Received: 3 February 2020 | Revised: 9 August 2022 | Accepted: 8 March 2023

DOI: 10.1111/stan.12291

ORIGINAL ARTICLE

WILEY

A partial posterior *p* value test for multilevel mediation

Kyle Cox[1] | **Benjamin Kelcey**[2]

[1]Educational Research, Measurement, and Evaluation, University of North Carolina at Charlotte, Charlotte, North Carolina, USA

[2]Quantitative and Mixed Methods Research Methodologies, University of Cincinnati, Cincinnati, Ohio, USA

Correspondence
Kyle Cox, Educational Research, Measurement, and Evaluation, University of North Carolina at Charlotte, 266 Cato College of Education, Charlotte, NC 28223, USA.
Email: kyle.cox@uncc.edu

Funding information
National Science Foundation, Grant/Award Number: 1552535

Abstract

A variety of inferential tests are available for single and multilevel mediation but most come with notable limitations that balance tradeoffs between power and Type I error. We extend the partial posterior *p* value method (p_3 method) to test multilevel mediation. This contemporary resampling-based composite approach is specifically suited for complex null hypotheses. We develop the p_3 method and investigate its performance within the context of two-level cluster-randomized multilevel mediation studies. Similar to its performance in single-level studies, we found that the p_3 method performed well relative to other mediation tests suggesting it provides a judicious balance between Type I error rate and power. While bias-corrected bootstrapping achieved the best overall performance, the p_3 method serves as an alternative tool for researchers investigating multilevel mediation that is especially useful when conducting a priori power analyses. To encourage utilization, we provide R code for implementing the p_3 method.

KEYWORDS

mediation, mediation test, multilevel mediation, partial posterior *p* value

1 | INTRODUCTION

Mediation analysis captures the mechanisms and pathways through which an independent variable acts upon an outcome (e.g., MacKinnon, 2008). The literature provides a framework from which to draw inferences regarding mediation, delineated the structure and decomposition of mediation effects, and outlined the assumptions that support mediation (e.g., Pituch & Stapleton, 2012; VanderWeele, 2015). The literature also details a broad range of inferential tests designed to determine the statistical significance of mediated effects (MacKinnon, Fairchild, & Fritz, 2007; MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002).

However, mediated effects are often quantified using the product of two path coefficients resulting in a sampling distribution that depends on the specific population values of the individual coefficients. This complicates inferential testing as the sampling distribution of and resulting inferences regarding a mediated effect concurrently depend on the population values of the two path coefficients. The dependence creates a complex null hypothesis or one in which many path coefficient combinations result in the null hypothesis being true. Under a complex null hypothesis, the Type I error rate and power associated with a test varies depending on the population parameter values. Put differently, a single observed mediated effect can align with various p values (Biesanz, Falk, & Savalei, 2010).

Past inferential test literature details a variety of approximation techniques that account to varying degrees for this dependence under the complex null hypothesis of no mediation such as: the Sobel test (Sobel, 1982), the joint test (MacKinnon et al., 2002); Monte Carlo (MC) interval test (Preacher & Selig, 2012), bias-corrected (BC) bootstrap, and parametric percentile bootstrap (Pituch, Stapleton, & Kang, 2006). An alternative approach developed by Bayarri and Berger (2000) and Robins, van der Vaart, and Ventura (2000) uses a partial posterior p value distribution method (p_3 method) that formulates inferences more theoretically aligned with the complex and composite nature of mediation effects. First, the p_3 method takes a composite approach by employing two subordinate tests, one for each path coefficient comprising the mediated effect. Second, the p_3 method accounts for the complex null hypothesis of no mediated effect through a resampling technique. The subordinate tests each produce a distribution of p values assuming one null path coefficient and a distribution of values for the remaining path coefficient. By incorporating each path coefficient in the subordinate test they are explicitly formulated to obtain p values under a complex null hypothesis in which the sampling distribution is dependent on another population parameter (Bayarri & Berger, 2000; Robins et al., 2000). The p values can then be used to make inferential decisions about the mediated effect. The p_3 method is applicable in many contexts with a complex null hypothesis (e.g., Bayarri & Berger, 2000; Robins et al., 2000) but we focus on its application with testing mediated effects and a novel extension to multilevel mediation.

Previous work has evaluated numerous tests available to determine the significance of a multilevel mediation effect (e.g., Krull & MacKinnon, 2001; Pituch et al., 2006; Pituch & Stapleton, 2008; Pituch, Whittaker, & Stapleton, 2005) and identified several persistent problems including inaccurate Type I error rates in the null condition and inadequate power to detect a mediated effect in small sample sizes (Kelcey, Dong, Spybrook, & Cox, 2017; Kelcey, Dong, Spybrook, & Shen, 2017; Pituch et al., 2006). Biesanz et al. (2010) extended this work to include the p_3 method for testing single-level mediation effects but a multilevel adaptation has yet to be developed. For testing multilevel mediation, we hypothesize the p_3 method will better track the asymmetric sampling distributions that arise when considering the product of two random

variables (e.g., each path coefficient utilized when estimating a mediated effect) leading to more accurate inferences and greater statistical power. To investigate this hypothesis, we extend the p_3 method to multilevel mediation settings and conduct simulation studies to assess its performance with two common types of multilevel mediation.

We begin with two pertinent multilevel mediation models and then detail and extend the p_3 method for testing multilevel mediation effects. Following these sections, we present two simulation studies comparing power and Type I error rates of the p_3 method to the Sobel test, joint test, MC interval test, and two bootstrapping methods—parametric percentile and BC—in group-randomized studies with group and individual level mediators (i.e., 2-1-1 and 2-2-1). Simulation conditions include two mediated effects, three cases of the null condition, two sample sizes typical for these types of multilevel experiments (e.g., Schochet, 2011; Spybrook, Shi, & Kelcey, 2016), and normal and nonnormal data. These simulation studies provide an assessment of the p_3 method's potential in multilevel mediation settings and further develop understanding inferential testing of multilevel mediation. We conclude by discussing implications of the results, study limitations, and related future inquiry.

2 | MULTILEVEL MEDIATION MODELS

Our description of the p_3 method to multilevel mediation and our simulations involve two analytic models. The first model reflects a two-level cluster randomized trial with a cluster-level treatment, cluster-level mediator, and individual-level outcome (i.e., 2-2-1 mediation) such that (e.g., Zhang, Zyphur, & Preacher, 2009)

$$\text{Mediator model (Level 2)} \qquad M_j = \pi_0 + aT_j + \varepsilon_j^M \qquad \varepsilon_j^M \sim N\left(0, \sigma^2_{M|}\right). \tag{1}$$

$$\text{Outcome model (Level 1)} \qquad Y_{ij} = \beta_{0j} + \varepsilon_{ij}^Y \qquad \varepsilon_{ij}^Y \sim N\left(0, \sigma^2_Y\right). \tag{2}$$

$$\text{(Level 2)}\ \beta_{0j} = \gamma_{00} + bM_j + c'T_j + u_{0j} \qquad u_{0j} \sim N\left(0, \tau^2_{Y|}\right).$$

Under the mediation model (Equation 1), we use M_j as the mediator for cluster j, T_j as the treatment assignment coded as $\pm 1/2$ with associated coefficient a, capturing the relationship between the treatment and mediator, and ε_j^M as the error term with conditional normal distribution $\varepsilon_j^M \sim N\left(0, \sigma^2_{M|}\right)$. In the outcome equations (Equation 2), we use a multilevel model with Y_{ij} as the outcome for individual i in cluster j, and ε_{ij}^Y as the normally distributed level one error term $\varepsilon_{ij}^Y \sim N\left(0, \sigma^2_Y\right)$. At the cluster-level, we use b as the conditional relationship between the mediator and the outcome, c' as the direct effect of the predictor, and u_{0j} as the cluster-level random effects with conditional normal distribution $u_{0j} \sim N\left(0, \tau^2_{|}\right)$.

