example, only the mediator of interest can be affected by the manipulation, and no other potential mediating variable should be affected (e.g., Bullock, Green, & Ha, 2010). From a purely statistical perspective, alternative approaches to traditional mediation analysis have been proposed that may serve as a remedy when the mediator–outcome path is prone to confounding. For example, Zheng, Atkins, Zhou, and Rhew (2015) focused on the rank-preserving model (RPM; see also Small, 2012; Ten Have et al., 2007), which allows consistent estimation of mediation effects in the presence of confounders. However, the RPM also rests on unique assumptions. For example, one key requirement is that at least one covariate exists that has a strong interaction effect with the intervention predicting the mediator; that is, at least one moderator of the causal effect of the intervention on the mediator is needed. Alternatively, instrumental variables (IVs; see, e.g., Angrist & Krueger, 2001; Angrist & Pischke, 2009) can be used to generate consistent parameter estimates under confounding. In essence, IVs are used to isolate that part of the variation in the explanatory variable that is uninfluenced by the confounder. For an IV to be reliable, two conditions must be met (cf. Pearl, 2009): The IV must be (1) independent of the error term of the model (representing exogenous factors that affect the outcome when the explanatory variable under study is held constant; this is known as the *exclusion restriction*) and (2) not independent of the explanatory variable (often called the “strength” of an IV). Both conditions are crucial for valid results. Bound, Jaeger, and Baker (1995) showed that “weak” IVs (i.e., IVs that explain little of the variation in the explanatory variable) lead to biased effect estimates. The exclusion restriction condition cannot be tested directly in just-identified models (i.e., models with as many IVs as explanatory variables), and thus, a strong substantial rationale is usually needed to justify the status of a variable as a reliable IV.

When these alternatives are not feasible, researchers are advised to use sensitivity analysis (e.g., Cox, Kisbu-Sakarya, Miočević, & MacKinnon, 2013; Imai, Keele, & Yamamoto, 2010; Imai et al., 2011; Mauro, 1990) or significance tests to evaluate the unconfoundedness assumption (in econometrics, these are often referred to as tests of *exogeneity*; see, e.g., Blundell & Horowitz, 2007; Caetano, 2015; de Luna & Johansson, 2014; Donald, Hsu, & Lieli, 2014; Hausman, 1978). Although sensitivity analysis is useful to assess the robustness of the empirical conclusions drawn from a mediation model to potential confounding, in the present study we focus on significance tests to detect whether influential confounding is present in an estimated mediation model. Common tests of unconfoundedness, again, require the availability of IVs (e.g., Blundell & Horowitz, 2007; de Luna & Johansson, 2014; Donald, Hsu, & Lieli, 2014; Hausman,

1978; Wooldridge, 2015). Caetano (2015) proposed a discontinuity test of exogeneity for a single predictor in a multivariate model that does not require an IV. However, this procedure depends on the continuity of the causal effect of the explanatory variable on the outcome, and it detects the presence of confounders by means of discontinuities in the expected outcomes conditional on all variables. Thus, the test depends on data situations in which discontinuities can unambiguously be attributed to the existence of confounders.

The present study focused on testing the unconfoundedness assumption without requiring IVs or discontinuities in the expected outcome. Instead, the proposed method makes use of higher-than-second moments of variables. In other words, the approach presented here assumes that variables are nonnormally distributed. Higher-than-second moment information of variables has been used in the past in the development of causal discovery algorithms (Mooij, Peters, Janzing, Zscheischler, & Schölkopf, 2016; Shimizu, Hoyer, Hyvärinen, & Kerminen, 2006; Shimizu et al., 2011), confirmatory methods to test the direction of dependence in linear models (Wiedermann & Li, 2018; Wiedermann & von Eye, 2015), estimation algorithms in independent component analysis (Hyvärinen, Karhunen, & Oja, 2001), and search algorithms for covariate selection in linear models (Entner, Hoyer, & Spirtes, 2012). In the present study, we discuss similar principles for the development of unconfoundedness tests in mediation analysis and evaluate their performance in detecting potential confounders in mediator–outcome relations with randomized treatment.

The remainder of this article is structured as follows: First, we review the assumptions about mediation models that need to be made when endowing indirect-effect parameters with causal meaning, and discuss the consequences of violated unconfoundedness assumptions concerning the mediator–outcome relation. Second, we show that, in the presence of a confounder, nonindependence of a mediator and regression errors can be detected when the latter is nonnormally distributed, and we propose a simple, two-step regression approach to evaluating whether unconfoundedness holds for the mediator–outcome component of a mediation model. Third, we introduce the Hilbert–Schmidt independence criterion (Gretton et al., 2008) as a kernel-based measure of independence that is able to detect nonindependence in linearly uncorrelated variables, and then discuss related asymptotic and resampling-based significance tests. Fourth, results from an extensive Monte Carlo simulation experiment are presented that (1) quantify the magnitude of bias of the indirect effect that can be expected due to confounding, and (2) evaluate the performance (i.e., size and statistical power) of independence tests when continuous or categorical confounders are present. Fifth, a real-world data example is presented that demonstrates how the proposed approach can be used to minimize the risk of erroneous conclusions concerning mediation processes due

to confounding. We close the article with a discussion of data requirements to guarantee best-practice applications, and we provide a Monte Carlo–based power analysis tool for Type II error control of unconfoundedness tests.

Confounders in mediation models

Many previous studies have described the issue of biased parameter estimates in mediation models when confounders are present (see, e.g., Bullock et al., 2010; Fritz, Kenny, & MacKinnon, 2016; Greenland & Morgenstern, 2001; Imai, Keele, & Tingley, 2010; Imai, Keele, & Yamamoto, 2010; Judd & Kenny, 1981, 2010; MacKinnon, 2008; MacKinnon, Krull, & Lockwood, 2000; MacKinnon & Pirlott, 2015; Vanderweele, 2010, 2015). Assuming that the mediation mechanism can validly be characterized by the linear model and, without loss of generality, that continuous variables are standardized to exhibit zero means and unit variances, the standard (simple) mediation model with a randomized exposure (x ; e.g., 0 = control, 1 = treatment), a continuous mediator (m), and a continuous outcome (y) is

$$\begin{aligned} m &= ax + e_m, \\ y &= cx + bm + e_y, \end{aligned} \quad (1)$$

where c is the direct effect of x on y , ab defines the indirect effect of x on y through m , and $c + ab$ denotes the total effect of x on y . Furthermore, e_m and e_y are mutually independent error terms with zero means and variances $\sigma_{e_m}^2$ and $\sigma_{e_y}^2$. Parameter estimates $\hat{a}$, $\hat{b}$, and $\hat{c}$ are usually obtained using ordinary least squares (OLS) regression or structural equation models (SEMs) under maximum likelihood or weighted least squares loss function (see, e.g., Iacobucci, Saldanha, & Deng, 2007; MacKinnon, 2008). The following assumptions are necessary to ensuring that $\hat{c}$ and $\hat{ab}$ constitute unbiased estimates for the direct and indirect effects (cf., e.g., Imai, Keele, Tingley, & Yamamoto, 2014; Loeys, Talloen, Goubert, Moerkerke, & Vansteelandt, 2016; Pearl, 2014):

- (A1) no unmeasured confounder of the relation between x and m ,
- (A2) no unmeasured confounder of the relation between x and y ,
- (A3) no unmeasured confounder of the relation between m and y , and
- (A4) no common causes m and y that are affected by x .

Although assumptions A1 and A2 can be expected to hold when the predictor x is randomized, assumptions A3 and A4 are never guaranteed to be satisfied even when x is under experimental control. In particular, A3 states that the absence of common causes is required, which implies that the

covariance of e_m and e_y is assumed to be zero, $\text{cov}(e_m, e_y) = 0$ —that is, unobserved causal influences of m are uncorrelated with the unobserved causal influences of y .

In contrast, when an unobserved confounder (u) is present, Model (1) extends to

$$\begin{aligned} m &= ax + d_mu + e_m, \\ y &= cx + bm + d_yu + e_y, \end{aligned} \quad (2)$$

where d_m and d_y quantify the magnitude of the confounding effects. Erroneously using Model (1) to obtain OLS estimates for c and b leads to the biased estimates (see Appendix A of Loeys et al., 2016, for a proof)

$$\begin{aligned} \hat{c} &= c - ad_md_y \frac{\sigma_u^2 \sigma_x^2}{\sigma_x^2 \sigma_m^2 - \text{cov}(x, m)^2}, \\ \hat{b} &= b + d_md_y \frac{\sigma_u^2 \sigma_x^2}{\sigma_x^2 \sigma_m^2 - \text{cov}(x, m)^2}, \end{aligned} \quad (3)$$

From the equations above, it follows that the magnitude of biases increase with the magnitude of confounding effects. In other words, even under randomization of x spurious effects of the m – y relation can occur (Fritz et al., 2016; Judd & Kenny, 1981; Loeys et al., 2016; MacKinnon, 2008; MacKinnon & Pirlott, 2015). In the following section, we show that by making use of variable information beyond means, variances, and covariances—that is, variable information that is accessible when variables deviate from the normal distribution, the unconfoundedness assumption of the mediation model becomes testable.

