Introducing a Comprehensive Sensitivity Analysis Tool for Meta-Analytic Reviews

Authors

James Field, West Virginia U., jamesfield6912@gmail.com
Frank A. Bosco, Virginia Commonwealth U., aom@frankbosco.com
Sven Kepes, Virginia Commonwealth U., skepes@vcu.edu
Michael A McDaniel, Virginia Commonwealth U., mamcdani@vcu.edu
Sheila List, Virginia Commonwealth U., lists@vcu.edu

James Field, West Virginia U., jamesfield6912@gmail.com
Frank A. Bosco, Sven Kepes, Michael A McDaniel, Sheila List

Introducing a Comprehensive Sensitivity Analysis Tool for Meta-Analytic Reviews

ABSTRACT

Meta-analytic reviews are considered the primary means for generating cumulative scientific knowledge and their results are often used by practitioners to inform evidence-based practice. However, the robustness of meta-analytic summary estimates is rarely examined. Consequently, the results of published meta-analyses may be misestimated and, thus, untrustworthy. Outliers can inflate the amount of residual heterogeneity in meta-analytic datasets, which can lead to biased meta-analytic and publication bias analysis results. We introduce a tool that will help researchers to conduct a meta-analysis that adheres to recommended reporting standards and best practices. Specifically, we describe and demonstrate a comprehensive sensitivity analysis tool that can assist in accounting for outlier-induced heterogeneity when performing a meta-analysis and the corresponding publication bias analyses. In addition, we use a dataset from a recently published meta-analysis to illustrate the functionality of the comprehensive sensitivity analysis tool and assess the robustness of our cumulative scientific knowledge regarding the validity of personality as a predictor of employee performance. We also describe how the range of estimates returned by the comprehensive sensitivity analysis tool can be used to produce more trustworthy recommendations for practice. We conclude with consumer-centric science implications, limitations, and future directions.

Keywords:

Meta-analysis; publication bias, outliers

James Field, West Virginia U., jamesfield6912@gmail.com
Frank A. Bosco, Sven Kepes, Michael A McDaniel, Sheila List

Introducing a Comprehensive Sensitivity Analysis Tool for Meta-Analytic Reviews

Meta-analytic reviews are the primary way to summarize, integrate, and synthesize areas of research, which allows for the generation of cumulative knowledge (Schmidt & Hunter, 2015). Our current understanding of phenomena as well as their effects and relations, however, rests on the assumption that our cumulative scientific knowledge is robust (Kepes, Bennett, & McDaniel, 2014). Unfortunately, the trustworthiness of many of our literatures has been questioned (e.g., Bettis, 2012; Kepes & McDaniel, 2013) and recent evidence suggests that small-study effects, which can be due to the "file-drawer" problem or study characteristic heterogeneity, "are the most important source of bias in meta-analysis" (Fanelli, Costas, & Ioannidis, 2017, p. 3,717). Indeed, outliers and publication bias (PB) pose serious threats to the accuracy of meta-analytic results and conclusions and, thus, our cumulative scientific knowledge. Given that organizational scientists have come to rely heavily on meta-analyses to provide the building blocks for knowledge creation and theory building (Bosco, Uggerslev, & Steel, 2017) and practitioners often use their results to guide evidence-based practice (Kepes et al., 2014), any threat to the robustness of meta-analytic summary estimates should be worrisome. As such, it is important for meta-analysts to conduct sensitivity analyses, which allow researchers and practitioners to assess if meta-analytic conclusions and recommendations for practice are trustworthy.

SENSITIVITY ANALYSES

Sensitivity analyses generally address the following question: "What happens [to the results] if aspects of the data or analyses are changed?" (Greenhouse & Iyengar, 2009, p. 418). In the meta-analytic context, sensitivity analyses take as input a collection of primary study data (e.g., correlations and sample sizes) and address whether or not the results are influenced by, for instance, extreme values (i.e., outliers) or distribution irregularities (i.e., asymmetry, skew), the former potentially being a sign of study characteristic and latter of the "file drawer" problem (Fanelli et al., 2017). Meta-analytic results and conclusions are considered to be more trustworthy if they do not "noticeably" change (i.e., differ by less than 20%; Kepes, Banks, McDaniel, & Whetzel, 2012) after aspects of the data or analyses are altered. Two types of sensitivity analyses for meta-analytic studies concerns examining the effect of outliers and PB on the obtained results.

Outliers

An *outlier* is an observation that appears "to deviate markedly from other members of the sample in which it occurs" (Grubbs, 1969, p. 1). The potential causes of outliers in the meta-analytic context are numerous. Table 1 contains a taxonomy of causes of outliers, which, similar to Kepes, Banks, McDaniel, and Whetzel's (2012) taxonomy of causes of PB, is differentiated between outcome-level and sample-level causes. Outcome-level causes of outliers refer to the role played by a sample's effect size magnitude and/or *p*-value in determining whether or not it is categorized as an outlier. For instance, samples that have an effect size and/or *p*-value that diverges from (i.e., is much larger or smaller than) all other samples in the dataset may need to be removed before performing a meta-analysis as they could introduce residual heterogeneity that may threaten its results (Kepes & McDaniel, 2015). With regard to sample-level causes of outliers, an effect size's corresponding sample size may play an important role in determining whether or not it is an outlier (see Table 1). Given that both the Hedges and Olkin (1985; see also Hedges & Olkin, 2014) and Schmidt and Hunter (2015) approaches to meta-analysis estimate the meta-analytic mean by giving more precise studies more weight, large samples can have an undue influence on the meta-analytic results and conclusions. Outliers may also be caused by study characteristic heterogeneity (Fanelli et al., 2017). For example, an effect size that differs from all other effect sizes in regard to some sample type characteristic (e.g., incumbents vs. applicants, employees vs. students) may need to be removed before performing a meta-analysis as it could introduce residual heterogeneity that may threaten its results and conclusions. This may be especially true if theoretical evidence suggests the sample characteristic is a boundary condition.

Taken together, outlier-induced heterogeneity presents a central challenge to conducting a meta-analysis as it can distort meta-analytic summary estimates (e.g., the mean estimate and the associated standard deviation) and, thus, the validity of conclusions from meta-analytic reviews (Ada, Sharman, & Balkundi, 2012; Viechtbauer & Cheung, 2010). Given the importance of meta-analytic reviews for establishing a cumulative knowledge (Schmidt & Hunter, 2003), new theoretical developments (Viswesvaran & Sanchez, 1998), and evidence-based practice (Kepes et al., 2014), it is important for researchers and practitioners to be able to assess the effect of outliers on meta-analytic results and conclusions. Indeed, failing to detect and, if present, remove outliers from meta-analytic datasets may lead to poor evidence-based practice recommendations, which, if implemented by practitioners, could yield unexpected results and, thus, widen the science-practice gap (Rousseau, 2012).

Insert Table 1 about here

Publication bias

Publication bias (PB) occurs when there is a systematic suppression of research findings, which causes the available literature to be unrepresentative of all completed research on a relation of interest (Begg & Mazumdar, 1994). Kepes et al. (2012) suggested that author

James Field, West Virginia U., jamesfield6912@gmail.com
Frank A. Bosco, Sven Kepes, Michael A McDaniel, Sheila List

decisions, the editorial review process, and organizational constraints are contributing factors to outcome-level and sample-level causes of PB. For example, an author may choose not to write up or report certain outcomes (e.g., statistically nonsignificant findings) when submitting a study to a journal. Such a selection process increases the prevalence of Type I error and limits efforts to assess the body of knowledge on a particular topic because null results, typically from small sample studies, tend to be suppressed from the scholarly community (Fanelli et al., 2017; Franco, Malhotra, & Simonovits, 2014).

Combined Effect of Outliers and Publication Bias

Although evidence suggests that outliers and PB can have independent adverse downstream effects for research and practice (Kepes et al., 2014), there appears to be some degree of interdependence between the causes of outliers and the causes of PB. For instance, an effect size may be removed from a manuscript before being submitted to a journal (i.e., author decision, outcome-level cause of PB; Kepes et al., 2012) because its corresponding *p*-value (i.e., outcome-level cause of outliers; see Table 1) was greater than the conventional statistical significance threshold ($p < .05$). In this case, an outlier-related phenomenon causes PB. Yet, to date, and to the best of our knowledge, sensitivity analyses of published meta-analytic results have failed to examine the combined effect of these phenomena (except for Kepes & McDaniel [2015], the only exception in the organizational sciences that we are aware of).

Furthermore, research from the medical sciences indicates that heterogeneity, which increases when outliers are included in meta-analytic datasets (Viechtbauer & Cheung, 2010), may limit the efficacy of PB detection methods for assessing the robustness of meta-analytic findings (Peters, Sutton, Jones, Abrams, & Rushton, 2007; Terrin, Schmid, Lau, & Olkin, 2003). As such, outliers and PB can have an interdependent effect as well as independent effects. Put differently, outlier-induced heterogeneity presents another central challenge to conducting a meta-analysis that adheres to recommended standards (e.g., *American Psychological Association's* [APA] Meta-Analytic Reporting Standards [2010]) and best practices (Kepes, McDaniel, Brannick, & Banks, 2013) as it can distort PB results (e.g., the meta-analytic mean effect size estimate adjusted for PB). Given the strong influence of meta-analytic reviews on research agendas and evidence-based practice decisions (Kepes et al., 2014), this should be worrisome as it suggests that previous attempts to assess the trustworthiness of our cumulative scientific knowledge (e.g., PB detection analyses) may themselves be untrustworthy.

The purpose of our manuscript is to introduce a comprehensive sensitivity analysis tool (CSAT) that will help scholars overcome the two aforementioned challenges to conducting a meta-analysis: accounting for the effect of outlier-driven heterogeneity when estimating meta-analytic parameters and performing the corresponding PB detection analyses. The remainder of the manuscript is arranged as follows. First, we explain why a CSAT is needed and how it will benefit organizational researchers and practitioners. Next, we provide an overview of the CSAT by briefly reviewing its instructions for use, the sensitivity analyses it performs, and the output it provides. Following this, we demonstrate the utility of the CSAT. Specifically, using a dataset from a recently published meta-analysis, we illustrate how the CSAT can be used to easily determine the degree to which meta-analytic and publication bias analysis results change after removing outlier-driven heterogeneity. We conclude with a discussion of customer-centric science (Aguinis, Werner, Abbott, Angert, Park, & Kohlhausen, 2010) and science-practice gap implications, limitations, and future directions for platforms like the CSAT.