The second analytic model again employs a cluster-level treatment and individual-level outcome but utilizes an individual-level mediator (e.g., 2-1-1 mediation). For this type of multilevel mediation multilevel models are necessary for both the mediator and outcome such that (Pituch & Stapleton, 2012; Raudenbush & Bryk, 2002; VanderWeele, 2010)

$$\begin{aligned} M_{ij} &= \pi_{0j} + \varepsilon_{ij}^M \quad \varepsilon_{ij}^M \sim N\left(0, \sigma^2_M\right) \\ \pi_{0j} &= \zeta_{00} + aT_j + u_{0j}^M \quad u_{0j}^M \sim N\left(0, \tau^2_{M|}\right). \end{aligned} \tag{3}$$

Here, the mediator, M_{ij}, is measured at the individual level with individual- and group-level variance $\sigma^2_{M|}$ and $\tau^2_{M|}$, respectively. Other terms retain similar meaning to those described for Equations (1) and (2). At the cluster level the mediator model includes u^M_{0j} as the normally distributed mediator group-level random effects with a true residual variability of $\tau^2_{M|}$.

The multilevel model of the outcome is also adjusted to accompany the added complexities of 2-1-1 mediation (e.g., Pituch & Stapleton, 2012). Now,

$$\begin{aligned} Y_{ij} &= \beta_{0j} + b_1\left(M_{ij} - \overline{M}_j\right) + \varepsilon^Y_{ij} \quad \varepsilon^Y_{ij} \sim N\left(0, \sigma^2_{Y|}\right) \\ \beta_{0j} &= \gamma_{00} + B\overline{M}_j + c'T_j + u^Y_j \quad u^Y_j \sim N\left(0, \tau^2_{Y|}\right). \end{aligned} \tag{4}$$

with most terms retaining similar meaning to those in Equation (2). Additional terms include b_1 as the path coefficient capturing the individual-level conditional relationship between the mediator (M_{ij}) and the outcome and $\overline{M}_j$ as the cluster-level mean of the mediator with coefficient B capturing the conditional overall (cluster- and individual-level) relationship between the mediator and outcome. Studies with an individual-level mediator should allow for an individual outcome to be influenced by the individual-level mediator value (M_{ij}) and the mean mediator value of the cluster ($\overline{M}_j$).

Across each of these models the product of the treatment-mediator and mediator-outcome path coefficients provide a point estimate of the mediated effect such that 2-2-1 mediated effects can be estimated with

$$\text{ME} = ab, \tag{5}$$

and the 2-1-1 overall or cumulative mediation effect is estimated by (e.g., VanderWeele, 2010)

$$\text{ME} = aB. \tag{6}$$

3 | TESTING MEDIATION EFFECTS

There are a variety of well-established tests for mediation effects (i.e., ME). We include five mediation tests along with the p_3 method in our simulation studies. First, the Sobel test or z-test which has a legacy of assessing mediation but has recently come under scrutiny due to low power and inaccurate Type I error rates (Hayes & Scharkow, 2013). Second, the joint test which determines the significance of the a and b paths separately with inferences regarding the mediation effect determined based on each sub-test rejecting the null hypothesis of a zero path coefficient. The joint test has performed well in terms of power and Type I error rates for both single and multilevel mediation but it does not provide a summative descriptor of the mediated effect such as a confidence interval or p value (e.g., Hayes & Scharkow, 2013; Kelcey, Dong, Spybrook, & Cox, 2017; Kelcey, Dong, Spybrook, & Shen, 2017; Pituch et al., 2006).

The third test included is the MC interval test. It represents a resampling based alternative to the Wald-like approaches in the Sobel and joint test (Preacher & Selig, 2012). The test typically assumes a multivariate normal sampling distribution of path coefficients with means, variances, and covariances based on maximum likelihood estimates (Preacher & Selig, 2012). It then employs the primary path coefficient estimates and their error variances to simulate draws

from the posterior distribution of the mediation effect. Confidence intervals from this estimated sampling distribution allow inferential decisions regarding the mediated effect. MC interval test performance is similar to bootstrapping approaches but its use of the estimated path coefficients and their error variances allows for this resampling-based approach to be available in the study design phase (Hayes & Scharkow, 2013; MacKinnon, Lockwood, & Williams, 2004; Preacher & Selig, 2012).

Finally, we include the BC bootstrap and parametric percentile bootstrap and follow the bootstrapping procedures and methods set forth in Pituch et al. (2006). Bootstrapping represents a class of resampling based methods that determine the significance of a mediated effect by approximating its sampling distribution through repeated sampling of observed data. As with the MC interval test, resampling to approximate the distribution of the statistic (e.g., ab) avoids any a priori assumption about the statistic. Across both single level and multilevel mediation, bootstrap methods generally perform well in terms of power and Type I error rates. A significant body of the literature recommends the BC bootstrap or parametric percentile bootstrap (Hayes & Scharkow, 2013; MacKinnon et al., 2004; Pituch et al., 2006; Pituch & Stapleton, 2008) but concerns have been raised about inflated Type I error rates in BC bootstrap methods (Biesanz et al., 2010; Hayes & Scharkow, 2013). A more fundamental drawback of all bootstrapping approaches is a reliance on observed data for resampling. This precludes bootstrapping from consideration during the design phase of a study (i.e., before data have been collected).

Bootstrapping methods along with the Sobel test and MC interval test determine the significance of the mediated effect through confidence intervals based on point and precision estimates. These approaches disregard the complex null hypothesis of no mediation and open the approaches to inaccuracies in Type I error rate and possibly power rate limitations in smaller sample sizes (Biesanz et al., 2010). The joint test uses a composite null hypothesis approach that avoids the complex null of no mediation but fails to provide an effect size, p value, or confidence interval related to the mediated effect. This combination of issues and limitations has prevented broad and strong recommendations for any one method to test multilevel mediation effects. The p_3 method has the potential to address these concerns because it employs a composite approach (i.e., test the a path and b path separately), is formulated for complex null hypotheses, and produces a single summative p value for the mediated effect.

4 | PARTIAL POSTERIOR PREDICTIVE DISTRIBUTION TEST

Justification for developing the p_3 method for multilevel mediation stems from the difficulties of testing mediated effects. These difficulties often begin with the complexity of the null hypothesis of no mediation. Recall, a mediated effect is typically quantified using the product of the treatment-mediator and mediator-outcome path coefficients creating a null mediation effect under three different conditions (a) both the treatment-mediator and mediator-outcome path coefficients are zero (i.e., $a = 0$ and $b = 0$), (b) the treatment-mediator is zero but the mediator-outcome path coefficient is nonzero (i.e., $a = 0$ and $b \neq 0$), or (c) the treatment-mediator is nonzero but the mediator-outcome path coefficient is zero (i.e., $a \neq 0$ and $b = 0$). Note, we conceptually use the a and b variables here to represent any of the different treatment-mediator or mediator-outcome path coefficients represented in the different analytic models described above (e.g., B). In each of the null cases the mediation effect is zero but each scenario produces different sampling distributions. Effective tracking of the sampling distribution of the mediation effect

and subsequent determination of p values is only possible after evaluating the non-null path. For example, when $a = 0$ the sampling distribution of the mediated effect is heavily influenced by the magnitude of the b path and, conversely, when the mediation effect is zero because $b = 0$ the sampling distribution of the mediated effect is heavily influenced by the magnitude of the a path.

