Running head: Publication bias in test vendor manuals

Last revised: June 23, 2006

Publication Bias: A Case Study of Four Test Vendors

Michael A. McDaniel

Virginia Commonwealth University

Hannah Rothstein

Baruch College

Deborah L. Whetzel

Work Skills First, Inc.

Authors' Note: An earlier version of this report was presented at the 2005 Academy of Management meeting. This paper has substantially benefited from feedback from the editor and reviewers.

Publication Bias: A Case Study of Four Test Vendor Manuals

Abstract

This paper has two goals. First, we discuss publication bias and explain why it presents a potential problem for industrial and organizational psychology. After reviewing the traditional failsafe N, or file drawer analysis, we introduce a more sophisticated method of publication bias analysis (trim and fill) which has been developed in the medical literature but is largely unfamiliar to industrial and organizational psychology researchers. Second, we demonstrate trim and fill by applying it to validity information reported in the technical manuals of four test vendors. In doing so, we assess the likelihood that criterion-related validity information provided by test publishers may overestimate test validity. In our analysis of 18 validity distributions, we found evidence of either no or minimal bias for two of the vendors' distributions and evidence of moderate to severe bias in at least one distribution from each of the other two vendors. In both cases in which publication bias was found, we noted instances in which the publishers tended to report only statistically significant correlations and that this practice was detected using publication bias methodology.

Publication bias is the term used to refer to the possibility that not all completed studies on a topic are published in the literature, and that these studies are systematically different from published studies. Such bias may lead readers and reviewers to draw incorrect conclusions that can have substantive consequences, particularly when an ineffective practice is viewed as effective because of selective publication of results.

In the industrial and organizational psychology literature, the consequences of publication bias can be quite serious. Consider the research on the validity of employment interviews. In 1994, McDaniel, Whetzel, Schmidt and Maurer published a meta-analysis showing that structured interviews were more valid than unstructured interviews (.27 vs. .19 uncorrected). Publication bias analyses (Duval, 2005) revealed that, in the absence of publication bias, the validity of structured interviews likely would be lower (.21) and closer to the validity of unstructured interviews. The problem here is two-fold. First, many practitioners relied on these findings and likely created fairly laborious structured interviews to select employees thinking that they were substantially more valid than unstructured interviews. Second, the number of research studies comparing the two types of interviews decreased after the meta-analysis was published, reducing the potential for contradictory findings.

Publication bias also has been reported in the literature on black/white mean differences in job performance (McDaniel, McKay, & Rothstein, 2006). Results showed that the magnitude of black/white mean differences in job performance (favoring whites) are underestimated in journals, possibly because authors have discretion to report or not report mean racial differences. It appears that primary study authors are more likely to report differences when they are small than when they are large.

Consider also the employment test validity literature. Specifically, there are important negative consequences for both employers and job applicants if an invalid test or a test with low validity is viewed as having a higher level of validity. The potential problems posed by publication bias have not been a topic of substantial research in industrial and organizational psychology. One publication bias method, the failsafe N method, also known as the “file drawer problem” method has been used in industrial and organizational psychology (Bertua, Anderson, Salgado, 2005; Brewer & Shapard, 2004; Jenkins, Mitra, Gupta and Shaw, 1998; Mitra, Jenkins and Gupta, 1992; Parker, Baltes, Young, Huff, Altmann, Lacost, Roberts, 2003; Rhoades and Eisenberger, 2002; Roth, Bevier, Bobko, Switzer, & Tyler, 2001; Salgado, Anderson, Moscoso, Burtua, de Fruyt, Rolland, 2003). However, this method has been shown to be less effective for assessing publication bias (Becker, 2005) than other more accurate and powerful approaches that have been offered (Duval, 2005; Hedges & Vevea, 2005; Sterne & Egger, 2005; Sutton & Pigott, 2005).

The current paper has two purposes. First, we describe two methods for assessing publication bias. After reviewing and critiquing the failsafe N, or file drawer analysis method, we describe a recently introduced procedure to assess publication bias. This procedure, trim and fill (Duval & Tweedie, 2000a, 2000b), has been used widely in the healthcare literature and is increasingly used in psychological meta-analyses (e.g., *Psychological Bulletin* where editorial policy encourages publication bias analyses [Cooper, 2003]). Second, we illustrate the usefulness of the trim and fill procedure to explore the possibility of publication bias in the criterion-related validity information provided in test publisher manuals. Specifically, we apply trim and fill analyses to 18 distributions of employment test validities drawn from the manuals of four test publishers. For comparison purposes, we also conduct failsafe N analyses on the same data. We close with recommendations for test publishers as well as implications of publication bias for meta-analyses in I/O psychology.

Publication Bias and Procedures for Its Detection

Publication bias is defined as the tendency to publish studies depending on the magnitude, direction, or statistical significance of the results. When studies are missing from the literature available for meta-analytic research, we will refer to the studies as being “suppressed” consistent with its use in the medical publication bias literature. We use this term neutrally, without negative attributions as to the reasons for the “missingness” of the studies. There are several possible causes for the suppression of research results. There is considerable evidence that one cause of bias is that researchers are not likely to submit negative results for publication (Dickersin, 2005). There also is some evidence that editorial policy, at least in some journals, favors the publication of significant results (Dickersin, 2005; Greenwald, 1975). Below, we describe both failsafe N and trim and fill approaches to the detection of publication bias.