TRANSPARENCY IN REPORTING META-ANALYTIC RESULTS NEEDS TO IMRPOVE

Although it is recommended that researchers examine datasets for the presence of outliers and, if present, remove them prior to conducting a meta-analysis (Ada et al., 2012; Viechtbauer & Cheung, 2010), only about 3% of published meta-analytic reviews in the organizational sciences conduct empirical assessments of outliers (Aguinis, Dalton, Bosco, Pierce, & Dalton, 2011). Concomitantly, several reporting standards (e.g., APA's Meta-Analytic Reporting Standards [2010], Preferred Reporting Items for Systematic Reviews and Meta-Analyses [PRISMA; Moher, Liberati, Tetzlaff, & Altman, 2009]) highlight the importance of conducting publication bias analyses to assess the robustness and, thus, trustworthiness of meta-analytic findings and conclusions. Given that Ferguson and Brannick (2011) reported that 40% of all published meta-analyses in psychology are affected by PB, one would assume that examining the extent of PB and the degree to which it threatens meta-analytic results and conclusions would be common practice. However, less than 4% of published meta-analytic reviews use the recommended techniques to assess the effect of this phenomenon (Banks, Kepes, & McDaniel, 2012).

Furthermore, the combined effect of outliers and PB, or how PB results change after outlier-driven heterogeneity is removed from the meta-analytic dataset, has not been examined by organizational scientists (see Kepes & McDaniel [2015] for the sole exception that we are aware of). Therefore, the extent to which outlier-induced heterogeneity threatens the validity of meta-analytic *and* PB results and conclusions is largely unknown. This uncertainty has the potential to become a major problem for organizational scientists. Given that scientific output is growing at an exponential rate (Bornmann & Mutz, 2015), the need for quantitative

James Field, West Virginia U., jamesfield6912@gmail.com
Frank A. Bosco, Sven Kepes, Michael A McDaniel, Sheila List

methodologies, like meta-analyses, that integrate and synthesize research areas is unlikely to diminish. As the collection of scientific findings grows, meta-analysts are faced with an increasing pressure to deliver trustworthy cumulative knowledge summaries.

Given that meta-analytic relations often serve as proxies for the "building blocks of theory" (Schmidt, 1992, p. 1177), failing to deliver trustworthy cumulative knowledge summaries could lead result in meta-analysis being used to perpetuate pseudotheories, "the scientific equivalent of fool's gold … [and] the complete opposite of what other fields require for a theory" (Cucina & McDaniel, 2016, p. 1117). This will likely have damaging downstream effects for both science and practice. With regard to science, the promotion of relatively unimportant theories complicates the theoretical landscape unnecessarily (Leavitt, Mitchell, & Peterson, 2010), making it difficult to separate signal from noise and to build a trustworthy cumulative scientific knowledge. For practitioners, an overabundance of inconsequential theory inhibits their ability to assess the generalizability of scientific findings and, thus, adds credence to the notion that organizational researchers are unable to leverage meta-analytic evidence to bridge the science-practice gap (Rynes, Giluk, & Brown, 2007). As such, there appears to be a need for a tool that takes a comprehensive approach to sensitivity analyses, one that accounts for outlier-induced heterogeneity when performing a meta-analysis *and* the corresponding publication bias analyses.

Unfortunately, as previously mentioned, outlier and PB analyses are rarely conducted (Aguinis et al., 2011; Banks et al., 2012) and assessments which take both of these phenomena into account are almost completely nonexistent (see Kepes & McDaniel [2015] for the sole exception that we are aware of). There are likely many contributing factors that help to explain why outlier and/or PB detection analyses are conducted so infrequently. One possible explanation is that meta-analysts lack the expertise to conduct sensitivity analyses and are not motivated to augment their knowledge of meta-analytic procedures because it would take too long to do so. Going forward, sensitivity analysis reporting rates may improve if meta-analysts can access an open-source user-friendly tool that removes these types of barriers. Indeed, such a tool will help meta-analysts to determine the range of estimates in which the "true" meta-analytic mean effect size can be found, which can be used by practitioners to inform lower and upper bound utility analysis estimates (e.g., Hancock, Allen, Bosco, McDaniel, & Pierce, 2013) and, thus, yield more trustworthy return on investment expectations for practitioners.

DESCRIPTION OF THE COMPREHENSIVE SENSITIVITY ANALYSIS TOOL

Sensitivity Analysis Techniques

The CSAT uses a battery of recommended methods for the empirical assessment of outliers and PB. In addition to estimating meta-analytic parameters using the Hedges and Olkin (1985; see also Hedges & Olkin, 2014) approach to meta-analysis, the CSAT performs two outlier detection assessments (one-sample removed analysis [Borenstein, Hedges, Higgins, and Rothstein, 2009] and Viechtbauer and Cheung's [2010, see also Viechtbauer, 2015] multivariate, multidimensional influence diagnostics) and five PB detection assessments (contour-enhanced funnel plots [Peters, Sutton, Jones, Abrams, & Ruston, 2008], Duval and Tweedie's [2000; 2005] trim and fill models, cumulative meta-analysis (CMA) by precision [Kepes et al., 2012], a priori selection models [Vevea & Woods, 2005], and precision-effect test-precision effect estimate with standard error analysis [PET-PEESE; Stanley & Doucouliagos, 2014]). Importantly, the CSAT returns meta-analytic and PB analysis results before and after outlier removal[1]. This is

[1] Viechtbauer and Cheung's [2010, see also Viechtbauer, 2015] influence diagnostics procedure is conducted in an iterative fashion to ensure that all potential outliers from the respective meta-analytic distribution are identified and removed. Following their removal, meta-analytic and PB results are re-estimated.

advantageous as it allows users to assess the effect of outlier-driven heterogeneity on the range of meta-analytic mean estimates and, thus, determine if a greater threat to the trustworthiness of their results and conclusions arises from outliers or PB.

The CSAT uses the meta-analytic approach developed by Hedges and Olkin (1985; Hedges & Vevea, 1998) as most sensitivity analysis techniques have not be developed for psychometrically-adjusted effect sizes (Schmidt & Hunter, 2015). As such, most analyses are conducted using Fisher's *z* transformed Pearson correlation coefficients. This is advantageous because it creates a symmetrical sampling distribution (Kepes & McDaniel, 2015). Before reporting, all obtained results are back-transformed into Pearson's *r* for interpretation purposes when analyses were conducted using *z*. The PET-PEESE and one-sample removed analyses are conducted using untransformed correlation coefficients. All analyses rely on the R Statistics package "metafor" and the DerSimonian and Laird estimation method (Viechtbauer, 2015), except for a priori selection model analyses, which are conducted using R syntax developed by Field and Gillett (2010).

Insert Table 2 about here

A detailed account of the methods employed by the CSAT is beyond the scope of this manuscript. However, Kepes et al. (2012) provided detailed descriptions of contour-enhanced funnel plots, both fixed effects (FE) and random effects (RE) trim and fill models, CMA by precision, and a priori selection models. We direct the reader to Stanley and Doucouliagos (2014) for a description of the PET-PEESE analysis, to Borenstein et al. (2009) for an overview of the one-sample removed analysis, and to Viechtbauer and Cheung (2010; see also Viechtbauer, 2015) for a discussion of the influence diagnostics method. We also note that the

CSAT follows established recommendations for trim and fill (Kepes et al., 2012), CMA by precision (Kepes et al., 2012), and a priori selection models (Vevea & Woods, 2005). Specifically, it employs the fixed-effects (FE) model and L_0 estimator to implement trim and fill and assesses the robustness of these results by also examining the random-effects (RE) model with the L_0 estimator (Moreno, Sutton, Turner, Abrams, Cooper, Palmer et al., 2009). Following recommendations by Stanley, Jarrell and Doucouliagos (2010), the CSAT reports the meta-analytic mean of the five most precise effect sizes. In addition, it uses a priori selection models with the *p*-value cut-points to model moderate and severe instances of PB as recommended by Vevea and Woods (2005). Table 2 provides a list of all analyses performed by the CSAT.

User Instructions and Features

In this section, we demonstrate the functionality of the CSAT. Meta-analysts can access the preliminary graphical user interface (GUI) at http://meta-analysis.shinyapps.io/sensitivityShiny/. Figure 1 displays the preliminary CSAT GUI, which currently relies on an RShiny framework. An inspection of Figure 1 reveals that user instructions are provided on the landing page (see “A” in Figure 1). It is strongly recommended that individuals read the provided instructions before utilizing the CSAT as they specify requirements for use (e.g., the dataset must a column named “r” [lowercase; represents the raw correlation coefficient]). In addition, Figure 1 shows where the user can browse for and upload a meta-analytic dataset (see “B”) as well as three tabs, which display the comprehensive sensitivity analysis results before and after outlier removal (see “C”), the uploaded dataset with outlier classification (see “D”), and the corresponding PB analysis plots (see “E”).

Insert Figure 1 about here

The CSAT process unfolds as follows. We note that a sample CSV input data file (see "F" in Figure 2) is provided to help users replicate the process description that follows. After reading the user instructions (see "A" in Figure 1), the user uploads a meta-analytic dataset. A progress bar notifies the user when their meta-analytic dataset is uploaded. After successfully uploading the meta-analytic dataset, the CSAT echoes back the filename of the uploaded file above the progress bar and, thus, alerts the user if an incorrect file was uploaded (see "G" in Figure 2). Following this, the user will be able to initiate the comprehensive sensitivity analysis by pressing on the "Run analyses" button (see "H" in Figure 2), which does not appear in the interface until a data file uploads successfully. Upon clicking the "Run analyses" button, a progress bar will appear to inform the user the status of the analyses (see "I" in Figure 2). The comprehensive sensitivity analyses are completed and plots are generated when the progress bar disappears. At this point, the user can move to the "Results" tab (see "C" in Figure 1) where the parameter labels (e.g., fixed effects trim and fill: adjusted meta-analytic mean effect size estimate; see "J" in Figure 3) and the corresponding results before (see "K" in Figure 3) and after (see "L" in Figure 3) outlier removal are reported. In addition, the "Results" tab allows the user to export a CSV file that contains the comprehensive sensitivity analysis results table by clicking on the "Download results" button (see "M" in Figure 3).

Insert Figures 2 and 3 about here

After the analyses have been successfully conducted, the user can move to the "Raw data" tab, where they can view which, if any, effect size(s) in their meta-analytic dataset were identified by Viechtbauer and Cheung's 2010, see also Viechtbauer, 2015] influence diagnostics as being an outlier (see "N" in Figure 4). Effect sizes identified as not being an outlier (i.e.,

James Field, West Virginia U., jamesfield6912@gmail.com
Frank A. Bosco, Sven Kepes, Michael A McDaniel, Sheila List

marked with a “No” in the last column on the right-hand side of this table) are used to reevaluate the meta-analytic and sensitivity analysis parameters after outlier-driven heterogeneity is removed from the meta-analytic dataset and are reported in “After Outlier Removal” column of the “Results” tab (see “L” in Figure 3). Finally, Figure 5 shows the full view of the “Plots” tab, which displays the following figures, before (top panel) and after (bottom panel) outlier removal: FE trim and fill model funnel plot (see “O” and “P”), RE trim and fill model funnel plot (see “Q” and “R”), cumulative meta-analysis by precision forest plot (see “S” and “T”), and contour-enhanced funnel plot (see “U” and “V”). Importantly, each figure can be saved individually.