Under these conditions it is challenging to test the null hypothesis of mediated effects because there are many combinations of a and b that result in a true null hypothesis (i.e., mediation has a complex null hypothesis; MacKinnon et al., 2002). Most mediation tests use a point estimate and an estimate of precision to test the null hypothesis of no mediation, the joint test being a notable exception (Biesanz et al., 2010; MacKinnon et al., 2002). However, there are consequences to disregarding the complex nature of the null hypothesis. Under different sampling distributions it is the possible to draw different inferences about the significance of an observed mediated effect depending on the specific null hypothesis under consideration (Biesanz et al., 2010). Put differently, it is possible to find a variety of p values from a single observed mediated effect because of the possibility of many different sampling distributions of the mediated effect under a complex null hypothesis.

The p_3 method is especially well suited for testing mediated effects because it takes a composite approach with subordinate tests formulated to identify p values under complex null hypotheses. Specifically, the p_3 method incorporates one subordinate test assuming the treatment-mediator relationship is zero (without specifying the mediator-outcome relationship) and a second test assuming the mediator-outcome relationship is zero (without specifying the treatment-mediator relationship). Path-specific inferences for each subtest are generated by treating the unspecified relationship as a nuisance parameter, resampling this nuisance parameter, and calculating a p value for each draw. Inferences involving the overall test of no mediation are determined using the maximum p value of the two subordinate tests. The inferential approach of the p_3 method requires both subordinate tests to be significant in order to reject the null hypothesis of no mediation effect with the largest p value identified by the two subordinate tests a conservative summative value of the overall inferential test.

When implementing this method, the evaluation or assignment of a reasonable value to the nuisance parameter (i.e., the path not currently being assessed or the unspecified path) for each subordinate null hypothesis is required. One obvious choice is the maximum likelihood point estimate of the nuisance parameter sometimes referred to as the plug-in method. Unfortunately, this estimate assumes the nuisance parameter is known and therefore disregards any uncertainty in the estimate. It is possible to incorporate this uncertainty by considering the full posterior distribution of the nuisance parameter (e.g., posterior predictive p value; Bayarri & Berger, 2000). The inferences under this approach can be generated using weights proportional to the posterior density.

The posterior predictive approach is an improvement compared to the plug-in method as it incorporates the uncertainty in the nuisance parameter but it has its own limitation. Inferences using the posterior predictive approach are weakened by the dependence introduced through the methods twofold use of the observed data. The posterior predictive approach first estimates the mediation effect and then the posterior predictive distribution of the nuisance parameter. This is the specific weakness addressed by the partial posterior predictive approach. The p_3 method eliminates the dependency between the mediation effect and nuisance parameter by adjusting the posterior predictive distribution of the nuisance parameter by the density of the observed mediation effect under the null hypothesis. As a result, for each subordinate null hypothesis, the p_3 method asymptotically provides the true probability of observing the mediation effect given the null hypothesis (Robins et al., 2000).

4.1 | Testing multilevel mediation effects

Having conceptually outlined the p_3 method, we detail the extension of the test to multilevel mediation. We begin with the test statistics that inform the two subtests of the composite approach. For our implementation we used the noncentral t-test statistics (e.g., t_a and t_b). Here, t_a is the test statistic for the treatment-mediator path and is used to determine inferences regarding the a path coefficient. Conversely, t_b is the test statistic for the mediator-outcome path and is used to determine inferences regarding the b path coefficient. We formed the t-test statistics as

$$t_a = a/\sqrt{\sigma_a^2} \quad \text{and} \quad t_b = b/\sqrt{\sigma_b^2}, \tag{7}$$

with σ_a^2 and σ_b^2 as the error variances of the respective paths. For inference, we use as the referent distributions a pair of t-distribution with degrees of freedom based on cluster sample size (n_2) and number of cluster level predictors such that $df = n_2 - q - 1$ (Kelcey, Dong, Spybrook, & Shen, 2017; Kenny & Judd, 2014; Raudenbush & Bryk, 2002). The estimated error variances of each path (σ_a^2 and σ_b^2) are obtained using the diagonal of the observed information matrix associated with the maximum likelihood estimates of the parameters.

The p_3 method requires significance testing of each t-statistic to make inferential decisions about the separate paths and overall mediated effect. The complex null of no mediation indicates proper testing of one path requires accounting for the value of the other path (i.e., the nuisance parameter). More formally, the subordinate complex null hypotheses of the composite approach of the p_3 method are (a) $t_a = 0$ with t_b as the nuisance parameter and (b) $t_b = 0$ with t_a as the nuisance parameter.

In order to track the nuisance parameter under each subordinate hypothesis, the multilevel extension of the p_3 method resamples values to approximate p values under each subordinate null hypothesis. Recall, that changes in the nuisance parameter lead to differences in the sampling distribution of the mediated effect and therefore different p values under the null hypothesis. Capturing these different p values creates a distribution of p values allowing us to incorporate the uncertainty of our estimated nuisance parameter and eliminate the dependency between the mediation effect and nuisance parameter.

The resulting structures of the subordinate tests are.

$$p_{a=0,b=b_k} \approx \frac{\sum_{k=1}^{K} \frac{p\left(t_{\hat{a}}t_{\hat{b}}|t_a=0,t_b=t_{b_k}\right)}{f\left(t_{\hat{a}}t_{\hat{b}}|t_a=0,t_b=t_{b_k}\right)}}{\sum_{k=1}^{K} \frac{1}{f\left(t_{\hat{a}}t_{\hat{b}}|t_a=0,t_b=t_{b_k}\right)}} \text{ and } p_{a=a_k,b=0} \approx \frac{\sum_{k=1}^{K} \frac{p\left(t_{\hat{a}}t_{\hat{b}}|t_a=t_{a_k},t_b=0\right)}{f\left(t_{\hat{a}}t_{\hat{b}}|t_a=t_{a_k},t_b=0\right)}}{\sum_{k=1}^{K} \frac{1}{f\left(t_{\hat{a}}t_{\hat{b}}|t_a=t_{a_k},t_b=0\right)}}. \tag{8}$$

Let $p_{a=0,b=b_k}$ and $p_{a=a_k,b=0}$ be the respective partial posterior p values under the complex null hypothesis $t_a = 0$ with t_b as the nuisance parameter and $t_b = 0$ with t_a as the nuisance parameter. Further, let K be the number of draws from the posterior distribution of the respective nuisance parameter and k be a specific draw. Last, $f\left(t_{\hat{a}}t_{\hat{b}}|t_a = 0, t_b = t_{b_k}\right)$ and $f\left(t_{\hat{a}}t_{\hat{b}}|t_a = t_{a_k}, t_b = 0\right)$ are the densities of the observed mediation effect under the respective subordinate hypotheses. Adjustment—on the basis of these densities—addresses the partial association between the mediated effect and the nuisance parameter (i.e., t_b and t_a, respectively, in the above formulations) that arises from the dual use of the nuisance parameter to compute the posterior distribution and

estimate the mediated effect. Densities and p values are empirically estimated using the product of the appropriate t-statistics under, for example, 10,000 or more draws.

The final result of this composite test yields two p values, the largest value from each of the subordinate tests in Equation (8). We adopt the conservative approach and draw inferences regarding the mediated effect from the larger of the two p values. The formulation in Equation (8) utilizes path notation from the 2-2-1 analytic model (t_a and t_b) but it is applicable to a wide range of multilevel mediation effects (e.g., ME = aB with t_a and t_B) as long as the correct path estimate, error variance, and degrees of freedom are utilized in the formulation.