Detection of confounders under nonnormality

In the present study, detecting influential confounders under nonnormality essentially relies on properties of (in)dependence of sums of random variables summarized in the Darmois–Skitovich (DS) theorem (Darmois, 1953; Skitovich, 1953). This theorem states that if two linear functions v_1 and v_2 of the same independent continuous random variables w_i ($i = 1, \dots, k$ with $i \geq 2$), $v_1 = \sum_i^k \alpha_i w_i$, and $v_2 = \sum_i^k \beta_i w_i$ with α_i and β_i being constants, are independent, then all component variables w_i where $\alpha_i \beta_i \neq 0$ follow a normal distribution. The reverse corollary, however, implies that if a common w_i exists that is *nonnormal*, then v_1 and v_2 must be *nonindependent*. It is easy to show that this reverse corollary applies in the context of the mediator–outcome relation whenever the confounder u or the error term e_m deviates from normality.

We start with partialing out the effect of the randomized intervention x that, in the present study, is assumed to be represented by a binary variable. However, the same approach can be

used when the treatment variable is polytomous (here $j - 1$ dummy variables are used to partial out the effect of j study groups) or continuous. This can be done by extracting the estimated residuals of two auxiliary regressions in which m and y are regressed on x ; that is, $m' = m - a'x$ and $y' = y - c'x$ can be conceptualized as “purified” measures of m and y , with a' being the regression coefficient when regressing m on x , and c' being the regression coefficient when regressing y on x . The mediator–outcome part of Model (2) can then be expressed as

$$\begin{aligned} m' &= d_m u + e_m, \\ y' &= b m' + d_y u + e_y, \\ &= (b d_m + d_y) u + b e_m + e_y. \end{aligned} \quad (4)$$

According to the regression anatomy formula, first described by Frisch and Waugh (1933; see Angrist & Pischke, 2009, for further details), it follows that the parameters b , d_m , and d_y , and the regression errors in Eq. (4) are identical to the ones obtained through Model (2). In other words, the outcome variation that can be explained by independent variables used in a two-step process is identical to the variation explained by the multiple regression model that considers all independent variables simultaneously. Therefore, it follows that the residual terms (i.e., the unexplained variation in the outcome) will also be identical for the two approaches, which implies that either set of residuals conveys exactly the same information about the unknown population errors (cf. Lovell, 2008).

Assuming that the confounder u is erroneously ignored, the model $y' = b' m' + e'_y$ gives the biased OLS estimate of the mediator–outcome relation (for simplicity, now denoted with b') described in Eq. (3). Furthermore, the error term e'_y associated with the misspecified partial regression model can be rewritten as

$$\begin{aligned} e'_y &= y' - b' m' \\ &= b m' + d_y u + e_y - b' (d_m u + e_m) \\ &= [(b - b') d_m + d_y] u + (b - b') e_m + e_y. \end{aligned} \quad (5)$$

Because $(b - b') \neq 0$ will hold when the mediator–outcome path is confounded (i.e., when $d_m d_y \neq 0$), it follows from Eq. (5) that m' and e'_y consist of the same independent (continuously distributed) random variables u and e_m when a confounder is present. Thus, assuming that at least one of the two common components deviates from normality, e'_y and m' will be nonindependent according to the DS theorem. In contrast, if the mediator–outcome relation is unconfounded (i.e., if either d_m or d_y is zero), $(b - b')$ is zero, and Eq. (5) reduced to $e'_y = e_y$. In other words, in the confounder-free case, e'_y and m' will then be independent due to the independence of u , e_m , and e_y . The magnitude of the dependence of e'_y with m' increases with the sizes of d_m and d_y and the magnitude of nonnormality of the involved component variables. Furthermore, when the reverse

corollary of the DS theorem holds, nonnormality of e_y can also be expected to affect the dependence of e'_y and m' due to its impact on the distribution of e'_y (cf. Eq. (5)). Statistical inference methods to evaluate nonindependence beyond linear uncorrelatedness can be used to test the presence of confounders. In the following section, we introduce such a method, the Hilbert–Schmidt independence criterion (HSIC; Gretton et al., 2008).

Testing for the presence of confounders

As we showed in the previous section, under nonnormality of errors, confounders can be detected through evaluating the independence of linearly uncorrelated residuals (r'_y) and “purified” mediator scores (m'). Detecting nonindependence structures in linearly uncorrelated data is extensively discussed in the area of blind source separation (Hyvärinen et al., 2001). Formally, stochastic independence of two variables, v_1 and v_2 , is defined as $E[f(v_1)g(v_2)] - E[f(v_1)]E[g(v_2)] = 0$ (with E being the expected value operator) for any absolutely integrable functions f and g . Two immediate consequences emerge from this definition: First, uncorrelatedness can be considered a special case of stochastic independence (when using the identity functions $f(v_1) = v_1$ and $g(v_2) = v_2$). Although stochastic independence implies uncorrelatedness, the reverse statement does not hold; that is, uncorrelatedness does not necessarily imply independence. Second, tests of stochastic independence can, in principle, be constructed by inserting functions f and g and testing whether $\text{cov}(f(v_1)g(v_2)) = 0$ holds. Of course, inserting all possible functions is not feasible in practice, which implies that such nonlinear correlation tests induce additional Type II errors (Wiedermann, Artner, & von Eye, 2017). In the present study, we, thus, focus on the HSIC as a kernel-based measure of independence that can be shown to be an omnibus measure for detecting any form of dependence in the large sample limit (Gretton et al., 2008).

For notational simplicity, we present the HSIC for the two original variables, v_1 and v_2 with sample size n . Let $H = I - n^{-1} \mathbf{1} \mathbf{1}^T$ with I being an identity matrix of order n , and $\mathbf{1}$ being a $n \times 1$ vector of ones and $\mathbf{1}^T$ being the transpose of $\mathbf{1}$. Furthermore, let K and L be $n \times n$ matrices with cell entries $k_{ij} = k(v_{1(i)}, v_{1(j)})$ and $l_{ij} = l(v_{2(i)}, v_{2(j)})$, where k and l define Gaussian kernels of the form $k(v_1, v_1^T) = \exp(-\sigma^{-2} \|v_1 - v_1^T\|^2)$, and $\|v_1 - v_1^T\|^2$ denotes the squared Euclidean distance of v_1 and v_1^T (l follows the same definition, replacing v_1 with v_2 ; cf. Sen & Sen, 2014). Here, σ represents a bandwidth parameter. It is well-known that the performance of kernel-based methods depends on the calibration of the bandwidth parameter (Schölkopf & Smola, 2002).

While σ is often chosen to be 1 (cf. Sen & Sen, 2014), the so-called *median heuristic*—that is, using the median of all pairwise Euclidian distances (Sriperumbudur, Fukumizu, Gretton, Lanckriet, & Schölkopf, 2009)—constitutes a popular and powerful alternative (see, e.g., Garreau, 2017).

The HSIC is defined as

$$HSIC = n \hat{T}_n, \quad (6)$$

where $\hat{T}_n$ is based on the trace of the matrix product $KHLH$, or, more specifically,

$$\hat{T}_n = 1/n^2 \text{trace}(KHLH). \quad (7)$$

When v_1 and v_2 are stochastically independent, $\hat{T}_n$ approximates zero. If the HSIC significantly deviates from zero, the null hypothesis of independence of v_1 and v_2 can be rejected. Gretton et al. (2008) recommended approximating the null distribution of T_n as a two parameter gamma-distribution (from now on denoted as gHSIC test). In the context of OLS regressions, one has to make use of the estimated residuals instead of “true” (unobservable) errors. Sen and Sen (2014) showed that replacing “true” errors with estimated residuals alters the limiting distribution of the test statistic and suggested a bootstrap alternative for approximation of the distribution of $n \cdot T_n$ under the null hypothesis (from now on denoted as bHSIC test).

In the context of testing unconfoundedness of the mediator–outcome path of a mediation model, we suggest the following stepwise approach (the corresponding hypotheses, statistical decisions, and implications for data analysis are also summarized as a flowchart in Fig. 1):

1. Regress the mediator m and the outcome y on the treatment indicator x and use the estimated residuals of the two models (i.e., $m' = m - \hat{a}'x$ and $y' = y - \hat{c}'x$) as treatment-adjusted (“purified”) measures of m and y .
2. Regress the “purified” outcome y' on the “purified” mediator m' and extract the corresponding regression residuals $r'_y = y' - \hat{b}m'$.
3. Evaluate the independence of r'_y and m' using the HSIC. If the HSIC test rejects the null hypothesis of independence, confounders of the m – y path are likely to be present (cf. Fig. 1).

