The Relative Trustworthiness of Inferential Tests of the Indirect Effect in Statistical Mediation Analysis: Does Method Really Matter?

Andrew F. Hayes and Michael Scharkow

Psychological Science published online 16 August 2013
DOI: 10.1177/0956797613480187

Psychological Science OnlineFirst, published on August 16, 2013 as doi:10.1177/0956797613480187

Research Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
XX(X) 1–10

DOI: 10.1177/0956797613480187
pss.sagepub.com
SAGE

The Relative Trustworthiness of Inferential Tests of the Indirect Effect in Statistical Mediation Analysis: Does Method Really Matter?

Andrew F. Hayes[1,2] and Michael Scharkow[3]
[1]School of Communication, The Ohio State University; [2]Department of Psychology, The Ohio State University; and [3]Institute of Communication Science, University of Hohenheim

Abstract

A content analysis of 2 years of *Psychological Science* articles reveals inconsistencies in how researchers make inferences about indirect effects when conducting a statistical mediation analysis. In this study, we examined the frequency with which popularly used tests disagree, whether the method an investigator uses makes a difference in the conclusion he or she will reach, and whether there is a most trustworthy test that can be recommended to balance practical and performance considerations. We found that tests agree much more frequently than they disagree, but disagreements are more common when an indirect effect exists than when it does not. We recommend the bias-corrected bootstrap confidence interval as the most trustworthy test if power is of utmost concern, although it can be slightly liberal in some circumstances. Investigators concerned about Type I errors should choose the Monte Carlo confidence interval or the distribution-of-the-product approach, which rarely disagree. The percentile bootstrap confidence interval is a good compromise test.

Keywords

mediation analysis, indirect effects, bootstrapping, Sobel test, statistical analyses, hypothesis testing

Received 12/18/12; Revision accepted 2/1/13

Although one can make a name for oneself as a psychological scientist by establishing whether some variable *X* causally affects variable *Y*, the better scientists go further by identifying not only whether *X* affects *Y* but also the mechanism by which that causal effect operates. Such mechanisms can be established in a number of ways, but strong theoretical argument combined with good measurement and research design is king.

Secondary to theory and design are a variety of statistical approaches to buttressing a causal story. One approach is *moderation analysis*, which establishes whether an effect exists under conditions in which the mechanism is allowed to operate but not under conditions in which it is disrupted (see, e.g., Spencer, Zanna, & Fong, 2005). A second approach, the topic of this article, is *mediation analysis* (Baron & Kenny, 1986; Hayes, 2013; MacKinnon, 2008). In this type of analysis, evidence supporting a mechanism is found in the existence of an *indirect effect* of *X* on *Y* through the proposed mediator variable *M*. For example, Minson and Mueller (2012) tested whether collaborative versus individual judgment (*X*) indirectly influenced susceptibility to outside influences (*Y*) through the mechanism of decision confidence (*M*). The causal chain of effects from *X* to *M* to *Y*—the indirect effect—represents the mechanism through which *X*'s effect on *Y* operates.

The principles of statistical mediation analysis are illustrated by the mediation model presented in Figure 1. In a typical mediation analysis, following the measurement (or manipulation) of *X* along with measurement of *M* and *Y*, the paths labeled *a, b, c*, and *c′* are estimated

Corresponding Author:
Andrew F. Hayes, The Ohio State University, School of Communication and Department of Psychology, 3016 Derby Hall, 154 N. Oval Mall, Columbus, OH 43210
E-mail: hayes.338@osu.edu

using a set of ordinary-least-squares regression analyses or simultaneously using a structural equation modeling program. When *X, M*, and *Y* are observed variables, *M* and *Y* are modeled as continuous variables, and effects are modeled as linear, then *a, b, c*, and *c′* can be estimated with the models represented in the following equations:

$$M = i_1 + \hat{a}X + e_1 \quad (1)$$

$$Y = i_2 + \hat{c}'X + \hat{b}M + e_2, \quad (2)$$

where i_1 and i_2 are regression intercepts, and e_1 and e_2 are residuals.

The indirect effect of *X* on *Y* through *M* is quantified as $\hat{a}\hat{b}$. The estimate of the *direct effect* of *X* on *Y* is path $\hat{c}'$ and represents the influence of *X* on *Y* independent of the mechanism through *M*. The estimated direct and indirect effects sum to yield $\hat{c}$, the *total effect* of *X* on *Y* and an estimator of *c* in Figure 1. That is, $\hat{c} = \hat{c}' + \hat{a}\hat{b}$. Of course, to interpret these statistical indices of effect in causal terms, the conditions of causality must also be met through proper research design, measurement, or convincing theoretical argument (see, e.g., Judd & Kenny, 2010; Mathieu & Taylor, 2006).

Because the indirect effect is so important in a mediation process, there is a large literature on approaches to testing hypotheses about indirect effects (e.g., MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; MacKinnon, Lockwood, & Williams, 2004; Preacher & Hayes, 2004, 2008; Shrout & Bolger, 2002). These articles have had a big influence on practice and are among the more highly cited methodology papers published this century, no doubt because of the popularity of statistical mediation analysis in general.