Rosenthal Failsafe N (the file drawer problem)

Rosenthal (1979) introduced what he called the “file drawer problem.” His concern was that some statistically non-significant studies may be missing from an analysis (i.e., placed in a file drawer) and that these studies, if included, would nullify the observed effect. By “nullify,” he meant to reduce the effect to a level not statistically significantly different from zero. Rosenthal suggested that rather than speculate on whether the file drawer problem existed, the actual number of studies that would be required to nullify the effect could be calculated. Cooper (1979) called this number the failsafe sample size or failsafe N. If this number is relatively small, then there is cause for concern. If this number is large, one might be more confident that the effect, while possibly inflated by the exclusion of some studies, is nevertheless not zero.

This approach is limited in two important ways (Becker,1994; 2005). First, it assumes that the correlation in the hidden studies is zero, rather than considering the possibility that some of the studies could have an effect in the reverse direction or an effect that is small but not zero. Therefore, the number of studies required to nullify the effect may be different than the failsafe N, either larger or smaller. Second, and more fundamentally, this approach focuses on statistical significance rather than practical or substantive significance (effect sizes). That is, it may allow one to assert that the mean correlation is not zero, but it does not provide an estimate of what the correlation might be (how it has changed in size) after the missing studies are included.

Consider an employer choosing between two tests, A and B, of the same construct offered by different test publishers. The validity information for test A suggests that the test has a mean validity of .25 while the validity information for test B shows a mean validity of .20. If there is no publication bias, the employer would choose test A, all other things (cost, ease of administration, etc.) being equal. However, if publication bias is suspected, one would like to know the validity of tests A and B in the absence of bias. Knowing that it takes 80 file drawer studies to nullify the validity of test A and 100 file drawer studies to nullify the validity of test B does not help to determine the validity of the tests in the absence of publication bias.

Orwin (1983) extended the idea of the failsafe N to effect sizes and reformulated the question as: “How many effect sizes averaging a particular value would be needed to reduce an observed mean effect size to a level at which it was no longer theoretically or practically significant?” Orwin’s variant also has been used in industrial and organizational psychology meta-analyses, (McNatt, 2000). Although Orwin’s method is an improvement on the original Rosenthal method, in that it incorporates information about effect size, it still does not estimate the likely magnitude of the population effect, taking into account the studies that may exist, but that are missing from an analysis.

Trim and Fill

To understand the trim and fill method of publication bias detection, one needs to be conversant with the concept of a funnel plot (Light & Pillemer, 1984). A funnel plot, as shown in Figure 1, plots the correlations from a set of studies. The correlations are represented by open circles. The X axis plots the magnitudes of the correlations. Thus, the correlations of large magnitude fall to the right of the graph, and the correlations of lower magnitude are to the left side of the graph. The Y axis plots the sample size of the studies. Correlations based on large sample sizes have smaller confidence intervals. Put another way, correlations from large samples have smaller standard errors. On average then, correlations from large samples will be closer to the population correlation than correlations from small samples. Thus, correlations from large samples will be similar to each other and cluster near the center line of the funnel. Conversely, correlations based on small sample sizes have large confidence intervals (large standard errors). This means that correlations from small samples will often overestimate or underestimate the population correlation. A collection of correlations from small sample studies will vary substantially around the population correlation causing the funnel to be wide at the bottom.

The concept of precision is relevant to funnel plots. Correlations based on large samples have small standard errors. Standard errors are influenced by the magnitude of the population correlation in addition to the sample size. Thus, while one could use the sample size as an indicator of the precision of the correlation, a more exact precision measure would be the inverse of the standard error, that is, 1 divided by the standard error. Precision is often used instead of sample size for the Y axis of the funnel plot (Sterne & Egger, 2005).

Trim and fill defines asymmetry as evidence of publication bias (Duval & Tweedie, 2000a, 2000b). Unbiased distributions of correlations become more symmetrical when they are transformed to Fisher *z*. This simple transformation does little to small magnitude correlations, but increases the value of large magnitude correlations. For example, a correlation of .20 has a Fisher *z* value of .203, while a correlation of .80 has a Fisher *z* value of 1.099 (note that Fisher *z* values can exceed the absolute value of 1.0). In the range of correlation values of test validities, the transformation does not have much of an impact on the underlying validities, except for improving symmetry in the absence of bias. Because trim and fill defines asymmetry as evidence of publication bias, correlations need to be transformed to Fisher *z* prior to analysis*.* Figure 2a shows a symmetrical funnel plot of correlations, expressed as Fisher *z* as a function of precision.