In Step 3, either the gHSIC or the bHSIC test can be applied. Although the bHSIC test has the advantage that the empirical size of the test can be expected to be close to the nominal significance level when using estimated residuals instead of the (unknown) “true” errors, the gHSIC test is computationally less demanding. Because a direct comparison of the performance of the two HSIC procedures is still missing in the literature, we considered both approaches in the present

study. In addition, we studied the impact of bandwidth selection on the statistical performance of the gHSIC test.

Monte Carlo simulation study

To quantify the bias of the indirect-effect estimates due to confounding and to evaluate the performance of HSIC tests to detect confoundedness under nonnormality, a simulation experiment was performed using the R statistical programming environment (R Core Team, 2019). Data were generated according to the confounded mediation model given in Eq. (2)—that is, x is a binary variable reflecting control ($x = 0$) and treatment assignment ($x = 1$; the study is restricted to equal group sizes), m is a continuous mediator, and y is a continuous outcome. The mediator–outcome relation is affected by the confounder u . Model intercepts were fixed at zero and regression coefficients d_m and d_y were selected to account for zero, small (2% of the variance of the dependent variable), medium (13% of the variance), and large confounding effects (26% of the variance; Cohen, 1988, pp. 412–414). Regression coefficients of the mediation model (c , a , and b) were selected to reflect medium effect sizes. The two error terms e_y and e_m were drawn from either the normal distribution or from various skewed (gamma-distributed) populations. Blanca, Arnau, López-Montiel, Bono, and Bendayan (2013) evaluated 693 empirically observed distributions from various psychological variables and observed a skewness range from -2.49 to 2.33 . More recently, Cain, Zhang, and Yuan (2017) evaluated 1,567 univariate distributions and reported skewness estimates for the empirically observed 1st and 95th percentiles of -2.08 and 2.77 across all distributions. Because the sign of the skewness has no impact on the performance of the approach, we considered positively skewed populations with skewnesses of 0.75, 1.5, and 2.25, which can be considered representative for nonnormal variables observed in psychological research.

Two different types of confounders were considered. In one half of the simulation study, the confounder was continuous with skewnesses $\gamma_u = 0$ (standard normal), 0.75, 1.5, and 2.25. In the other half, the confounder was binary with group proportions $P = .5, .32, .20$, and $.13$, which is in line with skewnesses of a Bernoulli variable of $\gamma_u = (1-2P)/\sqrt{(1-P)P} = 0, 0.75, 1.5$, and 2.25 . The sample sizes were 400 and 800, which correspond to minimum detectable effect sizes (MDES) of 0.249 and 0.176 for the total effect $x \rightarrow y$ of a randomized controlled trial with individual random assignment (assuming equal group sizes, a nominal significance level of 5%, and statistical power of 80%; cf. Dong & Maynard, 2013). The simulation factors were fully crossed and 500 samples were generated for each of the 4 (effect size of d_m) $\times 4$ (effect size of d_y) $\times 4$ (distribution of u) $\times 4$ (distribution of e_m) $\times 4$ (distribution of e_y) $\times 2$ (sample size) $\times 2$ (type of confounder) = 4,096 simulation conditions.

Fig. 1 Hypothesis, statistical decisions, and implications for the proposed confounder detection approach

For each variable triplet $\{x, m, y\}$, we first estimated the (biased) indirect effect ignoring the confounder u . Two outcome measures were computed for the indirect effect, the mean bias $\hat{\theta} - \theta$ and the mean percent bias $(\hat{\theta} - \theta) / \theta \times 100$ with $\hat{\theta} = \hat{a}\hat{b}$ being the OLS estimate and θ being the true indirect effect ab . In line with previous studies (see, e.g., Ames 2013; Collins, Schafer, & Kam, 2001; Wiedermann, Merkle, & von Eye, 2018), absolute biases $\geq 40\%$ were considered significant. Next, the effect of x was partialled out of the outcome and the mediator and the “purified” outcome (y') was regressed on the “purified” mediator (m') and the independence of m' and the extracted residuals $r'_y = y' - \hat{b}'m'$ was evaluated using the HSIC tests. Three different versions of the HSIC test were applied: (1) the bHSIC test with 200 resamples and (following Sen & Sen’s, 2014, recommendation) unit bandwidths, (2) the asymptotic gHSIC test with unit bandwidths (gHSIC₁), and (3) the asymptotic gHSIC test in which bandwidths were selected using the median heuristic (gHSIC_{Med}).¹ All significance tests were applied under a nominal significance level of 5%.

It is important to note that, in the present case, Type I error scenarios (i.e., rejecting the true null hypothesis of independence of predictors and residuals) only exist when all variables are normally distributed. The reason for this is that uncorrelatedness implies independence only in the

multivariate normal case (cf. Hyvärinen et al., 2001). To quantify the Type I error robustness of the three HSIC tests, we used Bradley’s (1978) liberal robustness criterion—that is, a test is considered robust if the empirical Type I error rates fall within the interval 2.5%–7.5%.

Magnitude of bias

In general, the bias of indirect-effect estimates is not affected by the magnitude of asymmetry of the confounder and the error terms. Thus, we only report the results for data scenarios in which all variables are normal. Table 1 gives the mean bias and mean percent bias for the indirect effect in the presence of a (continuous or binary) confounder as a function of sample size and magnitude of the confounder effects d_m and d_y . In general, the measurement level of the confounder does not affect the magnitude of bias. Thus, when discussing the results, we focus on the case of a continuous confounder. As expected, no biases (i.e., mean percent biases ranging from -1.24% to 2.35%) occur when at least one of the paths involving the confounder is zero. Furthermore, when at least one of the confounder effects is small, observed biases are still within a tolerable range of 4.71% to 21.86%. In general, biases systematically increase with the magnitude of confounding effects. Large biases (i.e., absolute biases larger than 40%) only occur for medium and large confounder effects (i.e., in six out of 32 conditions). Here, biases range from 43.96% to 66.61%. Within the parameter space of Cohen’s (1988) effect size conventions, we may conclude that biases of the indirect effect tend to be small for a broad range of effect size scenarios.

¹ Simulations were performed in R 3.3.1 using a high-performance computing cluster. The package dHSIC (version 1.1; Pfister & Peters, 2017) was used for the gamma-approximated HSIC test. The bootstrap HSIC approach was applied using the source code provided by Sen and Sen (2014).

Table 1 Mean bias and mean percent bias of indirect effect estimates as a function of sample size and magnitude of the confounding effects for continuous and binary confounders

n	$u \rightarrow m$	$u \rightarrow y$			
		Zero	Small	Medium	Large
Continuous Confounder					
400	Zero	− 0.001 (− 0.40%)	0.003 (2.18%)	0.002 (1.04%)	0.001 (0.74%)
	Small	− 0.001 (− 0.85%)	0.007 (4.71%)	0.022 (14.71%)	0.033 (21.86%)
	Medium	− 0.002 (− 1.09%)	0.020 (13.14%)	0.051 (33.47%)	0.075 (49.41 %)
	Large	0.004 (2.35%)	0.024 (15.67%)	0.067 (43.96 %)	0.099 (65.00 %)
800	Zero	0.000 (0.26%)	0.000 (0.18%)	0.000 (0.15%)	0.000 (− 0.05%)
	Small	− 0.001 (− 0.92%)	0.008 (5.33%)	0.019 (12.73%)	0.032 (21.33%)
	Medium	− 0.002 (− 1.24%)	0.017 (10.99%)	0.049 (32.38%)	0.076 (49.82 %)
	Large	0.000 (0.31%)	0.024 (15.98%)	0.067 (44.10 %)	0.101 (66.61 %)
Binary Confounder					
400	Zero	− 0.001 (− 0.96%)	0.004 (2.91%)	− 0.001 (− 0.95%)	0.005 (2.98%)
	Small	0.004 (2.59%)	0.007 (4.69%)	0.024 (16.07%)	0.032 (21.16%)
	Medium	− 0.002 (− 1.52%)	0.016 (10.67%)	0.047 (30.98%)	0.076 (49.76 %)
	Large	− 0.002 (− 1.26%)	0.026 (17.31%)	0.063 (41.74 %)	0.105 (68.91 %)
800	Zero	0.000 (0.75%)	0.001 (0.87%)	− 0.001 (− 0.71%)	0.001 (0.67%)
	Small	0.002 (1.12%)	0.010 (6.33%)	0.022 (14.61%)	0.034 (22.55%)
	Medium	0.001 (0.97%)	0.020 (12.87%)	0.052 (34.03%)	0.079 (52.11 %)
	Large	0.000 (0.20%)	0.024 (15.49%)	0.066 (43.66 %)	0.101 (66.63 %)

Cases of biases > 40% are marked bold

Type I error results

Next, we focus on the Type I error robustness of the HSIC tests—that is, data scenarios in which both error terms, e_y and e_m , follow a normal distribution and the confounder is continuous and normal. In these cases, HSIC tests should not be able to detect the presence of a confounder, and rates of rejecting the null hypothesis should be close to the nominal significance level of .05. The upper panel of Table 2 gives the empirical Type I error rates for the tests as a function of sample size and magnitude of confounder effects. In general, the gHSIC₁ is overly conservative in statistical decisions; that is, in all cases, the empirical Type I error rate is below Bradley's (1978) liberal robustness interval .025–.075. Conservatism is also observed for the gHSIC_{Md} test—however, to a lesser extent. Here, in ten out of 32 conditions, empirical Type I error rates of approximate HSIC test are below Bradley's liberal robustness interval. Of course, overly conservative decisions under H_0 do not invalidate the results of the gHSIC tests per se—conservative Type I error rates usually lead to lower statistical power. In contrast, the Type I error rates for the bHSIC test are close to 5%, independent of the sample size and the magnitude of confounding effects.