In summary, the final p value of the p_3 method is the largest p value from the two posterior distributions of p values for the a path and b path. This is conceptually similar to the joint test but the p_3 method appropriately incorporates the complex null hypothesis of mediation effects. Consider the p values from other common mediation tests. The p-value under the Sobel test reflects the location of a z-score formed using the estimated mediation effect and its SE on a standard normal distribution. The joint test does not have a p value directly associated with the mediation effect but two associated p values, one each reflecting the location of t statistics formed with the a path and b path coefficients and their SEs and compared to a t-distribution. The MC interval test and bootstrapping methods do not produce a specific p value as inferential decisions are based on a 95% confidence interval from the distribution of sampled mediation effects.

The literature has noted the p_3 method as the most advantageous for determining p values when faced with a complex null hypothesis (Bayarri & Berger, 2000; Biesanz et al., 2010; Robins et al., 2000) but its performance when applied to determining the significance of multilevel mediation effects is unknown.

5 | SIMULATION

To provide an assessment as to the relative and absolute utility of the p_3 method, we probed its performance in simulated two-level cluster-randomized multilevel mediation studies (Schochet, 2011; Spybrook et al., 2016). These simulations demonstrate the accuracy of our p_3 method derivations and help us assess its performance in terms of power and Type I error rate. Additionally, the inclusion of five other mediation tests allows us to gauge the p_3 method in more relative terms.

We generated 500 datasets for each condition of the fully crossed design (see Table 1) based on the analytic models above. The cluster-level treatment indicator was coded as ±0.5 with continuous mediator and outcome variables for both the 2-2-1 and 2-1-1 designs. Next, we analyzed each of the data sets using the aforementioned mediation tests and tracked the number of times a test found a statistically significant mediation effect. In the null condition (a and/or $b = 0$), any statistically significant result represented a Type I error. Conversely, when the alternative hypothesis was true (i.e., ME > 0), the proportion of significant results represented the power of the test to detect the mediated effect.

We systematically varied the a and b (or B) path coefficient values to create three null and two nonnull conditions (see Table 1) matching previous simulation studies (e.g., Pituch & Stapleton, 2008). We included two sample size conditions by pairing clusters of 40 and 80 with individuals per cluster of 20 and 40, respectively ($n_2 = 40$ and $n_1 = 20$; $n_2 = 80$ and $n_1 = 40$). These sample sizes represent small but typical two-level multilevel mediation studies (e.g., Schochet, 2011; Spybrook et al., 2016). Limited sample sizes are also appropriate based on prior

TABLE 1 Simulation conditions summary table.

		Conditions			
Test	**Outcome**	**Model**	**Data (residuals)**	**Sample size**	**Mediated effect (*ab*)**
Sobel	Power	2-2-1	Normal	$n_2 = 40\ n_1 = 20$	$a = 0.6\ b = 0.4$
Joint	Type I error	2-1-1	Nonnormal	$n_2 = 80\ n_1 = 40$	$a = 0.3\ b = 0.1$
MC					$a = 0\ b = 0$
% boot					$a = 0\ b = 0.4$
BC boot					$a = 0.6\ b = 0$
P3					

Note: c' is held constant at 0.1 across conditions. For 2-2-1 mediation, $\sigma^2_M = 0.9$, $\sigma^2_Y = 0.8$, and $\tau^2_Y = 0.2$. For 2-1-1 mediation, $\sigma^2_M = \sigma^2_Y = 0.8$, $\tau^2_M = \tau^2_Y = 0.2$, and the b notation above represents the B used in the 2-1-1 analytic model.

theoretical and simulation literature demonstrating methods converge in larger sample sizes (e.g., Hayes & Scharkow, 2013).

The remaining parameters influencing the power to detect mediation effects were held constant throughout the simulations. The chosen theoretical Type I error rate was set at $\alpha = .05$ and the conditional direct effect, c' was held at 0.1 because it has little influence on mediation test power (Kelcey, Dong, Spybrook, & Cox, 2017; Kelcey, Dong, Spybrook, & Shen, 2017) and this value reflects previous simulation studies (e.g., Pituch et al., 2006; Tofighi, West, & MacKinnon, 2013). Finally, the variance components for each mediation type varied slightly across models but were held constant within each type. For 2-2-1 mediation, $\sigma^2_M = 0.9$, $\sigma^2_Y = 0.8$, and $\tau^2_Y = 0.2$ resulting in an unconditional intraclass correlation coefficient for the outcome of $\rho_Y = 0.2$. For 2-1-1 mediation, $\sigma^2_M = \sigma^2_Y = 0.8$ and $\tau^2_M = \tau^2_Y = 0.2$ resulting in an unconditional intraclass correlation coefficient for the outcome and mediator of $\rho_Y = 0.2$ and $\rho_M = 0.2$, respectively.

5.1 | Nonnormal data

An additional consideration when testing mediation effects is the possibility of nonnormal data (i.e., skewness and kurtosis in the mediator and outcome variables). While often disregarded, these types of data are common in applied research (e.g., Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013) and when considered demonstrate a substantial and detrimental influence on mediation test performance (Biesanz et al., 2010; Pituch & Stapleton, 2008). For example, Pituch and Stapleton (2008) found increased inaccuracies in Type I error rate across several types of mediation tests in a multilevel context with nonnormal data. Additionally, they found mediation tests specifically designed to accommodate nonnormal data (e.g., BC nonparametric bootstrap) still experienced performance issues (e.g., inflated Type I error) and differences among tests in terms of power rates appear the most pronounced under critical conditions (e.g., small sample sizes). The p_3 method has not been examined in multilevel settings with nonnormal data but Biesanz et al. (2010) did consider nonnormal data when they applied the p_3 method to single-level mediation. Under those conditions the p_3 method performed relatively well in terms of Type I error and power in comparison to other mediation tests included in the study.

We replicated the simulation study described above with nonnormal data generated by transforming the residuals of the mediator and dependent variable to have skewness of 2 and kurtosis of 7 (Biesanz et al., 2010; Fleishman, 1978). This type of nonnormal data is a common occurrence in applied research (e.g., Blanca et al., 2013; Micceri, 1989). We expect the estimated mediation effect (e.g., *ab*) to remain unbiased even when the normality assumption fails to hold (Little & Rubin, 2002). However, SEs and therefore the methods employed in this study that utilize the SEs will be affected (e.g., Sobel test).

To conclude, we evaluate the performance of six mediation tests aimed at detecting two types of multilevel mediation effects from a cluster-randomized study. The evaluation includes three null and two nonnull conditions that vary by treatment-mediator and mediator-outcome path coefficient values. In each condition we generate 500 datasets and track the rejection of the null hypothesis for each test. The MC interval test and novel p_3 method uses 10,000 draws to determine the significance of the mediated effect. The bootstrapping methods utilize 1000 resampled datasets. We then repeated this investigation using nonnormal data. All data generation and analyses were completed in *R* using the *lme4* package and author created scripts (see Appendix A for p_3 method code and Data S1 for complete R code).

6 | RESULTS

The primary outcomes of interest were Type I error and power rate. Table 2 presents results for the simulated cluster-randomized study of 2-2-1 mediation and Table 3 presents results for the simulated cluster-randomized study of 2-1-1 mediation. Type I error rates were captured when mediation tests were applied to data generated under the null condition and the test—incorrectly—identified a significant mediated effect. To improve the presentation and interpretation of these results we apply two criteria utilized in similar studies (e.g., Biesanz et al., 2010). First, we bold values that fall outside of the 2.5%–7.5% range identified by Bradley's (1978) liberal criterion. Second, we italicize values that fall outside the 3.5%–6.5% range identified by Serlin (Serlin, 2000; Serlin & Lapsley, 1985). To further highlight inflated Type I error rates those cells that exceed upper levels of the criteria stated are shaded.

