Short Methodological Report

Organizational Research Methods
2022, Vol. 25(3) 591–617

DOI: 10.1177/1094428121999096
journals.sagepub.com/home/orm


A Robust Bootstrap Test for Mediation Analysis

Andreas Alfons[1], Nüfer Yasin Ateş[2], and Patrick J. F. Groenen[1]

Abstract

Mediation analysis is central to theory building and testing in organizational sciences. Scholars often use linear regression analysis based on normal-theory maximum likelihood estimators to test mediation. However, these estimators are very sensitive to deviations from normality assumptions, such as outliers, heavy tails, or skewness of the observed distribution. This sensitivity seriously threatens the empirical testing of theory about mediation mechanisms. To overcome this threat, we develop a robust mediation method that yields reliable results even when the data deviate from normality assumptions. We demonstrate the mechanics of our proposed method in an illustrative case, while simulation studies show that our method is both superior in estimating the effect size and more reliable in assessing its significance than the existing methods. Furthermore, we provide freely available software in R and SPSS to enhance its accessibility and adoption by empirical researchers.

Keywords

mediation analysis, robust statistics, linear regression, bootstrap

Organizational research scholars are often interested in developing a thorough understanding of the processes that produce an effect, and thereby investigate the mechanisms relating to how one phenomenon exerts its influence on another. This is called a mediation analysis (Kenny, 2008). Mediation, in its simplest form, explains how or by what means an independent variable (X) affects a dependent variable (Y) through an intervening variable, called a *mediator* (M) (Baron & Kenny, 1986). Several methods have been proposed for testing mediation (see MacKinnon et al., 2002, for a review) where the most widely adopted technique is regression analysis (Wood et al., 2008).[1] The statistical performance of these methods has been vastly tested via simulation studies (e.g., MacKinnon et al., 2002; MacKinnon et al., 2004). The tests considered in those studies are based on normal-theory maximum likelihood estimators (MLEs). However, data in organizational research frequently show deviations from normality. Examples include many

[1]Econometric Institute, Erasmus University Rotterdam, Rotterdam, Netherlands
[2]Sabancı Business School, Sabancı University, Tuzla, Istanbul, Turkey

Corresponding Author:
Nüfer Yasin Ateş, Sabancı Business School, Sabancı University, Orta Mahalle, 34956 Tuzla, Istanbul, Turkey.
Email: nufer.ates@sabanciuniv.edu

individual- and firm-level constructs: job anxiety (Mannor et al., 2016), counterproductive work behavior, workplace incivility, conflict (Penney & Spector, 2005), research and development intensity (Tatarynowicz et al., 2016), managerial network centrality (Tarakci et al., 2018), and firm network centrality (Zhelyazkov & Gulati, 2016), to name a few.

Deviations from normality, such as *outliers* (i.e., data points that deviate markedly from others; Aguinis et al., 2013), *heavy tails* of the observed distribution (i.e., values farther from the mean occurring much more often than under the assumed normal distribution),[2] or *skewness* (i.e., an asymmetric distribution of the observed values), pose a serious threat to the reliability and validity of mediation analysis. Outliers create bias in a normal-theory MLE due to their strong influence on the estimator (Cohen et al., 2003; Hunter & Schmidt, 2004). Skewness and heavy tails cause a normal-theory MLE to become biased and inefficient, as it maximizes the wrong likelihood. Moreover, deviations from normality are argued to have a particularly severe effect on mediation analysis, because the mediated effect itself is a multiplication of two regression coefficients (Zu & Yuan, 2010).

Researchers often resort to nonlinear transformations (NLTs) to deal with nonnormality. Nevertheless, NLTs not only induce interpretation, validity, and generalization problems, but also prevent discovery of substantive information about variables and their distributions (Becker et al., 2019). NLTs are criticized for fostering flawed hypothesis testing (i.e., the misalignment between the hypotheses and tests) (Becker et al., 2019), and for masking real relationships while revealing spurious ones (Cohen et al., 2003). Similarly, scholars employ various outlier treatment techniques against nonnormality. However, the common practices about outliers in organizations research are found to be vague, nontransparent, and even inconsistent in outlier definition, identification, and treatment (Aguinis et al., 2013). More importantly, the empirical literature suffers from the omission of proper reporting for NLTs and outlier treatment (Aguinis et al., 2013; Becker et al., 2019), where such negligence threatens the base of empirically built organization theories.

Despite the importance of nonnormal distributions and outliers, so far no clear guidelines have been developed for mediation methods to deal with these issues properly. Existing literature often tackles these issues separately and does not address mediation analysis specifically (e.g., Aguinis et al., 2013; Aguinis et al., 2019; Becker et al., 2019; Gibbert et al., 2020). For mediation analysis, Zu and Yuan (2010) focus on outliers and propose procedures based on data cleaning, while Yuan and MacKinnon (2014) propose a procedure based on median regression and study various nonnormal error distributions. However, both methods can result in considerable bias and unreliable significance tests (cf. our simulations). While both studies stress the need for robust mediation methods, they are not optimal from a robustness point of view.

We introduce a novel procedure for mediation analysis, ROBMED, that is robust against deviations from normality including outliers, heavy tails, or skewness. ROBMED is an integrated set of procedures that builds upon the widely used bootstrap test for mediation (Bollen & Stine, 1990; MacKinnon et al., 2004; Preacher & Hayes, 2004, 2008; Shrout & Bolger, 2002). ROBMED utilizes the robust MM-regression estimator (Salibián-Barrera & Yohai, 2006; Yohai, 1987) instead of the ordinary least squares (OLS) estimator for regression, and runs bootstrap tests with the fast-and-robust bootstrap methodology (Salibián-Barrera & Van Aelst, 2008; Salibián-Barrera & Zamar, 2002). We illustrate the use of ROBMED in an empirical case where the data show deviations from normality and compare the results with state-of-the-art methods for mediation analysis. Our simulation studies, which cover a wide range of situations, suggest that ROBMED systematically outperforms other methods in estimating the true effect size and reliably assessing its significance. Furthermore, we discuss how ROBMED improves and integrates current best-practice recommendations for outliers and nonnormality, and we provide researchers with freely available software for ROBMED in R and SPSS. As such, our novel method serves as a useful and accessible tool for scholars who engage in mediation analysis.

Mediation Analysis

The simple mediation model can be formulized by the following equations:

$$M = i_1 + aX + e_1, \tag{1}$$

$$Y = i_2 + bM + cX + e_2, \tag{2}$$

$$Y = i_3 + c'X + e_3, \tag{3}$$

where i_1, i_2 and i_3 are intercepts, a, b, c, and c' are weights, and e_1, e_2 and e_3 denote random error terms. Mediation is found if the product of the estimates of the $X \rightarrow M$ path's coefficient and the $M \rightarrow Y$ path's coefficient (i.e., the estimate $\widehat{ab}$ of the *indirect effect ab*) is significant.[3] Estimating the coefficients in the mediation model is typically done via normal-theory maximum likelihood procedures, with the most commonly used method being OLS regression (Wood et al., 2008).

Deviations from model assumptions pose serious threats to mediation testing based on normal-theory MLE (i.e., OLS regression), which is illustrated in Figure 1. The plot on the top left contains 100 simulated observations that follow the model assumptions, whereas the plot on the top right uses the same data except for one single outlier being added (indicated with an arrow). With the introduction of the outlier, the indirect effect $\widehat{ab}$ almost disappears for OLS estimation. That is, the solid regression lines corresponding to Equation (2) are pulled nearly flat by the outlier and no longer represent the main part of the data. Note that we chose this particular example with an outlier for ease of illustration. Skewness and heavy tails can also distort the estimates and standard errors, but the effect is difficult to visually capture.

Numerous methods have been proposed to test the significance of the indirect effect in the literature (for reviews, see MacKinnon et al., 2004; Wood et al., 2008). A comprehensive review of these methods is beyond the scope of this study, yet we note that the bootstrap—a computer-intensive resampling technique first introduced by Efron (1979)—is found to be superior to other methods. Traditional tests for mediation often make incorrect assumptions, such as a normal distribution of the indirect effect. Since the bootstrap makes fewer assumptions, it is applicable in a wider variety of situations, especially when analytical formulas for the standard errors are not available. As such, the bootstrap provides generic ways to reliably construct confidence intervals for the indirect effect (MacKinnon et al., 2007; Preacher & Hayes, 2004, 2008).

While the bootstrap is a nonparametric technique and can therefore handle nonnormal distributions, it is sensitive to outliers. Outliers may be oversampled, which can corrupt the obtained bootstrap distribution of the estimator (Salibián-Barrera & Zamar, 2002). Thus, the size and significance of the indirect effect can be severely influenced and may lead to incorrect conclusions regarding the mediation relationships between the variables. Consequently, mediation analysis that is robust against both nonnormal distributions and outliers requires not only a robust estimator of the mediation model, but also a robust bootstrap procedure.

Treatment of Nonnormality in the Empirical Literature

In empirical research, two-step procedures are frequently used for the treatment of nonnormality. For general deviations in the observed distributions, researchers first transform the data before applying traditional statistical methods (Becker et al., 2019). For outliers, researchers first identify and remove them from the data, then apply traditional methods to the cleaned data set (Aguinis et al., 2013). However, these two-step procedures have several drawbacks.