The lower panel of Table 2 gives the observed rates of rejecting the null hypothesis of independence when the

confounder is binary with equal group sizes (i.e., $P = .5$). Although these cases also show a skewness of zero, technically these observed rates do not necessarily correspond to Type I errors, because the confounder (which is, at the same time, a common component of m' and r'_y) deviates from the normal distribution. Thus, nonindependence between m' and r'_y exists, according to the reverse corollary of the DS theorem. However, when $P = .5$ the behavior of the independence tests equals the behavior in the normal case. That is, the portion of rejected null hypotheses is often close to zero for the gHSIC₁ and gHSIC_{Md} tests, and close to the nominal significance level for the bHSIC test. Cases of large confounder effects constitute the only exception. Here, empirical rejection rates are slightly elevated for all three tests.

Power results

To quantify the power of the HSIC tests, we next focus on nonnormal data scenarios (i.e., cases in which γ_u , γ_{e_m} , and γ_{e_y} are nonzero). Figure 2 gives the empirical power of the gHSIC₁, gHSIC_{Md}, and bHSIC tests for continuous and binary confounders based on the main effects of the simulation study (i.e., aggregating across all levels of the remaining simulation factors). Specifically, power curves are given for the

Table 2 Empirical rates of rejecting the null hypothesis of independence of the gHSIC and bHSIC tests, as a function of sample size and magnitude of the confounder effects $u \rightarrow m$ and $u \rightarrow y$

n	$u \rightarrow m$	$u \rightarrow y$											
		Zero			Small			Medium			Large		
		gHSIC ₁	gHSIC _{Md}	bHSIC	gHSIC ₁	gHSIC _{Md}	bHSIC	gHSIC ₁	gHSIC _{Md}	bHSIC	gHSIC ₁	gHSIC _{Md}	bHSIC
Normal Continuous Confounder													
400	Zero	.004	.024	.058	.008	.030	.058	.002	.032	.068	.012	.034	.046
	Small	.010	.028	.066	.018	.024	.056	.014	.040	.066	.018	.024	.054
	Medium	.006	.030	.060	.006	.018	.062	.006	.046	.060	.004	.026	.066
	Large	.010	.028	.050	.016	.016	.040	.010	.030	.034	.018	.016	.044
800	Zero	.010	.034	.052	.004	.032	.050	.016	.032	.086	.014	.040	.048
	Small	.006	.030	.050	.014	.022	.070	.004	.032	.036	.020	.022	.048
	Medium	.018	.030	.074	.006	.030	.054	.010	.030	.042	.006	.032	.046
	Large	.008	.030	.046	.016	.020	.034	.004	.028	.064	.014	.024	.052
Binary Confounder ($P = .5$)													
400	Zero	.010	.026	.064	.008	.024	.050	.008	.026	.046	.008	.028	.060
	Small	.006	.028	.058	.016	.020	.052	.012	.030	.056	.008	.034	.052
	Medium	.018	.034	.044	.010	.018	.066	.012	.028	.050	.016	.052	.076
	Large	.012	.018	.054	.016	.022	.062	.026	.032	.076	.042	.076	.120
800	Zero	.006	.036	.050	.020	.020	.050	.006	.032	.048	.008	.024	.064
	Small	.020	.032	.074	.010	.032	.054	.022	.026	.074	.006	.042	.058
	Medium	.004	.032	.046	.000	.016	.048	.012	.032	.066	.022	.052	.098
	Large	.012	.028	.062	.010	.026	.044	.046	.046	.094	.116	.202	.290

gHSIC₁ = gamma-approximated Hilbert–Schmidt independence criterion with unit bandwidths; gHSIC_{Md} = gHSIC using the median heuristic; bHSIC = bootstrap HSIC based on 200 resamples. In the case of a normal continuous confounder, rejections rates refer to empirical Type I errors. Type I error rates outside Bradley’s (1978) liberal robustness interval .025–.075 are marked bold

skewness of u , e_m , and e_y , the magnitude of the confounding effects (d_m and d_y), and the sample size. As expected, the power of all tests increases with the skewness of the error terms, the magnitude of the confounding effects, and the sample size. The gHSIC₁ test is less powerful than the two other procedures. The test is slightly more powerful when the confounders are continuous in nature. In addition, the bHSIC test for binary confounders has approximately the same power as the gHSIC₁ test for continuous confounders. The gHSIC_{Md} and bHSIC tests are almost indistinguishable with respect to detecting continuous confounders, except for highly asymmetric outcome-related error terms (i.e., $\gamma_{e_y} = 2.25$). There, the gHSIC_{Md} test outperforms the bHSIC test. Test performance is largely unaffected by the measurement level of the confounder, except for large effects of d_m and slightly unbalanced binary confounders (i.e., $P = .32$, which implies $\gamma_u = 0.75$). In both cases, a slight power advantage of the gHSIC_{Md} test is observed when the confounder is binary. Most importantly, the skewness of a continuous confounder has virtually no impact on the statistical power of the tests. In other words, both the distributional shape and the measurement level of the unobserved confounder are largely negligible when testing the unconfoundedness assumption.

Next, we focus on the performance of the tests with respect to Cohen’s (1988) widely used 80% power criterion. Figures 3, 4, and 5 give the statistical power of the tests as a function of those factors that have the largest impact on test performance—that is, the skewness of the error terms and the magnitude of the confounding effects (across all levels of γ_u). Because the gHSIC₁ test showed the lowest power, we only focus on the gHSIC_{Md} and bHSIC tests. Because, for $n = 400$, highly skewed error terms (i.e., skewnesses of 2.25) and large confounder effects are needed in order to achieve sufficient power, we focus on selected results for $n = 800$ (all remaining results are given in an online supplement). In general, observed power rates $\geq 80\%$ were most prevalent for the gHSIC_{Md} test in the presence of a binary confounder (i.e., in 20 out of the $3 (\gamma_{e_y}) \times 3 (\gamma_{e_m}) \times 3 (d_y) \times 3 (d_m) = 81$ data scenarios—i.e., 24.7%; cf. Fig. 4). In particular, focusing on data scenarios in which the percent biases due to confounding can be expected to be large (i.e., when both confounder effects are large or at least one confounder effect is large and the other one is medium-sized; see Table 1), confounders can be detected with sufficient power when the skewnesses of the error terms are ≥ 1.5 . When e_y is highly skewed (i.e., $\gamma_{e_y} = 2.25$), confounders can be detected with power larger than 80% even when the skewness of e_m is small ($\gamma_{e_m} = 0.75$).

Fig. 2 Observed power of gHSIC and bHSIC tests to detect binary and continuous confounders for the main effects of the skewness of the confounder (γ_u), the skewnesses of the error terms (γ_{em} and γ_{ey}), the magnitude of the confounding effects (d_m and d_y), and the sample size (n)

Figures 4 and 5 give the power for the gHSIC_{Md} and bHSIC tests when the unobserved confounder is continuous. Again, in the presence of influential confounding, the tests are able to detect unobserved confounders when both error terms are highly skewed or when at least one error term is highly skewed and the other error term is moderately skewed. Furthermore, for large confounder effects, the bHSIC test shows power values $\geq 80\%$ even when e_m is moderately skewed (i.e., $\gamma_{em} = 1.5$) and e_y is slightly asymmetric (i.e., $\gamma_{ey} = 0.75$).