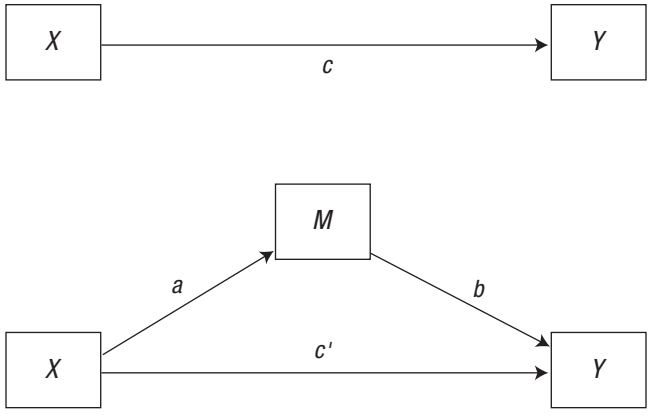

Fig. 1. Schematic illustrating a simple mediation model, in which the independent variable *X* influences the dependent variable *Y* directly (*c′*) and indirectly (*ab*) through a mediator *M*. The direct and indirect effects add to yield the total effect (*c*) of *X* on *Y*.

Indeed, statistical mediation analysis is alive and well in the pages of *Psychological Science*. We examined all 454 empirical articles published in the 2011 and 2012 volumes and identified 71 (15.6%) that included at least one mediation analysis. The number of articles in a single issue with such an analysis ranged between 1 and 6, with an average of 2.9 articles per issue. One cannot read an issue of *Psychological Science* without encountering at least one statistical mediation analysis.

But not all mediation analyses were conducted in the same manner, and it is these inconsistencies that are the topic of this article. To be sure, in every one of these 71 articles, a claim of mediation was based on the kind of path analysis described previously, and most included tests of significance of $\hat{a}$ and $\hat{b}$ as part of the analysis (known as the test of joint significance if the investigator insists on both nulls being rejected in order to claim mediation). Furthermore, in the majority of studies (59, or 83%), a claim of mediation was supported through the use of an inferential test of the indirect effect quantified as $\hat{a}\hat{b}$. However, that is where the consistencies ended, as investigators diverged in which test they used when they conducted such a test. In 30 (51%) of these 59 studies, investigators employed the Sobel test (e.g., Huang, Sedlovskaya, Ackerman, & Bargh, 2011; Neville, 2012), whereas a bootstrap confidence interval (CI) was used in 32 articles (54%; e.g., Johnson & Fujita, 2012; Na & Kitayama, 2011). In one case, inference was based on a Monte Carlo CI (Minson & Mueller, 2012). In a few instances (4, or 7%), multiple methods were used, such as the Sobel test as well as a bootstrap CI (e.g., Guendelman, Cheryan, & Monin, 2011; Oishi, Schimmack, & Diener, 2012). But some investigators (in 10 of the original 71 articles, or 14%) never quantified the indirect effect and relied exclusively on the statistical significance of estimates of *a* and *b*—the test of joint significance (e.g., Rindermann & Thompson, 2011; Sweeny & Vohs, 2012). A few articles were difficult to characterize because the description of the method used was ambiguous or absent.

In this article, we address how frequently different methods can yield different decisions. If they rarely can, there is little reason to be concerned about inconsistencies in approaches used by investigators. But if they can produce different answers with some nontrivial frequency, this raises the question as to which method is more trustworthy when the results of different tests disagree and, therefore, which test should be preferred over others. Thus, one of our goals is to give researchers some actionable advice they can apply across situations as to what test to use without requiring information not

usually available or that will be highly dependent on the study, such as whether certain effects in the causal chain are large or small in the population. We do so using the results of an extensive set of Monte Carlo simulations.

We are not the first to examine the performance of the methods we describe here (see, e.g., Biesanz, Falk, & Savalei, 2010; Fritz, Taylor, & MacKinnon, 2012; MacKinnon et al., 2002; MacKinnon et al., 2004; Preacher & Selig, 2012; Williams & MacKinnon, 2008). However, to our knowledge, no one has attempted to quantify how frequently different tests produce different answers or whether there is one preferred test when such disagreements arise—what we call a test's *relative trustworthiness*. Instead, researchers have focused on how tests perform on the aggregate relative to each other, ignoring the tests' tendency to agree or disagree in the decisions they render when applied to the same data.

An examination of disagreement in the outcomes of inferential tests in statistical mediation analysis can yield important information and actionable advice for the research community. First, anyone who follows the methods literature knows that there are many proposed solutions to the same inferential problem. Differences among the statistical properties (e.g., power, Type I error rate) of various tests are often marginal or dependent on information not available to the researcher, which makes that literature confusing to those who need concrete advice they can act on. So researchers may just shrug their shoulders and simply use whatever test is conveniently available to them.

Second, it is both tempting and easy to use more than one method, as software offers more and more tests or as methodologists write tools in the form of macros or procedures for popular programs such as SPSS, SAS, or R to implement advice they offer. For instance, INDIRECT for SPSS and SAS (Preacher & Hayes, 2008), Mplus (Muthén & Muthén, 2011), and MBESS for R (Kelley, 2007) all offer the Sobel test as well as percentile or bias-corrected bootstrap CIs (or both) for indirect effects; PROCESS for SPSS and SAS (Hayes, 2013) provides all methods that INDIRECT calculates as well as Monte Carlo CIs; SOBEL for SPSS and SAS (Preacher & Hayes, 2004) generates percentile bootstrap CIs and the Sobel test; PRODCLIN (MacKinnon, Fritz, Williams, & Lockwood, 2007; Tofighi & MacKinnon, 2011) generates the Sobel test as well as the distribution-of-the-product approach; and RMediate for R (Tofighi & MacKinnon, 2011) offers the Sobel test, distribution-of-the-product approach, and Monte Carlo CIs. Further, all statistical packages and most tools noted here provide p values for $\hat{a}$ and $\hat{b}$ used in the implementation of the test of joint significance. It is worth knowing how frequently tests can differ in the decisions they generate, and researchers need specific advice regarding which test to place more trust in when this occurs.