In the two conditions when data were generated to have a nonnull mediated effect power rates were determined using the number of times the test correctly identified the significant effect. To improve the readability of the power rate section of each table we bold the highest power rate in a specific condition and bold and italicize the second highest value (e.g., Biesanz et al., 2010).

Overall, the null condition created lower than expected Type I error rates. This was particularly true when both the treatment-mediator path and mediator-outcome path were null (i.e., $a = 0$ and $b = 0$). In this double null case, Type I error rates were well below the expected .05 value across all mediation types, tests, and conditions. The accuracy of Type I error rates varied in the two other null conditions depending on the type of multilevel mediation being considered but were still consistently lower than expected. Inflated Type I error rates were much less common. Across the 144 cells representing Type I error rates only five cells exceeded the criteria for inflated Type I error rate. Of note, four of the five cases were a result under the p_3 method.

In terms of overall results related to power, the selected mediation tests performed similarly across many of the conditions. The notable exception being the Sobel test which is consistently underachieved. Even though differences in power rates achieved by the tests were relatively close, the p_3 method and BC bootstrap method were consistently top performers. As noted in the previous literature, test performance converged in well-powered designs with larger samples sizes

TABLE 2 Power and Type I Error rate by mediation test for detecting 2-2-1 mediation in a cluster-randomized study.

	$n_2 = 40$ and $n_1 = 20$					$n_2 = 80$ and $n_1 = 40$				
	Power		**Type I error**			**Power**		**Type I error**		
	$a:b$		$a:b$			$a:b$		$a:b$		
Normal data										
Test	.6:.4	.3:.1	.0:.0	.6:.0	.0:.4	.6:.4	.3:.1	.0:.0	.6:.0	.0:.4
Sobel	0.526	0.026	**0.000**	**0.008**	0.060	0.858	0.170	**0.000**	**0.010**	0.058
Joint	0.528	0.070	**0.002**	**0.018**	0.060	0.854	0.322	**0.002**	0.032	0.056
MC	0.552	0.070	**0.002**	**0.016**	*0.034*	***0.866***	0.320	**0.002**	**0.018**	**0.024**
% boot	***0.562***	0.070	**0.002**	**0.008**	*0.032*	0.864	0.304	**0.002**	**0.018**	*0.026*
BC boot	**0.568**	**0.130**	**0.002**	*0.028*	0.038	**0.870**	**0.400**	**0.004**	0.036	*0.028*
P3	0.552	***0.086***	**0.004**	**0.024**	0.062	0.858	***0.368***	**0.004**	0.042	0.058
Nonnormal data										
Sobel	0.522	0.022	**0.000**	**0.004**	0.048	0.840	0.140	**0.000**	**0.008**	0.052
Joint	0.524	0.082	**0.002**	0.036	0.050	0.838	0.278	**0.000**	0.030	0.052
MC	***0.546***	0.084	**0.000**	**0.022**	*0.030*	0.844	0.276	**0.000**	**0.024**	*0.028*
% boot	0.544	0.078	**0.000**	**0.018**	*0.034*	***0.860***	0.272	**0.000**	**0.024**	*0.028*
BC boot	**0.556**	**0.130**	**0.002**	0.038	0.038	**0.868**	***0.374***	**0.004**	0.036	*0.028*
P3	0.544	***0.094***	**0.004**	0.040	*0.066*	0.840	***0.326***	**0.000**	0.052	0.052

Notes: For power, the largest power rate is bolded and the second largest is italicized. For type I error, values outside of 2.5%–7.5% are bolded, values outside of 3.5%–6.5% are italicized, and cells that exceed upper levels of the criteria stated are shaded. The conditions for this simulation include: $\alpha = .05$, $c' = 0.1$, $\sigma^2_{M|} = 0.9$, $\sigma^2_Y = 0.8$, and $\tau^2_| = 0.2$.

TABLE 3 Power and Type I Error rate by mediation test for detecting 2-1-1 mediation in a cluster-randomized study.

	$n_2 = 40$ and $n_1 = 20$					$n_2 = 80$ and $n_1 = 40$				
	Power		**Type I error**			**Power**		**Type I error**		
	$a:B\,(b1)$		$a:B\,(b1)$			$a:B\,(b1)$		$a:B\,(b1)$		
Normal data										
Test	.6:.4 (.2)	.3:.2 (.1)	.0:.0 (.0)	.6:.0 (.0)	.0:.4 (.2)	.6:.4 (.2)	.3:.2 (.1)	.0:.0 (.0)	.6:.0 (.0)	.0:.4 (.2)
Sobel	0.644	0.118	**0.000**	0.060	**0.004**	***0.940***	0.376	**0.000**	0.054	**0.018**
Joint	0.654	0.232	**0.000**	0.066	**0.024**	***0.940***	0.418	**0.004**	0.054	0.050
MC	***0.674***	0.228	**0.000**	0.042	**0.014**	***0.940***	0.434	**0.002**	**0.020**	*0.026*
% boot	0.646	0.216	**0.000**	0.042	**0.016**	0.934	0.432	**0.000**	**0.020**	**0.022**
BC boot	0.670	**0.272**	**0.002**	0.044	**0.022**	**0.944**	**0.464**	**0.004**	**0.022**	0.038
P3	**0.686**	***0.254***	**0.002**	**0.078**	0.038	0.938	***0.450***	**0.008**	0.054	0.056
Nonnormal data										
Sobel	0.628	0.100	**0.000**	*0.034*	**0.004**	**0.938**	0.418	**0.000**	0.064	**0.024**
Joint	0.634	0.184	**0.000**	0.038	**0.022**	**0.938**	0.442	**0.002**	0.064	0.046
MC	0.658	0.190	**0.000**	*0.030*	**0.006**	***0.936***	0.450	**0.000**	*0.034*	*0.034*
% boot	0.642	0.168	**0.000**	*0.026*	**0.006**	0.928	0.452	**0.000**	*0.028*	0.036
BC boot	**0.676**	**0.250**	**0.002**	*0.034*	*0.026*	***0.936***	**0.488**	**0.000**	0.036	0.058
P3	***0.670***	***0.214***	**0.000**	0.062	*0.032*	***0.936***	***0.478***	**0.002**	*0.070*	*0.068*

Notes: For power, the largest power rate is bolded and the second largest is italicized. For type I error, values outside of 2.5%–7.5% are bolded, values outside of 3.5%–6.5% are italicized, and cells that exceed upper levels of the criteria stated are shaded. The conditions for this simulation include: $\alpha = .05$, $c' = 0.1$, $\sigma^2_M = \sigma^2_{Y|} = 0.8$, and $\tau^2_{M|} = \tau^2_{Y|} = 0.2$.

(see Table 3, when $n_2 = 80$ and $n_1 = 40$ with $a : b = 0.6 : 0.4$). The remainder of the Section 6 is divided into two parts based on the mediated effect (i.e., 2-2-1 and 2-1-1). Within each section we describe more specific results involving Type I error and power with normal and nonnormal data.