Insert Figures 4 and 5 about here

Taken together, the CSAT can assist in tackling two central challenges to conducting a meta-analysis that adheres to recommended reporting standards (e.g., *APA*’s Meta-Analytic Reporting Standards [2010]) and best practices (Kepes et al., 2013). Specifically, the CSAT takes as input a meta-analytic dataset and returns two sets of meta-analytic and sensitivity analysis results, one with outliers included and the other without outliers included. As such, the CSAT allows users to assess the effect of outlier-driven heterogeneity on meta-analytic *and* sensitivity analysis results and, thus, has the potential to mitigate some of the biggest threats to building a robust cumulative scientific knowledge (Fanelli et al., 2017).

Reporting and Interpretation of Output

We urge caution when interpreting the CSAT’s results in isolation. Indeed, a *non causa pro causa* can be avoided if researchers do not rely on the result of any one sensitivity analysis technique alone as the conditions needed (e.g., specific level of heterogeneity, bias) for each sensitivity analysis technique to achieve optimal performance is still unknown (Macaskill, Walter, & Irwig, 2001; van Assen, van Aert, & Wicherts, 2015). Therefore, it is recommended that researchers employ a variety of methods, as the CSAT does, to inform their sensitivity analysis conclusions (Kepes & McDaniel, 2015). Indeed, estimating the possible range of meta-analytic mean effect size estimates instead of relying on a single one is aligned with the concept of triangulation, which refers to the use of "multiple reference points to locate an object's exact position (Jick, 1979, p. 602; see Orlitzky, 2012). Taken together, we do not advise users to "cherry-pick" sensitivity analysis results from the CSAT output. Instead, in the interest of scientific transparency and customer-centric science (Aguinis et al., 2010), we recommend that users report *all* meta-analytic and sensitivity analysis results returned by the CSAT. If the sensitivity analysis results converge on a mean that is noticeably different (i.e., by more than 20%; see Kepes et al., 2012) from the original meta-analytic mean effect size estimate (i.e., before outlier removal; see "K" in Figure 3), it can be concluded that the original meta-analytic mean estimate is likely non-robust and, thus, untrustworthy. As such, CSAT users should report their meta-analytic and sensitivity analysis results in terms of convergence on the originally obtained meta-analytic mean effect size estimate (i.e., before outlier removal).

DEMONSTRATION OF THE COMPREHENSIVE SENSITIVITY ANALYSIS TOOL

Methods

To illustrate the functionality of the CSAT, we examine whether or not outliers and/or PB threatens the trustworthiness of our cumulative scientific knowledge on the role played by personality in predicting employee performance. Specifically, we obtained Shaffer & Postlethwaite's (2012) data on the validity of the Big Five personalities traits. We note that Shaffer & Postlethwaite's (2012) meta-analytic dataset was selected for the purpose of demonstrating the CSAT because (1) it was available in an appendix with the published article,

James Field, West Virginia U., jamesfield6912@gmail.com
Frank A. Bosco, Sven Kepes, Michael A McDaniel, Sheila List

(2) the original authors did not examine if outlier-driven heterogeneity threatened the validity of their meta-analytic results, (3) the original authors did not perform recommended PB detection tests[2], and (4) the trustworthiness of our cumulative scientific knowledge on this literature (i.e., personality-employee performance) has never been assessed[3]. We do not repeat our description of the aforementioned CSAT process and, instead, only report the comprehensive sensitivity analysis results it produces.

In Table 3 we report the results of our reanalysis of the main effect distributions (e.g., "emotional stability-employee performance"). In addition, we report the CSAT results for the corresponding "noncontextualized" and "contextualized" distributions (e.g., "emotional stability: noncontextualized-employee performance) because "the purpose of [the original] meta-analysis was to examine the relative validity of contextualized and noncontextualized measures of self-report personality" (Shaffer & Postlethwaite, 2012, p. 464). However, due to space constraints, we only describe the CSAT results for the "emotional stability-employee performance" meta-analytic distribution. Although not described in the following sections, the results for the remaining distributions can also be found in Table 3.

We note that the original $\bar{r}_{oRE}$ (i.e., before outlier removal) for the "emotional stability-employee performance" distribution reported in Table 3 differed slightly from the one reported by Shaffer and Postlethwaite (2012) (.098 vs. .090; Δ = |.008|). Given that the *k* and *N* values returned by the CSAT matched the ones reported by Shaffer and Postlethwaite (2012), we

[2] Shaffer and Postlethwaite's (2012) meta-analysis was published after Rothstein et al.'s (2005) book on publication bias. It is reasonable to assume that recommended publication bias detection methods could have been used in their article. However, their meta-analysis was published around the same time as Kepes et al.'s (2012) introduction of publication bias methods to the organizational sciences. Therefore, it is possible that the authors may not have been privy to most of the publication bias methods used in our study because they were not yet explicitly introduced to the organizational sciences.
[3] Kepes and McDaniel (2015) examined the trustworthiness of the meta-analytic results for the "conscientiousness-employee performance" distribution that were originally reported by Shaffer and Postlethwaite (2012). However, they did not examine the trustworthiness of the validity of the other Big Five dimensions.

conclude that the observed difference in original $\bar{r}_{o_{RE}}$ can be explained by the fact that the original authors used psychometric meta-analyses (Schmidt & Hunter, 2015). In contrast, the CSAT employs the Hedges and Olkin (1985; Hedges & Vevea, 1998) approach to meta-analysis. A discussion of the differences between approaches to meta-analysis is beyond the scope of the current manuscript. However, we note that Kepes et al. (2013) provided an account of the differences between the Hedges and Olkin (1985; Hedges & Vevea, 1998) and psychometric meta-analyses (Schmidt & Hunter, 2015) approaches to meta-analysis.

Results

The meta-analytic and sensitivity analysis results returned by the CSAT for the main effect distributions (“conscientiousness-employee performance,” “agreeableness-employee performance,” “emotional stability-employee performance,” “openness-employee performance,” and “extraversion-employee performance”) as well as their corresponding “contextualized” versus “noncontextualized sub-distributions that were originally examined by Shaffer and Postlethwaite (2012) are reported in Table 3. Before and after outlier removal results are reported in the top and bottom panels of Table 3, respectively. The first two columns report the distribution label and its size (i.e., number of samples [k]). Columns three through nine contain the meta-analytic results, including the random-effects (RE) meta-analytic mean observed correlation ($\bar{r}_{o_{RE}}$), the 95% confidence interval (95% CI), the 90% prediction interval (90% PI), Cochran’s Q, I^2, tau (τ), and the results of the one-sample removed analysis (osr; minimum, maximum, and median mean estimates). Columns 10-17 contain the results from the trim and fill analyses (columns 10-13 contain the results for the recommended FE trim and fill model and columns 14-17 the RE trim and fill results). Columns 18 and 19 present the results from the one-tailed moderate ($\text{sm}_\text{m}\,\bar{r}_o$) and severe selection ($\text{sm}_\text{s}\,\bar{r}_o$) models, respectively. Column 20 shows the meta-analytic estimate for the five most precise samples (*pr* $\bar{r}_o$). The last column, column 21, reports the PET-PEESE adjusted mean effect size estimate (*pp* $\bar{r}_o$).

The CSAT results reported in Table 3 indicate that the originally reported meta-analytic mean effect size estimate ($\bar{r}_{o_{RE}}$ = .098, k = 86) for the "emotional stability-employee performance" relation is likely misestimated. Although the original $\bar{r}_{o_{RE}}$ was robust to the one-sample removed analysis before outlier removal (i.e., all three osr estimates were practically identical to the original $\bar{r}_{o_{RE}}$), the PB analyses indicated that it is likely to be untrustworthy. As an example, the FE trim and fill model imputed 17 samples on the left-hand side of the funnel plot before outlier removal, which yielded an adjusted mean estimate of .060, a difference of .038 or 39% to the original $\bar{r}_{o_{RE}}$. A similar pattern of results before outlier removal was observed for the RE trim and fill model (t&f$_{RE}$ $\bar{r}_o$ = .069), a priori selection model with moderate PB assumptions (sm$_m$ $\bar{r}_o$ = .072), the meta-analytic estimate based on the five most precise effects (*pr* $\bar{r}_o$ = .000), and the PET-PEESE analysis (*pp* $\bar{r}_o$ = .034), all of which indicated the original meta-analytic mean effect size estimate was overestimated. The result for the a priori selection model with severe PB assumptions was nonsensical due to extremely large variance and thus is not reported. Overall, the before outlier removal results returned by the CSAT provided somewhat conflicting results. Put differently, the one-sample removed analyses indicated that the "emotional stability-employee performance" relation is robust whereas the PB analyses suggested that the magnitude of the original $\bar{r}_{o_{RE}}$ is likely to be overestimated.

Insert Table 3 about here

Although the one-sample removed analysis before outlier removal indicated that the original $\bar{r}_{o_{RE}}$ is robust, Viechtbauer and Cheung's (2010; Viechtbauer, 2015) multivariate, multidimensional influence diagnostics identified four outliers in the original "emotional stability-employee performance" distribution. Interestingly, the $\bar{r}_{o_{RE}}$ changed only slightly after removing the identified outliers (.095, $\Delta = |.003|$). However, perhaps more importantly, removing the outliers reduced the degree of heterogeneity; the 90% PI (.022, .166) was narrower and Q, I^2, and τ were substantially smaller in their respective magnitude. Such reductions in heterogeneity should improve the performance of the PB detection techniques (Kepes & McDaniel, 2015; Terrin et al., 2003). After removing the identified outliers, the PB results returned by the CSAT still indicated that the originally reported $\bar{r}_{o_{RE}}$ is likely to be overestimated (see Table 3). Specifically, following outlier removal each meta-analytic mean estimate adjusted for the effect of PB is smaller in magnitude than the original $\bar{r}_{o_{RE}}$. However, the CSAT results indicate that the PB analysis results after outlier removal converged better with the original $\bar{r}_{o_{RE}}$ than the corresponding PB analysis results before outlier removal. Put differently, a comparison of the CSAT results, before and after outlier removal, indicates that the degree of PB may have been overestimated in the original analyses (i.e., when the four identified outliers were included in the meta-analytic dataset). Taken together, we can conclude that the validity of emotional stability is likely to be overestimated and that outlier-driven heterogeneity affected the performance of the PB detection techniques, causing them to overestimate the effect of this PB.