Regarding transformations, Becker et al. (2019) report serious problems with NLT selection, reporting, interpretation, and justification: scholars usually adopt NLTs without (a) any justification of their use (more than 50% of the surveyed articles), (b) reporting the effects on results (more than

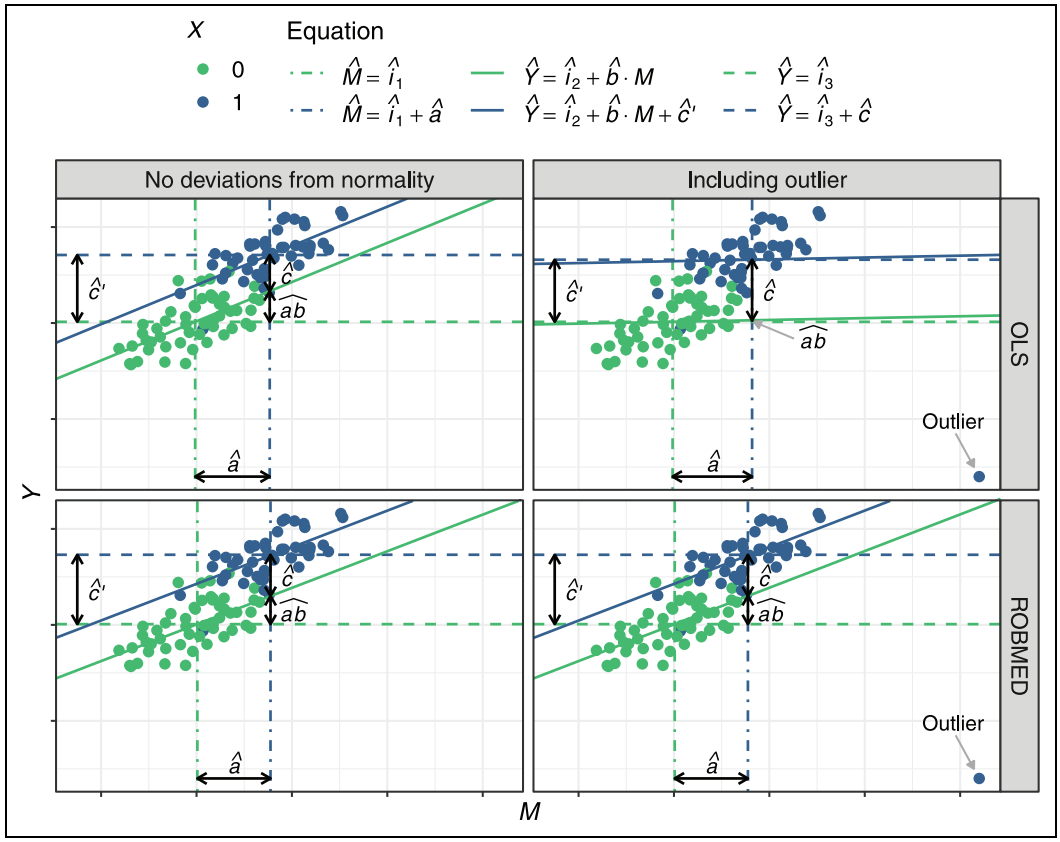

Figure 1. The effect of a single outlier on mediation analysis.
Note: The mediator *M* is depicted on the horizontal axis and the dependent variable *Y* on the vertical axis. The independent variable *X* is assumed to be dichotomous for simplicity in visual representation, i.e., the regression models in Equations (1)–(3) then correspond to two parallel fitted lines. Light green lines correspond to fitted regression lines for $X = 0$ (light green points), while dark blue lines correspond to fitted regression lines for $X = 1$ (dark blue points). The distance between the horizontal dashed regression lines represents the total effect $\hat{c}'$ of *X* on *Y*, and the distance between the vertical dash-dotted regression lines represents the effect $\hat{a}$ of *X* on *M*. The remaining solid regression lines describe the relation of *M* to *Y* within the groups of *X*. A change in M of $\hat{a}$ units (due to a change in *X* from 0 to 1) leads to an indirect change in *Y* of $\widehat{ab}$ units (i.e., the indirect effect). The figure also illustrates that the product of coefficients $\widehat{ab}$ is equal to the difference in coefficients $\hat{c}' - \hat{c}$ (MacKinnon et al., 2007). A color version of this figure is available online at: https://journals.sagepub.com/doi/figure/10.1177/1094428121999096

95%), and (c) alignment of hypotheses and tests in terms of the transformed variables (more than 90%). For instance, the log transformation, the most commonly used NLT,[4] changes the scale of a variable in a way that the relationship between the transformed variable and the dependent variables implies diminishing returns on the original scale of the variable (Becker et al., 2019). Such changes in the scale due to NLTs may also mask real relationships while revealing spurious ones (Cohen et al., 2003). Ideally, the use of NLTs should be motivated by theory rather than the observed distributions of variables. For example, if a log transformation of income is used as the dependent variable, then the theory should justify a relative change in income based on the independent variables, rather than a change in fixed amounts. Researchers sometimes apply NLTs that are designed to remove skewness (e.g., the log transformation) with the intention to reduce the effect

of outliers (Becker et al., 2019). However, when the main part of the data is already close to normal, this would introduce left skewness, thus actually making the data less normal.

Regarding outliers, the empirical use of outlier treatment techniques is documented to be ambiguous, inconsistent, and often dismissed in manuscripts (Aguinis et al., 2013). In addition, when statistical methods are applied to the cleaned data, the resulting standard errors do not include the additional uncertainty from the initial data-cleaning step. For instance, Chen and Bien (2020) show that OLS regression after outlier removal results in confidence intervals that are much too small. In some of their simulations, the coverage of the true parameter is as low as 75%, as opposed to the nominal coverage of 95%. Consequently, the p values from significance tests are too small and could incorrectly suggest significant results. Furthermore, outlier removal can result in a loss of stability in borderline situations (e.g., an observation close to the threshold of an outlier detection rule), as it requires to fully include or fully exclude the observation.

In practice, it is often unclear if deviating observations are produced by skewness or a heavy tail in the distribution, or whether they are outliers. Different treatment methods (NLTs or outlier removal) may lead to very different results and conclusions. Therefore, NLTs and outlier removal can also be abused as dangerous post hoc practices to increase the chances of finding what the researcher wants to find (Becker et al., 2019; Cortina, 2002), which threatens the base of empirically tested theory (Bettis, 2012).

Robust Statistics

Statistical methods are traditionally designed to be as efficient as possible under a certain model and assume that all data points strictly follow this model. However, the corresponding models typically make quite strong assumptions about the data, which are often violated in empirical settings. When this is the case, such methods can give unreliable results that may yield incorrect conclusions. The field of *robust statistics* aims to develop statistical methods that are less affected by model deviations and show good behavior in many situations. Robust methods therefore focus on the part of the data that is the most relevant for estimating the model parameters. In other words, robust methods exchange some statistical efficiency for wider applicability. This loss of efficiency is often small and can be seen as an insurance premium against failure under deviations from the model assumptions.

Modern robust methods typically aim for a *continuous downweighting* of deviating observations with weights between 0 and 1 that measure the degree of deviation. Moreover, these methods simultaneously downweight deviating observations while estimating the model. This approach solves the issues with current treatment of nonnormality, as discussed above. It is a unified approach for handling outliers and other deviations from normality (heavy tails and skewness), and continuous downweighting ensures stability of the results. The procedures for downweighting observations are extensively studied in the statistics literature and supported by statistical theory (e.g., Maronna et al., 2006), which improves research reproducibility compared to manual preprocessing of the data by the researcher. Finally, if there are no deviations from the model, observations receive a weight close to 1 such that the robust method yields similar results to the corresponding maximum likelihood method.[5]

In the following, we present a discussion on how various deviations from normality assumptions affect estimation, if and how downweighting deviating observations is a suitable treatment, and how downweighting aligns with current best-practice recommendations.

Heavy Tails, Skewness, and Their Effect on Estimation

Heavy tails are observed when values farther from the mean occur much more often than under the assumed normal distribution, and *skewness* refers to an asymmetric distribution of the observed

values. When the distribution is symmetric but has heavy tails, estimates of central tendencies are not much affected, yet their standard errors inflate. For instance, OLS estimates are still unbiased but very inefficient under heavy-tailed errors. This leads to large confidence intervals and a loss of power in significance tests. When the distribution is skewed, estimates become both inefficient and biased, hence significance tests may be less powerful (increased Type II error) and poorly calibrated (increased Type I error).

Empirically, skewness does not manifest itself much in the central part of the data but in the tails (Raymakers & Rousseeuw, 2020). By definition, the same holds for heavy tails. Therefore, when one is interested in central tendencies such as regression coefficients, the main issue is that deviating observations due to skewness or heavy tails have disproportional influence on estimators that assume normality. A gradual downweighting of these influential observations decreases their disproportional influence, resulting in more reliable significance tests. In case of skewed distributions, a gradual downweighting of the longer tail may introduce bias, for instance when interpreting regression coefficients as changes in the mean of the dependent variable. Nevertheless, this bias is often small (cf. our simulations).

In a regression setting, the gradual downweighting of deviations can be based on the residuals, leaving the observed variables untouched. This solves estimation issues due to nonnormality while still allowing for interpretation on the original scale of the variables—unlike NLTs, where the transformed model parameters are difficult to interpret (see also Becker et al., 2019). In that regard, the robust approach of downweighting deviations is in line with the best-practice recommendation that researchers should consider other treatments instead of assuming NLTs are the best option (Becker et al., 2019).

Outliers and Their Effect on Estimation

An important concept in robust statistics is that of *outliers*. Hawkins (1980, p. 1) defines an outlier as "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism."[6] Aguinis et al. (2013) further distinguish between *error outliers* (deviating observations as a result of inaccuracies), *interesting outliers* (deviating points that contain potentially valuable or unexpected knowledge), and *influential outliers* (deviating observations whose presence alters the model fit or parameter estimates).

Examples for error outliers are measurement or recording errors, or observations from a different population that is not of interest to the researcher. According to the best-practice recommendations of Aguinis et al. (2013), such outliers should be corrected if possible, or excluded from analysis. The latter is equivalent to assigning a weight of 0 for estimation. Examples for interesting outliers are a rare, extreme value that is part of the population of interest, or an observation from a subpopulation that is of interest to the researcher but that is otherwise not represented in the sample. Aguinis et al. (2013) strongly recommend studying those observations separately with appropriate qualitative or quantitative approaches.

Crucially, the last two examples above are also examples for influential outliers in a regression setting.[7] Consider a sample size of $n = 100$ and an extreme error that corresponds to a 1-in-10,000 event. In estimation, this observation would be treated as a 1-in-100 event rather than a 1-in-10,000 event, meaning that its value has a disproportionate influence on the estimates. Regarding an observation from a subpopulation that is otherwise not represented in the sample, such heterogeneities in the population would need to be modeled explicitly, which is not meaningful if the sample does not contain enough observations from the small subgroup. When using the same model for the entire sample, the researcher could end up with results that neither reflect the main part of the population nor the subgroup (cf. Figure 1). For influential outliers in regression, Aguinis et al. (2013)

recommend reporting the results both with treatment (e.g., robust estimation) and without treatment (e.g., OLS).

Modern robust methods allow to detect outliers by reporting observations with weights close to 0. The researcher should then further investigate the type of each detected outlier. For error and influential outliers, the robust method has already applied the correct treatment by downweighting them. Interesting outliers should be studied further in detail. For instance, an interesting outlier that turns out to be an extreme observation can be studied with statistical tools from extreme value theory (e.g., de Haan & Ferreira, 2006). If an interesting outlier is an observation from a subpopulation that is otherwise not represented in the sample, the researcher can study this observation qualitatively, collect more data to model the heterogeneities in the population, or design a follow-up study to analyze the small subpopulation. For a general roadmap on how to gain knowledge from outliers, we refer to Gibbert et al. (2020).

