A Robust Bootstrap Test for Mediation Analysis

Andreas Alfons

Econometric Institute, Erasmus University Rotterdam

Nüfer Yasin Ateş

Tilburg School of Economics and Management, Tilburg University

Faculty of Business Administration, Bilkent University

Patrick J.F. Groenen

Econometric Institute, Erasmus University Rotterdam

Corresponding author: Nüfer Yasin Ateş, Faculty of Business Administration, Bilkent Üniversitesi, 06800, Ankara, Turkey

Email: n.y.ates@bilkent.edu.tr

Abstract

Mediation analysis is central to theory building and testing in organizations research. Management 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 or heavy tails of the observed distribution. This sensitivity seriously threatens the empirical testing of theory about mediation mechanisms, as many empirical studies lack reporting of outlier treatments and checks on model assumptions. To overcome this threat, we develop a fast and robust mediation method that yields reliable results even when the data deviate from normality assumptions. 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. We illustrate the mechanics of our proposed method in three empirical cases and provide freely available software in R and SPSS to enhance its accessibility and adoption by researchers and practitioners.

Keywords: Mediation analysis, robust statistics, linear regression, bootstrap.

A Robust Bootstrap Test for Mediation Analysis

INTRODUCTION

Management 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; MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; Preacher & Hayes, 2008). For instance, Tost, Gino, & Larrick (2013) tested two mediation hypotheses in their study. They showed that a formal leader's power (X_1) reduces team communication (Y_1) through verbal dominance in team discussions (M_1), and this verbal dominance (X_2) leads to lower team performance (Y_2) due to the diminished communication within the team (M_2). Such mediation analyses are very popular and widely applied in management research (Wood, Goodman, Beckmann, & Cook, 2008).

Several methods have been proposed for testing mediation (see MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002, for a review) where the most widely adopted technique is regression analysis – in 63% of the studies (Wood, Goodman, Beckmann, & Cook, 2008). The statistical performance of these methods has long been tested via simulation studies (e.g., MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; MacKinnon, Lockwood, & Williams, 2004). The tests considered in those studies are based on normal-theory maximum likelihood estimators (MLE), which are the most efficient estimators under the assumption of normally distributed errors. However, data in management research frequently show deviations from normality such as *outliers* (i.e., data points that deviate markedly from others; Aguinis, Gottfredson, & Joo, 2013) or *heavy tails* of the observed distribution (i.e., values further from the mean occurring much more often than under the assumed normal distribution). These deviations 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, Cohen, West, & Aiken, 2003; Hunter & Schmidt, 2004). Other deviations from normality such as 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 more severe effect on mediation analysis as compared to multiple regressions, because the mediated effect itself is a multiplication of two regression coefficients (Zu & Yuan, 2010).

Despite the importance of outliers and deviations from normality in general, no clear guidelines have so far been developed for mediation methods for dealing with these issues properly. Unsurprisingly, a study on the treatment of outliers in organizations research found the common practices to be vague, non-transparent and even inconsistent in outlier definition, identification and treatment (Aguinis, Gottfredson, & Joo, 2013). To overcome these limitations, we introduce our procedure ROBMED for robust mediation analysis that yields reliable results even if there are outliers or heavy tails.

We build upon the state-of-the-art bootstrap test for mediation (Preacher & Hayes, 2004; Preacher & Hayes, 2008) and extend it by the fast and robust bootstrap methodology (Salibián-Barrera & Zamar, 2002; Salibián-Barrera & Van Aelst, 2008), which is well established within the literature on robust statistics. We compare ROBMED to available mediation testing methods through simulation studies and conclude that ROBMED is superior to others in terms of estimating the effect size and reliably assessing its significance. We also illustrate the use of ROBMED and compare it with the state-of-the-art bootstrap test on real data that show deviations from normality. Furthermore, we provide researchers and practitioners with freely available software for ROBMED.

MEDIATION ANALYSIS

Researchers often seek to develop a deeper understanding of the process that produces the effect of an independent variable (X) on a dependent variable (Y). This endeavor to comprehend the mechanism of how X exerts its influence on Y is frequently concerned with the identification of *mediators*. Baron & Kenny (1986) define a mediator M as a variable that partially accounts for the relation between X and Y. Figure 1 illustrates a simple mediation model. This simple mediation model can be formulized by the following equations:

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

$$Y = i_2 + c'X + e_2, \tag{2}$$

$$Y = i_3 + bM + cX + e_3, \tag{3}$$

where i_1, i_2 and i_3 are three intercepts, a, b, c, and c' are weights, and e_1, e_2 and e_3 denote random error terms. Mediation is said to occur if the product of the $X \rightarrow M$ path's coefficient and the $M \rightarrow Y$ path's coefficient (i.e., the *indirect effect* ab) is significant.[1] Estimating the coefficients in the mediation model is typically done via normal-theory maximum likelihood procedures, with the most commonly one used being ordinary least squares (OLS) regression (Wood, Goodman, Beckmann, & Cook, 2008).

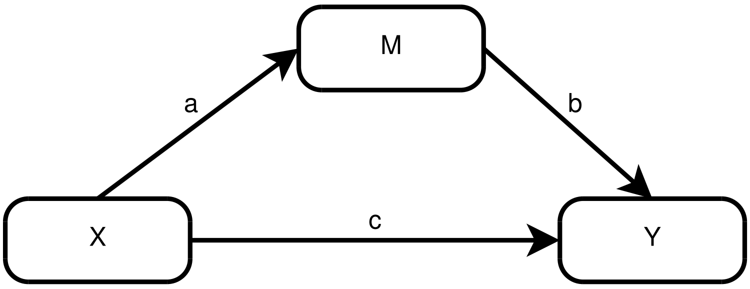

Figure 1. Illustration of a simple mediation model.

[1] This approach, called *product of coefficients*, is in many cases equivalent to the *difference in coefficients* approach that tests the significance of $c' - c$, where c' is the *total effect* of X on Y (i.e., not controlling for M). MacKinnon, Warsi, & Dwyer (1995) show that $ab = c' - c$ for ordinary least squares estimation. This equation, however, does not hold for multi-level models, logistic and probit regression, and survival models (MacKinnon, Fairchild, & Fritz, 2007), which are beyond the scope of our study. We acknowledge that our proposed method can easily be adjusted to bootstrap $c' - c$ without major change.

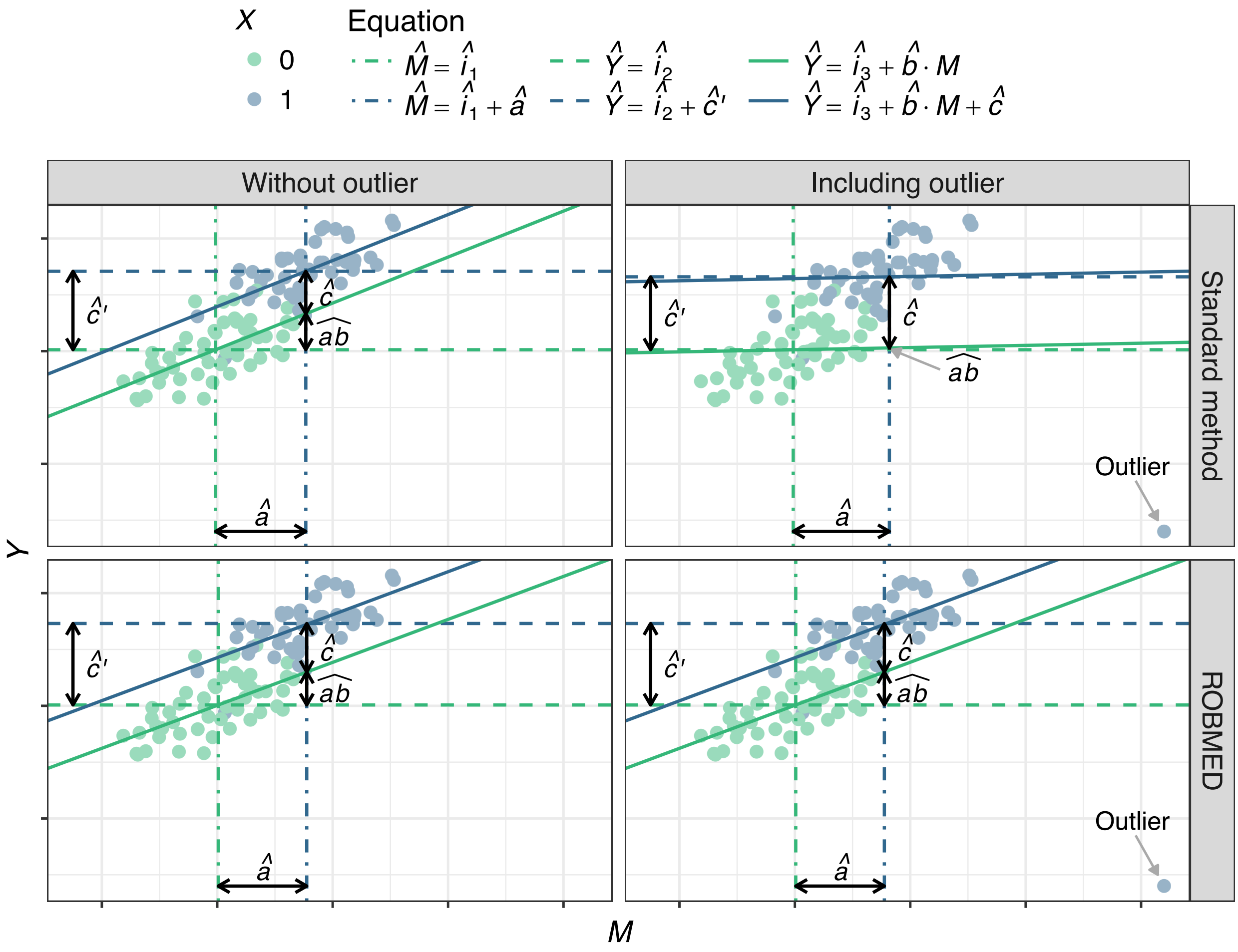

Figure 2. Illustration of the effect of a single outlier on mediation analysis for the case of a dichotomous independent variable X. Green lines correspond to fitted regression lines for $X = 0$ (green points), while blue lines correspond to fitted regression lines for $X = 1$ (blue points).