Assuming all the criterion-related validity studies conducted were reported, we expect the studies in the funnel plot to be distributed symmetrically around the estimated population correlation as in Figure 2a if sampling error is the only source of variance in the validities. When smaller or statistically non-significant correlations are suppressed, we expect an asymmetric funnel plot, with a relatively high number of small studies falling toward the right and relatively few falling toward the left. Figure 2b shows an asymmetric funnel that could indicate suppression of small effect, small sample size studies. Large sample size studies with relatively low validities are not as likely to be suppressed, because they are more likely to reach statistical significance than small sample studies with the same magnitude correlation.

Trim and fill first assesses whether, and to what degree, bias may be affecting the results of a meta-analysis. It then estimates how the effect (in our case, the validity) would change if the putative bias were to be removed. The key idea behind the funnel plot is that in the absence of bias, the plot would be symmetric about the mean correlation. If there were more small sample studies on the right of the plot than on the left, our concern is that there may be studies missing from the left. The trim and fill procedure imputes the missing studies, adds them to the analysis, and then re-calculates the effect size.

Assumptions of trim and fill. The trim and fill method assumes that in addition to the number of observed studies in a meta-analysis, there is an additional number of relevant studies that are not included, due to publication bias. The number of these studies, and the effect sizes (correlations) associated with them, is unknown but can be estimated. In addition, the uncertainty of these estimates has to be reflected in the (adjusted) meta-analytic result. Another key assumption underlying trim and fill is that the distribution of effect sizes in the population is homogeneous, that is, sampling error is the sole source of variation in a set of studies. Thus, in the application of trim and fill, the researcher should take reasonable steps to rule out moderators in the distributions. If, for example, a moderator were operating, one would conduct the analysis on a subset of the data where the moderator does not vary.

When searching for missing studies on the left side of the funnel, trim and fill uses an iterative procedure to remove the most extreme small studies from the positive (i.e., right) side of the funnel plot, (those without counterparts on the left) and re-computes the effect size at each iteration, until the funnel plot is symmetric about the (new) effect size. While this “trimming” yields the adjusted effect size, it also reduces the variance of the effects, yielding a confidence interval that is too narrow. Therefore, the algorithm then adds the original studies back into the analysis, and imputes a mirror image for each original study. The final estimate of the mean overall effect, as well as its variance, is based on the “filled” funnel plot (Duval & Tweedie, 2000a, 2000b). Figure 2c provides an example of a filled funnel plot. The clear circles are the original data and the dark circles are the imputed data.

The chief benefit of the trim and fill approach is that it yields an effect size estimate that is adjusted for the funnel plot asymmetry, something that the failsafe N method does not provide. We do not suggest that a mean validity estimated using imputed studies should serve as the best estimate of a test’s validity because the mean is estimated on imputed data points. However, it provides a useful sensitivity analysis that assesses the potential impact of missing studies on the meta-analysis. One can assess the degree of divergence between the original mean validity and the trim and fill adjusted mean validity. We suggest that the potential impact of publication bias could be regarded as “minimal” when the observed mean and the trim and fill adjusted mean are essentially the same. Publication bias impact can be called “moderate” when the observed and adjusted means differ meaningfully but the decision to use the test probably would not change. We label the impact of potential bias “severe” when the observed and adjusted means differ substantially and that the decision to use the test would likely change.

The results obtained from the failsafe N analysis and the trim and fill analysis may not be in agreement because they answer different questions. The trim and fill analysis defines publication bias as the difference between the original effect size and the recomputed effect size after the “missing” studies have been added to make the distribution symmetrical. Thus, trim and fill interprets effect size distribution asymmetry as evidence of publication bias. The failsafe N analysis defines publication bias as the number of studies obtaining no effect that it would take to completely nullify the observed mean effect size. Whenever there is a distribution with effect sizes far from zero and when the number of studies is relatively large, failsafe N analyses will yield a conclusion that there is no publication bias.

Consider a researcher who conducted 50 studies and reported only those 25 studies with validities over some value (e.g., validities above .20). The resulting distribution would be similar to the asymmetric funnel plot shown in Figure 2b. The mean of the distribution would be well above .20 and the distribution would contain 25 effect sizes. The failsafe N analyses would report little evidence of bias because many studies would need to be missing to move the distribution much closer to zero. In contrast, due to distribution asymmetry stemming from the missing smaller validities, trim and fill analyses would yield evidence of publication bias. We believe that trim and fill analysis is more accurate than the failsafe N analysis for locating publication bias in our validity data because it asks the question of interest to most researchers and practitioners, namely, how much has the effect shifted due to publication bias?

Method

Data source. We obtained test vendor technical manuals from four test vendors. They provided validity data on a total of 18 scales. The decision rules used for each data set are described.

Based on feedback from the editor, reviewers, and others, we make the test vendors anonymous. We did this to allow the reader to focus on the publication bias methods, and not on the specific tests or vendors. It also rebuts concerns that the authors, editor, or the journal seeks to stigmatize any test vendor whose data appear to reflect publication bias. We refer to the test vendors as Vendor, A, B, C, and D.