Robust Statistics and Mediation Analysis

Despite the common presence of deviations from model assumptions and the sensitivity of mediation results to such deviations, we could only find two articles on robust mediation analysis. Zu and Yuan (2010) focus on outliers and propose to clean the data via winsorization, which is neither as robust nor as efficient as modern robust regression methods (cf. our simulations). Yuan and MacKinnon (2014) propose a bootstrap procedure that replaces OLS regression with median regression, and study nonnormal error distributions and outliers in the errors. While median regression is robust in those settings, it is not robust even if a single outlier occurs in the explanatory variables (Koenker, 2005, p. 46). Yet outliers in the explanatory variables are considered to be the most harmful type of outliers in regression due to their high leverage effect on the estimates (cf. Figure 1).

Both studies pinpoint the need for robust methods for mediation analysis and propose valuable potential alternatives, but both suffer from the aforementioned disadvantages. In that sense, although these methods clearly are more robust than OLS-based procedures, they still need to be improved upon from a robustness point of view.

ROBMED: Robust Mediation Analysis

We propose a robust test for mediation, ROBMED, that builds on bootstrapping the indirect effect via linear regression. First, linear regression analysis is the most widely used mediation technique in empirical studies (Wood et al., 2008). Second, the bootstrap test is the state-of-the-art method for testing the indirect effect in mediation models, as its distribution is in general asymmetric.[8] Accordingly, ROBMED constitutes a combination of two essential building blocks.

The first building block is to use the robust MM-regression estimator (Salibián-Barrera & Yohai, 2006; Yohai, 1987) rather than the OLS estimator. Instead of the quadratic loss function of the OLS estimator, the MM-estimator uses a loss function that is quadratic for small residuals, but smoothly levels off for larger residuals. This ensures that the coefficient estimates are determined by the central part of the data and that the influence of deviations from normality is limited. The left panel in Figure 2 illustrates this loss function. The MM-estimator can be seen as a weighted least-squares estimator with data dependent weights. A compelling feature of the estimator is that the weights that are assigned to the data points can take any value between 0 and 1, where a lower weight indicates a higher degree of deviation. An illustration of this weight function is given in the right panel in Figure 2.

The second building block is to adopt the fast-and-robust bootstrap (Salibián-Barrera & Van Aelst, 2008; Salibián-Barrera & Zamar, 2002) instead of the standard bootstrap. There are two issues with the standard bootstrap. The first issue is that it is not robust to outliers. It draws so-called

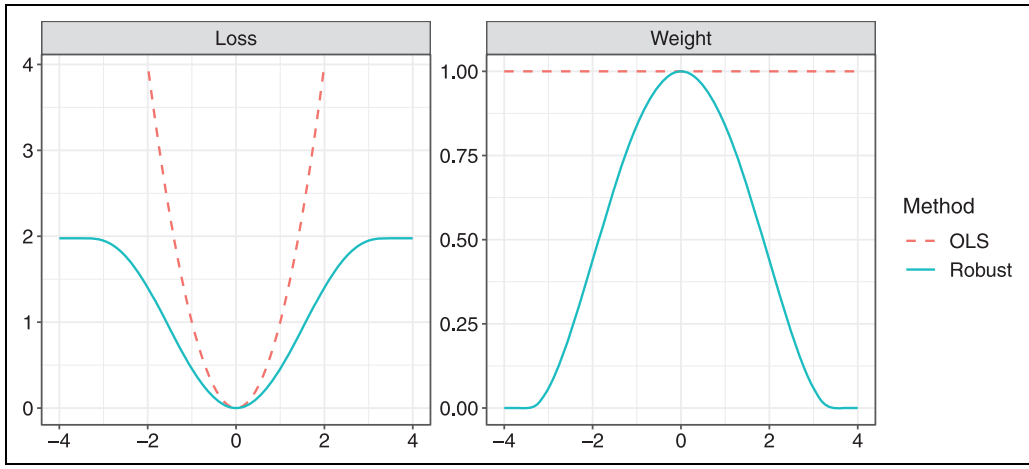

Figure 2. Loss function (left) and assigned weights (right) for OLS regression and the robust MM-regression estimator.

bootstrap samples of the same size as the original sample via random sampling with replacement and estimates the model on each of those bootstrap samples. Even if a robust method can reliably estimate the model in the original sample, it may happen that outliers are oversampled in some bootstrap samples. If those bootstrap samples contain more outliers than the robust method can handle, bootstrap confidence intervals become unreliable. The second issue is that robust methods typically come with increased computational complexity. While this is less of an issue in most applications due to modern computing power, there can be a noticeable increase in computing time compared to traditional methods, in particular when combined with computer-intensive procedures such as the bootstrap.

To solve the two issues, Salibián-Barrera and Zamar (2002) developed the fast-and-robust bootstrap. Keep in mind that the MM-regression estimator can be seen as weighted least-squares estimator, where the weights are dependent on how much an observation is deviating from the rest. The essence of the fast-and-robust bootstrap is that on each bootstrap sample, first a weighted least-squares estimator is computed (using the robustness weights from the original sample) followed by a linear correction of the coefficients. The purpose of this correction is to account for the additional uncertainty of obtaining the robustness weights.

The integration of the robust MM-regression estimator with the fast-and-robust bootstrap procedure allows us to construct a test for mediation analysis that follows the same principles as the widely used OLS bootstrap test. However, our proposed test is more reliable under deviations from the model assumptions such as outliers, heavy tails, and skewness. Technical derivations and a brief discussion of our software can be found in the appendix.

Revisiting Figure 1, the bottom plots depict the same mediation model estimated with ROBMED. Without the outlier (left), the estimated effects are nearly identical to those of OLS. When the outlier is included (right), the fitted regression lines remain virtually unchanged and all effects are accurately estimated, illustrating the merit of ROBMED.

Illustrative Empirical Case

In order to show the role of deviations from the model assumptions in mediation analysis and how ROBMED overcomes those challenges, we test an illustrative hypothesis. It is not our aim to build

Table 1. Descriptive Statistics of the Variables Used in the Illustrative Empirical Case.

Variable	Mean	Standard Deviation	Median	Median Absolute Deviation	Minimum	Maximum
Task conflict	1.761	0.392	1.688	0.371	1.125	2.938
Team commitment	3.822	0.448	3.875	0.371	2.125	4.688
Value diversity	1.676	0.344	1.587	0.366	1.105	2.548

Note: The median is a more robust measure of centrality than the mean, and the median absolute deviation is a more robust measure of dispersion than the standard deviation (e.g., Maronna et al., 2006).

Table 2. Correlation Table of the Variables Used in the Illustrative Empirical Case.

	Task conflict	Team commitment	Value diversity
Task conflict	1.000	–.297	.268
Team commitment		1.000	–.024
Value diversity			1.000

Note: The reported correlations are Spearman's rank correlations, transformed to be consistent with the Pearson correlation coefficient (Croux & Dehon, 2010). Those provide more robust estimates than the sample Pearson correlation.

and test theory with this example, therefore we do not interpret the indirect effect or discuss its effect size. The data contain information on $n = 89$ randomly assigned 4-person teams of senior business administration students who played a business simulation game as part of their capstone strategy course at a Western European university.[9] We investigate the following hypothesis:

> Illustrative Hypothesis: Task conflict (M) mediates the relationship between team value diversity (X) and team commitment (Y).

More information on data collection, scales, and the underlying theory is presented in the Online Appendix 1. Tables 1 and 2 contain descriptive statistics and correlations for the studied variables.

We focus on a comparison between ROBMED and the OLS bootstrap, but we also apply other state-of-the-art methods for mediation testing. Table 3 gives an overview of these methods and the abbreviations we use to refer to them.[10] All bootstrap tests report a bias-corrected and accelerated percentile-based confidence interval (Davison & Hinkley, 1997) for the indirect effect. Table 4 reports results on all coefficients in the mediation analyses. The estimate of the indirect effect ab is nearly twice as large in magnitude for ROBMED compared to the OLS bootstrap. In addition, the 95% confidence interval of ROBMED is strictly negative but that of the OLS bootstrap contains 0. For further insight, we estimate the p value for the indirect effect as the smallest significance level α where the $(1 - \alpha) \cdot 100\%$ confidence interval obtained from the bootstrapped distribution does not contain 0. We observe that ROBMED finds evidence against the null hypothesis of no mediation (p value $= .027$), whereas the OLS bootstrap finds no evidence (p value $= .158$). Other than the indirect effect, the main difference between the two methods is in the estimation of the a path, which is clearly not significant for the OLS bootstrap (p value $= .209$) but highly significant for ROBMED (p value $= .003$). Hence, we take a closer look at the relationship between the independent variable and the hypothesized mediator.

Figure 3 shows a scatterplot of task conflict (M) against value diversity (X) together with *tolerance ellipses*. The shape of such a tolerance ellipse is defined by the covariance matrix, and its size is determined such that a certain proportion of the data points is expected to lie within the ellipse under the assumption of a normal distribution (here 97.5%). The plot contains a tolerance ellipse based on the sample covariance matrix, which is closely linked to OLS regression,[11] as well

Table 3. Methods Included in the Illustrative Empirical Case and the Simulation Study, as Well as the Abbreviations Used to Refer to Them.

Abbreviation	Description
OLS bootstrap	The bootstrap test following OLS estimation (Bollen & Stine, 1990; MacKinnon et al., 2004; Preacher & Hayes, 2004, 2008; Shrout & Bolger, 2002).
OLS Sobel	The Sobel test following OLS estimation (Sobel, 1982), which assumes a normal distribution of the indirect effect. The indirect effect $\widehat{ab}$ is divided by (a first-order approximation of) the standard error of the indirect effect $s_{\widehat{ab}}$ to obtain a test statistic for which the p value is computed with the standard normal distribution. In the literature, the Sobel test has been criticized for the assumption of a normal distribution of $\widehat{ab}$, as the product of two normally distributed random variables—the coefficients $\hat{a}$ and $\hat{b}$—is not normally distributed (MacKinnon et al., 2002).
Box-Cox bootstrap	We first apply a Box-Cox transformation (Box & Cox, 1964) to each variable, then perform the OLS bootstrap test. Note that due to the transformations, the estimates are not comparable to those of the other methods.
SNT bootstrap	We perform regression with normal, skew-normal, *t*, or skew-*t* error distributions (Azzalini & Arellano-Valle, 2013) within the bootstrap procedure and select the best fitting error distribution via the Bayesian information criterion (BIC) (Schwarz, 1978). Note that a similar test was proposed by Asparouhov and Muthén (2016) using structural equation modeling.
Winsorized bootstrap	Zu and Yuan's (2010) bootstrap test following winsorization of the data.
Median bootstrap	Yuan and MacKinnon's (2014) bootstrap test using median regression.
ROBMED	Our proposed test using MM-estimation (Yohai, 1987) and the fast-and-robust bootstrap (Salibián-Barrera & Zamar, 2002).

as a robust tolerance ellipse based on the weighted covariance matrix using the weights from the robust regression of M on X.[12] The plot reveals that there are a small number of influential observations, but it is not so clear whether these observations are true outliers or the result of a heavy upper tail in task conflict. Only the three most far away points receive a weight of exactly 0, with two more points being assigned a weight < 0.01. The points close to the border of the robust tolerance ellipse are only partly downweighted and receive a weight in between 0 and 1. Overall, the robust tolerance ellipse better fits the main bulk of the data, as there is much more empty space in the nonrobust tolerance ellipse. The influence of the far away points is also visible in the OLS regression line, which is tilted to become more horizontal.