The top row in Figure 2 illustrates the potential threat of outliers to mediation testing based on normal-theory maximum likelihood estimation (i.e., OLS regression). It consists of two plots with the mediator M on the horizontal axis and the dependent variable Y on the vertical axis. The independent variable X is assumed to be dichotomous for a simpler visual representation, as each regression model in Equations (1), (2) and (3) then corresponds to two fitted lines that are parallel. The two plots on the left contain 100 simulated observations that follow the model assumptions, whereas the plots in the right column use the same data except for one single outlier being added. The distance between the horizontal dashed regression lines represents the total effect c' of X on Y, and the distance between the vertical dash-dotted regression lines represents the effect 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 a units (due to a change in X from 0 to 1) leads to an indirect change in Y of ab units (i.e., the indirect effect).[2] With the introduction of the outlier (top right plot of Figure 2), the indirect effect $\widehat{ab}$ almost disappears for OLS estimation, as the solid regression lines corresponding to Equation (3) are pulled almost flat by the outlier. Also note how those fitted lines no longer represent the main part of the data.

For testing the significance of the indirect effect, numerous methods have been proposed in the literature (see MacKinnon, Lockwood, & Williams, 2004; Wood, Goodman, Beckmann, & Cook, 2008, for reviews). A comprehensive review of these methods is beyond the scope of this study, yet we note that computer-intensive resampling methods (e.g., bootstrapping) are found to be superior to other methods for at least two reasons. First, computer-intensive resampling methods provide generic ways to construct confidence intervals for the indirect effect and test its significance. Therefore, they are applicable in a wider variety of situations than other mediation methods, especially when the analytical formulas for standard errors are not available. Second, they make fewer assumptions than other tests. This property makes them more reliable than traditional mediation analysis, as the latter often make incorrect assumptions such as a normal distribution of the indirect effect (MacKinnon, Fairchild, & Fritz, 2007; Preacher & Hayes, 2004; Preacher & Hayes, 2008).

Despite their superiority to traditional inference methods, computer-intensive resampling methods are also sensitive to outliers and other problems such as heavy tails. Outliers may be oversampled, and heavy tails may become even heavier in some of the subsamples, which of course decreases the reliability of resampling-based significance tests

[2] Note that the plots in Figure 2 also illustrate the product of coefficients ab is equal to the difference in coefficients $c' - c$ (MacKinnon, Fairchild, & Fritz, 2007).

even further. Thus, if the data exhibit deviations from the usual normality assumptions, 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. By applying state-of-art knowledge on robust statistics, we can diminish the sensitivity of mediation analysis to deviations from normality assumptions.

ROBUST STATISTICS

Statistical methods are traditionally designed to be as efficient as possible under a certain 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*, on the other hand, aims to develop statistical methods that are less affected by model deviations and show good behavior in many situations. An important concept in robust statistics is that of *outliers*. An outlier is an "observation which deviates so much from other observations as to arouse suspicions that is was generated by a different mechanism" (Hawkins, 1980). While much of the literature on robust statistics is focused on outliers, robust methods are also an effective tool against other model deviations such as heavy tails.

To illustrate the need for robust methods, consider the mean and the median, two measures of central tendency. The mean is efficient under normally distributed data but is easily distorted when some observations lie outside the main bulk of the data. In extreme situations, even a single outlying observation with a large value can drive the mean to take completely divergent values that do not represent the population. The median, on the other hand, does not make any assumptions about the distribution and focuses only on the central part of the data. Even if there are heavy tails or several distant outliers, its value does not change severely. Hence the median is a more robust measure of central tendency than the mean.

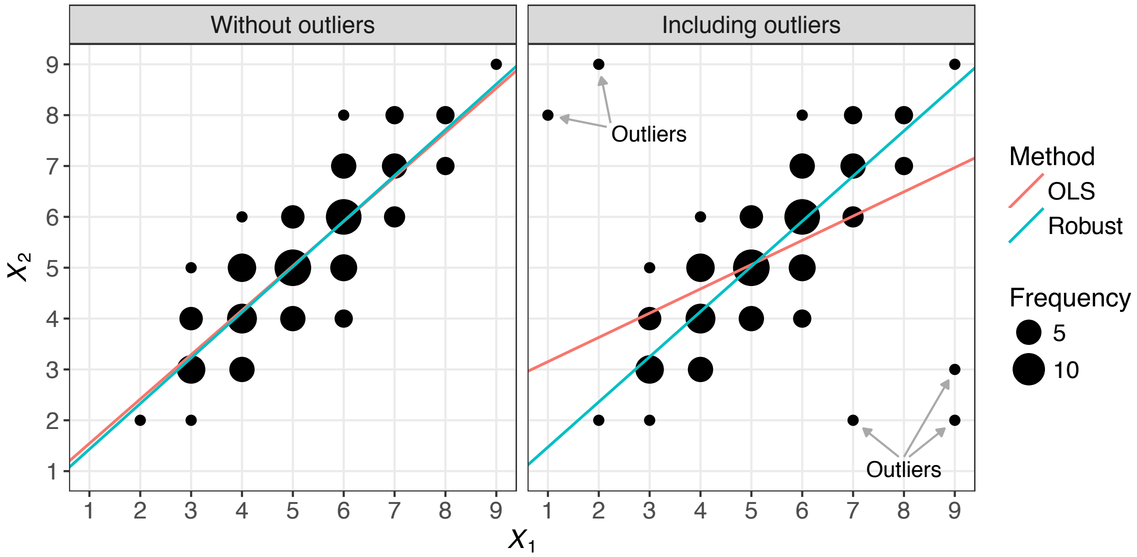

Figure 3. Illustration of the effect of correlation outliers on regression estimates in data with a limited range (here simulated data on a 9-point Likert scale). As these data are on a discrete grid, the size of the points reflects the number of observations with those values.

When analyzing multiple variables, it is important to note that outliers do not have to be extreme in any variable. Consider the illustrative example in Figure 3 with 100 simulated observations on two variables X_1 and X_2 on a 9-point Likert scale. Due to the discrete nature of the data, the size of the points reflects the number of observations with the corresponding values. The plot on the left does not contain any outliers. In this case, the regression lines obtained via OLS and a robust estimator (Yohai, 1987) are almost identical. In the plot on the right, a small number of outliers are added. While none of those outliers are extreme in the direction of either axis, they clearly deviate from the correlation structure of the main data cloud and tilt the OLS regression line such that it no longer represents the trend in the data. The robust estimator, on the other hand, is unaffected by the outliers. Robust methods are therefore necessary even if the data have a limited range such as responses on Likert items. Note that while it is easy to identify clear correlation outliers in a plot when there are only two variables, this is no longer possible when many variables are involved in the analysis.

Outliers are common and unavoidable in empirical data gathering. It does not come as a surprise that problems can arise when researchers go to the field to collect empirical data (e.g., experiments, surveys, interviews). For instance, consider a questionnaire consisting of Likert scale items. There may be several reasons for outlying cases. Apart from the possible data entry errors, respondents not taking the survey seriously may give inconsistent responses to survey items. Survey fatigue may cause the same problem in the later items of long surveys due to loss of attention. Some participants may inadvertently reverse the scales of the Likert items due to differences in cultural anchors (e.g., 1 is the highest grade in Germany and the lowest grade in the Netherlands). Even though careful survey designs can evade some of these problems, outlying cases may still arise even when participants answer correctly: certain individuals may simply behave or think differently from the majority of respondents, resulting in different response patterns for those individuals. There is of course nothing wrong with individuals who think differently, and they could be the most interesting observations in the data set leading to new insights about the phenomenon under investigation. Yet they should not influence statistical analysis in such a way that the results no longer reflect any part of the data (cf. the OLS regression lines in Figures 2 and 3 that represent neither the main cloud of data points nor the deviating observations).

Standard statistical methods assume that *all* data points follow the model and therefore cannot handle deviations such as outliers or heavier tails (compared to the assumed distribution). Robust methods assume that the *majority of the data* follow some model, but allow parts of the data to deviate from this model. In other words, they trade in some efficiency for being more widely applicable. This loss of efficiency is often small and should be seen as an insurance premium against failure under deviations from the model assumptions.

Traditional techniques for outlier treatment mainly consist of two-step procedures: first identify outliers and remove them from the data, then apply standard methods to the cleaned data set. While such ad-hoc robust techniques are still frequently used in empirical research (see Aguinis, Gottfredson, & Joo, 2013, for a review), this approach has its drawbacks that go beyond requiring an extra step in the analysis. When standard methods are applied to the cleaned data, the resulting standard errors do not include the uncertainty from the data-cleaning step, such that the standard errors of the two-step approach are underestimated. For instance, Chen & Bien (2017) show that OLS regression after outlier removal results in confidence intervals that are much too small as they do not possess the nominal coverage. Consequently, the *p*-values from significance tests are too small and could incorrectly suggest significant results. Another disadvantage of completely removing outliers is a certain loss of stability. In borderline situations, or if the data are showing a somewhat longer tail rather than containing clear outliers, the decision to fully include or fully exclude observations could have a considerable influence on the results of the analysis. Hence deletion of outliers must be approached with caution from a standpoint of research integrity. If the decision of whether or not to include an observation is taken by the researcher, it can be abused as a dangerous post-hoc practice to increase the chances of finding what the researcher wants to find (Cortina, 2002), which threatens the base of empirically tested theory (Bettis, 2012).