Empirical example

To illustrate the application of the proposed method to minimize the risk of obtaining biased estimates in a real-world data

example, we use data from the Job Search Intervention Study (JOBS II; cf. Vinokur, Price, & Schul, 1995). JOBS II is a randomized field study that evaluates the efficacy of a job-training intervention for unemployed workers. The subjects in the treatment condition participated in job search skills seminars, and the subjects in the control condition received a booklet that describes job search tips. Vinokur et al. (1995) and Vinokur and Schul (1997) reported that the subjects in the treatment condition showed better employment and mental health outcomes, due to the subjects' enhanced confidence in their job-searching skills. In a reanalysis, Imai, Keele, and Tingley (2010) found a small, albeit significant, negative mediation effect of the intervention on workers' depressive symptoms through workers' job search self-efficacy. The program increased perceived job search efficacy, which, in turn, decreased depressive symptoms. In the present study, we used

Fig. 3 Statistical power of the gHSIC_{Md} test to detect a binary confounder for $n = 800$. Black squares correspond to an empirical power $\geq 80\%$

the same data ($n = 1,193$; 373 in the control condition and 820 in the treatment condition) to evaluate the assumption of unconfoundedness inherent to the mediator (job search efficacy)–outcome (depression) part of the model. Both variables are measured on continuous scales. The job search self-efficacy composite measure is based on six items (each ranging from 1 = *not at all confident* to 5 = *a great deal confident*). The composite measure of depressive symptoms is based on the subscale of 11 items (ranging from 1 = *not at all* to 5 = *extremely*) of the Hopkins Symptom Checklist (Derogatis, Lipman, Rickles, Uhlenhuth, & Covi, 1974). We considered the covariates sex (52.8% female), age ($M = 36.8$ yrs., $SD = 10.6$), race (19.4% non-White), marital status (31.4% never married, 45.1% married, 3.4% separated, 17.8% divorced, and 2.2% widowed), education (7.0% did not complete high school, 32.1% completed high school, 36.0% completed some college, 14.8% had 4 years of college, and 10.1% had > 4 years of college), income (20.7% < \$15,000, 23.6% \$15,000–\$24,000, 24.1% \$25,000–\$39,000, 11.7% \$40,000–\$49,000, 20.0% \$50,000+), economic hardship at

baseline ($M = 3.1$, $SD = 1.0$), and occupation (17.6% professionals, 17.4% managerial, 23.6% clerical/kindred, 7.5% sales workers, 11.1% craftsmen/kindred, 11.6% operatives/kindred, 11.2% laborers/service workers). In addition, the mediation models were adjusted for depressive symptoms at baseline ($M = 1.9$, $SD = 0.6$), job-seeking efficacy prior to treatment ($M = 3.7$, $SD = 0.8$), level of anxiety prior treatment (measured using 11 items ranging from 1 to 5, where higher scores indicate more severe anxious symptomology; $M = 1.9$, $SD = 0.7$), and level of self-esteem (based on eight items from Rosenberg's, 1965, self-esteem scale ranging from 1 to 5, where higher scores indicate higher self-esteem; $M = 4.1$, $SD = 0.7$).

Overall, three different mediation models were estimated. All three models used treatment status as the predictor, job-seeking self-efficacy after 6 months as the mediator, and depressive symptoms after 6 months as the outcome. Nonparametric bootstrapping using 500 resamples was applied in order to evaluate the significance of the indirect effect. In Model I, we adjusted for demographic background

Fig. 4 Statistical power of the gHSIC_{Md} test to detect a continuous confounder for $n = 800$. Black squares correspond to an empirical power $\geq 80\%$

information (i.e., age, gender, race, marital status, and education). In Model II, we additionally incorporated information related to the subjects' financial situation (i.e., income, economic hardship at baseline, and occupation). In Model III, we added psychological background variables (i.e., depression at baseline, anxiety prior treatment, and job-seeking self-efficacy prior to treatment). For each model (I–III), we regressed the treatment-/covariate-purified outcome on the treatment-/covariate-purified mediator and extracted the residuals for further analyses. The gHSIC_{Md} and bHSIC tests (based on 500 resamples and applying the median heuristic for bandwidth selection) were used to evaluate the independence of the (treatment-/covariate-purified) mediator and the corresponding residuals. Again, a significant test indicated the presence of unmeasured confounders that would hamper causal interpretation of the obtained mediation effect.

Table 3 gives the estimated regression parameters, indirect-effect estimates, measures of skewness of residuals, and HSIC test results for the three models. In each mediation model, we

observed a small, albeit significant, negative mediation effect (ranging from -0.029 to -0.020), which is in line with the previous results of Imai, Keele, and Yamamoto (2010). The estimated model residuals deviated from symmetry, as indicated by D'Agostino's (1971) z values. The residual skewnesses of the mediator model range from -0.76 to -0.69 , and those of the outcome model range from 1.05 to 1.34 (all $ps < .001$) and thus fulfill the distributional requirement of error nonnormality. HSIC values systematically decrease from 0.651 (Model I) to 0.206 (Model III), indicating that the magnitude of error dependence decreases with every additional set of considered covariates. However, even after including demographic and financial background information, we obtain an HSIC of 0.488 and both HSIC tests reject the null hypothesis of independence. In other words, influential confounders may still be present. This is no longer the case, however, when adjusting for psychological background variables (Model III). Here, the HSIC drops to 0.206 and both HSIC tests retain the null hypothesis of independence. In other words, among the

Fig. 5 Statistical power of the bHSIC test to detect a continuous confounder for $n = 800$. Black squares correspond to an empirical power $\geq 80\%$

three candidate models, Model III is the one with the lowest risk of a biased indirect-effect estimate. Although the three indirect-effect estimates are not remarkably different from each other (i.e., additionally adjusting for financial and psychological background characteristics reduces the indirect effect by 0.007 points and leads to a slightly narrower confidence interval), and although all three models suggest that the treatment increased job-seeking self-efficacy, which in turn decreased depression, one has gained knowledge that adjusting the mediation model for the demographic, financial, and psychological background variables leads to sufficiently independent errors according to the two HSIC tests, which is a fundamental requirement for interpretation of the observed indirect effect as causal.

Discussion

In the present study, we introduced nonindependence properties of linearly uncorrelated nonnormal variables. These

properties can be used to evaluate unconfoundedness assumptions imposed on standard mediation models. Independence tests that detect nonindependence beyond first-order correlations were reviewed and their performance to identify confounding in the mediation setting was evaluated using an extensive simulation study. It is important to note that the magnitude of confounding as well as test performances in terms of Type I error and power rates were evaluated using cutoff criteria that have previously been applied in the field of quantitative psychological research. However, cutoff criteria are not universal; rather, they depend on research context. For example, statistical power of 80% may be sufficient for some research fields but not for others. Of course, ultimately, only the researcher can judge the practical relevance of the statistical results.

Overall, the bootstrap (bHSIC) test with unit bandwidth and the gamma-approximated (gHSIC) test with median heuristic-based bandwidth perform equally well in terms of statistical power, while the former approach shows better Type I error control. Considering, however, the fact that the

Table 3 Parameter estimates of three different mediation models to evaluate the impact of the intervention on depressive symptoms through job-seeking efficacy (t_0 = baseline, t_1 = prior to intervention, t_2 = 6 months after intervention)