To further investigate the deviations from normality, Figure 4 shows a diagnostic plot of the robust regression weights. For varying threshold on the horizontal axis, the vertical axis displays how many observations in each tail of the residual distribution have a weight below this threshold. For comparison, a reference line is drawn for the expected percentages under normal errors. Clearly, there are more downweighted observations with positive residuals than expected and fewer with negative residuals, confirming skewness with a heavy upper tail.

Based on the two plots, ROBMED better captures the main trend in the data and can be considered more reliable than the OLS bootstrap. The SNT bootstrap, which explicitly models skewness in the errors, yields similar results to ROBMED. Other methods give somewhat different results: the Box-Cox bootstrap and the winsorized bootstrap come to the opposite conclusions regarding weak or strong significance of the coefficients a and b, with the winsorized bootstrap also reporting only a weakly significant indirect effect ab. With the median bootstrap, neither the coefficient b nor the indirect effect ab is found to be significant. In order to decide which of these methods can be considered the most reliable and should be adopted as the current best practice in mediation analysis, we conduct simulation studies in the next section.

Table 4. Results from the Illustrative Empirical Case: Comparison of ROBMED to Various Methods.

	OLS Bootstrap				OLS Sobel			
Direct Effects	Estimate	Std. Error	*z* Statistic	*p* Value	Estimate	Std. Error	*t* Statistic	*p* Value
X→M (a path)	0.159	0.127	1.255	.209	0.155	0.121	1.283	.203
(X), M→Y (b path)	–0.370	0.160	–2.312	.021*	–0.364	0.118	–3.090	.003**
X,(M)→Y (c path)	–0.016	0.145	–0.113	.910	–0.021	0.134	–0.156	.877
X→Y (c′ path)	–0.076	0.159	–0.480	.631	–0.077	0.139	–0.555	.580
Indirect Effect	Estimate	95% Confidence Interval		*p* Value	Estimate	Std. Error	*t* Statistic	*p* Value
ab	–0.060	(–0.208, 0.025)		.158	–0.057	0.048	–1.185	.236
	Box-Cox Bootstrap				SNT Bootstrap			
Direct Effects	Estimate	Std. Error	*z* Statistic	*p* Value	Estimate	Std. Error	*z* Statistic	*p* Value
X→M (a path)	0.175	0.092	1.898	.058†	0.254	0.111	2.293	.022*
(X), M→Y (b path)	–10.951	4.253	–2.575	.010*	–0.325	0.174	–1.874	.061†
X,(M)→Y (c path)	0.964	3.119	0.309	.757	0.018	0.159	0.115	.908
X→Y (c′ path)	–0.923	3.368	–0.274	.784	0.001	0.167	0.005	.996
Indirect Effect	Estimate	95% Confidence Interval		*p* Value	Estimate	95% Confidence Interval		*p* Value
ab	–1.886	(–5.162, –0.055)		.044*	–0.082	(–0.243, –0.002)		.042*
	Winsorized Bootstrap				Median Bootstrap			
Direct Effects	Estimate	Std. Error	*z* Statistic	*p* Value	Estimate	Std. Error	*z* Statistic	*p* Value
X→M (a path)	0.197	0.111	1.772	.076†	0.290	0.135	2.144	.032*
(X), M→Y (b path)	–0.392	0.123	–3.185	.001**	–0.326	0.209	–1.563	.118
X,(M)→Y (c path)	0.013	0.137	0.093	.926	0.053	0.211	0.252	.801
X→Y (c′ path)	–0.063	0.148	–0.426	.670	–0.041	0.214	–0.191	.848
Indirect Effect	Estimate	95% Confidence Interval		*p* Value	Estimate	95% Confidence Interval		*p* Value
ab	–0.076	(–0.198, 0.001)		.052†	–0.094	(–0.285, 0.010)		.107

(continued)

Table 4. (continued)

ROBMED				
Direct Effects	Estimate	Std. Error	*z* Statistic	*p* Value
X→M (a path)	0.321	0.107	2.998	.003**
(X), M→Y (b path)	–0.344	0.178	–1.934	.053†
X,(M)→Y (c path)	0.065	0.186	0.350	.726
X→Y (c′ path)	–0.045	0.187	–0.241	.810
Indirect Effect	Estimate	95% Confidence Interval		*p* Value
ab	–0.110	(–0.294, –0.010)		.027*

Note: Variables are value diversity (*X*), task conflict (*M*), and team commitment (*Y*). Sample size = 89. Number of bootstrap samples = 5,000.
†$p < .1$. *$p < .05$. **$p < .01$. ***$p < .001$.

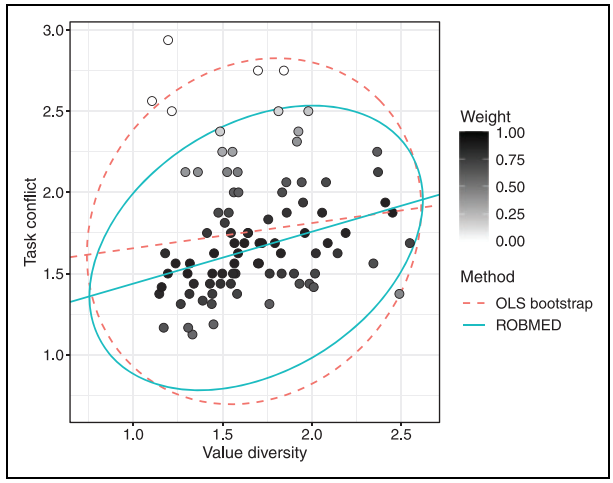

Figure 3. Scatterplot of value diversity and task conflict with tolerance ellipses.

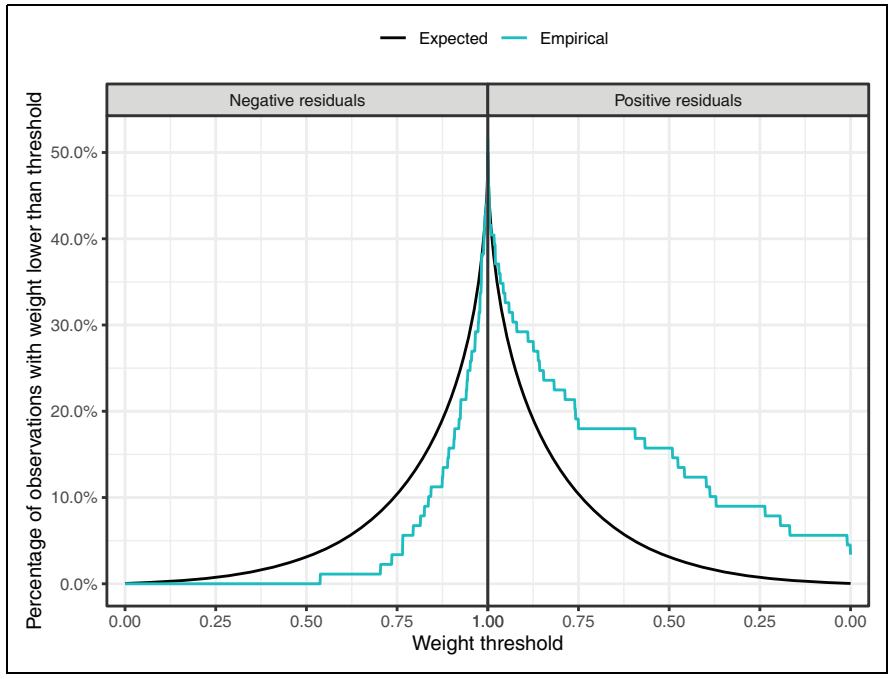

Figure 4. Diagnostic plot of weights from robust regression of task conflict on value diversity.
Note: The horizontal axis contains different weight thresholds, and the vertical axis displays the percentage of observations that have a weight below this threshold. A black reference line indicates the expected percentages under normally distributed errors. Observations with negative and positive residuals are shown separately to make it possible to distinguish between symmetric and asymmetric deviations from normality.

Table 5. Settings Regarding the Error Distributions and Outliers in the Simulations.

Setting	Details
Normal	The error terms e_1 and e_2 follow a standard normal distribution; no outliers.
Outliers	The error terms e_1 and e_2 follow a standard normal distribution. With probability 0.02, observations are turned into outliers by setting $M_i^* = M_i/10 - 3$ and $Y_i^* = Y_i/10 + 3$.
Skewness	The error terms e_1 and e_2 follow a skew-normal distribution with skewness equal to 0.995 as an example of a skewed distribution; no outliers. Note that a skewness of 0.995 is the maximum skewness that the skew-normal distribution can exhibit.
Heavy tails	The error terms e_1 and e_2 follow a t distribution with 2 degrees of freedom as an example of a distribution with heavy tails; no outliers.

Simulations

We simulate $n = 100$ observations according to the models $M = aX + \sigma_1 e_1$ and $Y = bM + cX + \sigma_2 e_2$. First, the explanatory variable X is generated from a standard normal distribution. We set $a = c = 0.4$, but we vary the value of b to investigate two different situations: one with mediation ($b = 0.4$, true indirect effect $ab = 0.16$), and one where mediation does not exist ($b = 0,$ indirect effect $ab = 0$). Moreover, we consider four settings regarding the error distributions and outliers, as described in Table 5. The parameters σ_1 and σ_2 are chosen such that M and Y have variance 1 in the setting with normally distributed errors. We apply the same methods as for the empirical case, except that the Box-Cox bootstrap uses a generalization of the Box-Cox transformation that allows for negative values (Hawkins & Weisberg, 2017). We perform two-sided tests with null hypothesis $H_0 : ab = 0$ against the alternative $H_a : ab \neq 0$. The whole process is repeated $K = 1,000$ times.

Simulations with Mediation

Figure 5 shows the average estimates of the indirect effect (top row), as well as a measure of realized power of the tests on the indirect effect (bottom row). This measure of realized power is taken as the rate of how often the methods reject the null hypothesis and the corresponding estimate of the indirect effect has the correct sign.[13] The columns of the figure correspond to the four investigated settings for error distributions and outliers.