We obtained data from Vendor A’s personality measure assessing the Big 5 (Digman,1989). Validity data are available for each of the Big 5 scales: extraversion, agreeableness, conscientiousness, (emotional) stability, and openness. These scales are labeled A1 through A5. The validity data for each scale were analyzed separately. We restricted our analyses to supervisory ratings because the trim and fill method assumes the correlations are homogeneous (i.e., a moderator-free distribution), and we did not want to contaminate the analyses with a potential criterion-type moderator. Moreover, supervisory ratings were the most frequently used criterion in these studies. The data set contains only one validity coefficient per sample. All studies were concurrent validity studies. All five scales of Vendor A's measure derive from the same personality instrument. Thus, each of the validity distributions is based on the same 14 studies reported by Vendor A.

Vendor B's technical manual provides criterion-related validity coefficients for three scales derived from item clusters of the Big 5 scales. We refer to these scales as B1, B2, and B3. Vendor B's validity studies use a variety of criteria; however, as with the data from Vendor A, we used validity data based solely on supervisor ratings. This makes the criterion data comparable across test publishers and avoids the possibility of a criterion type serving as a moderator in the data. The manual often listed more than one validity coefficient for each sample. We choose the supervisory rating of "overall performance" for inclusion in the analysis. When a validity coefficient was not reported for an overall performance measure, we averaged across the available validity coefficients (e.g., ratings of quality and quantity of performance). Similar to the situation with Vendor A, the three scales of Vendor B are derived from the same personality measure. However, not all the validity coefficients in each of the three validity distributions are reported. To the best of our knowledge, all studies are concurrent.

Vendor C's technical manual summarizes criterion-related validity data for four tests that we label C1 through C4. All studies were concurrent. Validity data for multiple criteria were used and consistent with the decision rules described above, we selected the overall performance validity coefficients for analysis. Some of the sample sizes were reported as ranges. In those cases, we used the midpoint of the range as the sample size. For Vendor C's scale C2, results are presented for four samples. For the four samples, the criterion was a supervisor rating collected from two raters. Validities were reported separately by rater and then averaged. Data from sample one used one of the rater's ratings as the criterion and used those data to key the instrument. Since the validity data for sample one using rater one is likely to be inflated because it was a keying sample, we used the validity coefficient based on the ratings from the second rater. For samples two through four, we used the validity coefficient for the criterion that combined the ratings of raters one and two.

Vendor C's manual also reports validity data on two additional tests (scales C3 and C4). Vendor C's scale C4 contains all items from Vendor C's scale C3 plus three sets of cognitive ability items. The validity data for these two scales are based on the same samples. Validities are reported for overall performance and for sub-scales of overall performance. We used the validities for the overall performance scale. These validities were corrected for measurement error in the criterion using .60 as the estimate of the reliability. We attenuated the validity coefficients and used the observed coefficients in our analysis to make them comparable to those from the other data sets.

Vendor C's scales C1 and C2 are not components of the same test. They are different tests and thus their validity distribution need not be based on the same number of samples. There were eight samples of validity data for scale C1 and four samples of validity data for scale C2. Vendor C's scale C3 is a subset of Vendor C's scale C4. Thus, one would expect that the two validity distributions would be based on the same number of validity coefficients. This is the case as both validity distributions are from the same seven samples.

We obtained data on five scales (D1 through D5) from three test vendor manuals for Vendor D. All criteria were supervisory ratings. When an overall rating was available, we used that validity coefficient. When an overall rating was not available we used the mean of the available validities. All studies were concurrent. All tests were tests of specific cognitive abilities that required us to consider cognitive complexity as a moderator (Hunter & Hunter, 1984). The validity of cognitive tests varies with the cognitive complexity of jobs. With cognitive ability tests, more complex jobs yield higher validities on average than less complex jobs. Vendor D reported validity data in tables, one per validity study. The tables were inconsistent in how data were reported. Specifically, some tables reported all data regardless of statistical significance, but other tables were clearly labeled to indicate that only statistically significant correlations were reported. Thus some tables had cells which, if present, would have contained lower magnitude correlations that were not statistically significant. The validity studies did not overlap across manuals. Thus, if two tests are from different manuals, the number of validity studies in the distributions is likely to differ.

The first Vendor D manual provided data on scales D1 and D2. The tests were designed for low complexity jobs and all validity data were collected from incumbents in low complexity occupations. Some tables listed all validities regardless of statistical significance and other tables listed only the statistically significant correlations. The restriction of some studies data to only statistically significant correlations caused the number of studies for scales D1 and D2 to differ.

The second manual from Vendor D contributed data for two additional scales, D3 and D4. All validity data for these tests were from high-complexity jobs. The tables in this manual only listed the statistically significant correlations. Also, some studies only used one of the two scales. The restriction of the data to only statistically significant correlations and some studies using only one of the tests caused the number of studies for scales D3 and D4 to differ.

The third test manual from Vendor D contributed an additional scale, D5. Concerns arose because the studies contributing data to the technical manual included data from incumbents in jobs that varied widely in complexity and thus differences across studies in job complexity were likely a source of variation. We used the DOT data code as a measure of complexity consistent with past research (Rivkin & McDaniel, 1990) and identified the majority of studies as having relatively high complexity (data codes of 1 or 2). We limited our analysis to those studies to reduce the variance of cognitive complexity in the samples. In reporting validities by study, some tables listed all validities regardless of statistical significance and other tables listed only the statistically significant correlations.