Figure 6, which displays funnel plots for the FE (see "W" and "X") and RE (see "Y" and "Z") trim and fill models, CMA by precision forest plots (see "AA" and "BB"), and contour-enhanced funnel plots (see "CC" and "DD") for the Shaffer and Postlethwaite's (2012) "emotional stability-employee performance" dataset, before and after outlier removal, adds credence to the claim that outlier-induced heterogeneity affected the performance of the PB detection techniques. For example, an inspection of the contour-enhanced funnel plot suggests

James Field, West Virginia U., jamesfield6912@gmail.com
Frank A. Bosco, Sven Kepes, Michael A McDaniel, Sheila List

that, before the removal of outliers, 77% (13/17) of the imputed samples were in the area of statistical insignificance (see "CC" in Figure 6). However, the contour-enhanced funnel plot after outlier removal suggests that the effect of PB is attenuated after removing outliers from the "emotional stability-employee performance" distribution. Specifically, the contour-enhanced funnel plot following outlier removal displays noticeably more symmetry and the trim and fill model imputed only eight samples to achieve symmetry, nine fewer than when outliers were included in the distribution. This indicates that the degree of asymmetry was reduced after outlier removal. Therefore, holding all else constant, outlier-driven heterogeneity affected the degree of symmetry in the funnel plot and, consequently, the meta-analytic and PB results.

Insert Figure 6 about here

Discussion

Although recent research indicates that PB and outliers can distort meta-analytic results (e.g., Ada et al., 2012; Banks, Kepes, & McDaniel, 2015; Kepes & McDaniel, 2015; Viechtbauer & Cheung, 2010), analyses are rarely conducted to assess the effects of these phenomena (Kepes et al., 2013). Furthermore, analyses that examine the combined effect of outliers and PB are practically nonexistent in the existing literature. Consequently, the potential non-robustness of meta-analytic results and their associated conclusions often goes undetected, which brings into question the trustworthiness of our cumulative knowledge. To address this concern, we introduced and demonstrated a comprehensive sensitivity analysis tool that can assist in accounting for outlier-driven heterogeneity when performing a meta-analysis and the corresponding PB analyses. Specifically, we described the features of the CSAT, an open-access online platform that performs a meta-analysis and a battery of outlier and PB detection analyses, which allows users to easily assess the range of estimates in which the "true" meta-analytic mean effect size may be found. In addition, we demonstrated the functionality of the CSAT by using it to assess the trustworthiness of our cumulative scientific knowledge on the validity of personality (e.g., emotional stability) as a predictor of employee performance. In the remainder of this section, we describe how adoption of the CSAT by producers, publishers, and consumers of science will serve the goals of customer-centric science (Aguinis et al., 2010), recommended standards (e.g., *APA's* Meta-Analytic Reporting Standards [2010]), and best practice guidelines (Kepes et al., 2013). In addition, we describe several limitations of the CSAT before concluding with a discussion of future opportunities for platforms like the CSAT.

Implications for Research and Practice

With regard to implications for science, we hope that researchers will utilize the CSAT when conducting future systematic reviews and meta-analyses. Indeed, PB has been identified as the potentially greatest threat to the trustworthiness of our cumulative knowledge (Rothstein, Sutton, & Borenstein, 2005) and has also been referred to as the "kryptonite of evidence-based practice" (Banks & McDaniel, 2011, p. 40). In addition, outliers can inflate the amount of residual heterogeneity into a meta-analytic dataset, which can lead to biased meta-analytic results and conclusions (Viechtbauer & Cheung, 2010). Yet, some have suggested that sensitivity analyses, especially regarding outliers (Aguinis, Pierce, Bosco, Dalton, & Dalton, 2011) and PB (Dalton, Aguinis, Dalton, Bosco, & Pierce, 2012) may be irrelevant. The CSAT can help researchers to examine the effect of outliers and PB, as well as their combined effect, on meta-analytic results. We encourage future researchers to incorporate the CSAT into their future meta-analytic studies as it will help to determine whether or not the aforementioned claims regarding the irrelevancy of sensitivity analyses are true or are merely urban myths.

We contend that the CSAT should be integrated into future meta-analyses as it will increase the transparency of scientific findings, which is aligned with the idea of customer-centric science (Aguinis et al., 2010). Indeed, the CSAT returns a range of meta-analytic mean effect size estimates that can be used collectively to triangulate the potentially best estimate of the "true" population effect size. Furthermore, such ranges may be used by practitioners to inform lower and upper bound utility analysis estimates (e.g., Hancock et al., 2013), which could yield more trustworthy return on investment expectations and, thus, help to narrow the science-practice gap. Indeed, an inspection of the results reported in Table 3 illustrates the efficacy of the CSAT for informing evidence-based practice recommendations.

For instance, with regard to emotional stability, the authors claimed in the original meta-analysis that "the magnitude of the validity of contextualized measures was a least twice that of noncontextualized measures" (Shaffer & Postlethwaite, 2012, p. 465). However, an examination of the sensitivity analysis results indicates that the magnitude of the difference may be much larger. For example, the FE trim and fill model result before outlier removal for the "emotional stability: noncontextualized-employee performance" distribution was .045. In contrast, it was .230 for the emotional stability: contextualized-employee performance" distribution, a difference of 411%, which is much large than the difference originally reported by Shaffer and Postlethwaite (2012). We note that similar differences were observed for the RE trim and fill model, a priori selection model with moderate PB assumptions, cumulative meta-analysis by precision, and PET-PEESE before outlier removal. This discrepancy should be worrisome as it might lead practitioners to implement ill-informed evidence-based practice recommendations and, thus, achieve unexpected returns on investment, which could widen the science-practice gap.

In addition, our results indicate that contextualized measures of emotional stability are the most valid predictor of job performance. Specifically, the t&f$_{FE}$ $\bar{r}_o$, t&f$_{RE}$ $\bar{r}_o$, *pr* $\bar{r}_o$, and *pp* $\bar{r}_o$ results before outlier removal for the “emotional stability: contextualized-employee performance” distribution suggest contextualized measures of emotional stability are better at predicting – have the strongest positive relation with – job performance than any other measure of the Big Five, even after the effect of outliers and PB is taken into consideration (see Table 3)[4]. This is surprising given that conscientiousness, not emotional stability, has been referred to as the “most important of the Big Five” (Dudley, Orvis, Lebiecki, & Cortina, 2006, p. 40), which may mean that practitioners are not using the optimal predictors of job performance when making personnel decisions.

To support this claim, we use Kepes and McDaniel’s (2015) utility formula to compute the dollar amount on using a suboptimal predictor of employee performance when making personnel selection decisions. Specifically, we use the FE trim and fill adjusted estimate before outlier removal for the “conscientiousness:contextualized” (t&f$_{FE}$ $\bar{r}_o$= .169) and “emotional stability: contextualized” (t&f$_{FE}$ $\bar{r}_o$= .230) distributions to estimate the potential cost of assuming that conscientiousness is a better predictor of job performance than emotional stability. Given the findings reported in Table 3, results from Kepes and McDaniel’s (2015) formula indicate that the utility value of the FE trim and fill adjusted estimate before outlier removal for the “emotional stability: contextualized” distribution is about $2,200,000 larger than the one for the “conscientiousness:contextualized” distribution[5]. Although the utility formula was originally used by Kepes and McDaniel (2015) to show that phenomena like outliers and PB may affect the

[4] We note that outliers were not detected in the “emotional stability: contextualized-employee performance” and “emotional stability: noncontextualized-employee performance” distributions, which is why the before outlier removal results are referenced for these distributions.
[5] We greatly appreciate Kepes and McDaniel’s willingness to share with us their utility formula.

James Field, West Virginia U., jamesfield6912@gmail.com
Frank A. Bosco, Sven Kepes, Michael A McDaniel, Sheila List

utility value of conscientiousness by $1,800,000, our sensitivity analysis results suggest that it may not be the optimal predictor of job performance among the Big Five. Indeed, this should be of major concern for organizations as it suggests that practitioners may be using flawed selection practices that are likely to yield weaker-than-expected results (Kepes and McDaniel, 2015). Taken together, these discrepancies illustrate why future meta-analyses should utilize the CSAT. Specifically, the tool introduced in this manuscript will help meta-analysts to account for the effect of outlier-induced heterogeneity on meta-analytic and PB results when making practical recommendations, which could help to provide more trustworthy return on investment estimates and, thus, narrow the science-practice gap (see Kepes & McDaniel, 2015).

We contend that it is high time for journals to play a more proactive role in helping to build more trustworthy cumulative knowledge. Although others (e.g., Banks et al., 2012; Kepes & McDaniel, 2015) have suggested that journals should make sensitivity analyses a prerequisite for the publication of meta-analytic reviews, the rate at which they are conducted remains low. To improve this state of affairs, we encourage publishers to require researchers who submit meta-analytic reviews to their journals to include a CSAT report in their manuscript. Alternatively, the CSAT report could be made available as supplementary material on the journal's website (Kepes & McDaniel, 2015). Such steps will increase the transparency of meta-analytic findings, which will help to improve the trustworthiness of our cumulative knowledge.

Limitations and Future Directions

Although the CSAT should satisfy one of Aguinis and Edwards' (2014, p. 143) methodological wishes for management research by furthering our "understanding [of] the nature and impact of outliers," a number of limitations must be shared. Currently the CSAT can conduct meta-analyses and sensitivity comprehensive sensitivity analyses using correlation coefficients as the effect size input. Indeed, the current version of the CSAT limits its utility as many different types of effect sizes (e.g., Cohen's *d*, odds ratios) can be meta-analyzed. Still, correlation coefficients are used across a variety of research areas, particularly in the organizational sciences (e.g., organizational behavior, human resource management, strategic management), to build cumulative scientific knowledge bases. As such, the CSAT will likely be of use to many researchers across a number of research areas as it addresses previous cautions regarding the effect of PB and outliers (e.g., Kepes & McDaniel, 2015). Still, we note that the CSAT developers are actively working to expand the functionality of the interface to allow users to conduct a comprehensive sensitivity analysis of all types of meta-analytic data.

The analyses performed by the CSAT rely on the Hedges and Olkin (1985) approach to meta-analysis, not the Schmidt and Hunter (2015) approach. The latter is the universal approach to meta-analysis in the organizational sciences as it allows for corrections due to artefactual variance (e.g., unreliability in the dependent variable), which may affect the performance of the sensitivity analysis. We note that the CSAT employs the Hedges and Olkin (1985) approach as most sensitivity analysis methods have not been developed for psychometrically-adjusted effect sizes. For example, the PB detection methods are not accommodating to psychometric meta-analytic perspectives on study weighting (sample size vs. inverse variance weighting), the lack of effect size transformations (i.e., Fisher z), and their approach to sampling error estimation (i.e., estimate of rho in sampling error estimates). However, from a practical perspective, we note that the Hedges and Olkin (1985) and Schmidt and Hunter (2015) approaches tend to yield very similar, if not virtually identical, meta-analytic mean effect size estimates (Harrison, Banks, Pollack, O'Boyle, & Short, 2014; Kepes et al., 2013). Indeed, the observed convergence between the meta-analytic mean effect size estimates reported in Table 3 and the ones originally reported

James Field, West Virginia U., jamesfield6912@gmail.com
Frank A. Bosco, Sven Kepes, Michael A McDaniel, Sheila List

by Shaffer and Postlethwaite (2012) highlight this point[6]. Still, the CSAT team plans to adapt the battery of sensitivity analysis methods to the psychometric meta-analytic context (Kepes & McDaniel, 2015). This enhanced functionality would be valuable to not only assess the robustness of meta-analytic results of observed correlations but also the robustness of correlations that have been corrected for measurement error and/or range restriction.