Third, it may be the case that the outcomes of different tests rarely differ or that some tests yield the same decision so frequently that they can be considered interchangeable. If so, this information can be useful not only to substantive researchers but also to editors and reviewers who can focus their critique on matters other than the selection of statistical test. But if differences are more common than is appreciated, this means that investigators should offer justification to readers, reviewers, and editors for why the test they use is preferred in the circumstance in which they have used it.

Method

Our procedure largely mirrored the method used in a related study of power and sample-size selection in mediation analysis published in *Psychological Science* (Fritz & MacKinnon, 2007). We generated multiple samples from a model defined by Figure 1. Population intercepts were set to zero, values of X were sampled from a standard normal distribution, and M and Y were generated from X and M using the population paths (a for the model of M, b and c' for the model of Y). A standard normal error was added to generate sampling discrepancy between parameters and estimates thereof. To simulate a wide range of indirect and direct effects, we used seven values for population a and b (−0.59, −0.39, −0.14, 0.00, 0.14, 0.39, 0.59) and five values for the total effect c (−0.70, 0.35, 0.00, 0.35, 0.70) in order to generate values of the direct effect c' by the formula $c' = c - ab$. Seven different samples sizes were also used (ns = 25, 50, 75, 100, 200, 500, 1,000). When crossed, this resulted in 1,715 different multivariate populations, with 455 (26.5%) defined by an absence of an indirect effect of X on Y (i.e., $ab = 0$) and the remaining with a nonzero indirect effect (i.e., $ab \neq 0$). For each combination of conditions, we generated 1,000 samples, which yielded a total of 1,715,000 data sets. In each data set, $\hat{a}$, $\hat{b}$, and their standard errors, $\hat{s}_{\hat{a}}$ and $\hat{s}_{\hat{b}}$, respectively, were calculated using two ordinary-least-squares regression analyses conforming to Equations 1 and 2. Standard errors were constructed based on the ordinary-least-squares standard estimator produced by default in most regression procedures in popular software used by psychological scientists. The simulation was programmed using GAUSS Version 12 (Aptech Systems, 2012) with additional computation and analysis using R Version 2.14 (R Development Core Team, 2012).

These methods included in the simulation are not exhaustive of all tests available. We restricted this study to the ones described here because they are frequently used

in the social-science literature, have appeared in the pages of *Psychological Science*, and are implemented either in software that is popular among substantive researchers (SPSS, SAS, or R) or can be calculated easily using tools available online.

Delta methods (Δ_1 and Δ_2)

The delta method is also known as the Sobel test. It has two variants (e.g., Aroian, 1947; Sobel, 1982), but researchers rarely specify which version they use (in our content analysis, we found that this information was never provided). The delta method requires the estimation of the standard error for use in the construction of a CI for *ab* or for null-hypothesis testing. For Δ_1 (first-order delta), the standard error estimator used was

$$\hat{s}_{\hat{a}\hat{b}} = \sqrt{\hat{a}^2\hat{s}_{\hat{b}}^2 + \hat{b}^2\hat{s}_{\hat{a}}^2}, \quad (3)$$

whereas for Δ_2, (second-order delta) the estimator was

$$\hat{s}_{\hat{a}\hat{b}} = \sqrt{\hat{a}^2\hat{s}_{\hat{b}}^2 + \hat{b}^2\hat{s}_{\hat{a}}^2 + \hat{s}_{\hat{a}}^2\hat{s}_{\hat{b}}^2}. \quad (4)$$

A 95% CI for *ab* was generated as $\hat{a}\hat{b} \pm 1.96\hat{s}_{\hat{a}\hat{b}}$. A 95% CI using Δ_1 or Δ_2 produces exactly the same decision as a null-hypothesis test that $ab = 0$ at the .05 level of significance, using a *p* value for the ratio of $\hat{a}\hat{b}$ to $\hat{s}_{\hat{a}\hat{b}}$ derived from the standard normal distribution.

Percentile and bias-corrected bootstrap CIs

There are several methods used to construct bootstrap CIs, but only two are popularly used: *percentile* and *bias-corrected* (see, e.g., MacKinnon et al., 2004; Preacher & Hayes, 2008). For each data set, a bootstrap sample of *n* cases was generated by drawing from the sample with replacement, and $\hat{a}$ and $\hat{b}$ were calculated in each bootstrap sample. This process was repeated a total of 5,000 times for each data set, yielding 5,000 bootstrap estimates of *ab*. A percentile-based 95% CI for *ab* was constructed by finding the two bootstrap estimates of *ab* in the sample of 5,000 defining the 2.5th and 97.5th percentiles of the distribution. Using this same distribution, a bias-corrected CI for *ab* was calculated using the bias adjustment described by Hayes (2013), MacKinnon (2008), and Preacher and Selig (2012).