6.1 | Cluster randomized studies of 2-2-1 mediation

We found very low Type I error rates in the double null case of 2-2-1 mediation but results were more nuanced when only one path was null. Overall, the Type I error rate was lower when the null condition was a result of a null mediator-outcome path (i.e., $b = 0$). Sample size also influenced Type I error rates but this influence varied based on the null condition. For example, Type I error rates were typically greater in the larger sample size condition when $b = 0$ but smaller in the larger sample size condition when $a = 0$. The influence of nonnormal data on Type I error rate when testing 2-2-1 mediation was test specific with increasing and decreasing rates observed across the six mediation tests. As far as relative test performance, the p_3 method, BC bootstrap, and joint test performed well with 2-2-1 mediation in terms of Type I error rate outside of the double null condition.

When considering power, all tests performed well in the large sample size and large mediated effect condition and few substantial differences were noted in the smaller sample size and large mediated effect condition. However, test power performance differed in the more difficult $a = 0.3$ and $b = 0.1$ condition. Here, power rates in the larger sample size and normal data condition ranged from 17% with the Sobel test to 40% with the BC bootstrap. Across all conditions with $a = 0.3$ and $b = 0.1$ the BC bootstrap and p_3 method achieved the greatest power rates. We also noted a general but minor decrease in power to detect 2-2-1 mediation effects when comparing tests across the normal and nonnormal data conditions.

6.2 | Cluster randomized studies of 2-1-1 mediation

For 2-1-1 mediation, all tests again failed to achieve a Type I error rate approaching the .05 level when both path values were 0 (i.e., double null condition). In the other null conditions for 2-1-1 mediation, accuracy of Type I error rates varied by sample size and mediation test but these results did not parallel the results involving 2-2-1 mediation. For example, we found some inflated Type I error rates in the smaller sample size condition but overall Type I error rate was more accurate in the $B = 0$ null condition. While not as conclusive, the opposite was observed in the larger sample size condition where Type I error rate was more accurate in the $a = 0$ null condition and less accurate in the $B = 0$ null condition. Additionally, inflated Type I error rates were more common when testing 2-1-1 mediation.

Unlike the results noted for 2-2-1 mediation, the effect of nonnormal data on Type I error rate varied by sample size for 2-1-1 mediation. In the smaller sample size condition, Type I error rates generally decreased in the nonnormal data condition while they increased in the larger sample size condition. As for the overall performance of specific tests, the p_3 method, BC bootstrap, and joint test again performed well but all demonstrated some cases of inaccurate Type I error rates. In terms of statistical power to detect 2-1-1 mediation, the tests performed very similarly with the greatest power rates achieved with BC bootstrapping, the p_3 method, and the MC interval test. We did see a similar minor decrease in power when conducting mediation tests with nonnormal data but not across all conditions (e.g., $n_2 = 80$, $n_1 = 40$, and $a : B = 0.3 : 0.2$).

To summarize, we investigated the power to detect a mediated effect and the accuracy of Type I error rates for a variety of mediation tests in simulated cluster-randomized studies with two different types of multilevel mediation effects. The primary purpose of this simulation was to serve as a case study of the newly developed p_3 method for testing multilevel mediation. For Type I error rates across the different null conditions considered here, the p_3 method performed admirably compared to the other established tests even the highly recommended BC bootstrap but did suffer from some inflated Type I error rates. While valuable, the method is far from a comprehensive solution to the problems associated with testing multilevel mediation effects. For example, all tests in this investigation suffered extremely low Type I error rates when both the treatment-mediator and mediator-outcome path were nonsignificant.

In terms of power rates, the p_3 method ranked first or second in 11 of the 16 conditions. It along with the BC bootstrap and the MC interval test performed consistently well. The p_3 method and BC bootstrap were particularly advantageous when data were nonnormal and in conditions with a smaller mediated effect. Interestingly, we noted some differences in mediation test performance with normal versus nonnormal data but relative test performance was fairly consistent (e.g., test performance rank remained the same in normal and nonnormal conditions). Overall, the newly developed p_3 method performed well compared to the current set of mediation tests but any advantages should also be considered from a practical standpoint (e.g., additional power of 1%–2%, more accurate Type I error rate by 1%–2%).

7 | DISCUSSION

The literature has found it difficult to develop sound inferential tests of mediated effects because their composite nature creates a complex null hypothesis (MacKinnon et al., 2002; MacKinnon et al., 2007). For example, inaccurate Type I error rates and low power are persistent issues when conducting inferential tests on multilevel mediation effects. In this study, we sought to address this limitation by developing and investigating the p_3 method. This new and promising test is designed for and sensitive to the composite nature of mediation. Our results suggest the p_3 method is an appropriate and effective test of multilevel mediation effects and in relative terms its performance was commensurate or better than currently available inferential tests.

The availability of the p_3 method to determine the significance of multilevel mediation effects is an important development for three reasons: (a) The p_3 method performs relatively well compared to commonly utilized mediation tests in terms of power. Increases in power to detect multilevel mediation effects increases the capacity of applied researchers to conduct studies aimed at multilevel mediation. (b) The p_3 method produces a single representative p value. This type of clear and concise summative value allows easy interpretation of results from the p_3 method and directly overcomes a disadvantage of using the joint test. (c) With properly formatted variance components, the p_3 method can be utilized for study design and planning (e.g., a priori power analysis).

Quickly reformulating the error variance of the path coefficients allows the p_3 method to be available before data collection. In the context of experimental designs this is a crucial feature and one that prevents the use of bootstrapping methods. As expressed, the estimates of σ_a^2 and σ_b^2 (or σ_B^2) are based on observed data, thus precluding the use of the p_3 method for study planning. However, it is possible to restructure these error variance formulations on the basis of the expected

information which can be tracked in terms of path coefficients, intraclass correlation coefficients, and, if applicable, variance explained by covariates (see Appendix B; Kelcey, Dong, Spybrook, & Cox, 2017; Kelcey, Dong, Spybrook, & Shen, 2017). The literature and other past research provide plausible estimates of these values allowing power analyses and other study planning activities to consider the p_3 method. Using these formulations, the newly extended p_3 method is accessible for power analyses or adequate sample size determination when planning group-randomized studies of multilevel mediation.

The p_3 method is not without limitations. In comparison to other mediation tests, it often avoids substantially underestimated Type I error rates but performs just as poorly as its peers under a double null condition (i.e., $a = 0$ and $b = 0$). When considering power, the p_3 method ranked among the best performers but was consistently less than BC bootstrapping and typically outperformed other tests by >5%. In many cases, power differences of >5% may not be practically meaningful. The p_3 method also has substantial computational demands requiring more time to complete inferential testing and as a newly developed test is less accessible and certainly less understood than other methods. Here, we provide R code to encourage utilization of the p_3 method and improve accessibility but recognize adoption of novel methods often takes substantial time and ongoing dissemination efforts.

Given the alignment between the p_3 method and inferential testing of multilevel mediation effects, it is worth asking why it did not substantially outperform the currently available tests. The strong overall performance of the p_3 method suggests that its composite null approach that explicitly addresses the complex null of no mediation is advantageous when considering the asymmetric sampling distributions of mediated effects. However, the strong results from other mediation tests indicate a robustness to any asymmetries and adequate accommodation of the complex null hypothesis of no mediation. Put differently, our results suggest, several confidence interval approaches (e.g., bootstrapping, MC interval test, and joint test) appropriately accommodate the asymmetric sampling distributions and complex null hypotheses of mediated effects even with smaller sample sizes. For example, the p_3 method performed well with nonnormal data but not well enough to distinguish itself from other tests as their relative rank remained fairly consistent across the normal and nonnormal data conditions.