Modern robust methods typically aim for a *continuous downweighting* of deviating observations with weights between 0 and 1 that measure the degree of outlyingness. In addition, robust methods simultaneously downweight deviating observations while estimating the model. To illustrate the benefit of continuous downweighting during estimation, consider the following simple example: Suppose that we have a sample of the height of five men. The first four observations are 174cm, 192cm, 184cm, and 179cm. The fifth observation is former professional basketball player and Hall-of-Famer Shaquille O’Neal with a height of 216cm. The average height of the five men is 189cm. While Shaquille O’Neal is part of the population and his height therefore carries some relevant information, it is not realistic to assume that 20% of the population are of similar height. Hence, such a large value has a disproportionately large influence and yields an unreliable estimate. It needs to be downweighted to more accurately reflect the expected proportion of men of such height.

The downweighting strategy solves the issues discussed above: there is no separate extra step in the analysis, standard errors are estimated accurately, and continuous downweighting ensures stability of the results. Moreover, the decision if and by how much a data point deviates is taken objectively by an algorithm, which improves research reproducibility compared to subjective outlier deletion by the researcher. In addition, such continuous downweighting is not only effective for outliers; it also allows for a gradual downweighting of heavy tails. Finally, if there are no observations deviating from the model, all observations receive a weight close to 1 such that the robust method yields approximately the same results as the corresponding standard method (cf. Figures 2 and 3).

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 & Ronchetti (2015). The interested reader can find detailed technical descriptions of commonly used robust statistical methods in Maronna, Martin, and Yohai (2006).

Robust Statistics and Mediation Analysis

Given the common presence of outliers and the sensitivity of mediation results to outliers and deviations from model assumptions, Zu & Yuan (2010) took a first (and so far the only) step towards a robust version of mediation analysis. They propose methods based on cleaning the data beforehand via local influence methods or Huberization, which are rather outdated approaches towards robustness. First, their local influence procedure involves examining a plot of the local influence measure to decide on the number of outliers to exclude. This approach is far from optimal, as it requires manual interaction and is a highly subjective decision by the researcher. Second, data cleaning via Huberization, although being a more objective procedure, is neither as robust nor as efficient as modern robust regression methods. Furthermore, Zu & Yuan simply plug in the cleaned data into the standard bootstrap procedure, which does not include the uncertainty from the data cleaning process and may therefore underestimate the true confidence intervals. Although they attend to an important problem, their proposed methods were not only far from being optimal, but also not easy to implement and they do not provide their code. As a result, their method is not widely adopted by empirical researchers as it is cited only 14 times (with only 2 cites from empirical articles and 12 from other methodological articles).[3]

We note two issues to overcome here: (i) the methodological shortcomings of the procedure of Zu & Yuan (2010) concerning robustness, and (ii) the inaccessibility of their mediation method to the wider audience of empirical researchers due to its technical complexity. To resolve the first issue, we propose our new method ROBMED, drawing on more advanced techniques from robust statistics. To avoid the second issue, we provide freely available software for ROBMED to make it easy to use for researchers and practitioners.

ROBMED: ROBUST MEDIATION ANALYSIS

We build our method on the linear regression model, since it is the most widely used mediation technique in empirical studies is regression analysis (Wood, Goodman, Beckmann, & Cook, 2008). Moreover, for testing the indirect effect in linear regression models, the bootstrap test of Preacher & Hayes (2004, 2008) is the state-of-the-art method, as the distribution of the indirect effect is in general asymmetric. Hence, we further build our

[3] This is compared to 6,072 citations of Preacher & Hayes (2004) and 10,029 citations of Preacher & Hayes (2008), who provide SPSS and SAS implementations of their procedure. Citation numbers are taken from the Web of Science, accessed on May 31, 2018.

method on bootstrapping the indirect effect. We achieve a robust test for mediation through two essential building blocks.

First, we replace the OLS estimator for regression with the robust MM-regression estimator (Yohai, 1987; Salibián-Barrera & Yohai, 2006). Instead of the quadratic loss function of the OLS estimator, this estimator uses a loss function that is quadratic for small residuals, but smoothly levels off for larger residuals (see Figure 4, left). This ensures that the coefficient estimates are determined by the central part of the data and that the influence of outliers or heavy tails is limited. It turns out that this 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 outlyingness. An illustration of this continuous weight function is given in Figure 4 (right). Technical details on how the weight function is derived from the loss function can be found in Maronna, Martin, & Yohai (2006).

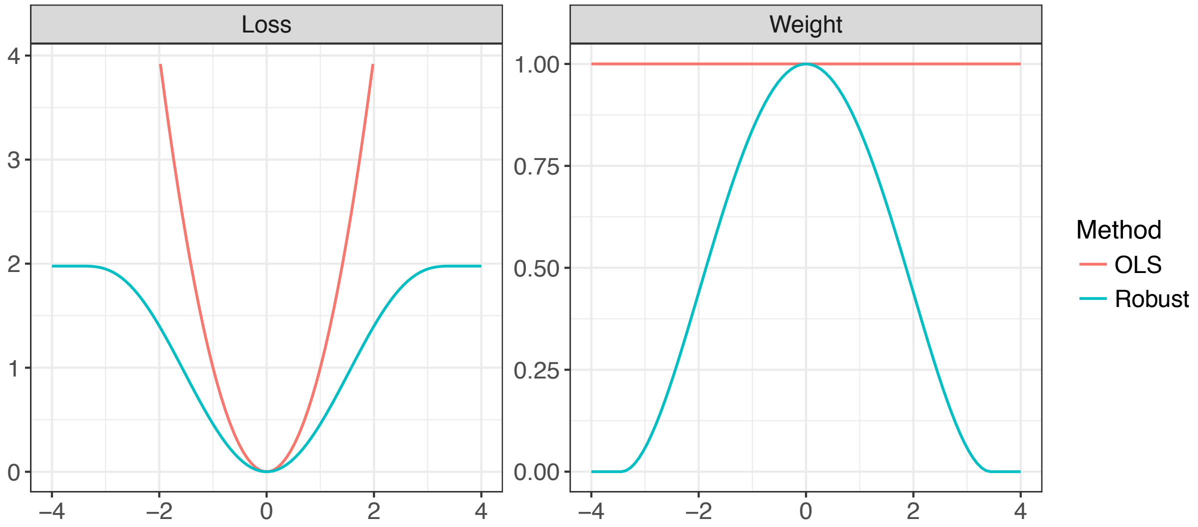

Figure 4. Loss function (left) and assigned weights (right) for OLS regression and the robust MM-regression estimator.

Second, we replace the standard bootstrap by the fast and robust bootstrap of Salibián-Barrera & Zamar (2002) and Salibián-Barrera & Van Aelst (2008). There are two issues with the standard bootstrap for our purposes. The first issue is that it is not robust. It draws so-called 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 of the bootstrap samples, or that heavy tails become even heavier. If some bootstrap samples exhibit more severe deviations from the model assumptions than the robust method can handle, bootstrap confidence intervals can become unreliable. The second issue is that robust methods typically come with increased computational complexity. While this is no longer an issue in most applications due to modern computing power, there can be a noticeable increase in computing time compared to standard methods, in particular when combined with computer intensive procedures such as the bootstrap.

To solve the two issues, Salibián-Barrera & 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 trick for 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. For full technical derivations of the fast and robust bootstrap, we refer to Salibián-Barrera & Zamar (2002) and Salibián-Barrera & Van Aelst (2008).

In short, combining the robust MM-regression estimator with the fast and robust bootstrap methodology allows us to construct a test for mediation analysis that follows the same principles as the state-of-the-art test of Preacher & Hayes (2004, 2008). However, our proposed test is more reliable than Preacher and Hayes' test under deviations from the model assumptions such as outliers and heavy tails.

Coming back to our earlier example in Figure 2 that illustrates the threat of outliers to mediation testing based on OLS regression, we re-ran the mediation analyses with ROBMED and depict the very same plots in the bottom row of Figure 2. Without any outliers (left column of Figure 2), the estimated effects are nearly identical for OLS estimation and ROBMED. When the outlier is introduced (right column of Figure 2), the fitted regression lines remain unchanged and all effects are still accurately estimated for ROBMED, while the indirect effect is substantially misrepresented for OLS estimation.

Software and further details

To facilitate the use of our methodology, we provide software that is freely available. For the open-source statistical computing environment R (R Core Team, 2018), our add-on package **robmed** (Alfons, 2018) 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 bootstrap test of Preacher & Hayes (2004, 2008) and the Huberized bootstrap test of Zu & Yuan (2010). A macro of ROBMED for SPSS (IBM Corp., 2017) is under development as well.[4]

Even though we cannot emphasize enough the importance of the contribution of Preacher & Hayes (2004, 2008) regarding testing the indirect effect, the output of their SPSS macro INDIRECT does have some inconsistencies. While they advocate to use the mean of the bootstrap replicates as point estimate for the indirect effect, for the remaining effects they only report the point estimates obtained on the full sample. We assume that they leave the bootstrap framework for those effects in order to use the standard *t*-tests based on statistical

[4] The SPSS macro will be available upon publication of this manuscript. Its development can be followed on https://github.com/aalfons/ROBMED-SPSS.

theory. However, a considerable drawback is that the advocated point estimate for the indirect effect no longer equals the product of the reported a and b coefficients.

We suggest to stay completely within the bootstrap framework. Therefore, 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 (i.e., to assume a normal distribution for those effects using the mean and standard deviation over the bootstrap replicates).[5] The significance of the indirect effect will still be 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, we report model summaries for Equation (3) that are the robust counterparts of the usual model summaries reported by Preacher & Hayes' (2004, 2008) INDIRECT macro. 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 $n \rightarrow \infty$. All computations in this article have been performed using R version 3.4.4 and our package **robmed** version 0.2.0.

SIMULATION STUDY

For a thorough evaluation of the performance of ROBMED, we perform simulation studies in this section. We simulate data in the same manner as in Zu & Yuan (2010), and we compare the following six methods: the standard bootstrap test of Preacher & Hayes (2004,

[5] Nevertheless, our software can also report t-tests for the robust coefficient estimates obtained from the full sample.