Distribution of the product

The distribution-of-the-product approach (see MacKinnon et al., 2007) derives estimates of the end points of a 95% CI for *ab* in reference to an analytical derivation of the distribution of the product of random, independent normal variables. These computations were completed using the RMediation package in R (Tofighi & MacKinnon, 2011) with $\hat{a}$, $\hat{b}$, $\hat{s}_{\hat{a}}$, and $\hat{s}_{\hat{b}}$ as input arguments and using the "type=prodclin" option. Although no researchers reported using the distribution-of-the-product method in an article published in *Psychological Science* during 2011 and 2012, this approach has appeared in research published in other well-regarded journals in psychology (e.g., Caprara, Alessandri, & Eisenberg, 2012; Wickens, Wiesenthal, Flora, & Flett, 2011), so we included it here.

Monte Carlo CI

Following procedures described in MacKinnon et al. (2004) and Preacher and Selig (2012), we constructed a Monte Carlo CI for *ab* by sampling 5,000 pairs of independent, random normal deviates V_a and V_b from populations with means $\hat{a}$ and $\hat{b}$ and standard deviations $\hat{s}_{\hat{a}}$ and $\hat{s}_{\hat{b}}$, respectively. V_a and V_b in each pair were multiplied together to produce a distribution of 5,000 values of V_aV_b. The end points of a 95% CI for *ab* were calculated as the values of V_aV_b defining the 2.5th and 97.5th values in the distribution.

Joint significance

The test of joint significance was conducted by testing the null hypotheses that $a = 0$ and $b = 0$ based on $\hat{a}$, $\hat{b}$, $\hat{s}_{\hat{a}}$, and $\hat{s}_{\hat{b}}$. Each null was tested by deriving the *p* value for the ratio of point estimate to standard error in reference to the $t(df)$ distribution, where *df* is the residual degrees of freedom ($n - 2$ for Equation 1, $n - 3$ for Equation 2).

Results

In the following description of the results, we collapse across all nonzero values of *ab* because the size of the population indirect effect or its constituent components (*a* or *b*) is not available to investigators, and thus, any advice we could offer that requires such knowledge is not actionable. Although the zero or nonzero status of the population indirect effect is also unknown, we describe the results separately for these two sets of conditions because the performance of a test in the presence versus the absence of a population effect is so fundamental to how scientists evaluate the quality of a statistical test. Tables that further break these results down by the size of the indirect effect and sample size are available on request, and we will provide the entire data file containing the outcomes of all tests in each of the 1,715,000 data sets to anyone interested.

Our first set of descriptions of the results focuses on whether a test made the correct inference, with correct

defined as an inference consistent with the population indirect effect (*ab*) as zero or not. With the exception of the test of joint significance, when $ab = 0$, the inference was considered correct if the CI contained zero and considered incorrect otherwise. When $ab \neq 0$, the inference was considered correct if the CI did not contain zero and was in the direction consistent with *ab*. Otherwise, it was incorrect. For the test of joint significance, the inference was considered correct if both null hypotheses that $a = 0$ and $b = 0$ were rejected when $ab \neq 0$, or if either null hypothesis was not rejected when $ab = 0$. Otherwise, the inference was incorrect.

The diagonals of Tables 1 and 2 tabulate the percentage of times, as a function of sample size (which we have trichotomized for the sake of presentation), that a test made the correct decision when $ab = 0$ (Table 1) and when $ab \neq 0$ (Table 2). These results replicate those of simulation studies showing that the Sobel test is quite conservative, and methods that account for nonnormality of the sampling distribution of $\hat{a}\hat{b}$ tend to be more powerful. In smaller samples, the bias-corrected bootstrap CI tends to be more powerful than other methods, and the difference in power diminishes with increasing sample size.

Table 1. Decision Accuracy of Tests (on the Diagonal), Pairwise Disagreement Between Tests (Below the Diagonal), and Relative Trustworthiness of Tests (Above the Diagonal) for Different Sample Sizes, When $ab = 0$