Our results further indicate that test selection for multilevel mediation studies is context dependent and researchers must weigh multiple factors to select the most advantageous test. Generally, we recommend BC bootstrapping for testing multilevel mediation effects. It garners default status through consistent performance in terms of power and Type I error rate accuracy. However, bootstrapping approaches are unavailable for study planning so we suggest employing the MC interval test or the p_3 method for this purpose. These approaches still achieve consistently high power rates and relatively accurate Type I error rates but only require estimated path coefficients and their error variances to predict power. Applied researchers should consider these guidelines but also weigh their tolerance for Type I errors, interpretability of test results, and ease of implementation when selecting a test for multilevel mediation. We would, however, make a final recommendation to avoid the Sobel test because it consistently and substantially underperformed compared to the other tests considered in this study.

To conclude, the p_3 method represents a continued push to better align mediation tests, inferential decisions, and mediated effects. While we echo general recommendations for BC bootstrapping, the p_3 method was consistently the most powerful test available for study planning. Ultimately, these general guidelines should be disregarded by researchers in favor of test selection based on their study specific conditions and mediated effect of interest. We believe our

extension of the p_3 method to multilevel mediation and assessment of its potential aids in these efforts. Given some promising findings and remaining questions future research is important especially considering different analytic models, sample allocations, and path coefficient values. Additionally, a comparison of the p_3 method against Bayesian approaches to mediation analysis (e.g., Yuan & MacKinnon, 2009) could further distinguish advantageous conditions for either approach. This work and these future studies build on the capacity of applied researchers to conduct adequately powered multilevel mediation studies and therefore improves the quality of research across substantive fields.

FUNDING INFORMATION

This article is based on work funded by the National Science Foundation (grant number 1552535). The opinions expressed herein are those of the authors and not the funding agency.

DATA AVAILABILITY STATEMENT

Data are available on request from the authors. The data and simulation study R code that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Kyle Cox https://orcid.org/0000-0002-7173-4701

REFERENCES

Bayarri, M., & Berger, J. (2000). P values for composite null models. *Journal of the American Statistical Association*, *95*, 1127–1142.

Biesanz, J., Falk, C., & Savalei, V. (2010). Assessing mediational models: Testing and interval estimation for indirect effects. *Multivariate Behavioral Research*, *45*, 661–701.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*, 78–84.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521–532.

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.

Kelcey, B., Dong, N., Spybrook, J., & Cox, K. (2017). Statistical power for causally-defined indirect effects in group-randomized trials with individual-level mediators. *Journal of Educational and Behavioral Statistics*, *42*, 499–530.

Kelcey, B., Dong, N., Spybrook, J., & Shen, Z. (2017). Statistical power for causally-defined mediation in group-randomized studies. *Multivariate Behavioral Research*, *52*, 699–719.

Kenny, D. A., & Judd, C. M. (2014). Power anomalies in testing mediation. *Psychological Science*, *25*, 334–339.

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, *36*, 249–277.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken: Wiley.

MacKinnon, D. (2008). *Introduction to Statistical Mediation Analysis*. Hoboken: Lawrence Erlbaum Associates.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, *58*, 593–614.

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.

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*, 37–67.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166.

Pituch, K. A., & Stapleton, L. M. (2008). The performance of methods to test upper-level mediation in the presence of nonnormal data. *Multivariate Behavioral Research*, *43*, 237–267.

Pituch, K. A., & Stapleton, L. M. (2012). Distinguishing between cross- and cluster-level mediation processes in the cluster randomized trial. *Sociological Methods & Research*, *41*, 630–670.

Pituch, K. A., Stapleton, L. M., & Kang, J. Y. (2006). A comparison of single sample and bootstrap methods to assess mediation in cluster randomized trials. *Multivariate Behavioral Research*, *41*, 367–400.

Pituch, K. A., Whittaker, T. A., & Stapleton, L. M. (2005). A comparison of methods to test for mediation in multisite experiments. *Multivariate Behavioral Research*, *40*, 1–23.

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, *6*, 77–98.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks: Sage.

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). Asymptotic distribution of *p* values in composite null hypotheses. *Journal of the American Statistical Association*, *95*, 1143–1156.

Schochet, P. Z. (2011). Do typical RCTs of education interventions have sufficient statistical power for linking impacts on teacher practice and student achievement outcomes? *Journal of Educational and Behavioral Statistics*, *36*, 441–471.

Serlin, R. C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods*, *5*, 230–240.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*, 73–83.

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

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. *International Journal of Research and Method in Education*, *39*, 255–267.

Tofighi, D., West, S. G., & MacKinnon, D. P. (2013). Multilevel mediation analysis: The effects of omitted variables in the 1-1-1 model: Multilevel mediation. *British Journal of Mathematical and Statistical Psychology*, *66*, 290–307.

VanderWeele, T. J. (2010). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods & Research*, *38*, 515–544.

VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. New York: Oxford University Press.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*, 301–322.

Zhang, Z., Zyphur, M. J., & Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organizational Research Methods*, *12*, 695–719.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Cox, K., & Kelcey, B. (2023). A partial posterior *p* value test for multilevel mediation. *Statistica Neerlandica*, *77*(4), 408–428. https://doi.org/10.1111/stan.12291