2008), the standard Sobel test[6] (Sobel, 1982), Zu & Yuan's (2010) versions of the bootstrap and Sobel test following Huberization of the data, a robust version of the Sobel test that replaces OLS regression with MM-regression, as well as our robust bootstrap test ROBMED.[7]

We compare the six methods in two situations: (i) when there is mediation, (ii) when there is no mediation. The data are simulated according to the models $M = aX + e_1$ and $Y = bM + cX + e_3$, see Equations (1) and (3), following the simulation design of Zu & Yuan (2010). The explanatory variable X and the error terms e_1 and e_3 follow a standard normal distribution. The sample size is $n = 250$. In addition to analyzing the clean data, we replace the first $1, \dots, 10$ observations, respectively, with outliers by setting $M_i^* = M_i - 6$ and $Y_i^* = Y_i + 6$. On each of those 11 data sets, two-sided tests are performed with null hypothesis $H_0\colon ab = 0$ against the alternative $H_a\colon ab \neq 0$. The whole process is repeated $R = 500$ times. For case (i) where mediation exists, we set $a = b = c = 0.2$, yielding a true indirect effect $ab = 0.04$. For case (ii) where mediation does not exist, we set $b = 0$, giving a true indirect effect $ab = 0$.

Simulations with mediation

Figure 5 shows the average estimates of the indirect effect (left) and the rate of how often the methods reject the null hypothesis and the corresponding estimate of the indirect effect has the correct sign (right) under an increasing percentage of outliers. Note that evaluating the methods by the rejection rate from the two-sided tests alone does not provide a

[6] The Sobel test provides a statistical test for the significance of the indirect effect by assuming that it follows a normal distribution. The indirect effect ab is divided by (a first-order approximation of) the standard error of the indirect effect s_{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 ab, as the product of two normally distributed random variables – the coefficients a and b – is not normally distributed (MacKinnon D. P., Lockwood, Hoffman, West, & Sheets, 2002).

[7] We did not include Baron & 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, Lockwood, Hoffman, West, & Sheets, 2002).

meaningful comparison in this simulation setting, because the outliers push the estimated indirect effect from a positive one towards a negative one. For higher number of outliers, this incorrectly estimated negative indirect effect is often 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 wrong 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 test in the presence of outliers.

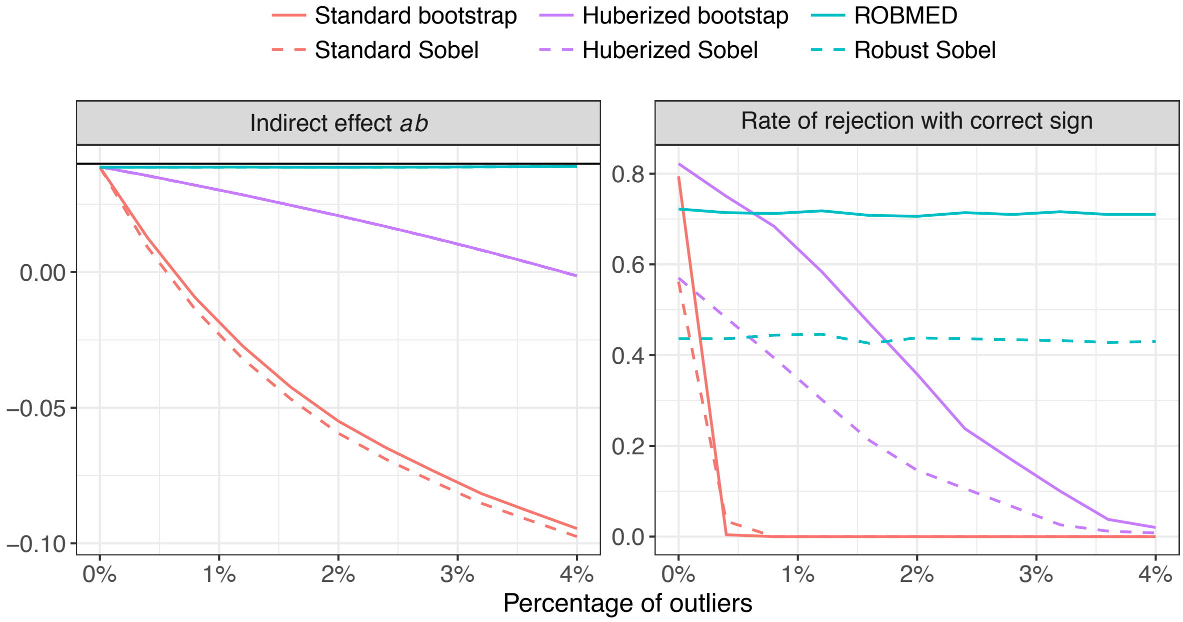

Figure 5. Results for the simulation setting with mediation from 500 simulation runs. The left hand side side shows the average estimates of the indirect effect and includes a horizontal reference line for the true indirect effect $ab = 0.04$. The right hand side displays the rate of how often the methods 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). ‘Standard bootstrap’ and ‘Standard Sobel’ denote the standard versions of the bootstrap test of Preacher & Hayes (2004, 2008) and the test of Sobel (1982), ‘Huberized bootstrap’ and ‘Huberized Sobel’ denote Zu & Yuan’s (2010) versions of those tests following Huberization of the data, ‘ROBMED’ is our proposed fast and robust bootstrap test, and ‘Robust Sobel’ is a version of the Sobel test replacing OLS regression with MM-regression.

The left panel of Figure 5 indicates that the standard methods perform the worst under the presence of increasing amounts of outliers. The Huberized methods of Zu & Yuan (2010) are also already affected by small numbers of outliers, with increasing effect as that number increases. However, ROBMED and the robust Sobel test remain stable in estimating the indirect effect. It is also worth noting that the estimated indirect effect for the standard bootstrap test and the standard Sobel test are not the same. That is, because bootstrap procedure reports the average of the indirect effect over the bootstrap samples rather than the value computed from the original data set and different bootstrap samples contain different numbers of outliers. Hence, the effect of the outliers on the bootstrap samples is different from the effect on the original sample. Consequently, the difference in the estimated indirect effect can be seen as the influence of the outliers on the standard bootstrap on top of their influence on the regression estimates. This difference further illustrates that both the estimation of the regression coefficients and the bootstrap procedure need to be robustified (although the effect here is small due to the small number of outliers).

The right panel of Figure 5 displays the rate of how often the methods reject the null hypothesis and the corresponding estimate of the indirect effect has the correct sign (our measure of realized power of the tests). Clearly, the results from the estimation of the indirect effect carry over. For the standard tests, the realized power drops to 0 when as little as two outliers are present (0.8% of the data). But also the Huberized tests of Zu & Yuan (2010) continuously lose power and the realized power eventually drops to about 0.05 for 10 outliers (4% of the data). Again, ROBMED and the robust Sobel test remain stable, with ROBMED being more powerful than its competitors for two or more outliers. In addition, all bootstrap tests have higher power than their Sobel test counterparts.

Simulations with no mediation

In the left panel of Figure 6, we observe that the outliers push the standard estimates towards a negative estimated effect. A similar effect, although to a lesser extent, is visible for the estimates according to Zu & Yuan's Huberized methods. ROBMED and the robust Sobel test, on the other hand, remain again stable and close to the true value of the effect.

The right panel of Figure 6 presents the rejection rate (i.e., the realized size of the tests). As expected, the rejection rate for the standard methods quickly rises, but interestingly it starts to fall again for higher percentages of outliers. This may be because of the estimated confidence intervals being even more affected by the outliers than the point estimates,

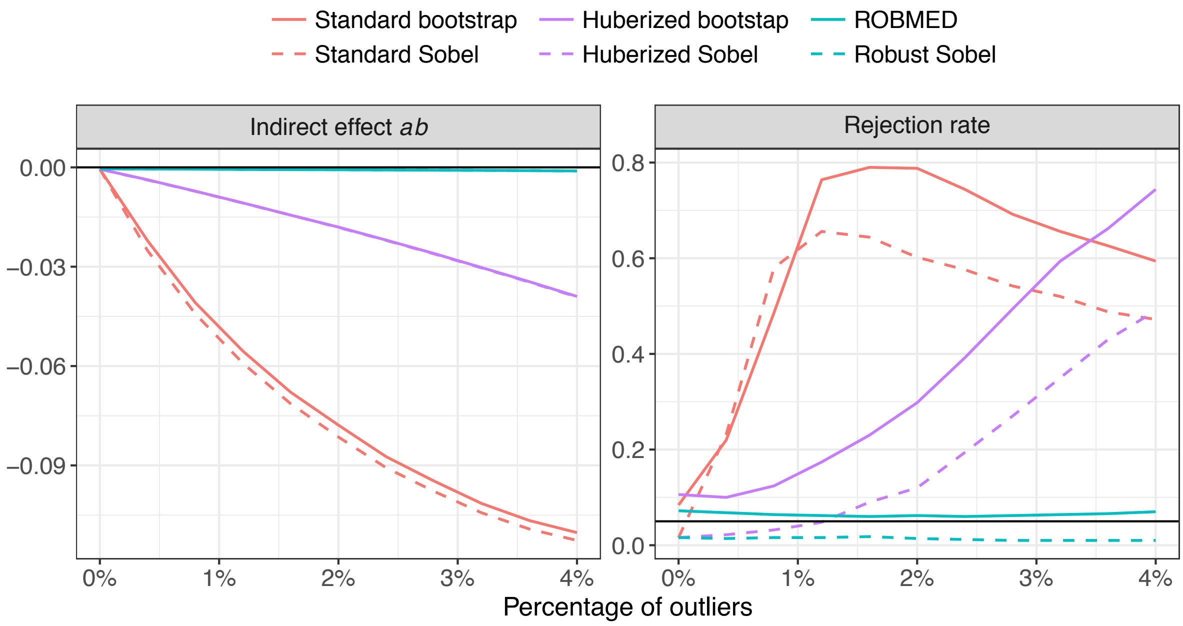

Figure 6. Results for the simulation setting with no mediation from 500 simulation runs. The left hand side side shows the average estimates of the indirect effect and includes a horizontal reference line for the true indirect effect $ab = 0$. The right hand side displays the rejection rate of the corresponding 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). 'Standard bootstrap' and 'Standard Sobel' denote the standard versions of the bootstrap test of Preacher & Hayes (2004, 2008) and the test of Sobel (1982), 'Huberized bootstrap' and 'Huberized Sobel' denote Zu & Yuan's (2010) versions of those tests following Huberization of the data, 'ROBMED' is our proposed fast and robust bootstrap test, and 'Robust Sobel' is a version of the Sobel test replacing OLS regression with MM-regression.