Sample size and test type	1	2	3	4	5	6	7
All sample sizes (455,000 data sets)							
1. First-order delta (Δ_1)	**98.1**	0.0	3.6	1.1	0.6	0.0	0.0
2. Second-order delta (Δ_2)	0.2	**97.9**	4.8	1.4	0.7	0.0	0.0
3. Percentile bootstrap CI	1.9	1.8	**96.3**	0.9	60.4	60.8	49.9
4. Bias-corrected bootstrap CI	3.5	3.3	1.7	**94.6**	92.4	92.8	90.0
5. Monte Carlo	1.5	1.4	1.2	2.2	**96.5**	51.0	22.5
6. Distribution of the product	1.5	1.4	1.2	2.2	0.3	**96.5**	0.1
7. Joint significance	1.8	1.6	1.3	2.0	0.4	0.3	**96.3**
$n < 100$ (195,000 data sets)							
1. First-order delta (Δ_1)	**99.1**	0.0	2.1	0.4	0.0	0.0	0.0
2. Second-order delta (Δ_2)	0.2	**98.9**	3.1	0.6	0.0	0.0	0.0
3. Percentile bootstrap CI	2.3	2.1	**96.9**	0.5	62.6	63.0	50.4
4. Bias-corrected bootstrap CI	4.2	4.1	2.1	**94.9**	93.6	93.7	89.9
5. Monte Carlo	1.8	1.6	1.5	2.8	**97.3**	53.1	13.8
6. Distribution of the product	1.8	1.6	1.5	2.8	0.2	**97.3**	0.0
7. Joint significance	2.1	2.0	1.6	2.6	0.5	0.4	**96.9**
$100 \leq n \leq 200$ (130,000 data sets)							
1. First-order delta (Δ_1)	**98.0**	0.0	3.8	1.1	0.0	0.0	0.0
2. Second-order delta (Δ_2)	0.2	**97.8**	5.0	1.4	0.1	0.0	0.0
3. Percentile bootstrap CI	1.9	1.8	**96.2**	0.3	59.5	60.4	50.7
4. Bias-corrected bootstrap CI	3.5	3.4	1.7	**94.5**	92.8	93.0	91.1
5. Monte Carlo	1.6	1.4	1.2	2.2	**96.4**	53.9	26.5
6. Distribution of the product	1.6	1.4	1.2	2.2	0.3	**96.4**	0.0
7. Joint significance	1.8	1.6	1.3	2.1	0.4	0.2	**96.2**
$n \geq 500$ (130,000 data sets)							
1. First-order delta (Δ_1)	**96.6**	0.0	7.1	3.1	2.7	0.0	0.0
2. Second-order delta (Δ_2)	0.1	**96.5**	8.8	3.7	3.5	0.0	0.0
3. Percentile bootstrap CI	1.3	1.2	**95.5**	3.0	55.2	53.9	46.9
4. Bias-corrected bootstrap CI	2.3	2.2	1.1	**94.5**	88.3	89.5	88.3
5. Monte Carlo	1.1	1.0	0.7	1.4	**95.6**	46.4	34.6
6. Distribution of the product	1.1	1.0	0.7	1.3	0.3	**95.5**	0.8
7. Joint significance	1.2	1.0	0.7	1.2	0.4	0.1	**95.5**

Note: Test accuracy is the percentage of times that a test produced the correct inference that *ab* was equal to zero (i.e., there was no indirect effect). Pairwise disagreement is the percentage of times the two tests produced different decisions. Relative trustworthiness is the percentage of times the test in the column produced the correct decision when its results disagreed with those of the test in the row. CI = confidence interval.

Table 2. Decision Accuracy of Tests (on the Diagonal), Pairwise Disagreement Between Tests (Below the Diagonal), and Relative Trustworthiness of Tests (Above the Diagonal) for Different Sample Sizes, When $ab \neq 0$

Sample size and test type	1	2	3	4	5	6	7
All sample sizes (1,260,000 data sets)							
1. First-order delta (Δ_1)	**55.0**	100.0	95.5	98.8	99.9	100.0	100.0
2. Second-order delta (Δ_2)	0.8	**55.9**	93.5	98.3	99.8	100.0	100.0
3. Percentile bootstrap CI	6.1	5.4	**60.6**	99.3	50.8	51.2	60.5
4. Bias-corrected bootstrap CI	9.1	8.3	3.4	**63.9**	11.9	11.7	16.7
5. Monte Carlo	5.6	4.8	3.1	4.4	**60.6**	51.9	81.6
6. Distribution of the product	5.6	4.8	3.1	4.3	0.5	**60.6**	100.0
7. Joint significance	6.2	5.4	3.4	4.0	1.0	0.6	**61.3**
$n < 100$ (540,000 data sets)							
1. First-order delta (Δ_1)	**27.2**	100.0	95.9	99.2	100.0	100.0	100.0
2. Second-order delta (Δ_2)	1.3	**28.4**	93.8	98.7	100.0	100.0	100.0
3. Percentile bootstrap CI	9.3	8.3	**35.7**	99.5	52.3	52.4	62.7
4. Bias-corrected bootstrap CI	14.2	13.1	5.5	**41.2**	11.9	11.8	18.2
5. Monte Carlo	8.8	7.5	5.1	6.8	**36.0**	50.5	86.4
6. Distribution of the product	8.8	7.6	5.0	6.8	0.7	**36.0**	100.0
7. Joint significance	10.0	8.7	5.5	6.3	1.6	1.2	**37.1**
$100 \leq n \leq 200$ (360,000 data sets)							
1. First-order delta (Δ_1)	**57.8**	100.0	95.1	98.5	> 99.9	100.0	100.0
2. Second-order delta (Δ_2)	0.8	**58.5**	93.0	98.0	99.8	100.0	100.0
3. Percentile bootstrap CI	5.5	4.9	**62.7**	99.4	46.6	46.9	54.0
4. Bias-corrected bootstrap CI	8.3	7.6	3.1	**65.8**	10.9	10.7	12.8
5. Monte Carlo	4.8	4.0	2.8	4.1	**62.6**	51.4	72.0
6. Distribution of the product	4.8	4.0	2.7	4.1	0.6	**62.6**	100.0
7. Joint significance	5.2	4.4	2.9	3.8	0.9	0.4	**63.0**
$n \geq 500$ (360,000 data sets)							
1. First-order delta (Δ_1)	**94.0**	100.0	93.6	96.8	98.8	100.0	100.0
2. Second-order delta (Δ_2)	0.3	**94.3**	91.8	96.0	98.2	100.0	100.0
3. Percentile bootstrap CI	1.7	1.5	**95.6**	96.7	50.9	55.1	61.5
4. Bias-corrected bootstrap CI	2.3	2.0	0.7	**96.2**	15.7	16.1	18.0
5. Monte Carlo	1.6	1.3	0.6	0.9	**95.6**	60.0	72.0
6. Distribution of the product	1.6	1.3	0.5	0.8	0.2	**95.6**	100.0
7. Joint significance	1.7	1.4	0.6	0.7	0.3	0.1	**95.7**

Note: Test accuracy is the percentage of times that a test produced the correct inference that *ab* was different from zero (i.e., there was an indirect effect). Pairwise disagreement is the percentage of times the two tests produced different decisions. Relative trustworthiness is the percentage of times the test in the column produced the correct decision when its results disagreed with those of the test in the row. CI = confidence interval.