Meta-analysis procedure. We conducted a meta-analysis of the observed validity coefficients and a meta-analysis of the distribution adjusted by the trim and fill procedure. For comparison with trim and fill, we also conducted a traditional Rosenthal file drawer (failsafe N) analysis.

Most meta-analyses of employment test validity data use the psychometric meta-analysis method (Hunter & Schmidt, 2004). However, statistical software for publication bias is not available for psychometric meta-analysis as it is for meta-analyses in the tradition of Hedges and Olkin (1985). For this reason, we conducted the meta-analysis using the Comprehensive Meta-Analysis (CMA) software (Borenstein, Hedges, Higgins & Rothstein, 2005) which follows procedures associated with Hedges and Olkin. We note that this procedure is similar to a “bare bones” psychometric meta-analysis in which observed validity coefficients are analyzed (i.e., correlations are not corrected for statistical artifacts such as measurement error and range restriction). Psychometric meta-analysis is a random effects model of meta-analysis, and we used the random effects implementation of the Hedges and Olkin approach. Both random effects meta-analysis methods yield similar and accurate results for correlation coefficients (Field, 2005).

Trim and fill procedures. As stated above, trim and fill estimates the number of missing effects (correlations) and the magnitude of these effects. It uses a non-parametric method based on the ranks of the absolute values of the observed effect sizes, and the signs of those effect sizes, and measures the imbalance in the set of ranked effects. In the original work on trim and fill (Duval and Tweedie, 2000a, 2000b) three estimators, *L*, *R,* and *Q* were proposed to estimate the “missing” studies. Duval advises against the use of *Q* is as it tends to overestimate the number of missing studies. Both *L* and *R* have desirable properties, as both have low bias. As the number of studies gets larger the estimator *R* becomes preferable to *L* in terms of having a relatively smaller variance. On the other hand, *R* is not robust to some situations, in particular when there is one isolated large "negative" effect size and then a gap. In our case, we had relatively small numbers of studies, and chose to use *L* as the estimator. Further technical details about trim and fill are beyond the scope of this article, but we refer the interested reader to the results of monte carlo simulations, worked examples and additional information in Duval and Tweedie (2000a, 2000b) and Duval (2005).

Rosenthal file drawer analysis. Like trim and fill, file drawer analysis assumes that studies with statistically significant results are more likely to be published than those with non-significant results. Apart from this common assumption, it operates entirely differently than trim and fill, mostly because it answers a different question. The file drawer analysis starts with a test of combined significance (*p*-value summary) often called the "Sum of *Z*s," which is based on the probability values for the effect (correlation) observed in each study. If the sum of *Z*s is significant, one can conclude that at least one of the studies has an effect that is different from zero. The file drawer analysis then asks, if the observed value of *Z is* above the critical value for significance, how many studies with z_i values averaging zero would need to be added to reduce the value of *Z*s to below the critical value at the desired probability level (e.g., p = .05). The probability value tested in this situation is not the same as the probability for the combined mean effect of the meta-analysis, While the failsafe N procedure computes a *p*-value for each study and then combines these *p*-values, the generally accepted approach is to compute an effect size for each study, combine the effect sizes, and then compute the *p*-value for the combined effect. The two approaches do not generally yield identical results.

The failsafe N is the number of studies needed to drop the *Z*s below the value required for statistical significance. According to Rosenthal, if this number is large relative to the number of observed studies, the results of the meta-analysis can be considered robust to publication bias, because it is highly unlikely that such a large number of additional unpublished results exist.

Results

Table 1 presents the results of the publication bias analyses for the 18 scales. The first column shows the scale studied. Each scale is labeled according to vendor (e.g., scales A1-A5 are from Vendor A). The next column shows the number of studies in each meta-analysis. The next two columns show the meta-analytic results (mean *r* and 95% confidence interval) before trim and fill. The next three columns show the trim and fill results, including number of studies needed to achieve symmetry, mean *r,* and 95% confidence interval after the "missing" studies have been imputed. The difference between the original meta-analytic results and the trim and fill results are shown in Column 8. The last two columns show the failsafe N results. Figure 3 shows the funnel plots for each distribution. The clear circles are observed studies. The dark circles are the imputed studies. They are provided to show a graphic representation of the distribution.

Trim and Fill Results

Table 1 presents the results for all 18 scales for all four vendors. We describe the results in detail for scale A1 and summarize the results for the remaining scales. For Vendor A's scale A1, the 14 correlations yielded a mean correlation of .06 with a confidence interval from -.01 to .13. The trim and fill analysis found that five additional studies would be needed to make the distribution symmetrical and the trim and fill adjusted mean correlation is .00, a difference of .06 from the observed mean of .06. In the case of Vendor A's scale A1, there may be some publication bias operating, but it does not change the conclusion about the test which is that it has extremely low validity. Note that if the means of the observed and trim and fill adjusted distribution were higher and showed a .06 difference (e.g., .20 vs. .14), we would be more concerned about the extent of publication bias. However, since the mean of the observed and trim and fill adjusted distributions are both very low, we argue that any publication bias would not change one's conclusion concerning using the test. Relying on either mean validity, one would be unlikely to recommend the scale for selection purposes.