yielding very large confidence intervals for higher number of outliers. The rejection rates of the Huberized tests of Zu & Yuan (2010) increase more slowly, but eventually even surpass the rejection rate of the standard methods. ROBMED and the robust Sobel test are the only ones unaffected by the outliers. Furthermore, while all tests are performed using significance level $\alpha = 0.05$, results without outliers show that only the standard bootstrap test and ROBMED actually achieve a realized size that is reasonably close to the nominal size $\alpha = 0.05$. The realized size of the bootstrap test of Zu & Yuan (2010) is higher than 0.1, which may be an indication that the standard error is underestimated by leaving out the variability from the Huberization step. On the other hand, all Sobel tests exhibit a realized size of about 0.01 when there is no contamination. This difference from the nominal size $\alpha = 0.05$ is an indication that the assumptions on the distribution of the indirect effect do not hold in general.

Concluding discussion of the simulation study

In the above simulations, ROBMED clearly outperformed the alternative methods. It remains stable when outliers are introduced and is the most powerful test when there are multiple outliers. In addition, ROBMED does not lose much power to the standard methods when there are no outliers, and it realizes the theoretical size of the test when there is no mediation.

It should also be noted that while Zu & Yuan's bootstrap test seemingly has the highest power for 0 or 1 outliers in the simulation with mediation, its rejection rate in the simulation with no mediation was twice the nominal size of $\alpha = 0.05$. Hence the power of Zu & Yuan's bootstrap test is not really comparable, as the test is not well-calibrated and over-rejects in general.

As robustness checks, we also ran simulations in other settings than Zu & Yuan's (2010) design, because the outliers are reasonably far away from the main bulk of the data in their specific setting. We investigated settings were the outliers were much closer to the main part of the data, and settings where the outliers were even farther away. ROBMED outperformed its competitors in those situations as well. To keep the paper at a reasonable length, we prefer to report the simulations only with Zu & Yuan's (2010) design since the results are representative for the different settings that we investigated.

ILLUSTRATIVE EMPIRICAL CASES

In this section, we illustrate three empirical cases in which we test established hypotheses from the management literature. After presenting brief rationales for each illustrative hypothesis, we focus on the comparison between ROBMED and the state-of-the-art bootstrap test of Preacher & Hayes (2004, 2008). Note that the aim of this section is to demonstrate the need for ROBMED and to show the mechanics of its application rather than to build and test management theory. The cases are selected to demonstrate the role that deviations from the model assumptions play in mediation analysis and clarify how the proposed method overcomes those challenges. The first case shows that both robust and standard methods give similar results when there are no issues with the data. The second case exemplifies a situation where the proposed robust method finds evidence for mediation, whereas the standard method fails to do so. The third case presents a situation where the proposed robust method finds no evidence for mediation, while the standard method is driven to report evidence suggesting mediation.

The data for the illustrative cases comes from a larger research program on team processes. Data were collected from 354 senior business administration students playing a 12 round business simulation game[8] (two separate games of 6 rounds) in randomly assigned 4-person teams (92 teams in total) as part of their capstone strategy course at a Western European University. The overall response rate was 93% (332 students). Leaving out teams with less than 50% response rate yields n = 89 teams for further analysis. Data on several individual- and team-level constructs were collected in three survey waves: prior to, during, and after the simulation game with different constructs being surveyed in the different waves. Previously established survey scales were used to measure constructs, and the reliability and validity of the scales were satisfactory. Further information on measures and reliability is presented in the Supplementary Material. Tables 1 and 2 contain descriptive statistics and correlations for the variables studied in the empirical cases.

Table 1. Descriptive statistics of the variables used in the illustrative empirical cases.

Variable	Mean	Standard deviation	Median	Median absolute deviation	Minimum	Maximum
Process conflict	1.368	0.302	1.250	0.247	1.000	2.167
Shared experience	89.854	14.815	91.000	14.826	57.000	111.000
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
Team performance	3.968	0.423	4.000	0.463	3.000	4.750
Transactive memory systems	3.367	0.262	3.367	0.272	2.767	4.089
Value diversity	1.676	0.345	1.587	0.366	1.105	2.548

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, Martin, & Yohai, 2006).

[8] Other researchers on team processes have published findings based on data from this game as well (e.g., Mathieu & Rapp, 2009; Boies, Lvina, & Martens, 2010).

Table 2. Correlation table of the variables used in the illustrative empirical cases.

	Process conflict	Shared experience	Task conflict	Team commitment	Team performance	Transactive memory systems	Value diversity
Process conflict	1.000	-0.052	0.542	-0.367	-0.336	-0.344	0.172
Shared experience		1.000	-0.178	0.340	0.445	0.253	0.021
Task conflict			1.000	-0.297	-0.294	-0.389	0.268
Team commitment				1.000	0.569	0.612	-0.024
Team performance					1.000	0.515	0.080
Transactive memory systems						1.000	-0.138
Value diversity							1.000

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, which are highly influenced by outliers or heavy tails.

Empirical Case 1

Transactive memory systems (TMS) are defined as "shared systems that people in relationships develop for encoding, storing, and retrieving information about different substantive domains" (Ren & Argote, 2011, p. 191). TMS comprise the knowledge of 'who knows what' in a team suggesting a cooperative division of labor in a team's mental tasks (Wegner, 1987) and consist of three dimensions: specialization, credibility and coordination (Lewis, 2003). TMS improve team performance, because they enable the team to search and locate required knowledge quickly and accurately, to match issues with appropriate expertise within the team, to coordinate group activities, and eventually to arrive at better decisions (Moreland, 1999). Shared group experience and training is considered to be a driver of TMS, because teams with higher levels of shared experience have more opportunity to interact with each other and observe other team members while performing tasks, thereby form accurate representations of expertise within the team (Moreland, Argote, & Krishnan, 1996). In sum, we test the following hypothesis:

*Illustrative Hypothesis 1: Transactive memory systems (*M*) mediate the relationship between shared group experience (*X*) and team performance (*Y*).*

Before comparing the results of our robust method with those of the standard method, we first take a look at the data at hand. Figure 7 shows the pairwise scatter plots between the studied variables. While those plots indicate that the point cloud is not perfectly elliptical (which would correspond to the usual normality assumptions), the data appear to form a compact cloud without any clear outliers or heavy tails. Therefore, we expect no major issues with the data.

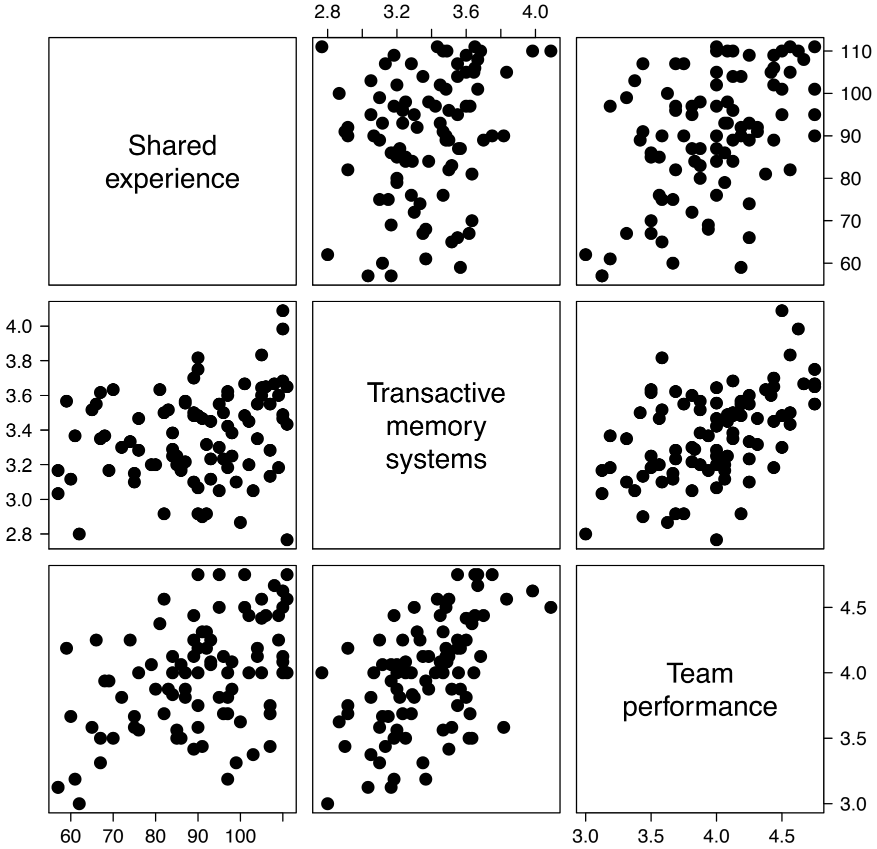

Figure 7. Scatter plots of study variables (Case 1). No clear outliers are observed although there are some deviations from normality.

Table 3. Comparison of the standard bootstrap method and ROBMED (Case 1).