More pertinent to our purpose is how frequently the tests disagreed and ascertaining the relative trustworthiness of the tests, meaning which test was more likely to have produced the correct decision when there was a disagreement in decision. The cells below the diagonal in Tables 1 and 2 provide the answer to the former question.[1] When there was no indirect effect (Table 1), the tests disagreed between 0.1% and 4.2% of the time, depending on sample size and which two tests were being compared. The two versions of the delta method rarely disagreed (never more than 0.2% of the time), and the Monte Carlo method, distribution-of-the–product approach, and test of joint significance also were largely indistinguishable in their decisions, as they too rarely disagreed with each other—0.5% of the time or less. Disagreements tended to occur more often between one of the bootstrap methods and any of the other tests, with disagreements varying between 0.7% and 4.2%. The takeaway from Table 1 is that when there is no indirect effect, you can expect the tests to agree most of the time. Rather rarely will certain pairs of tests produce different decisions.

When there is an indirect effect (Table 2), disagreements are more likely. The two versions of the delta method rarely disagreed with each other (between 0.3% and 1.3% of the time), nor did the Monte Carlo and

distribution-of-the-product methods (between 0.2% and 0.7% of the time). The delta methods disagreed most frequently with other tests, sometimes as much as 14% of the time, with moderate disagreements occurring between bootstrap methods and other tests (between 0.5% and 6.8%, excluding the delta methods). The most striking message from Table 2 is that when there is an indirect effect, the choice of test can matter. Although the probability that any two tests will disagree when a single investigator conducts a set of tests is relatively small, when considering the popularity of mediation analysis, it is safe to say that psychological scientists often find themselves in a situation in which two tests of an indirect effect in statistical mediation analysis disagree.

Is there a more trustworthy test when such disagreements arise? The cells above the diagonal in Tables 1 and 2 answer this question. Looking first at when there is no indirect effect (Table 1), it is clear that either version of the delta method would be preferred if power was of no concern, as it is most conservative. It is more likely than any other test examined here to produce the correct decision when there is no indirect effect and two tests disagree. But this conservatism comes at the price of the largest Type II error rate among these tests. When the less-conservative tests are pitted against each other, the evidence slightly favors the distribution-of-the-product approach or a Monte Carlo CI. The bias-corrected bootstrap CI was least likely of these less-conservative tests to be correct when there was a disagreement, with the chances of the percentile bootstrap CI slightly improved but still somewhat inferior. The Monte Carlo CI and distribution-of-the-product methods were about equally likely to be the one correct test relative to each other.

When there is an indirect effect (Table 2) and two tests disagree, either delta method almost never gets it right relative to any other test. Rather, the bias-corrected bootstrap CI is more likely to lead to the correct decision than any other method, regardless of sample size. Among the Monte Carlo CI, distribution-of-the-product approach, and test of joint significance, the test of joint significance tends to be more trustworthy. However, it is not as trustworthy as the bias-corrected bootstrap CI.

Although these results favor the bias-corrected bootstrap CI, two findings are noteworthy. Table 3 provides CI coverage when $ab \neq 0$, meaning that a CI contained the nonzero population indirect effect. (Table 1 can also be interpreted as a coverage table, in that if a test leads to the correct inference that $ab = 0$, this means the CI for that method contains zero.) As can be seen in Table 3, although disagreements are generally rarer than as evidenced in Table 2, in smaller samples, the bias-corrected CI does not fare as well by this standard. Rather, the Monte Carlo CI and distribution-of-the-product approaches are the more trustworthy of the four methods that do not assume normality of the sampling distribution of the indirect effect.

Additionally, recent simulation results led Fritz et al. (2012) to recommend that the bias-corrected bootstrap CI not be used because it purchases the power advantage observed in Table 2 at the expense of slight Type I error inflation when either a or b but not both is zero. Our simulations show some evidence of this as well (see Tables S1–S4 in the Supplemental Material available online). Although the tests still usually agree, the distribution-of-the-product approach, the Monte Carlo CI, and the percentile bootstrap CI were indeed more trustworthy than the bias-corrected bootstrap CI. But given that an investigator can never know whether one or another effect is zero or the size of the nonzero effect, it is hard to use these findings as a guide to decision making with one's own data.

Implications for Practice

Our results suggest that on the balance, two tests are much more likely to agree than disagree, regardless of sample size and whether or not an indirect effect exists. In the vast majority of conditions we simulated, two tests usually produced the same decision about the population indirect effect. Most of the time, especially in samples of 500 or more, it will not make any difference what you do.