Conclusion

Comprehensive sensitivity analyses are rarely conducted in the organizational sciences. In this manuscript, we introduce and demonstrate a comprehensive sensitivity analysis tool that can assist in accounting for outlier-induced heterogeneity when performing a meta-analysis and the corresponding publication bias analyses. We recommend that the tool be integrated into future meta-analytic reviews as it will help to assess the trustworthiness of their results and conclusions, which will fulfill the goals of customer-centric science (Aguinis et al., 2010) and best practice recommendations (Kepes et al., 2013).

[6] We observed an average difference of |.003| between the meta-analytic mean effect size estimates reported in Table 3 and the corresponding ones reported by Shaffer and Postlethwaite (2012).

REFERENCES

Ada, S., Sharman, R., & Balkundi, P. 2012. Impact of meta-analytic decisions on the conclusions drawn on the business value of information technology. ***Decision Support Systems***, 54(1): 521-533.

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. 2010. Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind ***Organizational Research Methods***, 13(3): 515-539.

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. 2011. Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. ***Journal of Management***, 37(1): 5-38.

Aguinis, H., Pierce, C. A., Bosco, F. A., Dalton, D. R., & Dalton, C. M. 2011. Debunking myths and urban legends about meta-analysis. ***Organizational Research Methods***, 14(2): 306-331.

Aguinis, H. & Edwards, J. R. 2014. Methodological wishes for the next decade and how to make wishes come true. ***Journal of Management Studies***, 51(1): 143-174.

American Psychological Association. 2010. ***Publication manual of the American psychological association***. Washington, DC: American Psychological Association.

Banks, G. C. & McDaniel, M. A. 2011. The kryptonite of evidence-based I–O psychology. ***Industrial and Organizational Psychology***, 4(01): 40-44.

Banks, G. C., Kepes, S., & McDaniel, M. A. 2012. Publication bias: A call for improved meta-analytic practice in the organizational sciences. ***International Journal of Selection and Assessment***, 20(2): 182-196.

Banks, G. C., Kepes, S., & McDaniel, M. A. 2015. Publication bias: Understanding the myths concerning threats to the advancement of science. In C. E. Lance & R. J. Vandenberg (Eds.), ***More statistical and methodological myths and urban legends.***: 36-64. New York, NY: Routledge.

Begg, C. B. & Mazumdar, M. 1994. Operating characteristics of a rank correlation test for publication bias. ***Biometrics***, 50(4): 1088-1101.

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

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. 2009. ***Introduction to meta-analysis***. West Sussex, UK: Wiley.

Bornmann, L. & Mutz, R. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. ***Journal of the Association for Information Science and Technology***, 66(11): 2215-2222.

Bosco, F. A., Uggerslev, K. L., & Steel, P. 2017. MetaBUS as a vehicle for facilitating meta-analysis. ***Human Resource Management Review***, 27(1): 237-254.

Cucina, J. M. & McDaniel, M. A. 2016. Pseudotheory proliferation is damaging the organizational sciences. ***Journal of Organizational Behavior***, 37(8): 1116-1125.

Dalton, D. R., Aguinis, H., Dalton, C. M., Bosco, F. A., & Pierce, C. A. 2012. Revisiting the file drawer problem in meta-analysis: An assessment of published and nonpublished correlation matrices. ***Personnel Psychology***, 65(2): 221-249.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. 2006. A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. ***Journal of Applied Psychology***, 91(1): 40-57.

Duval, S. & Tweedie, R. 2000. A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. ***Journal of the American Statistical Association***, 95(449): 89-98.

Duval, S. J. 2005. The “trim and fill” method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), ***Publication bias in meta-analysis: Prevention, assessment, and adjustments***: 127-144. West Sussex, UK: Wiley.

Fanelli, D., Costas, R., & Ioannidis, J. P. A. 2017. Meta-assessment of bias in science. ***Proceedings of the National Academy of Sciences***, 114(14): 3714-3719.

Ferguson, C. J. & Brannick, M. T. 2011. Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. ***Psychological Methods***: 120-128.

Field, A. P. & Gillett, R. 2010. How to do a meta-analysis. ***British Journal of Mathematical and Statistical Psychology***, 63(3): 665-694.

Franco, A., Malhotra, N., & Simonovits, G. 2014. Publication bias in the social sciences: Unlocking the file drawer. ***Science***, 345(6203): 1502.

Greenhouse, J. B. & Iyengar, S. 2009. Sensitivity analysis and diagnostics. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), ***The handbook of research synthesis and meta-analysis.***, 2nd ed.: 417-433. New York, NY: Russell Sage Foundation.

Grubbs, F. E. 1969. Procedures for detecting outlying observations in samples. ***Technometrics***, 11(1): 1-21.

Hancock, J. I., Allen, D. G., Bosco, F. A., McDaniel, K. R., & Pierce, C. A. 2013. Meta-analytic review of employee turnover as a predictor of firm performance. ***Journal of Management***, 39(3): 573-603.

Harrison, J., S., Banks, G. C., Pollack, J., M. , O'Boyle, E., H. , & Short, J. 2014. Publication bias in strategic management research. ***Journal of Management***, 43(2): 400-425.

Hedges, L. V. & Olkin, I. 1985. ***Statistical method for meta-analysis***. New York, NY: Academic press.

Hedges, L. V. & Vevea, J. L. 1998. Fixed- and random-effects models in meta-analysis. ***Psychological Methods***, 3(4): 486-504.

Jick, T. D. 1979. Mixing qualitative and quantitative methods: triangulation in action. ***Administrative Science Quarterly***, 24(4): 602-611.

Kepes, S., Banks, G. C., McDaniel, M. A., & Whetzel, D. L. 2012. Publication bias in the organizational sciences. ***Organizational Research Methods***, 15(4): 624-662.

Kepes, S. & McDaniel, M. A. 2013. How trustworthy is the scientific literature in industrial and organizational psychology? ***Industrial and Organizational Psychology: Perspectives on Science and Practice***, 6(3): 252-268.

Kepes, S., McDaniel, M. A., Brannick, M. T., & Banks, G. C. 2013. Meta-analytic reviews in the organizational sciences: Two meta-analytic schools on the way to MARS (the Meta-Analytic Reporting Standards). ***Journal of Business and Psychology***, 28(2): 123-143.

Kepes, S., Bennett, A. A., & McDaniel, M. A. 2014. Evidence-based management and the trustworthiness of our cumulative scientific knowledge: Implications for teaching, research, and practice. ***Academy of Management Learning & Education***, 13(3): 446-466.

Kepes, S. & McDaniel, M. A. 2015. The validity of conscientiousness is overestimated in the prediction of job performance. ***PLoS One***, 10(10): e0141468.

Leavitt, K., Mitchell, T. R., & Peterson, J. 2010. Theory pruning: Strategies to reduce our dense theoretical landscape. ***Organizational Research Methods***, 13(4): 644-667.

Macaskill, P., Walter, S. D., & Irwig, L. 2001. A comparison of methods to detect publication bias in meta-analysis. ***Statistics in Medicine***, 20(4): 641-654.

Moreno, S. G., Sutton, A. J., Turner, E. H., Abrams, K. R., Cooper, N. J., Palmer, T. M., & Ades, A. E. 2009. Novel methods to deal with publication biases: secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. ***British Medical Journal***, 339.

Orlitzky, M. 2012. How can significance tests be deinstitutionalized? ***Organizational Research Methods***, 15(2): 199-228.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. 2007. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. ***Statistics in Medicine***, 26(25): 4544-4562.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. 2008. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. ***Journal of Clinical Epidemiology***, 61(10): 991-996.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. 2005. Publication bias in meta-analyses. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), ***Publication bias in meta-analysis: Prevention, assessment, and adjustments***: 1-7. West Sussex, UK: Wiley.

Rousseau, D. M. 2012. Envisioning evidence-based management. In D. M. Rousseau (Ed.), ***The Oxford handbook of evidence-based management***. New York, NY: Oxford Universtiy Press.

Rynes, S. L., Giluk, T. L., & Brown, K. G. 2007. The very separate worlds of academic and practitioner periodicals in human resource management: Implications for evidence-based management. ***Academy of Management Journal***, 50(5): 987-1008.

Schmidt, F. L. 1992. What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. ***American Psychologist***, 47(10): 1173-1181.

Schmidt, F. L. & Hunter, J. E. 2003. History, development, evolution, and impact of validity generalization and meta-analysis methods, 1975-2001. In K. R. Murphy (Ed.), ***Validity generalization: A critical review.***: 31-65. Mahwah, NJ: Lawrence Erlbaum.

Schmidt, F. L. & Hunter, J. E. 2015. ***Methods of meta-analysis: Correcting error and bias in research findings.*** (3rd ed.). Newbury Park, CA: Sage.

Shaffer, J. A. & Postlethwaite, B. E. 2012. A matter of context: a meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. ***Personnel Psychology***, 65(3): 445-494.

Stanley, T. D., Jarrell, S. B., & Doucouliagos, H. 2010. Could it be better to discard 90% of the data? A statistical paradox. ***The American Statistician***, 64(1): 70-77.

Stanley, T. D. & Doucouliagos, H. 2014. Meta-regression approximations to reduce publication selection bias. ***Research Synthesis Methods***, 5(1): 60-78.

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. 2003. Adjusting for publication bias in the presence of heterogeneity. ***Statistics in Medicine***, 22(13): 2113-2126.

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. 2015. Meta-analysis using effect size distributions of only statistically significant studies. ***Psychological Methods***, 20(3): 293-309.

Vevea, J. L. & Woods, C. M. 2005. Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. ***Psychological Methods***, 10: 428-443.

Viechtbauer, W. & Cheung, M. W. L. 2010. Outlier and influence diagnostics for meta-analysis. ***Research Synthesis Methods***, 1(2): 112-125.

Viechtbauer, W. 2015. Meta-analysis package for R: Package ‘metafor.’ R package version 1.9-5.

Viswesvaran, C. & Sanchez, J. I. 1998. Moderator search in meta-analysis: A review and cautionary note on existing approaches. ***Educational and Psychological Measurement***, 58(1): 77-87.