For normal error terms, all methods estimate the indirect effect very accurately. In addition, all tests show a realized power close to 100%, except for the median bootstrap with a realized power slightly below 90%. In the presence of outliers, ROBMED is the only method that still gives accurate estimates of the indirect effect. The OLS-based methods are the most affected by the outliers, but the median bootstrap, the winsorized bootstrap, and the SNT bootstrap also show a considerable bias. The results from estimation clearly carry over to the realized power of the tests, with ROBMED remaining close to 100%. The winsorized bootstrap is the only method not too far behind with a realized power slightly below 90%, despite its bias in effect size. While the median bootstrap and the SNT bootstrap have realized power of about 60–65%, the remaining tests perform poorly.

For skew-normal error terms, all methods are very accurate in estimating the indirect effect and have realized power close to 100%. Clearly, the SNT bootstrap has the smallest variance in the estimates of the indirect effect. There are again more differences among the methods for t-distributed errors. ROBMED, the SNT bootstrap, the winsorized bootstrap, and the median bootstrap all estimate the indirect effect accurately and have high realized power of about 80–85%. The realized power of the Box-Cox bootstrap is also close to 80%. As expected, OLS Sobel and the OLS bootstrap still yield accurate estimates of the indirect effect, but perform rather poorly in terms of power.

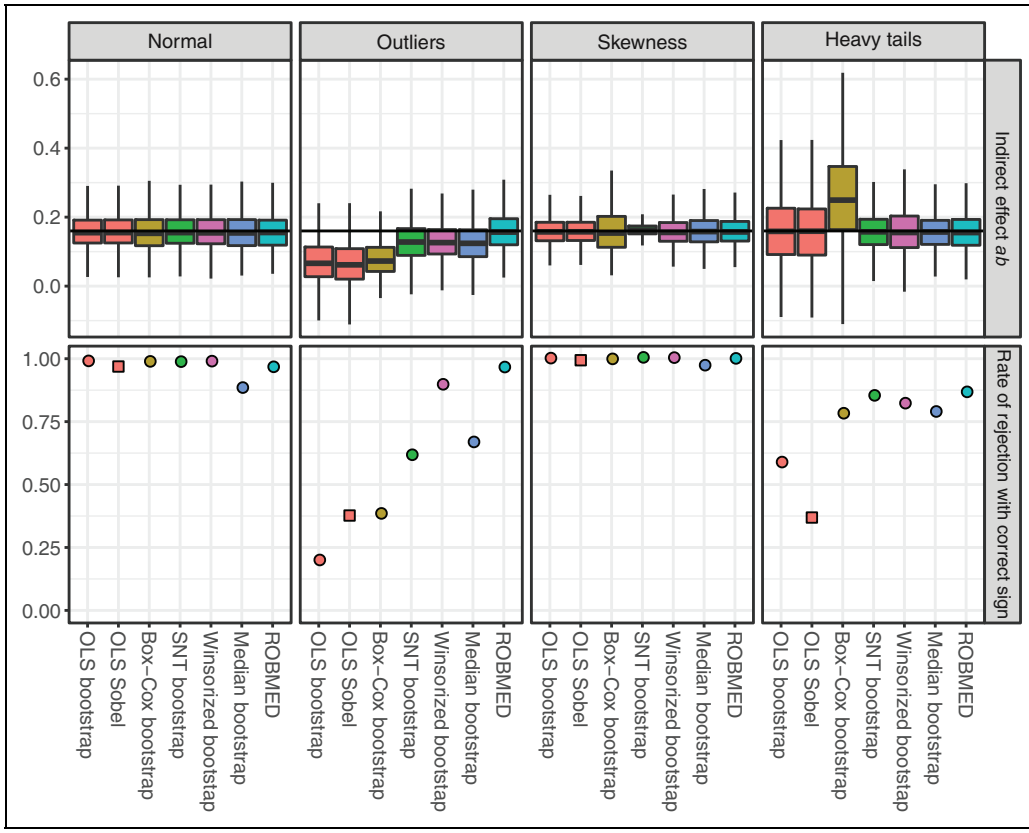

Figure 5. Results from 1,000 simulation runs for the simulation setting with mediation ($a = 0.4$, $b = 0.4$). Note: The top row contains box plots of the estimates of the indirect effect and includes a horizontal reference line for the true indirect effect $ab = 0.16$. Points outside the whiskers are not displayed for better readability. The bottom row displays the rate of how often the methods reject the null hypothesis and the corresponding estimate of *ab* has the correct sign (a measure of realized power of the tests in the presence of outliers; the higher this rate the better). The columns correspond to the four investigated settings for error distributions and outliers.

Simulations with No Mediation

The top row of Figure 6 again shows box plots of the estimates of the indirect effect, while the bottom row displays the rejection rate of the tests. Since the tests are performed with nominal size $\alpha = 0.05$, the rejection rate should be as close as possible to this value. It can be seen as the realized size of the test.

Under normal error terms, all methods again yield very accurate estimates of the indirect effect, and the rejection rates of all bootstrap tests are close to the nominal size. As expected, the OLS Sobel test has a slightly lower rejection rate than the bootstrap tests. When outliers are introduced, ROBMED again yields the most accurate estimates of the indirect effect, and its rejection rate is the closest to the nominal size. All other methods suffer from considerable bias, in particular the OLS-based methods. The median bootstrap is the only other method that is not too far off the nominal size with a rejection rate of 12%, whereas all other tests have too large rejection rates in the range of 20-35%.

As in the setting with mediation, all methods perform very well for skew-normal error terms. The most notable result is again that the SNT bootstrap has much lower variance in the estimates of the

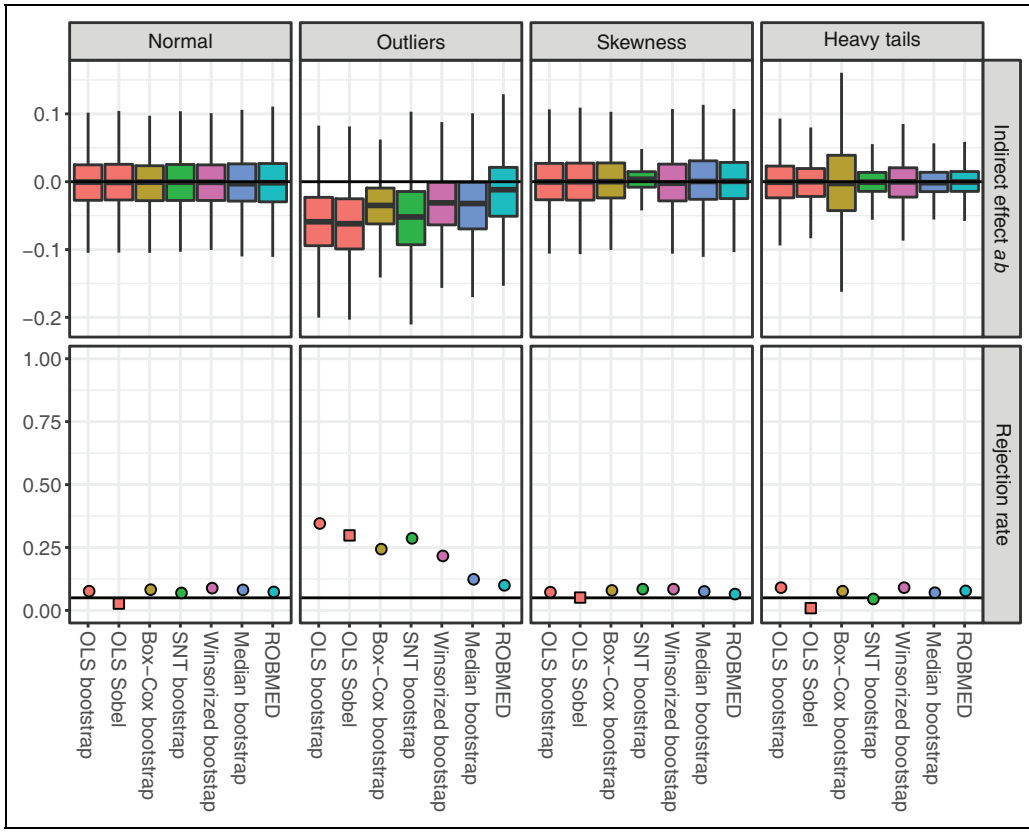

Figure 6. Results from 1,000 simulation runs for the simulation setting with no mediation ($a = 0.4$, $b = 0$). Note: The top row contains box plots of the estimates of the indirect effect and includes a horizontal reference line for the true indirect effect $ab = 0$. Points outside the whiskers are not displayed for better readability. The bottom row displays the rejection rate of the corresponding tests (i.e., the realized size), and a horizontal line is drawn for the nominal size $\alpha = 0.05$ (the closer to this line the better). The columns correspond to the four investigated settings for error distributions and outliers.

indirect effect. For *t*-distributed errors, all methods perform (reasonably) well, too. The main difference is in the variance of the estimates of the indirect effect, with the SNT bootstrap, the median bootstrap and ROBMED showing the smallest variances.

Additional Simulations and Concluding Discussion

We extended our simulation design with a wide range of sample sizes, effect sizes, outlier configurations, and error distributions with various levels of skewness and kurtosis. The results presented here are a representative selection, while all results from our extensive simulations with 700 different parameter settings can be found in the Online Appendix 2.[14] Overall, ROBMED clearly outperforms the other methods. It is the only method that remains accurate in estimating the indirect effect and powerful in hypothesis testing across all investigated deviations from normality. In addition, only ROBMED effectively protects against false mediation discoveries (inflated Type I errors) in the presence of outliers.