Source	Model I		Model II		Model III	
	Job-Seeking Efficacy	Depression	Job-Seeking Efficacy	Depression	Job-Seeking Efficacy	Depression
Treatment	0.15 (0.04) ^{***}	− 0.05 (0.04)	0.15 (0.04) ^{***}	− 0.05 (0.04)	0.20 (0.04) ^{***}	− 0.03 (0.04)
Job-seeking efficacy (t_2)	‡	− 0.18 (0.03) ^{***}	‡	− 0.19 (0.03) ^{***}	‡	− 0.10 (0.03) ^{***}
Gender: Female	0.02 (0.04)	0.08 (0.04) [*]	0.00 (0.05)	0.14 (0.04) ^{**}	0.03 (0.04)	0.07 (0.04) [†]
Age	0.00 (0.00)	0.00 (0.00)	0.00 (0.00) [†]	0.00 (0.00)	0.00 (0.00) [*]	0.00 (0.00)
Race: Non-White	0.04 (0.05)	− 0.01 (0.05)	0.08 (0.05)	− 0.04 (0.05)	− 0.04 (0.05)	0.08 (0.04) [†]
Marital Status: Married	0.15 (0.05) ^{**}	− 0.05 (0.05)	0.11 (0.06) [†]	− 0.05 (0.05)	0.04 (0.05)	− 0.01 (0.05)
Separated	0.26 (0.12) [*]	0.03 (0.11)	0.22 (0.12)	− 0.04 (0.11)	0.13 (0.11)	− 0.02 (0.09)
Divorced	0.28 (0.07) ^{***}	0.05 (0.06)	0.27 (0.07) ^{***}	− 0.05 (0.06)	0.20 (0.06) ^{***}	− 0.04 (0.05)
Widowed	0.07 (0.15)	0.23 (0.14)	0.09 (0.15)	0.15 (0.13)	0.05 (0.13)	0.15 (0.12)
Education: High school	0.33 (0.09) ^{***}	− 0.13 (0.08)	0.27 (0.09) ^{**}	− 0.10 (0.08)	0.20 (0.08) ^{**}	− 0.04 (0.07)
Some college	0.42 (0.09) ^{***}	− 0.10 (0.08)	0.30 (0.09) ^{***}	− 0.09 (0.08)	0.19 (0.08) [*]	− 0.07 (0.07)
4 years of college	0.56 (0.09) ^{***}	− 0.18 (0.09) [*]	0.37 (0.10) ^{***}	− 0.13 (0.09)	0.20 (0.09) [*]	− 0.07 (0.08)
> 4 years of college	0.53 (0.10) ^{***}	− 0.12 (0.09)	0.32 (0.11) ^{**}	− 0.07 (0.10)	0.17 (0.10) [†]	− 0.05 (0.09)
Income: \$15,000– \$24,000	–	–	0.06 (0.06)	− 0.10 (0.06)	0.02 (0.06)	− 0.11 (0.05) [*]
\$25,000– \$39,000	–	–	0.13 (0.07) [†]	− 0.08 (0.06)	0.10 (0.06) [†]	− 0.10 (0.05) [†]
\$40,000– \$49,000	–	–	0.12 (0.08)	0.00 (0.07)	0.11 (0.07)	− 0.06 (0.07)
\$50,000+	–	–	0.12 (0.08)	− 0.08 (0.07)	0.05 (0.07)	− 0.08 (0.06)
Economic hardship (t_0)	–	–	0.03 (0.02)	0.16 (0.02) ^{***}	0.06 (0.02) ^{**}	0.06 (0.02) ^{***}
Occupation: Managerial	–	–	0.10 (0.07)	0.06 (0.06)	0.06 (0.06)	0.04 (0.06)
Clerical/kindred	–	–	− 0.13 (0.07) [†]	− 0.06 (0.06)	− 0.08 (0.06)	− 0.04 (0.06)
Sales workers	–	–	0.01 (0.09)	0.02 (0.08)	0.02 (0.08)	0.00 (0.07)
Craftsmen/foremen	–	–	− 0.15 (0.09) [†]	0.05 (0.08)	− 0.09 (0.08)	0.02 (0.07)
Operatives/kindred	–	–	− 0.27 (0.09) ^{**}	− 0.01 (0.08)	− 0.18 (0.07) [*]	0.00 (0.07)
Laborers/service	–	–	− 0.24 (0.09) ^{**}	− 0.10 (0.08)	− 0.17 (0.08) [*]	− 0.09 (0.07)
Depression (t_0)	–	–	–	–	− 0.04 (0.04)	0.13 (0.04) ^{***}
Job-seeking efficacy (t_1)	–	–	–	–	0.34 (0.02) ^{***}	0.03 (0.02)
Anxiety (t_1)	–	–	–	–	− 0.04 (0.03)	0.28 (0.03) ^{***}
Self-Esteem (t_1)	–	–	–	–	0.15 (0.03) ^{***}	− 0.16 (0.03) ^{***}
Indirect effect [95% CI] ^(a)	− 0.027 [− 0.046; − 0.011]		− 0.029 [− 0.048; − 0.013]		− 0.020 [− 0.035; − 0.010]	
Skewness mediator error	− 0.755 (− 9.61) ^{***}		− 0.761 (− 9.48) ^{***}		− 0.691 (− 8.91) ^{***}	
Skewness outcome error	1.338 (14.75) ^{***}		1.228 (13.91) ^{***}		1.053 (12.45) ^{***}	
HSIC ^(b)	0.651 (p_b = .002; p < .001)		0.488 (p_b = .028; p < .001)		0.206 (p_b = .334; p = .220)	

[†] $p < .1$, ^{*} $p < .05$, ^{**} $p < .01$, ^{***} $p < .001$; ‡: Model-specific outcome variable. ^(a) Nonparametric bootstrap confidence intervals using 500 resamples. ^(b) p = gamma-approximated p value, p_b = bootstrap p value using 500 resamples

gamma-approximated test is computationally more efficient than the bootstrap procedure,² the gHSIC_{Md} is a reasonable compromise and performs well in various data scenarios. The gHSIC₁ test should generally be avoided, due to distorted Type I error rates and low statistical power.

² For example, computing the gHSIC_{Md} test for the JOBS II dataset (n = 1,193) takes 0.70 s, whereas the bHSIC test with 500 resamples (using four parallel cores) requires 1,011.3 s (about 17 min) on an Intel Core i7-4712HQ quad-core processor.

The proposed approach constitutes a diagnostic tool for the evaluation of the critical assumptions that are necessary to endow parameter estimates with causal meaning. However, both the proposed unconfoundedness tests and existing tests of exogeneity (e.g., Hausman-type procedures) must be applied with circumspection. Specifically, when applying these tests, one is usually not able to control the Type II error risk. In the present context, this implies that rival explanations exist for failing to reject the null hypothesis of independence: The null hypothesis might be retained because (1) one has adjusted for all influential confounders, and data requirements (i.e., the

nonnormality of errors) for valid application of the independence tests have been fulfilled; (2) one has adjusted for all influential confounders, and data requirements have *not* been fulfilled; and (3) one has *failed* to adjust for all influential confounders, and data requirements have *not* been fulfilled. Carefully evaluating the necessary data requirements before applying HSIC tests is, thus, indispensable for arriving at meaningful results. Here, two factors are of particular importance for the present approach: the magnitude of the *nonnormality of the error terms* and the *sample size*.

Nonnormality of error terms

The proposed method assumes that the errors of the mediation model are nonnormally distributed (i.e., exhibiting nonzero skewness and/or excess kurtosis). Although the present study focused on the case of skewed errors, simulation results suggest that the distributional asymmetry of a continuous confounder does not affect the power of the tests (the asymmetry of a binary confounder slightly reduced the power to detect confounding). In contrast, the asymmetry of error terms systematically increases the power to detect confounding, and thus constitutes the most important prerequisite to validly apply the method. Although the present study focused on asymmetric error distributions, it is important to note that the approach can also be applied when errors are symmetric but nonnormal (i.e., distributions with nonzero excess kurtosis). The reason for this is that the reverse corollary of the DS theorem (see above) does not make any statements about the type of nonnormality, and nonindependence of the mediator and the outcome-specific error term will also hold for symmetric nonnormal errors. Evaluating the statistical power of the HSIC tests under these distributional scenarios will be material for future work.

Various previous studies have repeatedly demonstrated that the normality assumption is often violated in real-world data (e.g., Blanca et al., 2013; Micceri, 1989). One theoretical explanation of why variables (and error terms) are likely to deviate from the normal distribution was given, for example, by Beale and Mallows (1959). The error term of any regression model usually captures factors outside the model in addition to measurement error. Even when the error term is defined as a mixture of independent and normally distributed variables with zero means, the resulting error will show elevated kurtosis whenever the involved component variables show unequal variances. Because unequal variances are likely to occur in practice, error terms are also likely to deviate from the normal distribution. Similarly, when normally distributed component variables occur in segments of the domain of the error (so-called constrained mixtures of distributions), the resulting distribution will be skewed when the mixing weights are unequal (cf. Miranda & von Zuben, 2015). Despite these theoretical justifications, it is important to keep in mind that not every

form of error nonnormality makes variables eligible for testing unconfoundedness. For example, independence tests are likely to give biased results when error nonnormality emerges due to outliers or ceiling/floor effects. Thus, carefully evaluating the distributional properties of the estimated model residuals must be a part of applying unconfoundedness tests. Here, data visualizations (e.g., Seier & Bonett, 2011), omnibus normality tests (Jarque & Bera, 1980; Shapiro & Wilk, 1965), or more specific tests of skewness (cf. D'Agostino, 1971; Randles, Fligner, Policello, & Wolfe, 1980) and kurtosis (Anscombe & Glynn, 1983) are available to evaluate distributional requirements.

Previous studies on mediation analysis under nonnormality almost exclusively focused on quantifying potential biases in statistical inference (e.g., Kisbu-Sakarya, MacKinnon, & Miočević, 2014; Ng & Lin, 2016) and the development of robust mediation models (e.g., Yuan & MacKinnon, 2014; Zhang, 2014; Zu & Yuan, 2010). Although the normality of errors constitutes an assumption that is routinely made in OLS regression modeling (White & MacDonald, 1980), it is important to note that normality is not needed for OLS estimates to be unbiased, consistent, and most efficient among all other linear estimators³ (Fox, 2008). As a consequence, extracted regression residuals are unbiased estimates of the “true” errors, and significance tests (such as the HSIC procedures) based on these residuals can be expected to give valid results under error nonnormality. Since nonnormality may affect the standard errors of OLS estimates, bootstrapping techniques can be used to guarantee valid statistical inference about the indirect effect (MacKinnon, Lockwood, & Williams, 2004). Thus, error nonnormality should not be prematurely dismissed as a source of bias. Instead, nonnormality might carry important information that can be used to gain deeper insight into the data-generating mechanism (cf. Wiedermann & von Eye, 2015).

Required sample size

The present simulation results suggest that large sample sizes are preferable in order to achieve acceptable power to detect confounding. The present simulation study took a perspective that may be predominant in practice—that is, determining the sample size for a randomized controlled trial with a focus on the total effect of the intervention, and using mediation analysis to investigate a secondary hypothesis that evaluates why the intervention effect occurs. For this setup, the gHSIC_{Md} and bHSIC tests are reasonably powerful for those scenarios in which indirect-effect biases due to confounding can be considered substantial, as long as the error distributions are

³ This result follows from the Gauss–Markov theorem, which depends “only” on the assumption of linearity, constant error variance, and independence (i.e., the error is independently distributed with zero expectation), but does not require error normality.

moderately or highly asymmetric. In other words, even when the existence of biases due to confounding is assumed or known and one asks questions concerning the extent (not the existence) of confounding, the HSIC tests are able to detect influential confounding.