TABLE 1

Taxonomy of Causes of Outliers

Cause of outliers	Explanation
Outcome-level causes	
Effect size magnitude	Samples that have an effect size that diverges from the effect sizes of all other samples in the dataset may need to be removed before performing a meta-analysis as they could introduce residual heterogeneity that may threaten its results and conclusions.
P-value	An effect size may be labelled as an outlier if its corresponding *p*-value deviates noticeably from the other *p*-values in the dataset. Failing to remove such effect sizes may increase the degree of heterogeneity observed in a dataset and thus threaten its meta-analytic results.
Sample-level causes	
Sample size	Sample size is a characteristic that may determine whether or not an effect size is labelled as an outlier because both the Hedges and Olkin (1985; see also Hedges & Olkin, 2014) and Schmidt and Hunter (2015) approaches to meta-analysis estimate the meta-analytic mean by giving more precise studies more weight. Thus, relatively large samples can have an undue influence on the meta-analytic mean.
Sample type	In the context of a meta-analysis, an effect size that differs from all other effect sizes in regard to some sample type characteristic (e.g., incumbents vs. applicants, employees vs. students) may need to be removed before performing a meta-analysis as it could introduce residual heterogeneity that may threaten its results and conclusions. This may be especially true if theoretical evidence suggests the sample characteristic is a boundary condition.

.

TABLE 2

ANALYSES PERFORMED BY THE CSAT

Analysis/parameter
Meta-analysis
k (number of independent samples)[a]
N (sum of independent sample sizes)[a]
$\bar{r}_{o_{RE}}$ (random effects meta-analytic mean effect size estimate)[a]
95% confidence interval[a]
90% prediction interval[a]
Q (weighted sum of squared deviations from the mean)[a]
I^2 (ratio of true heterogeneity to total variation)[a]
Tau (between-sample standard deviation)[a]
Outlier detection
One-sample removed[a]
Minimum, maximum, and median weighted mean observed correlation
Influence diagnostics[b]
Publication bias detection
Fixed-effects trim and fill model[a]
Side imputed
Number of imputed samples
Adjusted meta-analytic mean effect size estimate
Adjusted lower bound of 95% confidence interval
Random effects trim and fill model[a]
Side imputed
Number of imputed samples
Adjusted meta-analytic mean effect size estimate
Adjusted lower bound of 95% confidence interval
A priori selection model[a]
Moderate publication bias assumption
z score
Variance
z score
Back transformed adjusted meta-analytic mean effect size estimate
Severe publication bias assumption[a]
z score
Variance
Back transformed adjusted meta-analytic mean effect size estimate
Precision-effect test-precision effect estimate with standard error (PET-PEESE)[a]
Weighted least squares approach
PET estimate and corresponding one- and two-tailed p-values
PEESE estimate and corresponding one- and two-tailed p-values
Final adjusted meta-analytic mean effect size estimate (one-tailed test)
Final adjusted meta-analytic mean effect size estimate (two-tailed test)
Random effects meta-analysis (metafor; Viechtbauer [2015]) approach
PET estimate and corresponding one- and two-tailed p-values
PEESE estimate and corresponding one- and two-tailed p-values
Final adjusted meta-analytic mean effect size estimate (one-tailed test)
Final adjusted meta-analytic mean effect size estimate (two-tailed test)
Cumulative meta-analysis by precision[a]

Note: CSAT = comprehensive sensitivity analysis tool. [a] = estimated before and outlier removal; [b] = performed iteratively until all identified outliers are removed

TABLE 3

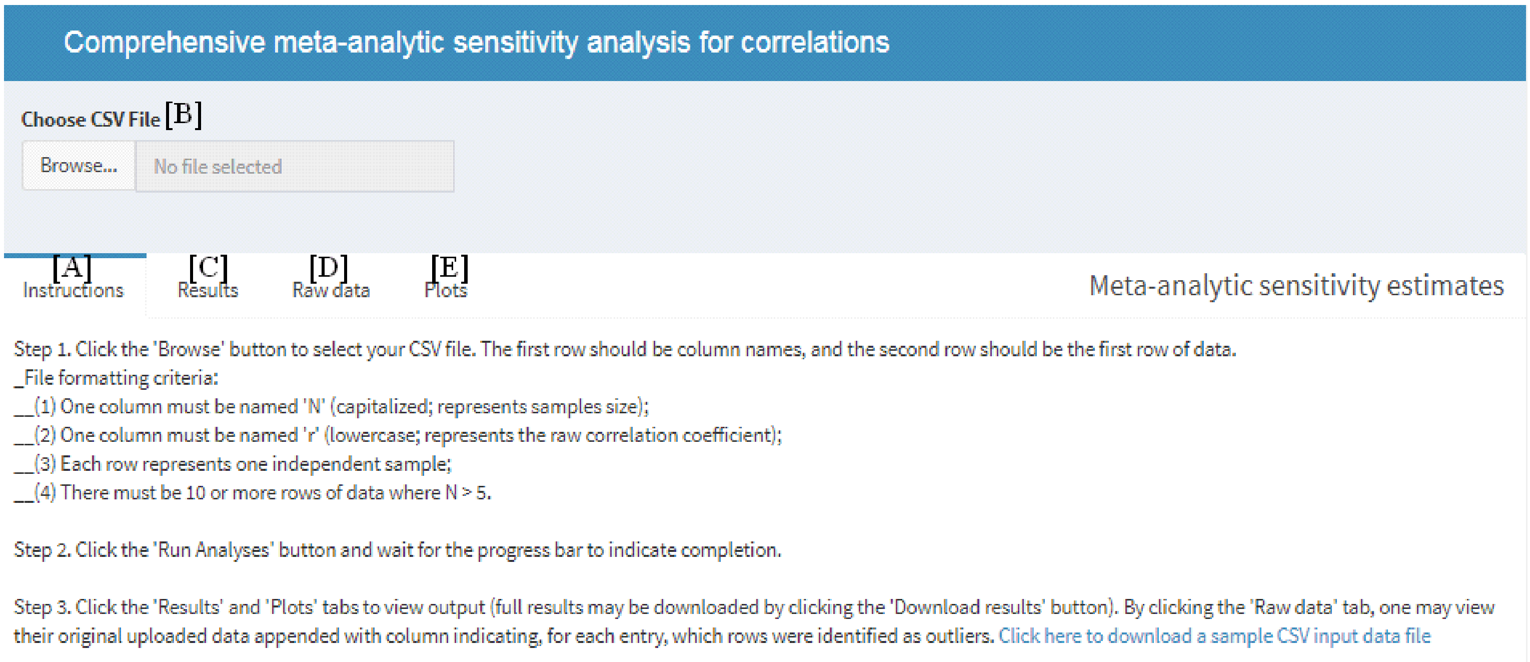
META-ANALYTIC AND SENSITIVITY ANALYSIS RESULTS FOR SHAFFER AND POSTLETHWAITE (2012)

	Meta-analysis								Publication bias analyses											
									Trim and fill								Selection models		CMA	PET-PEESE
Outliers included	k	$\bar{r}_{oRE}$	95% CI	90% PI	Q	I^2	τ	osr	FPS	ik	t&f$_{FE}$ $\bar{r}_o$	t&f$_{FE}$ 95% CI	FPS	ik	t&f$_{RE}$ $\bar{r}_o$	t&f$_{RE}$ 95% CI	sm$_m$ $\bar{r}_o$	sm$_s$ $\bar{r}_o$	$pr\ \bar{r}_o$	$pp\ \bar{r}_o$
Before outlier removal																				
Conscientiousness	113	.159	.138, .180	.026, .287	236.516	52.646	.081	.157, .162, .159	L	22	.126	.104, .148	L	16	.137	.116, .159	.142	.115	.102	.121
Noncontextualized	91	.151	.127, .175	.006, .290	210.056	57.154	.088	.148, .155, .151	L	15	.120	.095, .146	L	10	.132	.107, .157	.131	.095	.102	.073
Contextualized	22	.190	.158, .222	.163, .217	19.005	.000	.000	.185, .200, .190	L	5	.169	.136, .203	L	5	.169	.136, .203	.184	.177	.173	.113
Emotional stability	86	.098	.073, .124	-.044, .236	180.683	52.956	.085	.092, .103, .098	L	17	.060	.032, .088	L	13	.069	.041, .097	.072	NA	.000	.034
Noncontextualized	68	.074	.051, .098	-.017, .164	97.642	31.382	.054	.072, .078, .074	L	14	.045	.020, .070	L	11	.052	.028, .077	.053	NA	.000	-.016
Contextualized	18	.179	.110, .247	-.031, .374	53.74	63.366	.124	.159, .199, .180	R	4	.230	.156, .302	R	4	.231	.157, 303	.153	NA	.198	.246
Extraversion	90	.076	.049, .103	-.089, .237	228.851	61.11	.100	.072, .078, .076	L	2	.071	.043, .098	R	1	.078	.051, .105	.044	NA	.012	.054
Noncontextualized	72	.057	.028, .085	-.088, .199	157.513	54.924	.087	.052, .059, .056	L	1	.055	.026, .083	-	0	.057	.028, .085	.028	NA	.012	.016
Contextualized	18	.152	.089, .213	-.030, .323	44.831	62.080	.106	.139, .167, .151	-	0	.152	.089, .213	-	0	.152	.089, .213	.125	NA	.173	.229
Agreeableness	94	.084	.060, .109	-.061, .226	205.063	54.648	.087	.080, .087, .084	L	1	.082	.058, .107	L	1	.082	.058, .107	.057	NA	.101	.079
Noncontextualized	73	.063	.039, .086	-.035, .160	111.199	35.251	.059	.060, .066, .063	L	8	.046	.021, .070	L	8	.046	.021, .070	.041	NA	.050	.028
Contextualized	21	.152	.087, .215	-.060, .351	68.484	70.796	.125	.137, .168, .150	-	0	.152	.087, .215	-	0	.152	.087, .215	.018	NA	.180	.197
Openness	80	.023	-.002, .048	-.101, .146	148.876	46.936	.075	.019, .026, .023	R	1	.024	-.001, .049	R	1	.024	-.001, .049	-.003	NA	.034	.047
Noncontextualized	66	.009	-.017, .035	-.101, .119	110.274	41.056	.066	.004, .012, .009	R	4	.017	-.010, .044	R	4	.017	-.010, .044	-.015	NA	.034	.045
Contextualized	14	.089	.026, .152	-.063, .238	27.83	53.287	.087	.072, .105, .089	-	0	.089	.026, .152	-	0	.089	.026, .152	.062	NA	.101	.070
After outlier removal																				
Conscientiousness	112	.162	.142, .182	.042, .279	211.424	47.499	.074	.160, .165, .162	L	17	.138	.117, .159	L	15	.142	.121, .163	.147	.125	.102	.133
Noncontextualized	*Outliers were not detected and, thus, analyses were not performed*																			
Contextualized	*Outliers were not detected and, thus, analyses were not performed*																			
Emotional stability	82	.095	.074, .115	.022, .166	103.028	21.38	.043	.092, .097, .095	L	8	.082	.061, .104	L	3	.091	.070, .112	.078	.044	.036	.049
Noncontextualized	*Outliers were not detected and, thus, analyses were not performed*																			
Contextualized	*Outliers were not detected and, thus, analyses were not performed*																			
Extraversion	*Outliers were not detected and, thus, analyses were not performed*																			
Noncontextualized	*Outliers were not detected and, thus, analyses were not performed*																			
Contextualized	*Outliers were not detected and, thus, analyses were not performed*																			
Agreeableness	93	.080	.056, .104	-.053, .210	184.757	50.205	.080	.077, .083, .080	L	5	.071	.046, .095	L	5	.071	.046, .095	.054	.054	.101	.074
Noncontextualized	*Outliers were not detected and, thus, analyses were not performed*																			
Contextualized	*Outliers were not detected and, thus, analyses were not performed*																			
Openness	78	.014	-.008, .037	-.080, .109	115.688	33.442	.056	.012, .017, .014	-	0	.014	-.008, .037	-	0	.014	-.008, .037	-.007	NA	.034	.029
Noncontextualized	65	.004	-.020, .027	-.079, .087	88.902	28.011	.049	.000, .007, .004	R	2	.007	-.017, .031	-	0	.004	-.020, .027	-.016	NA	.034	.039
Contextualized	*Outliers were not detected and, thus, analyses were not performed*																			