But sometimes it does matter. From these results, we offer the following advice to researchers about choosing an inferential method. First, avoid the Sobel test (i.e., either delta version). It is the least powerful and least trustworthy of all methods when there is an indirect effect in the population. It is striking that this test is still popular in *Psychological Science* even though methodologists have warned about its low power for years. Perhaps this reflects a belief that a significant effect using a conservative test is hard to dispute. If so, we worry about all the indirect effects missed and mediation analyses not reported as a result of this mind-set.

Second, if power is at the forefront of concerns, a bias-corrected bootstrap CI is the best test, as it is most trustworthy in the conditions we simulated when an indirect effect exists and the focus is on detecting a nonzero effect rather than on interval estimation. But this superior power comes at the price of an elevation of false positives in some circumstances and worse coverage in smaller samples. The Monte Carlo CI or the distribution-of-the-product approaches rarely disagree, and one is no more trustworthy than the other. They do offer good Type I error protection, but they are less powerful and less trustworthy when an indirect effect exists than the bias-corrected bootstrap CI. We therefore recommend these methods to researchers worried about the liberal

Table 3. Confidence Interval Coverage (on the Diagonal), Pairwise Disagreement Between Tests (Below the Diagonal), and Relative Trustworthiness of Tests (Above the Diagonal) for Different Sample Sizes, When $ab \neq 0$

Sample size and test type	1	2	3	4	5	6
All sample sizes (1,260,000 data sets)						
1. First-order delta (Δ_1)	**95.4**	0.0	35.4	33.0	44.8	45.5
2. Second-order delta (Δ_2)	1.1	**94.3**	57.0	47.8	68.7	69.3
3. Percentile bootstrap CI	2.3	2.9	**94.7**	34.2	64.2	65.1
4. Bias-corrected bootstrap CI	3.7	4.1	1.9	**94.1**	71.9	72.5
5. Monte Carlo	1.7	2.4	1.8	2.5	**95.2**	53.5
6. Distribution of the product	1.7	2.4	1.7	2.5	0.3	**95.2**
$n < 100$ (540,000 data sets)						
1. First-order delta (Δ_1)	**95.8**	0.0	29.7	26.1	38.9	39.5
2. Second-order delta (Δ_2)	1.8	**94.0**	56.9	44.3	69.8	70.4
3. Percentile bootstrap CI	3.1	4.0	**94.6**	28.2	65.7	66.4
4. Bias-corrected bootstrap CI	5.2	5.9	2.8	**93.3**	77.9	78.4
5. Monte Carlo	2.2	3.4	2.5	3.6	**95.3**	54.7
6. Distribution of the product	2.1	3.4	2.5	3.6	0.3	**95.4**
$100 \leq n \leq 200$ (360,000 data sets)						
1. First-order delta (Δ_1)	**95.1**	0.0	38.7	38.1	49.7	50.3
2. Second-order delta (Δ_2)	0.9	**94.2**	57.9	51.6	69.9	70.4
3. Percentile bootstrap CI	2.2	2.7	**94.6**	39.8	65.0	65.9
4. Bias-corrected bootstrap CI	3.4	3.6	1.5	**94.3**	68.3	68.8
5. Monte Carlo	1.6	2.3	1.6	2.2	**95.1**	52.9
6. Distribution of the product	1.6	2.3	1.6	2.2	0.3	**95.1**
$n \geq 500$ (360,000 data sets)						
1. First-order delta (Δ_1)	**94.9**	0.0	51.0	53.7	56.0	57.0
2. Second-order delta (Δ_2)	0.1	**94.8**	55.6	56.8	61.1	62.1
3. Percentile bootstrap CI	1.2	1.3	**94.9**	57.2	55.6	57.2
4. Bias-corrected bootstrap CI	1.8	1.9	0.8	**95.1**	49.4	50.2
5. Monte Carlo	1.0	1.1	0.9	1.1	**95.0**	52.3
6. Distribution of the product	1.0	1.1	0.8	1.1	0.3	**95.1**

Note: Confidence interval (CI) coverage is the percentage of times that a CI contained the nonzero population indirect effect (i.e., there was an indirect effect). Pairwise disagreement is the percentage of times one method generated a CI that covered *ab* when the other method did not. Relative trustworthiness is the percentage of times the confidence interval for the method in the column covered *ab* when its results disagreed with those of the method in the row. The test of joint significance does not produce a CI estimate and therefore is not shown here.

nature of the bias-corrected bootstrap CI in some circumstances. The test of joint significance also performed well and was even more trustworthy than the other tests we recommend, but it does not provide a CI estimate, and it does not generalize to modern analytical approaches that integrate mediation analysis with a moderation component (e.g., Edwards & Lambert, 2007; Hayes, 2013; Preacher, Rucker, & Hayes, 2007). This makes it hard to recommend.

We conclude with a caveat. With the exception of either bootstrap method, all methods we simulated require estimates of the standard errors of $\hat{a}$ and $\hat{b}$. In principle, anything that could influence the accuracy of those standard errors, such as heteroscedasticity (see e.g., Hayes & Cai, 2007; Long & Ervin, 2000), could influence the performance of these methods. None of the computations that generate a bootstrap CI require a standard error of any of the paths. Thus, investigators who prefer to make fewer assumptions could consider the percentile bootstrap CI a good compromise test, as it is more powerful than either Sobel test, it shows less Type I error inflation in smaller samples when one path is zero, and it offers better coverage than the bias-corrected bootstrap CI. However, it is less trustworthy than competing tests (at least when the homoscedasticity assumption is met, as in our simulations). This advice is also consistent with that of Fritz et al. (2012), who recommended the percentile over the bias-corrected bootstrap CI.