When the evaluation of potential mediation mechanisms constitutes the primary study goal, larger sample sizes are usually needed. Fritz and MacKinnon (2007), for example, showed that the sample sizes needed for the detection of mediation effects can be incredibly large, in particular when the direct effect of a model is close to zero. This result is less surprising if we consider the power characteristics of product coefficients. For example, making use of Cohen, Cohen, West, and Aiken's (2003, p. 92) approach for computing power for regression coefficients, one concludes that a sample size of 2,636 would be required in order to detect an indirect effect when one of the paths involved in the indirect effect has a small effect size and the other path has a medium-sized effect.⁴ Even when one path shows a large effect, a sample size of 1,153 is required when the effect of the other path is small. The presented simulation results suggest that the HSIC tests can be expected to perform reasonably well for such “large-scale” scenarios.

Power analysis tool

To estimate a priori the statistical power to detect confounding, we provide a Monte Carlo–based power tool (see the [online supplement](#)) implemented in R (R Core Team, 2019). For reasons of computational efficiency, power is estimated for the gHSIC_{MD} procedure. To estimate a priori the required sample size, one needs to estimate the values for direct and indirect effects and the magnitude of the error asymmetry. The magnitude of the confounding effects is usually not known and can be treated as a sensitivity parameter when determining the minimum confounding scenarios for which a statistical power—for example, $\geq 80\%$ —is plausible. Ideally, parameter values for the direct/indirect effects and error asymmetries should be estimated using a representative sample from a pilot study. If pilot data are not available, parameter values may be retrieved from previous investigations. While the parameter estimates involved in a mediation effect are regularly reported in practice, error skewness estimates may be obtained indirectly from sample descriptive statistics. For example, in the

bivariate regression case, the skewness of y can be expressed as $\gamma_y = \rho_{my}^3 \gamma_m + (1 - \rho_{my}^2)^{3/2} \gamma_{e_y}$, with ρ_{my} being the Pearson correlation of m and y (cf. Dodge & Yadegari, 2010), and an estimate for the error skewness is available through $\gamma_{e_y} = (\gamma_y - \rho_{my}^3 \gamma_m) / (1 - \rho_{my}^2)^{3/2}$. Alternatively, the magnitude of error asymmetry can also be considered as a sensitivity parameter. Of course, the power tool can also be used in a post-hoc fashion—that is, by analyzing the power of previous studies, one can obtain important information for planning a future study (cf. Hox, 2010).

Research on mediation analysis has made tremendous progress by way of combining statistical modeling with the counterfactual framework of causation. From this line of research, it is known that, even in randomized experiments, strong assumptions are required in order to derive causal statements from mediation models. The presented method serves as a diagnostic tool to critically evaluate these assumptions and to reduce the risk of erroneous conclusions concerning the mechanisms behind interventions.

References

- Ames, J. A. (2013). Accuracy and precision of an effect size and its variance from a multilevel model for cluster randomized trials: A simulation study. *Multivariate Behavioral Research*, 48, 592–618. <https://doi.org/10.1080/00273171.2013.802978>
- Angrist, J., & Krueger, A. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15, 69–85. <https://doi.org/10.1257/jep.15.4.69>
- Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.
- Anscombe, F. J., & Glynn, W. J. (1983). Distribution of the kurtosis statistic b_2 for normal samples. *Biometrika*, 70, 227–234. <https://doi.org/10.1093/biomet/70.1.227>
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182. <https://doi.org/10.1037/0022-3514.51.6.1173>
- Beale, E. M. L., & Mallows, C. L. (1959). Scale mixing of symmetric distributions with zero means. *Annals of Mathematical Statistics*, 30, 1145–1151. <https://doi.org/10.1214/aoms/1177706099>
- Blanca, M. J., Armau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*, 9, 78–84. <https://doi.org/10.1027/1614-2241/a000057>
- Blundell, R., & Horowitz, J. L. (2007). A non-parametric test of exogeneity. *Review of Economic Studies*, 74, 1035–1058. <https://doi.org/10.1111/j.1467-937x.2007.00458.x>
- Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90, 443–450. <https://doi.org/10.1080/01621459.1995.10476536>
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144–152. <https://doi.org/10.1111/j.2044-8317.1978.tb00581.x>

⁴ Following Cohen et al. (2003), sample size (n) can be determined using the expression $n = L/f^2 + k + 1$, where L is a tabled value corresponding to the number of predictors, the desired nominal significance level, and the statistical power (e.g., for a Type I error of 5%, a power of 80%, and one predictor variable, one obtains $L = 7.85$; cf. Table E.2 in Cohen et al., 2003); f is an effect size measure for linear regression coefficients (in the present context, f corresponds to the values 0.14, 0.39, and 0.59, specifying small, medium, and large effects); and k is the number of predictors. Thus, considering, for example, the case of $a = 0.14$, $b = 0.39$, and $c = 0$, one obtains a total effect of $0.14 \times 0.39 = 0.0546$ and a required sample size of $n = 7.85/0.0546^2 + 1 + 1 = 2,635.2$.

- Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology*, 98, 550–558. <https://doi.org/10.1037/a0018933>
- Caetano, C. (2015). A test of exogeneity without instrumental variables in models with bunching. *Econometrica*, 83, 1581–1600. <https://doi.org/10.3982/ecta11231>
- Cain, M. K., Zhang, Z., & Yuan, K. H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, 49, 1716–1735. <https://doi.org/10.3758/s13428-016-0814-1>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Collins, L. M., Schafer, J. L., & Kam, L. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6, 330–351. <https://doi.org/10.1037/1082-989X.6.4.330>
- Cox, M. G., Kisbu-Sakarya, Y., Miočević, M., & MacKinnon, D. P. (2013). Sensitivity plots for confounder bias in the single mediator model. *Evaluation Review*, 37, 405–431. <https://doi.org/10.1177/0193841X14524576>
- D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika*, 58, 341–348. <https://doi.org/10.1093/biomet/58.2.341>
- Darmois, G. (1953). Analyse generale des liaisons stochastiques [General analysis of stochastic links]. *Review of the International Statistical Institute*, 21, 2–8. <https://doi.org/10.2307/1401511>
- de Luna, X., & Johansson, P. (2014). Testing for the unconfoundedness assumption using an instrumental assumption. *Journal of Causal Inference*, 2, 187–199. <https://doi.org/10.1515/jci-2013-0011>
- Derogatis, L. R., Lipman, R. S., Riekles, I. C., Uhlenhuth, E. H., & Covi, L. (1974). The Hopkins Symptom Checklist (HSCL). In P. Pichot (Ed.), *Psychological measurements in psychopharmacology: Modern problems in pharmacopsychiatry* (Vol. 7, pp. 79–110). New York, NY: Karger.
- Dodge, Y., & Yadegari, I. (2010). On direction of dependence. *Metrika*, 72, 139–150. <https://doi.org/10.1007/s00184-009-0273-0>
- Donald, S. G., Hsu, Y. C., & Lieli, R. P. (2014). Testing the unconfoundedness assumption via inverse probability weighted estimators of (L)ATT. *Journal of Business and Economic Statistics*, 32, 395–415. <https://doi.org/10.1080/07350015.2014.888290>
- Dong, N., & Maynard, R. (2013). PowerUp! A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental studies. *Journal of Research on Educational Effectiveness*, 6, 24–67. <https://doi.org/10.1080/19345747.2012.673143>
- Entner, D., Hoyer, P. O., & Spirtes, P. (2012). Statistical test for consistent estimation of causal effects in linear non-Gaussian models. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 22, 364–372.
- Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Thousand Oaks, CA: Sage.
- Frisch, R., & Waugh, F. (1933). Partial time regressions as compared with individual trends. *Econometrica*, 1, 387–401. <https://doi.org/10.2307/1907330>
- Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18, 233–239. <https://doi.org/10.3758/BRM.40.1.55>
- Fritz, M. S., Kenny, D. A., & MacKinnon, D. P. (2016). The combined effects of measurement error and omitting confounders in single-mediator models. *Multivariate Behavioral Research*, 51, 681–697.
- Garreau, D. (2017). *Asymptotic normality of the median heuristic*. arXiv preprint. arXiv:1707.07269
- Gottfredson, D. C., Cook, T. D., Gardner, F. E., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, 16, 893–926. <https://doi.org/10.1007/s11121-015-0555-x>
- Greenland, S., & Morgenstern, H. (2001). Confounding in health research. *Annual Review of Public Health*, 22, 189–212. <https://doi.org/10.1146/annurev.publhealth.22.1.189>
- Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., & Smola, A. J. (2008). A kernel statistical test of independence. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in neural information processing systems* (Vol. 20, pp. 585–592). La Jolla, CA: Neural Information Processing Systems Foundation.
- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46, 1251–1271. <https://doi.org/10.2307/1913827>
- Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? *Psychological Science*, 24, 1918–1927. <https://doi.org/10.1177/0956797613480187>
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.
- Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York, NY: Wiley & Sons.
- Iacobucci, D., Saldanha, N., & Deng, X. (2007). A meditation on mediation: Evidence that structural equations models perform better than regressions. *Journal of Consumer Psychology*, 17, 139–153. [https://doi.org/10.1016/S1057-7408\(07\)70020-7](https://doi.org/10.1016/S1057-7408(07)70020-7)
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15, 309–334. <https://doi.org/10.1037/a0020761>
- Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, 25, 51–71. <https://doi.org/10.1214/10-STS321>
- Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105, 765–789. <https://doi.org/10.1017/S0003055411000414>
- Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2014). Commentary: Practical implications of theoretical results for causal mediation analysis. *Psychological Methods*, 19, 482–487. <https://doi.org/10.1037/met0000021>
- Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6, 255–259. [https://doi.org/10.1016/0165-1765\(80\)90024-5](https://doi.org/10.1016/0165-1765(80)90024-5)
- Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5, 602–619. <https://doi.org/10.1177/0193841X8100500502>
- Judd, C. M., & Kenny, D. A. (2010). Data analysis in social psychology: Recent and recurring issues. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., Vol. 1, pp. 115–139). New York, NY: Wiley.
- Keele, L. (2015). Causal mediation analysis: Warning! Assumptions ahead. *American Journal of Evaluation*, 36, 500–513. <https://doi.org/10.1177/1098214015594689>
- Kisbu-Sakarya, Y., MacKinnon, D. P., & Miočević, M. (2014). The distribution of the product explains normal theory mediation confidence interval estimation. *Multivariate Behavioral Research*, 49, 261–268. <https://doi.org/10.1080/00273171.2014.903162>
- Loeys, T., Talloen, W., Goubert, L., Moerkerke, B., & Vansteelandt, S. (2016). Assessing moderated mediation in linear models requires fewer confounding assumptions than assessing mediation. *British*