The results for Vendor A's scales A2-A5 show evidence of no or minimal publication bias. Results for scales A2 and A5 show that no studies were missing, thus there was no evidence of publication bias. Scales A3 and A4 showed .01 and.03 difference between the observed mean validity and validity after missing studies were imputed, respectively. These differences seem unlikely to alter a decision about whether to use the test.

The results for Vendor B showed evidence of publication bias in two of the three scales. For B1, six studies were missing, resulting in a difference of .13 between the observed mean validity and the mean validity after the six studies were imputed by the trim and fill procedure. For scale B2, four studies were missing and the difference between the observed mean validity and the trim and fill imputed validity was .12. For scale B3, there appears to be an outlier in the distribution, so we analyzed the data separately with and without the outlier. The nine correlations (including the outlier) yielded a mean correlation of .38. The trim and fill analysis found no evidence of publication bias. We also note that the average validity is of high magnitude (.38). This distribution includes a validity coefficient of .71 based on a sample size of 130. Because .71 is an extraordinarily large validity coefficient, particularly for a personality measure, we tried to verify it by consulting the original study. Because we were unable to verify the coefficient, we re-analyzed the data for this scale with the coefficient removed. The second row for scale B3 shows that the original mean validity is .29. Trim and fill found that three studies were needed to achieve symmetry and the validity of the trim and fill adjusted distribution was .26, a value .03 lower than the mean of the observed distribution (.29). We conclude that any publication bias is minimal and does not affect conclusions about the test.

The results for Vendor C, scales C2 and C4, showed that no studies were missing, thus there was no evidence of publication bias. Scale C3 is a subset of scale C4 and minimal evidence of publication bias was present in scale C3. Similarly scale C1 showed minimal evidence of publication bias (.03 difference between the observed mean and the trim and fill imputed mean).

The results for Vendor D showed little or no evidence of publication bias for four of their five scales. For Scale D2, however, four studies were found to be missing resulting in a difference of .08 between the observed mean (.24) and the trim and fill imputed mean (.16). Because it is likely that different conclusions would be reached about a test with validity of .16 than a validity of .24, we suggest there is a moderate to severe amount of publication bias in estimating the validity of scale D2.

In summary, the scales from Vendors A and C showed little or no evidence of publication bias. Two of the scales from Vendor B showed moderate to severe publication bias and the distribution of validities from the third scale included an outlier. When the outlier was removed from the analysis, there was little evidence of publication bias. One of the five scales from Vendor D showed moderate to severe publication bias, the remaining four scales showed little or no evidence of publication bias.

File Drawer Analysis Results

All 18 file drawer analyses, shown in the last two columns of Table 1, found that a fairly large number of “null” studies (11 to 574 studies) would be needed for the combined 2-tailed *p*-value to exceed .05. Rosenthal suggested that if the failsafe N is relatively small, then there is cause for concern that publication bias might be responsible for the observed results, but if this number is large, we can have greater confidence that although the observed treatment effect might have been inflated by the exclusion of some studies, it is nevertheless not nil. While Rosenthal did not provide specific guidance as to what number of studies might be considered “large” enough to give us confidence that the results have not been nullified by publication bias, he offered a general guideline that a failsafe N equal to or greater than 5 times the number of studies in the original meta-analysis, plus 10 studies (5K + 10) would indicate that the meta-analytic results were robust to the threat of publication bias. Mullen, Muellerleile, and Bryant (2001) proposed Rosenthal’s guideline as a formal rule, and several recent psychology meta-analyses (Del Vecchio & O’Leary 2004; Ma & Kishor 1997; Rhoades & Eisenberger, 2002) used this formula to assess the results of their file-drawer analyses. We used this number as one means of assessing publication bias based on the file drawer analysis results. Using 5K+ 10 as a criterion of robustness to the threat of total nullification of the effect due to publication bias, for Vendor A only scale A3 met this criterion, while for Vendors B, C, and D all scales met the criterion of 5K + 10. Another, more “lenient” rule is to consider that it is unlikely to expect there to be more missing studies than located studies. Using this as the criterion, all scales from all vendors, except for Vendor A’s scale A1, are unlikely to be totally nullified by missing studies. Thus, the results of Rosenthal's file drawer analyses are quite dissimilar to those following from the application of trim and fill.

Discussion

The results of our trim and fill analyses indicate that no or minimal bias is operating for any of the scales of Vendor A's Big 5 measure or for the scales presented in Vendor C's technical manual. However, there was evidence of moderate to severe bias for two of the three measures of Vendor B and in one of the five measures provided by Vendor D. In trying to locate sources of the possible bias in these data, we believe it is informative to examine the reporting practices of the test publishers for clues about the reasons for the asymmetry in the distributions of the scale validities of Vendors B and D. Unlike Vendors A and C, Vendor B reports only statistically significant correlations in the desired direction. This reporting practice is not explicitly stated in the manual, but the vendor confirmed that this was their reporting practice (personal communication, January 10, 2005). Vendor D sometimes reports all validity data, but for most studies reports only the statistically significant correlations in the desired direction. Vendor D's manual makes it clear that this is their reporting practice.