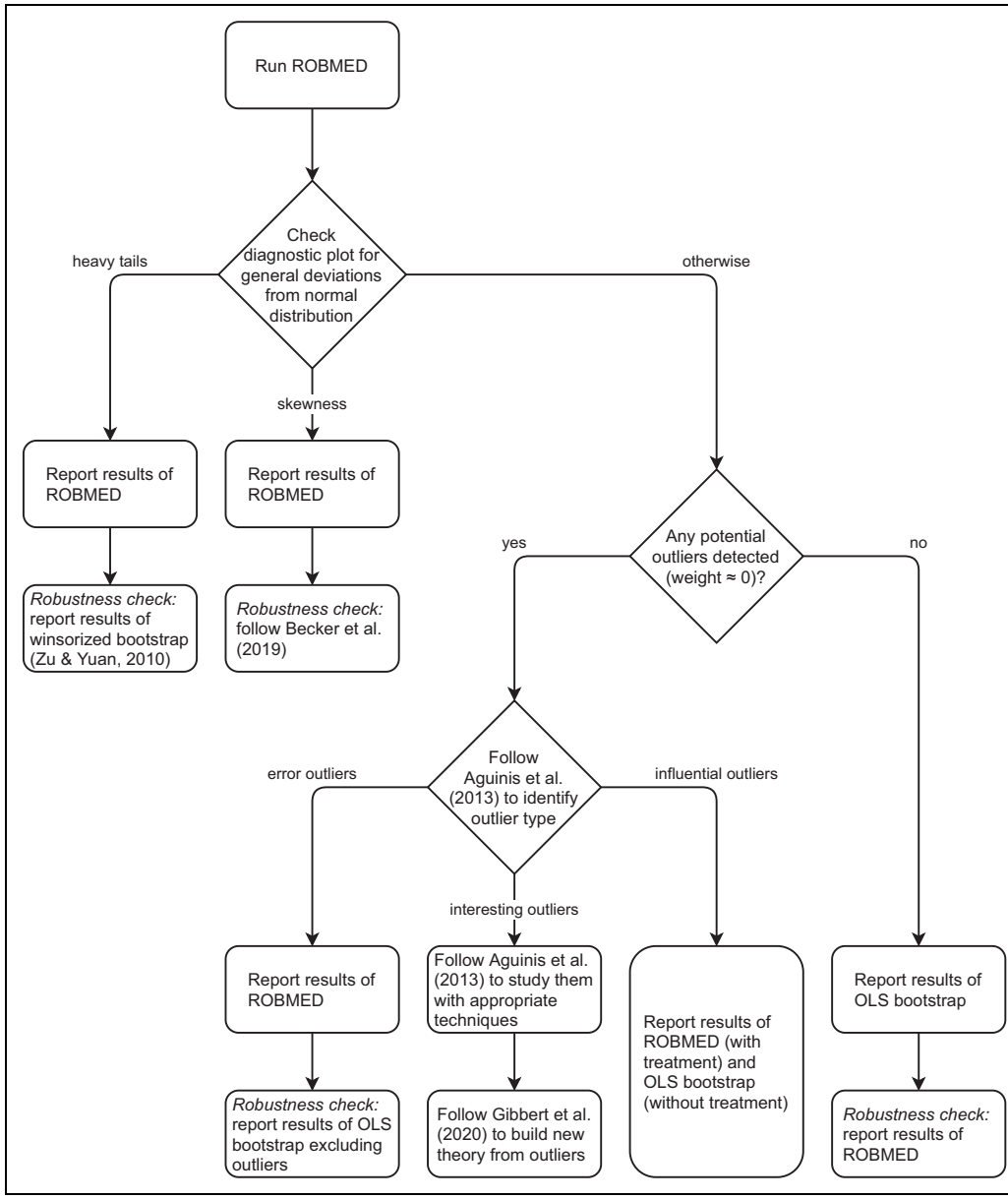

Figure 7. Flowchart of practical guidelines for using ROBMED.

Practical Guidelines for Using ROBMED

ROBMED is robust to deviations from normality, which makes it a useful tool to detect such deviations in the first place. We recommend researchers to estimate their hypothesized model with ROBMED, and to investigate the diagnostic plot of the weights from the robust regressions (cf. Figure 4). Depending on detected deviations from normality, we recommend taking the actions discussed below. A detailed flowchart for our practical guidelines regarding the use of ROBMED is given in Figure 7.

1. Heavy tails and skewness:
   a. If the diagnostic plot reveals more downweighted observations than expected, but roughly the same amounts in both tails, the distribution is symmetric but with heavy tails. We recommend reporting the results of ROBMED, since ROBMED outperformed the other methods in our simulations. The winsorized bootstrap (Zu & Yuan, 2010) could be used as an additional robustness check.
   b. If the diagnostic plot shows that observations in one tail are downweighted more heavily than those in the other tail, the distribution is skewed. We advise verifying the findings of ROBMED by following the recommendations of Becker et al. (2019) as a robustness check.
2. **Outliers:** Observations that receive a weight of (close to) 0 are potential outliers, and researchers should follow Aguinis et al. (2013) to identify the type of outliers.
   a. *Error outliers*: ROBMED already applies the correct outlier treatment following Aguinis et al. (2013) by excluding them, and we recommend reporting the results of ROBMED. As a robustness check, researchers could apply the OLS bootstrap with the error outliers excluded.
   b. *Interesting outliers*: Researchers should follow the protocols of Aguinis et al. (2013) and study them with appropriate techniques. We further recommend following the suggestions of Gibbert et al. (2020) to possibly build new theory from interesting outliers, and adjusting the model accordingly.
   c. *Influential outliers*: Robust estimation is a valid form of outlier treatment in line with Aguinis et al.'s (2013) suggestions. Their best-practice advice is to report both the results with outlier treatment (we recommend ROBMED) and without outlier treatment (we recommend the OLS bootstrap).
3. **No deviations:** If the diagnostic plot indicates that the percentages of downweighted observations are close to what is expected, and no observations receive a weight close to 0, no deviations from normality are detected. We recommend reporting the results of the OLS bootstrap, together with those of ROBMED as a robustness check.

Finally, we strongly recommend against the automatic treatment of outliers as harmful data points. Outliers may help clarifying inconsistencies in emerging theories, provide chances to integrate theoretical predictions with real-life observations (Lieberson, 1992), and may reveal essential contingency factors and boundary conditions to theories (Gerring, 2007). We refer readers to the established body of knowledge on outlier identification and treatment (Aguinis et al., 2013; Lewin, 1992; Nair & Gibbert, 2016; Pearce, 2002) and theory building by using outliers (Gibbert et al., 2020; for empirical examples, also see Gittell, 2001; Hitt et al., 1998; Pisano, 1994).

Discussion and Conclusion

Existing methods for mediation analysis are sensitive to nonnormality. The proposed procedure ROBMED integrates the robust MM-estimator (Yohai, 1987) and the fast-and-robust bootstrap (Salibián-Barrera & Zamar, 2002) in a mediation setting to overcome the widespread problem of deviations from normality assumptions. Indeed, ROBMED is shown to be more reliable than established methods for testing mediation under a variety of deviations. The key technical property that gives ROBMED its edge is that it continuously downweights deviating data points. Not only does this result in robust estimates, but also in a stable procedure, as it does not require different approaches for different deviations from normality, or any decision to fully include or exclude a data point. Instead, the weights indicate the degree of deviation of an observation. Downweighting observations based on the residuals avoids transformations of the variables that are not supported

by theory, ensuring interpretability of the coefficients and correct alignment of the analysis with the hypotheses.

We stress that ROBMED should not be viewed as a tool that absolves researchers from verifying model assumptions or checking for outliers. Instead, researchers should view ROBMED as a tool that allows to reliably estimate the model while simultaneously detecting outliers and deviations from the model assumptions. It is crucial to follow up on any detected deviations. This last step must not be skipped, and findings should be transparently described. As such, ROBMED plays an integral part in ensuring robust findings in empirical research—and therefore reproducibility.

While ROBMED is designed to handle deviations from normality, one cannot expect ROBMED to work under all possible distributions. Deviations from normality are often characterized by skewness (a measure of asymmetry) and kurtosis (a measure of tail-heaviness). ROBMED can effectively compensate for various levels of skewness and kurtosis, but it is not suitable for extremely heavy tails (e.g., tails that should be modeled with extreme value distributions) or extreme skewness (e.g., if not even a log transformation would suffice to remove right-skewness). Furthermore, as mediation analysis is concerned with the central tendencies in a population, namely the coefficients in Equations (1)–(3), ROBMED is not suitable if extreme values are of primary interest. However, it does report deviating observations that, if appropriate, should be further analyzed with statistical tools from extreme value theory (e.g., de Haan & Ferreira, 2006). It is also not intended for dynamic time series analysis, where it is often of interest to study how extreme shocks travel through systems over time.

While we focus on the simple regression model in this paper, our robust approach can be used for any other mediation model that can be estimated via linear regressions, such as models including multiple mediators.[15] Furthermore, ROBMED can easily be extended to cover moderated mediation or mediated moderation models (see, e.g. Muller et al., 2005, for an overview). Granting all this, ROBMED currently focuses on mediation models with continuous dependent variables and mediators. Developing robust methods for mediation models with binary, nominal, ordinal or count variables (e.g., Huang et al., 2004; Preacher, 2015; VanderWeele & Vansteelandt, 2010) is a fruitful venue for further research.

On a final note, our R implementation for ROBMED and the R extension bundle for SPSS are freely available from https://cran.r-project.org/package=robmed and https://github.com/aalfons/ROBMED-RSPSS, respectively, making ROBMED easily accessible to empirical researchers. Users can run our code by following simple steps in the accompanying documentation and code examples. Given its technical strengths and practicality, we strongly encourage scholars to adopt ROBMED to test mediation.

Appendix

Technical details

MM-regression estimator. Consider the linear regression model

$$y = \boldsymbol{x}^T\boldsymbol{\beta} + e,$$

where x is a $(p + 1)$-dimensional random vector with the first component fixed at 1 to account for an intercept, and a normally distributed error term $e \sim N(0, \sigma^2)$. With observations $(y_i, \boldsymbol{x}_i^T)^T, i = 1, \ldots, n$, the MM-estimate (Yohai, 1987) of the regression coefficients is defined as

$$\hat{\boldsymbol{\beta}}_n = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}_n}\right),$$

where ρ is a loss function, $r_i(\boldsymbol{\beta}) = y_i - \sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{\beta},\ i = 1,\ \ldots,\ n$, denotes the residuals, and $\hat{\sigma}_n$ is the residual scale estimate from a highly robust but inefficient initial robust regression estimator. Typically, $\hat{\sigma}_n$ comes from an S-estimator of regression (Rousseeuw & Yohai, 1984; Salibián-Barrera & Yohai, 2006), and we use Tukey's bisquare loss function given by

$$\rho(x) = \begin{cases} \dfrac{x^6}{6c^4} - \dfrac{x^4}{2c^2} + \dfrac{x^2}{2}, & \text{if } |x| \leq c, \\ \dfrac{c^2}{6}, & \text{if } |x| > c. \end{cases}$$

This loss function behaves like a quadratic loss for small values of the scaled residuals, but flattens out and becomes constant for larger squared residuals, depending on the tuning constant c (see Figure 2, left). By taking the derivative with respect to $\boldsymbol{\beta}$ and equating to $\mathbf{0}$, the MM-estimator is the solution of

$$\sum_{i=1}^{n} \rho' \left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}_n} \right) \boldsymbol{x}_i = \mathbf{0}. \tag{4}$$

With

$$w_i = \frac{\rho' \left(r_i(\boldsymbol{\beta}) / \hat{\sigma}_n \right)}{r_i(\boldsymbol{\beta}) / \hat{\sigma}_n}, \qquad i = 1, \ldots, n, \tag{5}$$

the system of equations in (4) can be rewritten as

$$\sum_{i=1}^{n} w_i r_i(\boldsymbol{\beta}) \boldsymbol{x}_i = \mathbf{0}, \tag{6}$$

which is a weighted version of the normal equations. Hence, the MM-estimator can be viewed as a weighted least-squares estimator with data-driven weights. Due to their definition in (5), those weights lie between 0 and 1 and indicate to what extent an observation is deviating. That is, observations with small residuals will receive a weight close to 1, while observations with large enough residuals will receive weight 0. An illustration of this continuous downweighting based on Tukey's bisquare loss can be found in the right panel in Figure 2. Then the solution of the system of equations in (6) can be written as

$$\hat{\boldsymbol{\beta}}_n = \left(\sum_{i=1}^{n} w_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \sum_{i=1}^{n} w_i \boldsymbol{x}_i y_i.$$

Yohai (1987) proved that the MM-estimator $\hat{\boldsymbol{\beta}}_n$ inherits its robustness from the initial residual scale estimate $\hat{\sigma}_n$, and that the efficiency of the estimator can be tuned with the constant c of Tukey's bisquare loss function. In this way, high robustness and high efficiency can be achieved at the same time. Large values of c will increase efficiency, yet even though bias due to model deviations can increase to some extent, such a bias will remain bounded and cannot become arbitrarily large. Throughout the paper, we use $c = 3.443689$, which results in 85% efficiency compared to the OLS estimator (at the model with normally distributed errors).