- Journal of Mathematical and Statistical Psychology*, 69, 352–374. <https://doi.org/10.1111/bmsp.12077>
- Lovell, M. C. (2008). A simple proof of the FWL theorem. *Journal of Economic Education*, 39, 88–91. <https://doi.org/10.3200/JECE.39.1.88-91>
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY: Taylor & Francis.
- MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding, and suppression effect. *Prevention Science*, 1, 173–181. <https://doi.org/10.1023/A:1026595011371>
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83–104. <https://doi.org/10.1037/1082-989X.7.1.83>
- MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39, 99–128. https://doi.org/10.1207/s15327906mbr3901_4
- MacKinnon, D. P., & Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis. *Personality and Social Psychology Review*, 19, 30–43. <https://doi.org/10.1177/1088868314542878>
- Mauro, R. (1990). Understanding L.O.V.E. (left out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin*, 108, 314–329. <https://doi.org/10.1037/0033-2909.108.2.314>
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156–166. <https://doi.org/10.1037/0033-2909.105.1.156>
- Miranda, C. S., & von Zuben, F. J. (2015). Asymmetric distributions from constrained mixtures. ArXiv preprint. arXiv:1503.06429
- Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., & Schölkopf, B. (2016). Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17, 1103–1204.
- Ng, M., & Lin, J. (2016). Testing for mediation effects under non-normality and heteroscedasticity: A comparison of classic and modern methods. *International Journal of Quantitative Research in Education*, 3, 24–40. <https://doi.org/10.1504/ijqre.2016.073643>
- Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence* (pp. 411–420). San Francisco, CA: Morgan Kaufmann.
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). New York, NY: Cambridge University Press.
- Pearl, J. (2012). The causal mediation formula—A guide to the assessment of pathways and mechanisms. *Prevention Science*, 13, 426–436. <https://doi.org/10.1007/s11121-011-0270-1>
- Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods*, 19, 459–481. <https://doi.org/10.1037/a0036434>
- Pfister, N., & Peters, J. (2017). dHSIC: Independence testing via Hilbert Schmidt independence criterion (R package version 1.1). Retrieved from <https://cran.r-project.org/web/packages/dHSIC/index.html>
- Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology*, 66, 29–38. <https://doi.org/10.1016/j.jesp.2015.09.012>
- R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Randles, R. H., Fligner, M. A., Policello, G. E., & Wolfe, D. A. (1980). An asymptotically distribution-free test for symmetry versus asymmetry. *Journal of the American Statistical Association*, 75, 168–172.
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press.
- Seier, E., & Bonett, D. G. (2011). A polyplot for visualizing location, spread, skewness, and kurtosis. *American Statistician*, 65, 258–261. <https://doi.org/10.1198/tas.2011.11012>
- Sen, A., & Sen, B. (2014). Testing independence and goodness-of-fit in linear models. *Biometrika*, 101, 927–942. <https://doi.org/10.1093/biomet/asu026>
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611. <https://doi.org/10.1093/biomet/52.3-4.591>
- Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7, 2003–2030.
- Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., ... Bollen, K. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12, 1225–1248.
- Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and non-experimental studies: New procedures and recommendations. *Psychological Methods*, 7, 422–445. <https://doi.org/10.1037/1082-989X.7.4.422>
- Skitovich, W. P. (1953). On a property of the normal distribution. *Doklady Akademii Nauk SSSR [Reports of the Academy of Sciences USSR]*, 89, 217–219.
- Small, D. S. (2012). Mediation analysis without sequential ignorability: Using baseline covariates interacted with random assignment as instrumental variables. *Journal of Statistical Research*, 46, 91–103.
- Sriperumbudur, B., Fukumizu, K., Gretton, A., Lanckriet, G. R. G., & Schölkopf, B. (2009). Kernel choice and classifiability for RKHS embeddings of probability distributions. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (eds.), *Advances in neural information processing systems* 22 (pp. 1750–1758). La Jolla, CA: Neural Information Processing Systems Foundation.
- Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A., & Beck, A. T. (2007). Causal mediation analyses with rank preserving models. *Biometrics*, 63, 926–934. <https://doi.org/10.1111/j.1541-0420.2007.00766.x>
- VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*, 21, 540–551. <https://doi.org/10.1097/EDE.0b013e3181df191c>
- VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford, UK: Oxford University Press.
- Vinokur, A. D., Price, R. H., & Schul, Y. (1995). Impact of the JOBS intervention on unemployed workers varying in risk for depression. *American Journal of Community Psychology*, 23, 39–74. <https://doi.org/10.1007/BF02506922>
- Vinokur, A. D., & Schul, Y. (1997). Mastery and inoculation against setbacks as active ingredients in the JOBS intervention for the unemployed. *Journal of Consulting and Clinical Psychology*, 65, 867–877. <https://doi.org/10.1037/0022-006X.65.5.867>
- White, H., & MacDonald, G. M. (1980). Some large-sample tests for nonnormality in the linear regression model. *Journal of the American Statistical Association*, 75, 16–28. <https://doi.org/10.1080/01621459.1980.10477415>
- Wiedermann, W., Artner, R., & von Eye, A. (2017). Heteroscedasticity as a basis of direction dependence in reversible linear regression models. *Multivariate Behavioral Research*, 52, 222–241. <https://doi.org/10.1080/00273171.2016.1275498>
- Wiedermann, W., Merkle, E. C., & von Eye, A. (2018). Direction of dependence in measurement error models. *British Journal of Mathematical and Statistical Psychology*, 71, 117–145. <https://doi.org/10.1111/bmsp.12111>
- Wiedermann, W., & Li, X. (2018). Direction dependence analysis: Testing the direction of effects in linear models with an

- implementation in SPSS. *Behavior Research Methods*, 50, 1581–1601. <https://doi.org/10.3758/s13428-018-1031-x>
- Wiedermann, W., & von Eye, A. (2015). Direction dependence analysis: A confirmatory approach for testing directional theories. *International Journal of Behavioral Development*, 39, 570–580. <https://doi.org/10.1177/0165025415582056>
- Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, 50, 420–445. <https://doi.org/10.3368/jhr.50.2.420>
- Yuan, Y., & MacKinnon, D. P. (2014). Robust mediation analysis based on median regression. *Psychological Methods*, 19, 1–20. <https://doi.org/10.1037/a0033820>
- Zhang, Z. (2014). Monte Carlo based statistical power analysis for mediation models: Methods and software. *Behavior Research Methods*, 46, 1184–1198. <https://doi.org/10.3758/s13428-013-0424-0>
- Zheng, C., Atkins, D. C., Zhou, X. H., & Rhew, I. C. (2015). Causal models for mediation analysis: An introduction to structural mean models. *Multivariate Behavioral Research*, 50, 614–631. <https://doi.org/10.1080/00273171.2015.1070707>
- Zu, J., & Yuan, K. H. (2010). Local influence and robust procedures for mediation analysis. *Multivariate Behavioral Research*, 45, 1–44. <https://doi.org/10.1080/00273170903504695>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.