The *Principles for the Validation and Use of Selection Procedures* (2003) and the *Standards for Educational and Psychological Testing* (1999) state that researchers should report **all** the validity data for a test that are available to them, even if the correlations are low, not statistically significant, or are in a direction opposite to those expected. We understand that publication bias may be present in test publisher data through no fault of the test publisher (e.g., when validity data are published or released to the test publisher when the validities are of high magnitude but not when they are of low magnitude). However, test publishers, and other researchers, have a role in preventing publication bias by publicly reporting all relevant known validity data.

We note that clear reporting of test validity data is an exception and not common practice. Some test publishers provide narrative summaries of past validity studies but sample sizes and validity coefficients are often not provided. Other test publishers provide copies of primary validity studies but do not have a technical manual that summarizes the data. Through the authors' personal experience, we know some test vendors make claims of validity in marketing materials but will not release any validity results in response to inquiries. As researchers in personnel selection, we find it disheartening that so few test publishers offer any validity data or claim to have validity results but are unwilling or unable to provide them.

We are not suggesting that publication bias is a widespread problem in industrial and organizational psychology. In fact, not all publication bias analyses have shown there to be such bias. Whetzel (2006) reanalyzed Frei and McDaniel's (1998) meta-analysis of the validity of customer service tests and found that there was no apparent publication bias in their dataset. McDaniel, Hurtz, and Donovan (2006) reanalyzed a subset of the Hurtz and Donovan (2000) meta-analysis and found no publication bias. Vevea, Clements, Hedges (1993) analyzed the validity data for the General Aptitude Test Battery validities and found no evidence of publication bias that would alter conclusions about the validity. However, with the exception of these studies, little attention has been paid to this important issue.

Implications of findings for test vendors

We have noted that the *Principles for the Validation and Use of Selection Procedures* and the *Standards for Educational and Psychological Testing* state that researchers should report all the validity data for a test. Some might take exception to this guidance. For example some will note that the *Principles* emphasize that its recommendations are aspirational and may not need to be followed even when practical. Others will note that all validity data may not be relevant to making decisions about a test. For example, many personality instruments have multiple scales that may not be applicable to the prediction of all criteria. Some will argue that it is not reasonable for test vendors to update their test manuals each time new data become available. Also, some test vendors have a large amount of validity data and the reporting of new validity data has little incremental informational value. Others may question the extent to which test vendors should seek out data on their tests collected by others. Still others might argue that test vendors are a business and should be allowed to follow any practice that is legal. Some might encourage professional associations, such as the Society of Industrial and Organizational Psychology, to play an active role in identifying and addressing publication bias in test vendor manuals or exposing false marketing claims of test vendors. Some might question the reasonableness of using test vendor data as part of a legal defense for test use unless it is clear that no publication bias exists in the manual. We anticipate that others may identify additional issues.

We do not presume to know the best answers to these issues. We do offer the following recommendations as reasonable. We encourage test publishers to report all validity data in their possession that are judged relevant to the use of the tests as well as characteristics of studies that may affect validity (e.g., substantial range restriction or estimates of validity based on use of samples that only included sample members with low or high criterion scores). We encourage test publishers to state explicitly the decision rules they use in determining if a result is relevant. Test publishers should apply the decision rules consistently. Thus, if the correlation between test X and criterion Y is judged relevant in study 1, the same predictor-criterion combination should be judged relevant in study 2. The decision rules should not include considerations regarding the magnitude of the validity coefficient or its direction. We also encourage test publishers to conduct and report publication bias analyses. Publishers should consider explaining their reporting policies, including the frequency of updates to their manuals. We believe these recommendations would increase the confidence that test users can place in test vendor manuals.

Implications of test vendor publication bias for meta-analyses in industrial and organizational psychology

One important implication of publication bias is the effect it has on the conclusions of some meta-analyses. The documentation of publication bias in some test vendor's data calls into question past meta-analyses that have relied primarily on test vendor data. For example, meta-analyses of integrity tests (Ones, Viswesvaran & Schmidt, 1993; 2003) have been primarily based on data from test vendors. Previously, Camara and Schneider (1994) had raised concerns about the credibility of integrity test validity data obtained from test vendors and noted that most integrity test validity studies are conducted by test vendors (See Ones, Viswesvaran & Schmidt, 1995 for a commentary and rebuttal). The concerns we raise about integrity test meta-analyses are clearly speculative and are best resolved by conducting publication bias analyses of the data.

Our concern is not limited to integrity test meta-analyses or test vendor data. As mentioned earlier in this paper, Duval (2005) concluded that there was evidence of substantial publication bias in the McDaniel, et al. (1994) data on the validity of structured interviews. The data were drawn from many published and unpublished sources. Because these data were drawn from very diverse sources, the biased results cannot be attributed to specific suppliers of data. Rather the bias likely lies with the decisions of primary study authors or editors. These sources of publication bias can affect the meta-analytic results of any topic.