	Standard Method					**ROBMED**				
Direct Effects	**Estimate**	**Std. Error**	***t*-Statistic**	***p*-Value**		**Estimate**	**Std. Error**	***z*-Statistic**	***p*-Value**	
X→M (a path)	0.004	0.002	2.331	0.022	*	0.005	0.002	2.143	0.032	*
(X),M→Y (b path)	0.674	0.140	4.819	<0.001	***	0.663	0.171	3.875	<0.001	***
X,(M)→Y (c path)	0.011	0.002	4.299	<0.001	***	0.011	0.003	3.263	0.001	**
X→Y (c' path)	0.014	0.003	5.030	<0.001	***	0.014	0.003	4.165	<0.001	***
Indirect Effect	**Estimate**	**95% Confidence Interval**				**Estimate**	**95% Confidence Interval**			
ab	0.003	(0.0004, 0.0064)				0.003	(0.0004, 0.0070)			
Model Summary										
Residual standard error	0.334	d.o.f. (86)				0.372	d.o.f. (86)			
R-squared	0.390					0.361				
Adjusted R-squared	0.376					0.346				
F-statistic	27.489	*** d.o.f. (2, 86)				9.803	*** d.o.f. (2, ∞)			

Variables are shared group experience (X), transactive memory systems (M), and team performance (Y). Sample size = 89, Number of bootstrap samples = 5000, significance levels: '***' 0.001, '**' 0.01 and '*' 0.05.

The outlyingness weights from our robust method agree with this assessment – although several observations do get partly downweighted, no observation receives a weight of 0. A frequently used threshold to define potential outliers is 0.25. There is one observation that falls below this threshold with a weight of 0.139.[9]

Table 3 shows the mediation analyses with the standard method of Preacher & Hayes (2004, 2008) and ROBMED. The two methods agree on the significance of the effects, and the coefficient estimates are comparable. In particular, both methods report a strictly positive 95% confidence interval for the indirect effect. To get a better insight into the evidence found against the null hypothesis of no mediation, we estimate the p-value as the smallest significance level α where the $(1-\alpha)\cdot 100\%$ confidence interval obtained from the bootstrapped distribution of the indirect effect ab does not contain 0. Also here ROBMED

[9] The corresponding observation has a standardized residual in the regression of M on X of -2.727. Under normally distributed errors, the probability of an observation having a standardized residual > 2.727 in absolute value is roughly 0.8%. With n = 89 observations, it is not unlikely that such a large residual – and thus such a low weight in the robust regression – is due to chance. Further investigation of the observation in question may provide more insight.

(p-value $= 0.027$) is comparable with the standard method (p-value $= 0.025$). Hence, this example indicates that in the absence of any major issues with the data, ROBMED yields similar results as the standard method.

Empirical Case 2

Values are standards that guide thought and action (Schwartz, 1992); they predispose individuals to favor one ideology over another, determine how one judges one self and others, cause taking certain positions on social issues (Rokeach, 1973). Schwartz's value theory proposes 10 distinct universal values that are theoretically derived from human nature; these ten values are power, achievement, hedonism, stimulation, self-direction, universalism, benevolence, tradition, conformity and security. When team members possess different set of values – leading to value diversity in teams – teams can experience higher levels of conflict in executing their tasks (Jehn, 1994), because the variety of worldviews may cause different prioritizations of actions that need to be coherently conducted. Conflict on the task content triggered by a difference in values can be detrimental to team outcomes (Jehn, Northcraft, & Neale, 1999). Therefore, we investigate the following hypothesis:

*Illustrative Hypothesis 2: Task conflict (*M*) mediates the relationship between value diversity in teams (X) and team commitment (Y).*

Table 4 reports the comparison of the standard method and ROBMED. The estimate of the indirect effect ab is nearly twice as large in magnitude for ROBMED compared to the standard method. In addition, the 95% confidence interval of ROBMED is strictly negative but that of the standard method contains 0. Considering p-values as well, we observe that ROBMED finds evidence against the null hypothesis of no mediation (p-value $= 0.027$), whereas the standard method finds no evidence (p-value $= 0.158$). 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 standard method (p-value $= 0.203$) but highly significant for

ROBMED (p-value = 0.003). Hence we take a closer look at the relationship between the independent variable and the hypothesized mediator.

Table 4. Comparison of the standard bootstrap method and ROBMED (Case 2).

	Standard Method					ROBMED				
Direct Effects	**Estimate**	**Std. Error**	***t*-Statistic**	***p*-Value**		**Estimate**	**Std. Error**	***z*-Statistic**	***p*-Value**	
X→M (a path)	0.155	0.121	1.283	0.203		0.321	0.107	2.998	0.003	**
(X),M→Y (b path)	-0.364	0.118	-3.090	0.003	**	-0.344	0.178	-1.934	0.053	.
X,(M)→Y (c path)	-0.021	0.134	-0.156	0.877		0.065	0.186	0.350	0.726	
X→Y (c' path)	-0.077	0.139	-0.555	0.580		-0.045	0.187	-0.241	0.810	
Indirect Effect	**Estimate**	**95% Confidence Interval**				**Estimate**	**95% Confidence Interval**			
ab	-0.060	(-0.208, 0.025)				-0.110	(-0.294, -0.010)			
Model Summary										
Residual standard error	0.430	d.o.f. (86)				0.390	d.o.f. (86)			
R-squared	0.103					0.090				
Adjusted R-squared	0.082					0.069				
F-statistic	4.944 **	d.o.f. (2, 86)				1.497	d.o.f. (2, ∞)			

Variables are value diversity (X), task conflict (M), and team commitment (Y). Sample size = 89, Number of bootstrap samples = 5000, significance levels: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1.

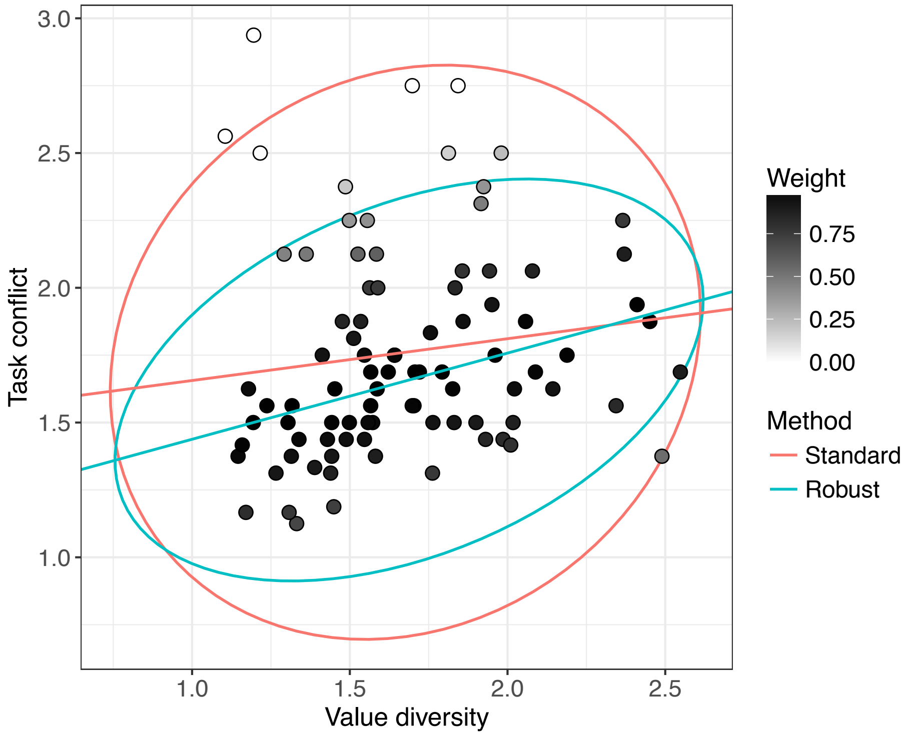

Figure 8. Scatter plot of value diversity and task conflict with tolerance ellipses (Case 2).

Figure 8 shows a scatter plot 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 standard covariance matrix, which is closely linked to OLS regression[10], as well as a tolerance ellipse based on the weighted covariance matrix using the outlyingness weights from the robust regression of M on X.

The plot reveals that there are a small number of influential observations due to 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 a lot of empty space in the standard tolerance ellipse below the data cloud. The influence of the far away points is also visible in the standard regression line, which is tilted to become more horizontal.

As the deviations are due to a heavy upper tail rather than clear outliers, we also applied a Box-Cox transformation (Box & Cox, 1964) to each variable and then performed the standard bootstrap test. After this transformation, the standard method finds evidence against the null hypothesis of no mediation (p-value $= 0.041$). While a simple transformation seems to solve the issue here, we note two points: (i) transformations can make it difficult to interpret the results of mediation analysis; and (ii) transformations are not widespread in the organizations studies, and empirical researchers often do not check if they are necessary to satisfy model assumptions. ROBMED, on the other hand, is able to handle model deviations

[10] 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 $\hat{\imath}_1$ and $\hat{a}$, the sample covariance $\hat{\sigma}_{MX}$ and the sample variance $\hat{\sigma}_X$.

such as heavy tails. In this case, it was not necessary to apply transformations and the results can be interpreted in the usual way, which underlines the benefits of ROBMED for empirical researchers. To summarize, while we emphasize that the deviating data points should be investigated further, ROBMED better captures the main trend in the data and can therefore be considered more reliable.

Empirical Case 3

When team members are diverse in their values, they may also experience relational conflict. Process conflict is defined as "conflict about how task accomplishment should proceed in the work unit, who's responsible for what, and how things should be delegated" (Jehn, 1997, p. 540). Because values serve as criteria for evaluating and selecting among policies and actions (Schwartz, 2006), value diversity within a team indicates different guidelines in deciding how to conduct the task. This may lead to higher process conflict in the team. Process conflict is found to hold negative consequences for team performance (Jehn, 1997; Greer & Jehn, 2007). Hence, we test the following hypothesis:

Illustrative Hypothesis 3: Process conflict (M) mediates the relationship between value diversity in teams (X) and team performance (Y).