Fast-and-robust bootstrap. While we present a brief technical overview of the fast and robust bootstrap, we refer to Salibián-Barrera & Zamar (2002) and Salibián-Barrera & Van Aelst (2008) for complete derivations. For a bootstrap sample $(y_i^*, \boldsymbol{x}_i^{*T})^T, i = 1, \ldots, n$, we can compute

$$\hat{\boldsymbol{\beta}}_n^* = \left(\sum_{i=1}^{n} w_i^* \boldsymbol{x}_i^* \boldsymbol{x}_i^{*T} \right)^{-1} \sum_{i=1}^{n} w_i^* \boldsymbol{x}_i^* y_i^*,$$

where $w_i^* = \rho'(r_i^*/\hat{\sigma}_n)/(r_i^*/\hat{\sigma}_n)$ and $r_i^* = y_i^* - \boldsymbol{x}_i^{*T}\hat{\boldsymbol{\beta}}_n$ for $i = 1, \ldots, n$. Here it is important to note that the MM-estimates $\hat{\boldsymbol{\beta}}_n$ and $\hat{\sigma}_n$ are only computed once on the original sample, they are not recalculated on each bootstrap sample. Hence, only a weighted least-squares fit on each bootstrap sample is necessary to obtain $\hat{\boldsymbol{\beta}}_n^*$. The robustness weights w_i^* are inherited from the original sample, hence oversampling of outliers in certain bootstrap samples is not an issue. However, there is a loss of variability by not recomputing the robustness weights on the bootstrap samples. To overcome this loss of variability, a linear correction of the coefficients is applied. With the correction matrix

$$\boldsymbol{K}_n = \left(\sum_{i=1}^{n} \rho''(r_i/\hat{\sigma}_n) \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \sum_{i=1}^{n} w_i \boldsymbol{x}_i \boldsymbol{x}_i^T$$

(which only needs to be computed once on the original sample), the fast-and-robust bootstrap replicates are given by

$$\hat{\boldsymbol{\beta}}_n^{R*} = \hat{\boldsymbol{\beta}}_n + \boldsymbol{K}_n(\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n).$$

As the MM-estimator $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ (Yohai, 1987), $\sqrt{n}(\hat{\boldsymbol{\beta}}_n^{R*} - \hat{\boldsymbol{\beta}}_n)$ has the same asymptotic distribution as $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ (Salibián-Barrera & Zamar, 2002; Salibián-Barrera & Van Aelst, 2008).

For the mediation model, we propose to use the MM-estimator of regression and the fast-and-robust bootstrap methodology to estimate Equations (1) and (2). In this way, we get consistent estimates of the respective regression coefficients, as well as the indirect effect *ab*, and we obtain accurate confidence intervals.

Software for ROBMED

To facilitate the use of our methodology, we provide scholars with freely available software. For the open-source statistical computing environment R (R Core Team, 2020), our add-on package **robmed** can be obtained from https://cran.r-project.org/package=robmed (including the user manual, examples and sample datasets). In addition to ROBMED, our R package also contains code for the OLS bootstrap test, Zu & Yuan's (2010) winsorized bootstrap test, Yuan & MacKinnon's (2014) bootstrap test based on median regression, and a bootstrap test based on Azzalini & Arellano-Valle's (2013) regression estimator with normal, skew-normal, *t*, or skew-*t* error distributions. An R extension bundle for SPSS (IBM Corp., 2019), which allows to use our R code for ROBMED from within SPSS, is available from https://github.com/aalfons/ROBMED-RSPSS (including instructions on how to install and how to use the extension bundle).

For reporting mediation results, we suggest to stay completely within the bootstrap framework. We advocate to use the means of the bootstrap replicates as point estimates for all effects (although our software reports the estimates obtained on the full sample as well). Consequently, to test significance of the effects other than the indirect effect, we propose normal-approximation bootstrap *z*-tests (using the mean and standard deviation over the bootstrap replicates). Nevertheless, our software can also report *t*-tests for the robust coefficient estimates obtained from the full sample. The significance of the indirect effect is assessed via a bias corrected and accelerated percentile-based confidence interval (Davison & Hinkley, 1997) to account for the asymmetry of its distribution.

In addition to the coefficient estimates and corresponding significance tests, the software reports potential outliers, as well as model summaries for Equations (1) and (2) that are the robust

counterparts of the usual OLS model summaries. Specifically, we provide a robust estimate of the residual standard error (Yohai, 1987), robust estimates of the R^2 and adjusted R^2 (Renaud & Victoria-Feser, 2010), as well as a robust F-test (Hampel, Ronchetti, Rousseeuw, & Stahel, 1986). Note that this robust F-test is an asymptotic test for $n \to \infty$. Moreover, our diagnostic plot (cf. Figure 4) allows to easily detect deviations from normality such as skewness and heavy tails.

All computations in this article have been performed using R version 3.6.3 and our package **robmed**. For reproducibility, our code for the illustrative empirical case and all our simulations is available at https://github.com/aalfons/ROBMED-Reproducibility.

Acknowledgments

We would like to thank Editor Paul Bliese and Associate Editor Lisa Schurer Lambert and the three anonymous reviewers for their guidance and comments in the revision process. We thank Marius van Dijke, Richard Haans, and Murat Tarakci for their feedback on an earlier version of this article. We further thank the audiences of the Research Methods Division at Academy of Management Annual Meeting (AOM 2019) and the International Conference on Robust Statistics (ICORS 2019). The data used in the article have not been published elsewhere before. An earlier version of this article is available as an Erasmus Research Institute of Management (ERIM) research report at https://repub.eur.nl/pub/109594/.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work was partially supported by a grant from the Dutch Research Council (NWO), research program Vidi (Project No. VI.Vidi.195.141).

ORCID iD

Nüfer Yasin Ateş https://orcid.org/0000-0003-4572-4101

Supplemental Material

Supplemental material (Appendices 1, 2, 3) for this article is available online at http://journals.sagepub.com/doi/suppl/10.1177/1094428121999096

Notes

1. Wood et al. (2008) reported 63% of studies using OLS regression. We conducted a review of articles that empirically test mediation published in *Academy of Management Journal*, *Strategic Management Journal*, *Journal of Applied Psychology*, *Organization Science*, and *Administrative Science Quarterly* in 2019, which confirmed that OLS is still the most frequently used method for testing mediation (i.e., in 42 out of 86 studies). The review can be accessed in Online Appendix 3 at http://orm.sagepub.com.
2. There are several definitions of the term *heavy-tailed distribution* in the statistical literature. Our use of the term implies that the mean of the distribution still exists, but that the tails of the probability density function are thicker than those of the normal distribution. We acknowledge that this colloquial sense of *heavy tails* is more inclusive and does not adhere to the strict technical definitions of the term, which are mostly used in extreme value theory.
3. This approach, called *product of coefficients*, is in many cases equivalent to the *difference in coefficients* approach that tests the significance of $\hat{c}' - \hat{c}$, where c' is the *total effect* of X on Y (i.e., not controlling for M). MacKinnon et al. (1995) show that $ab = c' - c$ for the mediation model according to Equations (1)–(3), and the same holds for the estimates of ordinary least squares estimation. This identity, however, does not hold for multilevel models, logistic and probit regression models, and survival models (MacKinnon

et al., 2007), which are beyond the scope of our study. We acknowledge that our proposed method can easily be adjusted to bootstrap $\hat{c}' - \hat{c}$ without major change.

4. Becker et al.'s (2019) review reports that log transformations compose 88% of the NLTs they observed in their random sample of 324 articles from 6 top management journals from 2012 to 2017.
5. More information on the aims of robust statistics is given in an essay by Morgenthaler (2007) and in a more technical overview by Avella-Medina and Ronchetti (2015). The interested reader can find detailed technical descriptions of commonly used robust statistical methods in Maronna et al. (2006).
6. Note that this definition implies that based on the observed data alone, the data analyst typically cannot be certain whether an observation is an outlier or how an outlier was generated. This requires further investigation of the data collection procedure that goes beyond a quantitative analysis of the observations.
7. Error outliers and interesting outliers are of a conceptual nature and model-agnostic, while influential outliers are model-specific. In their original definitions, Aguinis et al. (2013) explicitly exclude error outliers and interesting outliers from being influential outliers. We take a divergent standpoint that the former two types can be influential outliers as well, as underlined by our two examples. Recognizing that error outliers and interesting outliers can be highly influential for traditional statistical methods is important for understanding that such methods are not suitable for outlier detection (see also chapter 4.3 of Maronna et al., 2006, or Rousseeuw & Hubert, 2018, for a recent overview).
8. In our earlier mentioned review of recently published articles that empirically test mediation, 72 of 86 articles used some form of a bootstrap test; see the Online Appendix 3 at http://orm.sagepub.com.
9. Other researchers on team processes have published findings based on data from this game as well (e.g., Boies et al., 2010; Mathieu & Rapp, 2009).
10. We did not include Baron and Kenny's (1986) causal steps approach because despite being conceptually appealing, it has been severely criticized for its shortcomings including increased Type I error (Holmbeck, 2002), and low statistical power (MacKinnon et al., 2002).
11. For the regression model $M = i_1 + aX + e_1$, it holds that $a = \sigma_{MX}/\sigma_X^2$ and $i_1 = \mu_M - a\mu_X$, where μ_M and μ_X denote the means of M and X, σ_{MX} is the covariance of M and X, and σ_X^2 is the variance of X. The same relationship holds for the OLS estimates $\hat{i}_1$ and $\hat{a}$, the sample covariance $\hat{\sigma}_{MX}$ and the sample variance $\hat{\sigma}_X$.
12. To obtain the robust tolerance ellipse, the weighted sample covariance matrix with robustness weights from the MM-estimator of regression is corrected for consistency under the model assumptions. This correction is necessary because the MM-estimator also partly downweights observations in a mediation model with normally distributed error terms. Consequently, the size of the tolerance ellipse would be underestimated without the correction.
13. Note that evaluating the methods by the rejection rate from the two-sided tests alone does not provide a meaningful comparison in these simulation settings, as deviations from normality can push the estimated indirect effect from a positive one toward a negative one. This incorrectly estimated negative indirect effect can be large enough in magnitude to reject the null hypothesis of a two-sided test. However, while the sign of the estimated effect is negative, the sign of the true effect is positive, which would result in an incorrect interpretation of the indirect effect. By taking into account the sign of the estimated indirect effect as well, we obtain a better measure of realized power of the tests.
14. We also ran simulations with various extensions of the design of Zu and Yuan (2010), and with the design of Yuan and MacKinnon (2014). Online Appendix 2 can be accessed at http://orm.sagepub.com.
15. In fact, models including multiple mediators and/or control variables are already implemented in our software. Additional models will be added in the future.