Note. $\bar{r}_{o_{RE}}$ = random-effects weighted mean observed correlation; 95% CI = 95% confidence interval; 90% PI = 90% prediction interval; Q = weighted sum of squared deviations from the mean; I^2 = ratio of true heterogeneity to total variation; τ = between-sample standard deviation; osr = one-sample removed, including the minimum and maximum effect size and the median weighted mean observed correlation; Trim and fill = trim and fill analysis; FPS = funnel plot side (i.e., side of the funnel plot where samples were imputed; L = left, R = right); *ik* = number of trim and fill samples imputed; t&f_{FE} $\bar{r}_o$ = fixed-effects trim and fill adjusted observed mean; t&f_{FE} 95% CI = fixed-effects trim and fill adjusted 95% confidence interval; t&f_{RE} $\bar{r}_o$ = random-effects trim and fill adjusted observed mean; t&f_{RE} 95% CI = random-effects trim and fill adjusted 95% confidence interval; sm_m $\bar{r}_o$ = one-tailed moderate selection model's adjusted observed mean; sm_s $\bar{r}_o$ = one-tailed severe selection model's adjusted observed mean; CMA = cumulative meta-analysis; *pr* $\bar{r}_o$ = meta-analytic mean estimate of the five most precise effects; PET-PEESE = precision-effect test-precision effect estimate with standard error (two-tailed weighted least squares approach); *pp* $\bar{r}_o$ = PET-PEESE adjusted observed mean. Dashes indicate that the corresponding trim and fill model did not impute any sample on either side of the funnel plot. NA = not applicable (because sm_s $\bar{r}_o$ presented nonsensical results due to inflated variance estimates).

FIGURE 1

Full View of Comprehensive Sensitivity Analysis Tool Graphical User Interface

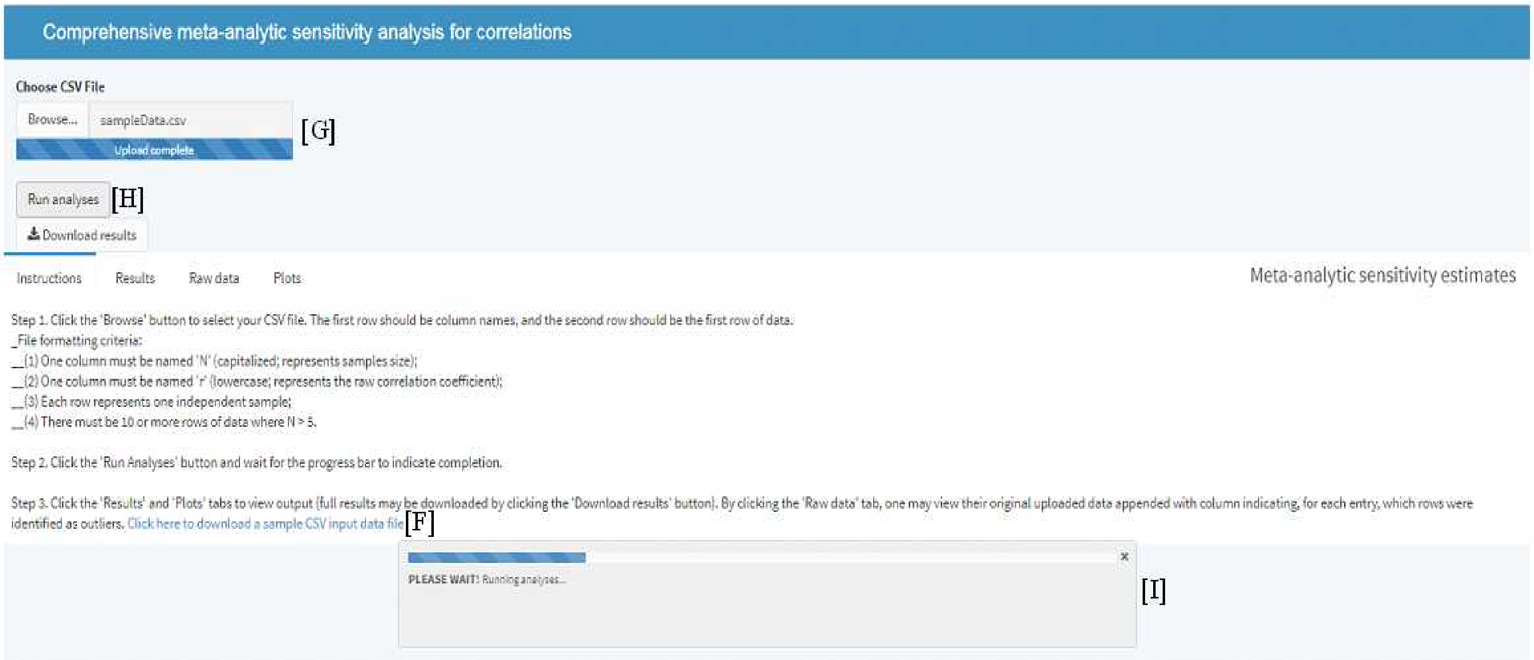


Note. Letters in brackets are referred to in the text and do not appear in the interface.

James Field, West Virginia U., jamesfield6912@gmail.com
Frank A. Bosco, Sven Kepes, Michael A McDaniel, Sheila List

FIGURE 2

Uploading a Meta-Analytic Dataset and Performing Comprehensive Sensitivity Analysis

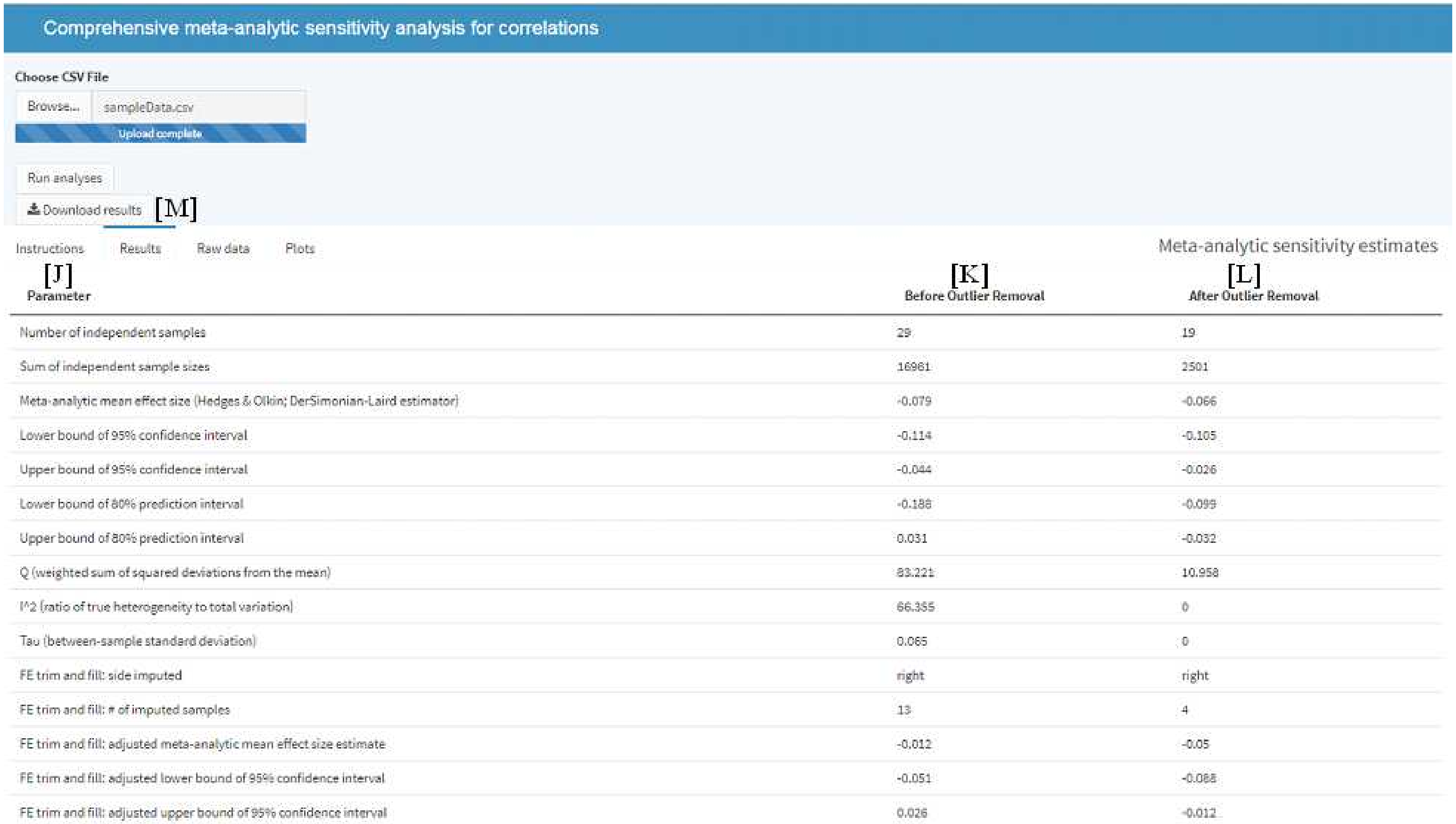


Note. Letters in brackets are referred to in the text and do not appear in the interface.