APPENDIX A

```
#————————————————————-#
# PPP Function: R Code
# Testing 2-2-1 and 2-1-1 Mediation

## Partial Posterior Method

## Explanation of function/method:
# This requires two functions: PathANull, PathBNull
# Take the maximum p-value from the output of these functions
# Partial Posterior method is first p-value from each function's output
# Mediation model (221 or 211) dictates ta and tb values

## Note: this code will take a while to run, please be patient

## Terms:
# n_nullsim Number of draws in PathANull and PathBNull for each grid value
# ngrid Number of grid points evaluated within PathANull and PathBNull
# ndraws Number of posterior values to draw from ta and tb.


# PPP Function- PathANull

PathANull <- function(ta, tb, dfa, dfb, n_nullsim=1000000, ngrid=200, ndraws=50000){
   test <- ta*tb
  posteriorB<- (rnorm(ndraws,mean=0,sd=1) + (tb)*(sqrt(rchisq(n= ndraws,df=dfb,ncp=0)/
(dfb))))/(sqrt(rchisq(n= ndraws,df=(dfb+1),ncp=0)/(dfb)))

   pvalues<-matrix(NA,ncol=3,nrow= ngrid+1)

   for (i in 1: (ngrid+1)){

        x <-(i-1)*((max(posteriorB)-min(posteriorB))/ngrid) + min(posteriorB)

      yp <- (rnorm(n=n_nullsim,mean=0,sd=1) + x*(sqrt(rchisq(n= n_nullsim,df=dfb,ncp=0)/
(dfb))))/sqrt(rchisq(n= n_nullsim,df=dfb,ncp=0)/dfb)*(rnorm(n= n_nullsim,mean=0,sd=1)/
sqrt(rchisq(n= n_nullsim,df=dfa,ncp=0)/dfa))
      yd <- (rnorm(n=n_nullsim,mean=0,sd=1) + x*(sqrt(rchisq(n= n_nullsim,df=dfb,ncp=0)/
(dfb))))/sqrt(rchisq(n= n_nullsim,df=dfb,ncp=0)/dfb)*(rnorm(n= n_nullsim,mean=0,sd=1)/
sqrt(rchisq(n= n_nullsim,df=dfa,ncp=0)/dfa))

        FN<- ecdf(yp)
        kernden <- density(yd, from=test, to= test+1)

        pvalues[i,1]<-x
        pvalues[i,2]<-1-FN(abs(test))+ FN(-1* abs(test))
        pvalues[i,3]<-kernden$y[kernden$x == test]
   }
```

```
##Note:
#If the largest p-value is very small (<.0001), then return 0.
#The spline function will fail for very large observed values of ta and tb
#as p-values and densities will be 0 under the null hypotheses.

   if (max(pvalues[,2]>.0001)){
        pvalues<-subset(pvalues, pvalues[,3]> 1e-8)
        pvalues.lo <- smooth.spline(y=pvalues[,2],x=pvalues[,1],tol=1e-6)
        dvalues.lo <- smooth.spline(y=pvalues[,3],x=pvalues[,1],tol=1e-6)
        pvalx <-predict(pvalues.lo, data.frame(x= posteriorB))
        dvalx <-predict(dvalues.lo, data.frame(x= posteriorB))

        postval <- cbind(posteriorB,pvalx$y,dvalx$y)

        #Replacing zero densities with the smallest positive value.
        postval[,3][postval[,3]<=0]<- min(subset(postval, postval[,3]>0))
        postval[,2][postval[,2]<=0]<- min(subset(postval, postval[,2]>0))

        pplug<- as.numeric(predict(pvalues.lo, data.frame(x= tb))$y)
        postvalue <- mean(postval[,2])
        ppvalue <- sum(postval[,2]/postval[,3])/sum(1/postval[,3])

        if (ppvalue<0){ppvalue<-0}
        if (postvalue<0){postvalue<-0}
        if (pplug<0){pplug<-0}

        c(ppvalue, postvalue, pplug)
   } else c(0,0,0)
}


# PPP Functions- PathBNull

PathBNull <- function(ta, tb, dfa, dfb, n_nullsim=1000000, ngrid=200, ndraws=50000){
   test <- ta*tb
     posteriorA<- (rnorm(ndraws,mean=0,sd=1) + (ta)*(sqrt(rchisq(n= ndraws,df=dfa,ncp=0)/
(dfa))))/(sqrt(rchisq(n= ndraws,df=(dfa+1),ncp=0)/(dfa)))
   pvalues<-matrix(NA,ncol=3,nrow= ngrid+1)

   for (i in 1: (ngrid+1)){

        x <-(i-1)*((max(posteriorA)-min(posteriorA))/ngrid) + min(posteriorA)

       yp <- (rnorm(n=n_nullsim,mean=0,sd=1) + x*(sqrt(rchisq(n= n_nullsim,df=dfa,ncp=0)/
(dfa))))/sqrt(rchisq(n=  n_nullsim,df=dfa,ncp=0)/dfa)*(rnorm(n=  n_nullsim,mean=0,sd=1)/
sqrt(rchisq(n= n_nullsim,df=dfb,ncp=0)/dfb))
```

```
        yd <- (rnorm(n=n_nullsim,mean=0,sd=1) + x*(sqrt(rchisq(n= n_nullsim,df=dfa,ncp=0)/
(dfa))))/sqrt(rchisq(n=  n_nullsim,df=dfa,ncp=0)/dfa)*(rnorm(n=  n_nullsim,mean=0,sd=1)/
sqrt(rchisq(n= n_nullsim,df=dfb,ncp=0)/dfb))

        FN<- ecdf(yp)
        kernden <- density(yd, from=test, to= test+1)

        #cbind(kernden$x,kernden$y)
        pvalues[i,1]<-x
        pvalues[i,2]<-1-FN(abs(test))+ FN(-1* abs(test))
        pvalues[i,3]<-kernden$y[kernden$x == test]
    }

#See note above regarding very small p values

    if (max(pvalues[,2]>.0001)){
        pvalues<-subset(pvalues, pvalues[,3]> 1e-8) ###changed from 8 to 6
        pvalues.lo <- smooth.spline(y=pvalues[,2],x=pvalues[,1],tol=1e-6)
        dvalues.lo <- smooth.spline(y=pvalues[,3],x=pvalues[,1],tol=1e-6)
        pvalx <-predict(pvalues.lo, data.frame(x= posteriorA))
        dvalx <-predict(dvalues.lo, data.frame(x= posteriorA))

        postval <- cbind(posteriorA,pvalx$y,dvalx$y)

        #Replacing zero densities with the smallest positive value.
        postval[,3][postval[,3]<=0]<- min(subset(postval, postval[,3]>0))
        postval[,2][postval[,2]<=0]<- min(subset(postval, postval[,2]>0))

        pplug<- as.numeric(predict(pvalues.lo, data.frame(x= ta))$y)
        postvalue <- mean(postval[,2])
        ppvalue <- sum(postval[,2]/postval[,3])/sum(1/postval[,3])

        if (ppvalue<0){ppvalue<-0}
        if (postvalue<0){postvalue<-0}
        if (pplug<0){pplug<-0}

        c(ppvalue, postvalue, pplug)
    } else c(0,0,0)
}

#————————————————————-#
# PPP Application

# 1) Run code above to create PPP functions
# 2) Determine ta and tb and dfa and dfb using mediation power literature
    # 211- Kelcey, Dong, Spybrook, & Cox, 2017
    # 221- Kelcey, Dong, Spybrook, & Shen, 2017
```

```
# 3) Get posterior p value for each path using code below
# 4) Determine significance of mediated effect using the largest of the two p # values

a_path_p3<-PathANull(ta=, tb=, dfa=, dfb=,
n_nullsim=1000000, ngrid=200, ndraws=50000)

b_path_p3<-PathBNull(ta=, tb=, dfa=, dfb=,
n_nullsim=1000000, ngrid=200, ndraws=50000)
```

APPENDIX B

B.1 Error variance of the path coefficients in 2-2-1 mediation

For the 2-2-1 model in a group-randomized study with a balanced design the expected error variance of the a path coefficient can be estimated as (Kelcey, Dong, Spybrook, & Shen, 2017),

$$\sigma_a^2 = \frac{4\left(1-\left(a^2+4\right)\right)}{n_2}. \tag{B1}$$

Similarly, for the b path coefficient the expected error variance is

$$\sigma_b = \frac{\left(\rho - \frac{(ab+c\prime)^2}{4} - b^2\left(1-\frac{a^2}{4}\right)\right) + (1-\rho)/n_1}{n_2\left(1-\frac{a^2}{4}\right)}, \tag{B2}$$

where n_1 and n_2 are the individual per cluster and cluster sample size, respectively, ρ is the intraclass correlation for the outcome such that $\rho = \tau_{Y|}^2/\sigma_Y^2 + \tau_{Y|}^2$, and a, b, c', σ_Y^2, and $\tau_{Y|}^2$ retain their interpretation from Equations (1) and (2).

B.2 Error variance of the path coefficients in 2-1-1 mediation

For the 2-1-1 model in a group-randomized study with a balanced design the expected error variance of the a path coefficient can be estimated as (Kelcey, Dong, Spybrook, & Cox, 2017),

$$\sigma_a^2 = \frac{\rho_M\left(1-\frac{a^2}{4\rho_M}\right) + \frac{(1-\rho_M)}{n_1}}{n_2}. \tag{B3}$$

For the b path coefficient the expected error variance is

$$\sigma_B^2 = \frac{\left(\rho_Y - \frac{(aB+c\prime)^2}{4} - \left[\frac{4B^2\rho_M+4B^2(1-\rho_M)/n_1-a^2B^2}{4}\right] + (1-\rho_Y)\left(1-\left[\left(\frac{1-\rho_M}{1-\rho_Y}\right)b_1^2\right] - \left[\frac{(1-\rho_M)}{4(1-\rho_Y)}\right]\right)/n_1\right)}{n_2\left(\rho_M\left(1-\frac{a^2}{4\rho_M}\right) + \frac{(1-\rho_M)}{n_1}\right)}. \tag{B4}$$

Terms retain a similar meaning as previously described but now we have an intraclass correlation coefficient (ρ) for the outcome and mediator.