References

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*(2), 270-301.

Aguinis, H., Hill, N. S., & Bailey, J. R. (2019). Best practices in data collection and preparation: Recommendations for reviewers, editors, and authors. *Organizational Research Methods*. doi:10.1177/1094428119836485

Asparouhov, T., & Muthén, B. (2016). Structural equation models and mixture models with continuous non-normal skewed distributions. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 1-19.

Avella-Medina, M., & Ronchetti, E. (2015). Robust statistics: A selective overview and new directions. *Wiley Interdisciplinary Reviews: Computational Statistics*, *7*(6), 372-393.

Azzalini, A., & Arellano-Valle, R. B. (2013). Maximum penalized likelihood estimation for skew-normal and skew-t distributions. *Journal of Statistical Planning and Inference*, *143*(2), 419-433.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173-1182.

Becker, T. E., Robertson, M. M., & Vandenberg, R. J. (2019). Nonlinear transformations in organizational research: Possible problems and potential solutions. *Organizational Research Methods*, *22*(4), 831-866.

Bettis, R. A. (2012). The search for asterisks: Compromised statistical tests and flawed theories. *Strategic Management Journal*, *33*(1), 108-113.

Boies, K., Lvina, E., & Martens, M. L. (2010). Shared leadership and team performance in a business strategy simulation. *Journal of Personnel Psychology*, *9*(4), 195-202.

Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, *20*, 115-140.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, *26*(2), 211-252.

Chen, S., & Bien, J. (2020). Valid inference corrected for outlier removal. *Journal of Computational and Graphical Statistics*, *29*, 323-334.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Erlbaum.

Cortina, J. M. (2002). Big things have small beginnings: An assortment of "minor" methodological misunderstandings. *Journal of Management*, *28*(3), 339-362.

Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications*, *19*(4), 497-515.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.

de Haan, L., & Ferreira, A. F. (2006). *Extreme value theory: An introduction*. Springer.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*(1), 1-26.

Gerring, J. (2007). *Case study research: Principles and practices*. Cambridge University Press.

Gibbert, M., Nair, L. B., Weiss, M., & Hoegl, M. (2020). Using outliers for theory building. *Organizational Research Methods*, *24*, 172-181.

Gittell, J. H. (2001). Supervisory span, relational coordination, and flight departure performance: A reassessment of postbureaucracy theory. *Organization Science*, *12*(4), 468-483.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. John Wiley.

Hawkins, D. M. (1980). *Identification of outliers*. Chapman & Hall.

Hawkins, D. M., & Weisberg, S. (2017). Combining the Box-Cox power and generalized log transformations to accommodate nonpositive responses in linear and mixed-effects linear models. *South African Statistical Journal*, *51*(2), 317-328.

Hitt, M., Harrison, J., Ireland, R. D., & Best, A. (1998). Attributes of successful and unsuccessful acquisitions of US firms. *British Journal of Management*, *9*(2), 91-114.

Holmbeck, G. N. (2002). Post-hoc probing of significant moderational and mediational effects in studies of pediatric populations. *Journal of Pediatric Psychology*, *27*(1), 87-96.

Huang, B., Sivaganesan, S., Succop, P., & Goodman, E. (2004). Statistical assessment of mediational effects for logistic mediational models. *Statistics in Medicine*, *23*(17), 2713-2728.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Sage.

IBM Corp. (2019). *IBM SPSS statistics, Version 26.0*. IBM Corp.

Kenny, D. A. (2008). Reflections on mediation. *Organizational Research Methods*, *11*(2), 353-358.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

Lewin, A. Y. (1992). On learning from outliers. In F. Y. Phillips & J. J. Rousseau (Eds.), *Systems and management science by extremal methods* (pp. 11-17). Kluwer.

Lieberson, S. (1992). Einstein, Renoir, and Greeley: Some thoughts about evidence in sociology: 1991 presidential address. *American Sociological Review*, *57*(1), 1-15.

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*(1), 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*(1), 99-128.

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, *30*(1), 41-62.

Mannor, M. J., Wowak, A. J., Bartkus, V. O., & Gomez-Mejia, L. R. (2016). Heavy lies the crown? How job anxiety affects top executive decision making in gain and loss contexts. *Strategic Management Journal*, *37*(9), 1968-1989.

Maronna, R. M., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. John Wiley.

Mathieu, J. E., & Rapp, T. L. (2009). Laying the foundation for successful team performance trajectories: The roles of team charters and performance strategies. *Journal of Applied Psychology*, *94*(1), 90-103.

Morgenthaler, S. (2007). A survey of robust statistics. *Statistical Methods & Applications*, *15*(3), 271-293.

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, *89*(6), 852-864.

Nair, L. B., & Gibbert, M. (2016). Analyzing inconsistent cases in management fsQCA studies: A methodological manifesto. *Journal of Business Research*, *69*(4), 1464-1470.

Pearce, L. D. (2002). Integrating survey and ethnographic methods for systematic anomalous case analysis. *Sociological Methodology*, *32*(1), 103-132.

Penney, L. M., & Spector, P. E. (2005). Job stress, incivility, and counterproductive work behavior (CWB): The moderating role of negative affectivity. *Journal of Organizational Behavior*, *26*(7), 777-796.

Pisano, G. P. (1994). Knowledge, integration, and the locus of learning: An empirical analysis of process development. *Strategic Management Journal*, *15*(S1), 85-100.

Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, *66*, 825-852.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Bahavior Research Methods, Instruments, & Computers*, *36*(4), 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*(3), 879-891.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Raymakers, J., & Rousseeuw, P. J. (2020). Transforming variables to central normality. *Machine Learning*. https://arxiv.org/abs/2005.07946

Renaud, O., & Victoria-Feser, M.-P. (2010). A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference*, *140*(7), 1852-1862.

Rousseeuw, P. J., & Hubert, M. (2018). Anomaly detection by robust statistics. *WIREs Data Mining and Knowledge Discovery*, *8*(2), e1236.

Rousseeuw, P. J., & Yohai, V. J. (1984). Robust regression by means of S-estimators. In J. Franke, W. Härdle, & D. Martin (Eds.), *Lecture notes in statistics: Robust and nonlinear time series analysis* (Vol. 26, pp. 256-272). Springer.

Salibián-Barrera, M., & Van Aelst, S. (2008). Robust model selection using fast and robust bootstrap. *Computational Statistics & Data Analysis*, *52*(12), 5121-5135.

Salibián-Barrera, M., & Yohai, V. J. (2006). A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics*, *15*(2), 414-427.

Salibián-Barrera, M., & Zamar, R. H. (2002). Bootstrapping robust estimates of regression. *Annals of Statistics*, *30*(2), 556-582.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461-464.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, *7*(4), 422-445.

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

Tarakci, M., Ateş, N. Y., Floyd, S. W., Ahn, Y., & Wooldridge, B. (2018). Performance feedback and middle managers' divergent strategic behavior: The roles of social comparisons and organizational identification. *Strategic Management Journal*, *39*(4), 1139-11162.

Tatarynowicz, A., Stych, M., & Gulati, R. (2016). Environmental demands and the emergence of social structure: Technological dynamism and interorganizational network forms. *Administrative Science Quarterly*, *61*(1), 51-86.

VanderWeele, T. J., & Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, *172*(12), 1339-1348.

Wood, R. E., Goodman, J. S., Beckmann, N., & Cook, A. (2008). Mediation testing in management research. *Organizational Research Methods*, *11*(2), 270-295.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics*, *15*(20), 642-656.

Yuan, Y., & MacKinnon, D. P. (2014). Robust mediation analysis based on median regression. *Psychological Methods*, *19*(1), 1-20.

Zhelyazkov, P. I., & Gulati, R. (2016). After the break-up: The relational and reputational consequences of withdrawals from venture capital syndicates. *Academy of Management Journal*, *59*(1), 277-301.

Zu, J., & Yuan, K.-H. (2010). Local influence and robust procedures for mediation analysis. *Multivariate Behavioral Research*, *45*(1), 1-44.

Author Biographies

Andreas Alfons is an associate professor of Statistics at Erasmus School of Economics, Erasmus University Rotterdam. He obtained his doctorate from Vienna University of Technology. His research focuses on the development of statistical methods that are robust against outliers and deviations from model assumptions. He is interested in computational statistics, machine learning, statistical software, as well as empirical applications in the behavioral sciences. His work has been published in the *Journal of the American Statistical Association*, the *Journal of Statistical Software*, and the *Annals of Applied Statistics*.

Nüfer Yasin Ateş is an assistant professor of Strategy and Organization at Sabancı Business School, Sabancı University. He received his PhD from the Erasmus Research Institute of Management, Erasmus University Rotterdam. His research interests include managerial cognition, corporate entrepreneurship, and strategy process, with a special focus on top and middle managers. He is also interested in advancing quantitative research methods. Nufer's research has appeared in the *Strategic Management Journal*, the *Journal of Management, Strategic Organization, Journal of Business Ethics*.

Patrick J. F. Groenen is a professor of Statistics at the Erasmus School of Economics, Erasmus University Rotterdam. He also serves as the Dean of the School. Professor Groenen's work focuses on data science techniques and their numerical algorithms. He is the co-author of several textbooks on multidimensional scaling published by Springer and has published articles in the top peer-reviewed journals including, among others, the *Journal of Machine Learning Research*, the *Journal of Marketing Research, Psychological Methods, Psychometrika*, the *Journal of Classification, Computational Statistics and Data Analysis*, the *British Journal of Mathematical and Statistical Psychology*, and the *Journal of Empirical Finance*.