The results of mediation analysis with the standard method and ROBMED are shown in Table 5. Here the 95% confidence interval of the robust method contains 0, whereas that of the standard method is strictly negative. The p-values offer a more detailed picture, where we see that the robust method finds no evidence for mediation (p-value $= 0.147$), while the standard method does report evidence (p-value $= 0.041$).

As in the previous example, one of the main differences between the methods is in the estimation of the a path, which in this case is clearly not significant for the robust method (p-value $= 0.231$) but weakly significant for the standard method (p-value $= 0.087$). For further investigation, we visualize the relationship between the independent variable and the proposed mediator. Figure 9 contains a scatter plot with tolerance ellipses of process conflict (M) against value diversity (X). We observe again a small number of influential observations,

Table 5. Comparison of the standard bootstrap method and ROBMED (Case 3).

	Standard Method					ROBMED				
Direct Effects	**Estimate**	**Std. Error**	***t*-Statistic**	***p*-Value**		**Estimate**	**Std. Error**	***z*-Statistic**	***p*-Value**	
X→M (a path)	0.160	0.093	1.733	0.087	.	0.123	0.103	1.197	0.231	
(X),M→Y (b path)	-0.473	0.144	-3.273	0.002	**	-0.546	0.210	-2.602	0.009	**
X,(M)→Y (c path)	0.107	0.127	0.840	0.403		0.140	0.113	1.243	0.214	
X→Y (c' path)	0.031	0.131	0.234	0.816		0.073	0.142	0.513	0.608	
Indirect Effect	**Estimate**	**95% Confidence Interval**				**Estimate**	**95% Confidence Interval**			
ab	-0.077	(-0.243, -0.002)				-0.067	(-0.236, 0.024)			
Model Summary										
Residual standard error	0.403	d.o.f. (86)				0.403	d.o.f. (86)			
R-squared	0.111					0.134				
Adjusted R-squared	0.091					0.114				
F-statistic	5.387 **	d.o.f. (2, 86)				2.776 .	d.o.f. (2, ∞)			

Variables are value diversity (X), process conflict (M), and team performance (Y). Sample size = 89, Number of bootstrap samples = 5000, significance levels: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1.

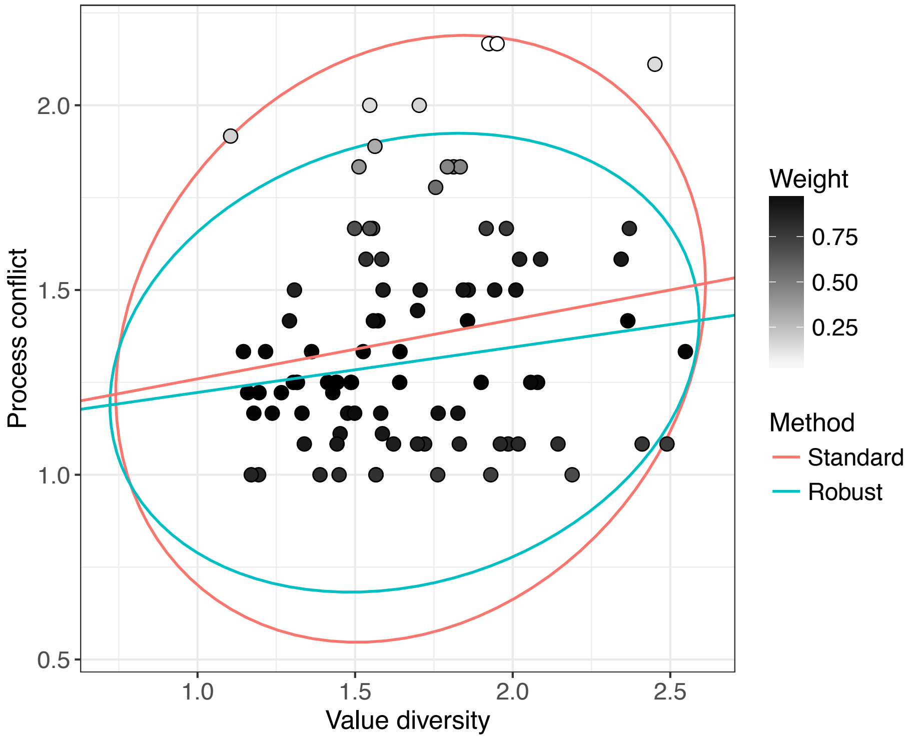

Figure 9. Scatter plot of value diversity and process conflict with tolerance ellipses (Case 3).

this time due to a heavy upper tail in process conflict. While no observation receives a weight of exactly 0, the points close to the border of the robust tolerance ellipse are partly downweighted. Two observations thereby receive a weight that is very close to 0 (0.024 and 0.026, respectively). Overall, the robust tolerance ellipse better fits the main bulk of the data, as there is less empty space below the data cloud compared to the standard tolerance ellipse. The plot further shows that the influential data points above the main data cloud tilt the standard regression line upwards, thus exaggerating the significance of the effect.

After a Box-Cox transformation of each variable and re-applying the standard bootstrap test to the transformed data, the standard test now reports only weak evidence against the null hypothesis of no mediation (p-value $= 0.070$), and significance of the a coefficient also diminishes (p-value $= 0.115$). Hence the data do not clearly support that value diversity increases process conflict, or that consequently process conflict mediates the relationship between value diversity and team commitment, at least not in our study context. Closer inspection of the downweighted observations, and possibly a replication study, are necessary for more definitive conclusions on the hypothesized mediation relationship.

DISCUSSION AND CONCLUSION

Mediation analyses are sensitive to deviations from model assumptions such as outliers, yet outliers are ubiquitous in empirical data collection. To overcome this widespread problem, we developed a new statistical procedure called ROBMED. ROBMED replaces the OLS estimator for regression with the robust MM-estimator (Yohai, 1987) and the standard bootstrap with the fast and robust bootstrap (Salibián-Barrera & Zamar, 2002; Salibián-Barrera & Van Aelst, 2008). This novel technical configuration to mediation analysis ensures that estimates of the indirect effect are not affected by outliers and heavy tails. Indeed, ROBMED was shown to be more reliable than standard methods for testing mediation. Our simulations demonstrated that ROBMED does not lose much power when there are no outliers but gains a lot more power in the presence of outliers.

There are a few key technical properties that give ROBMED its edge. First, ROBMED evaluates the outlyingness of each data point objectively and continuously downweights deviating points such as outliers and observations in heavy tails. As this downweighting is part of the estimation procedure, standard errors in ROBMED are accurate, whereas two-step procedures that use an extra outlier elimination step in the analysis yield underestimated standard errors due to omitting the uncertainty from the data-cleaning step. Second, ROBMED is more stable than two-step procedures because it does not require any decision to fully include or exclude a data point, and its gradual downweighting of deviating data points is more efficient in the case of heavy tails. Data points that do not deviate from the majority, on the other hand, receive a (nearly) full weight in the analysis. This means that results are similar to those of maximum likelihood procedures if there are no deviating data points, which is illustrated in one of our empirical cases.

As illustrated in our empirical cases as well, one way to deal with certain deviations from model assumptions is to transform the data. In our examples, mediation analysis with transformed variables gave qualitatively similar results to ROBMED. However, not only that the transformation introduces an additional step in the analysis which may be mishandled due to researchers' degree of freedom or negligence, but also transformations often make the interpretation of mediation results difficult. In addition, there is an epidemic for lack of reporting outlier and model assumption checks in organizations research (Aguinis, Gottfredson, & Joo, 2013). ROBMED relieves researchers from the burden of both transformations and outlier identification and treatment, since it is capable of dealing with them efficiently due to its unique technical properties.

By no means do we imply that the initial screening of data for potential problems is not valuable. On the contrary, ROBMED helps researchers to identify such anomalies for further investigation while simultaneously giving reliable results. In that sense, our empirical examples show that deviations from model assumptions can influence results in very different ways. ROBMED, in those cases, helps in understanding how deviations from model assumptions affect the results. Hence ROBMED allows more informed and reliable decisions.

If the user is concerned with relying solely on ROBMED, a good strategy is to apply both the standard method and ROBMED as a robustness check. If the two methods agree, it indicates that there are no severe data problems and the user can go ahead and use the results of the standard method. If the two methods disagree, however, the results of the standard method are unlikely to be reliable and the user needs to investigate what causes the disagreement of the two methods. As such, ROBMED plays an integral part in ensuring robust findings in empirical organizations research – and therefore reproducibility.

On a final note, the implementation of ROBMED could be inaccessible to those who are not handy with coding. To overcome this limitation and to increase the adoption of our method among empirical researchers, we make our R and SPSS implementations for ROBMED freely available. Researchers can download the code and run it by following the simple steps in the accompanying documentation and code examples. Given its technical strengths and practicality, we strongly encourage researchers and practitioners to adopt ROBMED to test mediation.

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.

Alfons, A. (2018). *robmed: (Robust) Mediation Analysis.* R package version 0.2.0.

Avella-Medina, M., & Ronchetti, E. (2015). Robust Statistics: A Selective Overview and New Directions. *Wiley Interdisciplinary Reviews: Computational Statistics , 7* (6), 372–393.

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.

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.

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. (2017). *Valid Inference Corrected for Outlier Removal.* arXiv:1711.10635.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Colquitt, J. A., & Zapata-Phelan, C. P. (2007). Trends in Theory Building and Theory Testing: A Five-Decade Study of the Academy of Management Journal. *Academy of Management Journal , 50* (6), 1281–1303.

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, UK: Cambridge University Press.

Edwards, J. R. (2010). Reconsidering Theoretical Progress in Organizational and Management Research. *Organizational Research Methods , 13* (4), 615–619.