James Field, West Virginia U., jamesfield6912@gmail.com
Frank A. Bosco, Sven Kepes, Michael A McDaniel, Sheila List

FIGURE 3

Short View of Results Tab Showing Meta-Analytic and Sensitivity Analysis Results Before and After Outlier Removal



Parameter	Before Outlier Removal	After Outlier Removal
Number of independent samples	29	19
Sum of independent sample sizes	16981	2501
Meta-analytic mean effect size (Hedges & Olkin; DerSimonian-Laird estimator)	-0.079	-0.066
Lower bound of 95% confidence interval	-0.114	-0.105
Upper bound of 95% confidence interval	-0.044	-0.026
Lower bound of 80% prediction interval	-0.188	-0.099
Upper bound of 80% prediction interval	0.031	-0.032
Q (weighted sum of squared deviations from the mean)	83.221	10.958
I^2 (ratio of true heterogeneity to total variation)	66.355	0
Tau (between-sample standard deviation)	0.065	0
FE trim and fill: side imputed	right	right
FE trim and fill: # of imputed samples	13	4
FE trim and fill: adjusted meta-analytic mean effect size estimate	-0.012	-0.05
FE trim and fill: adjusted lower bound of 95% confidence interval	-0.051	-0.088
FE trim and fill: adjusted upper bound of 95% confidence interval	0.026	-0.012

Note. Letters in brackets are referred to in the text and do not appear in the interface.

James Field, West Virginia U., jamesfield6912@gmail.com
Frank A. Bosco, Sven Kepes, Michael A McDaniel, Sheila List

FIGURE 4

Short View of Raw Data Tab Showing Uploaded Meta-Analytic Dataset and Outlier Classification

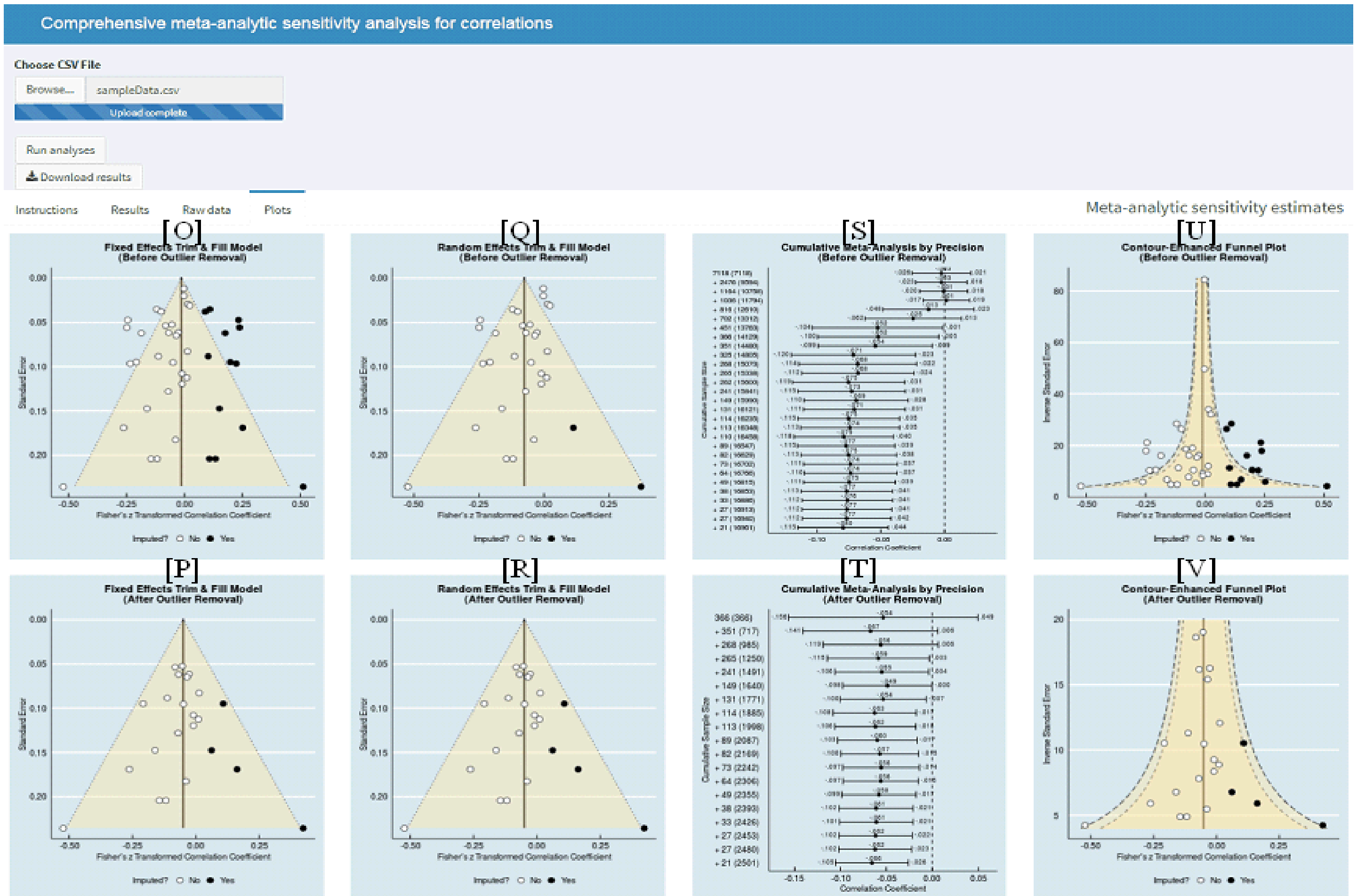
Art.ID	Sample.ID	Journal.name	Journal.code	Year	X1st.author	ESdata.source	N	r	outlier
1	1	Ppsych	10	2001	Koys	Correlation	27	-0.143375	No
2	1	AMJ	1	2006	Kacmar	Correlation	262	-0.184	Yes
3	1	AMJ	1	1994	Arthur	Correlation	27	-0.117735849	No
4	1	AMJ	1	1995	Huselid	Correlation	816	-0.12	Yes
8	1	AMJ	1	2004	Glebbeek	Correlation	110	-0.23	Yes

Note. Letter in brackets are referred to in the text and do not appear in the interface.

James Field, West Virginia U., jamesfield6912@gmail.com
Frank A. Bosco, Sven Kepes, Michael A McDaniel, Sheila List

FIGURE 5

Full View of Plots Tab Showing Sensitivity Analysis Results Before (Top Panel) and After (Bottom Panel) Outlier Removal

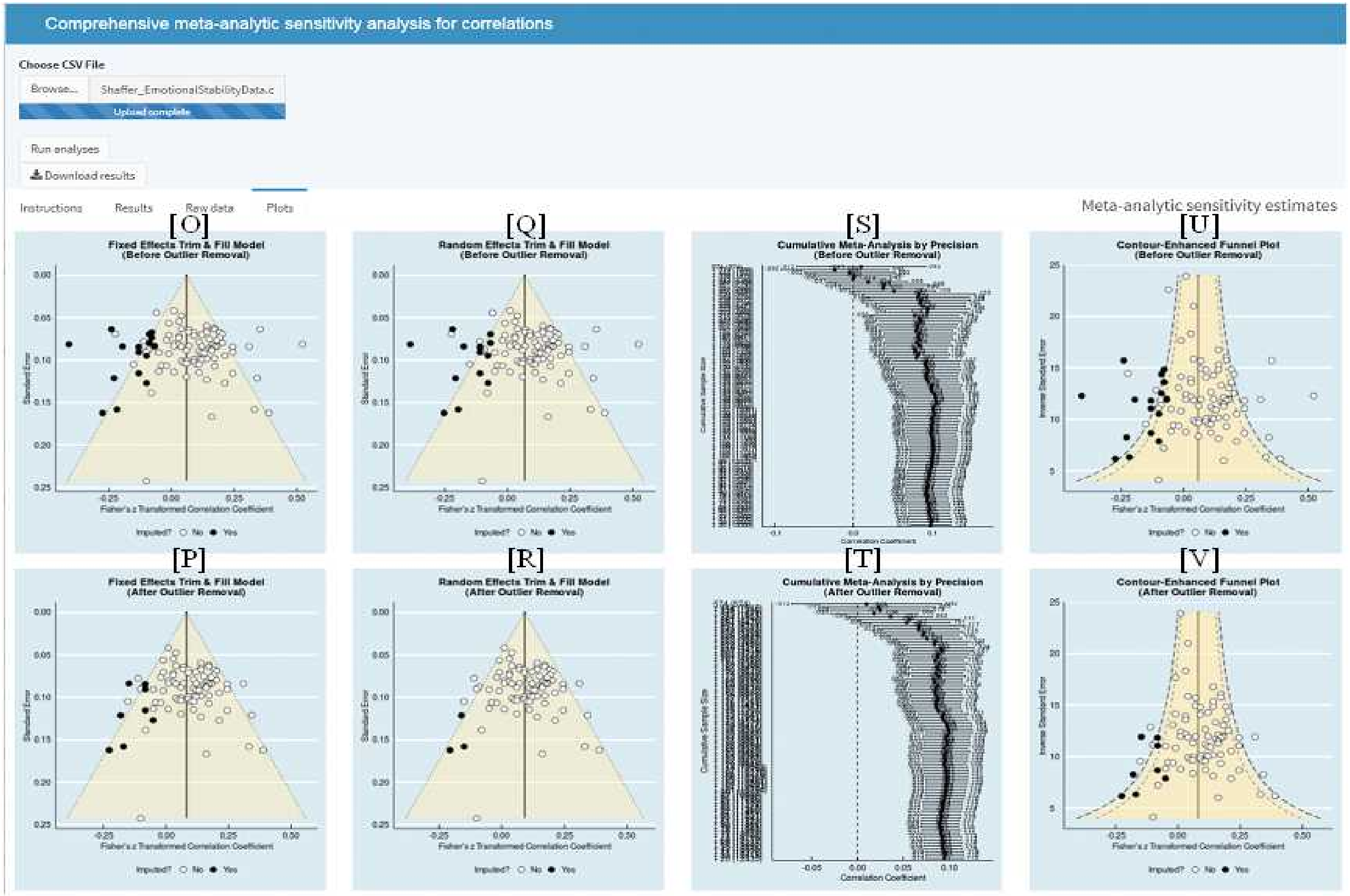


Note. Letters in brackets are referred to in the text and do not appear in the interface

James Field, West Virginia U., jamesfield6912@gmail.com
Frank A. Bosco, Sven Kepes, Michael A McDaniel, Sheila List

FIGURE 6

Full View of Plots Tab Showing Sensitivity Analysis Results Before (Top Panel) and After (Bottom Panel) Outlier Removal for Shaffer and Postlethwaite (2012)



Note. Letters in brackets are referred to in the text and do not appear in the interface