Author Contributions

A. F. Hayes developed the study concept, and both authors contributed to decisions about operationalization. M. Scharkow programmed the simulation and tabulated the results, with some editing of code by A. F. Hayes. A. F. Hayes wrote the majority of the first draft of the manuscript, but both authors contributed equally to editing and final decision making about emphasis and presentation of findings as well as responding to reviewers and framing the revision.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

Note

1. These cells in Tables 1 and 2 reflect *observed* disagreement. Two tests will sometimes disagree just by chance, with expected chance disagreement depending on the relative validity and power of the tests.

References

Aptech Systems. (2012). GAUSS (Version 12) [Computer software]. Black Diamond, WA: Aptech Systems. Available from http://www.statmodel.com

Aroian, L. A. (1947). The probability function of the product of two normally distributed variables. *Annals of Mathematical Statistics*, *18*, 265–271.

Baron, R., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychology: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

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

Caprara, G. V., Alessandri, G., & Eisenberg, N. (2012). Prosociality: The contribution of traits, values, and self-efficacy beliefs. *Journal of Personality and Social Psychology*, *102*, 1289–1303.

Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods*, *12*, 1–22.

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, *18*, 233–239.

Fritz, M. S., Taylor, A. B., & MacKinnon, D. P. (2012). Explanation of two anomalous results in statistical mediation analysis. *Multivariate Behavioral Research*, *47*, 61–87.

Guendelman, M. D., Cheryan, S., & Monin, B. (2011). Fitting in but getting fat: Identity threat and dietary choices among U.S. immigrant groups. *Psychological Science*, *22*, 959–967.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press.

Hayes, A. F., & Cai, L. (2007). Using heteroscedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, *39*, 709–722.

Huang, J. Y., Sedlovskaya, A., Ackerman, J. M., & Bargh, J. A. (2011). Immunizing against prejudice: Effects of disease protection on attitudes toward out-groups. *Psychological Science*, *22*, 1550–1556.

Johnson, I. R., & Fujita, K. (2012). Change we can believe in: Using perceptions of changeability to promote system-change motives over system-justification motives in information search. *Psychological Science*, *22*, 133–140.

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* (Vol. 1, pp. 115–139). New York, NY: Wiley.

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods*, *39*, 979–984.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*, 217–224.

MacKinnon, D. P. (2008). *An introduction to statistical mediation analysis*. Mahwah, NJ: Erlbaum.

MacKinnon, D. P., Fritz, M. S., Williams, J., & Lockwood, C. M. (2007). Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behavior Research Methods*, *39*, 384–389.

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*, 99–128.

Mathieu, J. E., & Taylor, S. R. (2006). Clarifying conditions and decision points for mediational type inferences in organizational behavior. *Journal of Organizational Behavior*, *27*, 1031–1056.

Minson, J. A., & Mueller, J. S. (2012). The cost of collaboration: Why joint decision making exacerbates rejection of outside information. *Psychological Science*, *23*, 219–224.

Muthén, L. K., & Muthén, B. O. (2011). *Mplus user's guide* (6th ed.). Los Angeles, CA: Author.

Na, J., & Kitayama, S. (2011). Spontaneous trait inference is culture-specific: Behavioral and neural evidence. *Psychological Science*, *22*, 1025–1032.

Neville, L. (2012). Do economic equality and generalized trust inhibit academic dishonesty? Evidence from state-level search-engine queries. *Psychological Science*, *23*, 339–345.

Oishi, S., Schimmack, U., & Diener, E. (2012). Progressive taxation and the subjective well-being of nations. *Psychological Science*, *23*, 86–92.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models.

Behavior Research Methods, Instruments, & Computers, *36*, 717–731.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*, 879–891.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Assessing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, *42*, 185–227.

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

R Development Core Team. (2012). R: A language and environment for statistical computing (Version 2.14). Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org

Rindermann, H., & Thompson, J. (2011). Cognitive capitalism: The effect of cognitive ability on wealth, as mediated through scientific achievement and economic freedom. *Psychological Science*, *22*, 754–763.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, *7*, 422–445.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhart (Ed.), *Sociological methodology* (pp. 290–312). San Francisco, CA: Jossey-Bass.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analysis in examining psychological processes. *Journal of Personality and Social Psychology*, *89*, 845–851.

Sweeny, K., & Vohs, K. D. (2012). On near misses and completed tasks: The nature of relief. *Psychological Science*, *23*, 464–468.

Tofighi, D., & MacKinnon, D. P. (2011). RMediation: An R package for mediation analysis confidence intervals. *Behavior Research Methods*, *43*, 692–700.

Wickens, C. M., Wiesenthal, D. L., Flora, D. B., & Flett, G. L. (2011). Understanding driving anger and aggression: Attributional theory in the driving environment. *Journal of Experimental Psychology: Applied*, *17*, 354–370.

Williams, J., & MacKinnon, D. P. (2008). Resampling and distribution of the product methods for testing indirect effects in complex models. *Structural Equation Modeling*, *15*, 23–51.