Greer, L. L., & Jehn, K. A. (2007). The Pivotal Role of Negative Affect in Understanding the Effects of Process Conflict on Group Performance. In E. A. Mannix, M. A. Neale, & C. P. Anderson (Eds.), *Research on Managing Groups and Teams* (Vol. 10, pp. 21–43).

Hackman, J. R. (1986). The Psychology of Self-Management in Organizations. In M. S. Pallack, & R. O. Perloff (Eds.), *Psychology and Work: Productivity, Change, and Employment* (pp. 89–136). Washington, DC: American Psychological Association.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Chichester, UK: John Wiley & Sons.

Hawkins, D. M. (1980). *Identification of Outliers.* London, UK: Chapman & Hall.

Hayes, A. F. (2013). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach.* New York, NY: The Guilford Pres.

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.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings* (2nd ed.). Thousand Oaks, CA: Sage.

IBM Corp. (2017). *IBM SPSS Statistics, Version 25.0.* Armonk, NY: IBM Corp.

Jehn, K. A. (1995). A Multi-Method Examination of the Benefits and Detriments of Intra-Group Conflict. *Administrative Science Quarterly , 40* (2), 256–285.

Jehn, K. A. (1997). A Qualitative Analysis of Conflict Types and Dimensions in Organizational Groups. *Administrative Science Quarterly , 42* (3), 530–557.

Jehn, K. A. (1994). Enhancing Effectiveness: An Investigation of Advantages and Disadvantages of Value-Based Intragroup Conflict. *International Journal of Conflict Management , 5* (3), 223–238.

Jehn, K. A., Northcraft, G. B., & Neale, M. A. (1999). Why Differences Make a Difference: A Field Study of Diversity, Conflict and Performance in Workgroups. *Administrative Science Quarterly , 44* (4), 741–763.

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

Lewis, K. (2003). Measuring Transactive Memory Systems in the Field: Scale Development and Validation. *Journal of Applied Psychology , 88* (4), 587–604.

Lindeman, M., & Verkasalo, M. (2005). Measuring Values With the Short Schwartz's Value Survey. *Journal of Personality Assessment , 85* (2), 170–178.

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., & 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., 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., Warsi, G., & Dwyer, J. H. (1995). A Simulation Study of Mediated Effect Measures. *Multivariate Behavioral Research , 30* (1), 41–62.

Maronna, R. M., Martin, R. D., & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods.* Chichester, UK: John Wiley & Sons.

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.

Moreland, R. L. (1999). Transactive Memory: Learning Who Knows What in Work Groups and Organizations. In L. L. Thompson, D. M. Messick, & J. M. Levine (Eds.), *Shared Cognition in Organizations: The Management of Knowledge* (pp. 3–31). Hillsdale, NJ: Erlbaum.

Moreland, R. L., Argote, L., & Krishnan, R. (1996). Socially Shared Cognition at Work: Transactive Memory and Group Performance. In J. L. Nye, & A. M. Brower (Eds.), *What's Social about Social Cognition? Research on Socially Shared Cognition in Small Groups* (pp. 57–84). Thousand Oaks, CA: Sage.

Morgenthaler, S. (2007). A Survey of Robust Statistics. *Statistical Methods & Applications , 15* (3), 271–293.

Mowday, R. T., Steers, R. M., & Porter, L. W. (1979). The Measurement of Organizational Commitment. *Journal of Vocational Behavior , 14* (2), 224–247.

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.

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.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Ren, Y., & Argote, L. (2011). Transactive Memory Systems 1985–2010: An Integrative Framework of Key Dimensions, Antecedents, and Consequences. *Academy of Management Annals , 5* (1), 189–229.

Renaud, O., & Victoria-Feser, M.-P. (2010). A Robust Coefficient of Determination for Regression. *Journal of Statistical Planning and Inference , 140* (7), 1852–1862.

Rokeach, M. (1973). *The Nature of Human Values.* New York, NY: The Free Press.

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. *The Annals of Statistics , 30* (2), 556–582.

Schwartz, S. H. (2006). Basic Human Values: Theory, Measurement, and Applications. *Revue Française de Sociologie , 47* (4), 929–968.

Schwartz, S. H. (1992). Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. *Advances in Experimental Social Psychology , 25*, 1–65.

Sobel, M. E. (1982). Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology , 13*, 290–312.

Tost, L. P., Gino, F., & Larrick, R. P. (2013). When Power Makes Others Speechless: The Negative Impact of Leader Power on Team Performance. *Academy of Management Journal , 56* (5), 1465–1486.

Wegner, D. M. (1987). Transactive Memory: A Contemporary Analysis of the Group Mind. In B. Mullen, & G. R. Goethals (Eds.), *Theories of Group Behavior* (pp. 185–208). New York, NY: Springer-Verlag.

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. *The Annals of Statistics , 15* (20), 642–656.

Zu, J., & Yuan, K.-H. (2010). Local Influence and Robust Procedures for Mediation Analysis. *Multivariate Behavioral Research , 45* (1), 1–44.

SUPPLEMENTARY MATERIAL

Study Context and Data

As part of their corporate strategy course, students played a business strategy game as an elaborate simulation experience, in which participants ran a virtual company and competed with other teams in a cyber-industry environment. Participants determined the best strategy for their "firm" to run a profitable business and outperform other teams taking part in the simulation. To this end, teams took decisions in areas such as operations, finance, marketing, sales, and human resources. The simulation format helped participants emerge into realistic experiential learning environment. Additionally, companies' performance in the industry determined 50 percent of the participants' grade, which provided them further incentive to perform well. The game thus created a realistic, engaging and challenging representation of a business environment for team members and aimed at inducing an environment similar to real companies (Chen, Katila, McDonald, & Eisenhardt, 2010). This game has been used in other research as well (Mathieu & Rapp, 2009; Boies, Lvina, & Martens, 2011). Students played the business simulation game in two separate games of 6 rounds. In addition, students took part in three surveys before, during and after the simulation game. Before the simulation game started, participants filled in the first online questionnaire that contained individual-level measures. During the game and after the game ended, participants filled in the second and third online questionnaires to assess several team level constructs. We assured students that their survey answers would remain confidential through anonymization procedures. The response rate for the first survey was 99% (351 respondents) and for the second survey 94% (334 students). The response rate of the third survey was the overall response rate reported in the main text (93%, 332 students).

Measures, validity and reliability

Empirical Case 1. We operationalized *Transactive memory systems (TMS)* by using Lewis' (2003) 15-item scale that measures the three sub-dimensions of TMS (i.e., credibility, specialization and coordination). Team members responded on a 5-point scale (1 = "strongly disagree"; 5 = "strongly agree"). The Cronbach's alpha values for credibility, specialization and coordination are 0.75, 0.81, and 0.71, respectively. Following Lewis (2003), we aggregated these three sub dimensions to form the TMS construct. Then we aggregated TMS to a team level construct (median $r_{WG} = 0.97$). Sample items include: "Different team members are responsible for expertise in different areas" (specialization), "I was confident relying on the information that other team members brought to the discussion" (credibility), "Our team worked together in a well-coordinated fashion" (coordination). TMS was measured in the second survey after the first game.

Since the teams are randomly formed, we expect no prior *shared group experience*.[11] Team members develop a shared group experience and training during the first game for the second game. Therefore, team performance in the first game is a good proxy for the level of shared group experience and training. That is, because higher team performance implies that team members were able to effectively coordinate, cooperate and deliver high performance compared to teams with lower performance. This variable is objectively determined by the simulation game based on two sets of objective performance measures: (i) the extent to which the teams meet the previously set performance criteria by investors (i.e., return on equity, earnings-per-share, stock price, credit rating, image rating), and (ii) the extent to which they outperform their competitors in the industry in these performance criteria.

To overcome common method bias and to cover the affective dimensions of performance, we measured the *team performance* in the second game with the team members' subjective perceptions of the team's functioning. We used Hackman's (1986; Holmbeck, 2002) 4-item scale. Sample items include "This team did a good job" and "This team performed poorly". The team members evaluated their team's performance on a 5-point Likert scale. Cronbach alpha is 0.89. We then aggregated team members' individual response to team level (median $r_{WG} = 0.92$) Team performance was measured in the third survey after the second game.

Empirical Case 2. We operationalized *task conflict* with the intra-group conflict scale of Jehn (1995). The five items on the presence of conflict were rated on a 5-point Likert scale (anchored by 1 = "None" and 5 = "A lot"). Examples of the scale measuring task conflict include the following: "How frequently are there conflicts about ideas in your work unit?" and "How often do people in your work unit disagree about opinions?". Cronbach's alpha is 0.86. We aggregated individual responses to team level (median $r_{WG} = 0.95$). We used the short version of Schwartz's Value Survey (SVS) to measure team members' individual values (Lindeman & Verkasalo, 2005). Then we operationalized *value diversity* with average of the coefficient of variations of each value dimension among team members. *Team commitment* is measured by four items based on Mowday, Steers, & Porter (1979). Sample items include "I feel proud to belong to this team" and "I am willing to exert extra effort to help this team succeed". Cronbach's alpha is 0.78. Individual responses were aggregated to team level (median $r_{WG} = 0.93$). Value diversity was measured in survey 1, task conflict in survey 2 and team commitment in survey 3.

Empirical Case 3. We measured *process conflict* with three items of Jehn (1995). Sample items included "How much conflict is there about delegation of tasks within your team?" and "How frequently do members of your team disagree about the way to complete a

[11] We controlled for familiarity of team members with each other prior to the simulation game. Team members assessed the extent to which they are familiar with each team member individually (1 = "not familiar at all", 5 = "very familiar"), and then these scores are averaged to represent the team familiarity.

group task?”. Process conflict was measured in survey 2, and Cronbach’s alpha is 0.90. The individual responses were aggregated to team level (median $r_{WG} = 0.94$). The operationalizations of *team performance* as subjective performance and *value diversity* have been described previously in illustrative case 1 and 2, respectively.