Meta-analytic studies have a substantial impact as judged by citation rates and researchers and practitioners often rely on meta-analytic results as the final word on research questions. Some speculate that meta-analyses can suppress new research in an area if there is a perception that the meta-analysis has largely settled all research questions. For example, following the employment interview meta-analyses of the 1990's, few primary studies on the validity of structured vs. unstructured employment interviews have been added to the literature. Given the evidence for publication bias in structured interviews and the publication bias found in the data reported in some test publishers' technical manuals, we believe that it would be prudent to conduct publication bias analyses of all past meta-analyses of validity data. We also recommend that publication bias analyses be a routine practice in all future meta-analyses in industrial and organizational psychology. This information is important, as it allows us to have confidence that the meta-analysis is likely to be accurate. In cases where publication bias analyses suggest that severe bias may exist, this can serve to avoid potentially serious mistakes such as recommending a policy, practice or intervention that could be useless or even harmful. We suggest that it is important to address bias, not only to ensure the integrity of individual meta-analyses, but also to reinforce the credibility of the meta-analytic method. By encouraging the application of publication bias analysis we hope to further the use, and usefulness, of meta-analysis.

It is important that competently prepared publication bias studies be published regardless of the outcomes (including those that show no publication bias operating). If the only articles that are published are those demonstrating bias problems, one may conclude falsely that publication bias is rampant in I/O psychology literature.

Limitations of trim and fill

The trim and fill procedure used in these analyses rests on the assumption that asymmetry is evidence of publication bias. This assumption, although reasonable, could be incorrect in specific applications of the method. For example, through random factors, all validity coefficients for a test may fail to form a symmetric funnel. Systematic factors, unrelated to publication bias, might also result in asymmetry. In the current study, two systematic factors, moderators and statistical artifacts, may exist in the data. To the extent that moderators of these validity data exist, the appropriateness of the trim and fill method may be called into question (Terrin, Schmid, Lau & Olkin, 2003). As Sterne and others (Sterne & Egger, 2005) have cautioned, the asymmetry detected by funnel plot based methods may be due to the fact that small sample studies may actually differ from large sample studies in important ways. For example, the small samples may be higher (or lower) on a moderator that might lead to these samples having disproportionately higher validities than larger studies, and cause asymmetry in the distribution. Likewise, the impact of measurement error, range restriction, and range enhancement, if different in the small studies as a group than in the large studies as a group, might distort the results.

The assumption that the missing studies are the most negative studies may be questionable in some situations. A reviewer suggested that some validities might be missing simply because they add “nothing new” to the literature. We believe that in the case of employment test validities, tests publishers will be highly motivated to keep track of, and to encourage publication of all positive results. In areas in which the literature is less practice-oriented, it may be possible that findings with unsurprising results are the ones that do not get published, but we do not think that is the case here. We also suggest that if the ”nothing new effect” is operating, it should either: 1) not selectively affect publication of studies on one side of the distribution, or 2) encourage the publication of studies that are counter to the prevailing results, which in this case could lead to the *increased* publication of negative results. It certainly would not lead to the pattern we see in this manuscript. Nevertheless, to be responsive to the reviewer’s concern we ran an additional analyses to see if there were missing “positive” studies.

When we applied trim and fill to look for missing positive studies, the results indicated that studies missing from the right side of the funnel (where larger positive validities would reside) was rare. In no case when there was evidence of some studies being missing from the left side of the funnel, were any studies also missing from the right. For Vendor A’s scale A2, trim and fill suggested that one study was missing from the right side of the funnel. While the mean of the observed distribution was .06, the mean of the trim and fill adjusted distribution was .08, a trivial difference. For Vendor C's scale C2, trim and fill suggests that one study can be imputed to the right. While the observed mean is .31, the mean of trim and fill distribution is .32, a trivial difference. For Vendor B's scale B3 that includes the outlier (*r* = .71), trim and fill suggests that three studies can be imputed to the right. The observed mean is .38 while the mean of trim and fill distribution is .48. Although this is a large difference, it is based on a distribution known to contain a questionable data point. No studies were missing from the right side of the distribution for any of Vendor D's scales. In sum, none of the 18 distributions showed any credible evidence of data suppression of large validity coefficients from the right side of the funnel. We note that even in analyses where we concluded that bias was minimal, the bias was almost always in the direction of overestimating the validity of the test.

Trim and fill defines publication bias as asymmetry in a distribution of correlations. Vendor B reported only statistically significant correlations and Vendor D usually reported only statistically significant correlations. Thus, the validities of many of their tests are likely to be overestimated in their technical manuals due to publication bias. However, there was no detected asymmetry in one of the three Vendor B tests and in four of the five Vendor D tests. Thus, trim and fill failed to detect publication bias in these five tests known to have suppressed statistically non-significant correlations.

An additional limitation of trim and fill is the need for moderator-free distributions. Thus, for scale D5, we had to restrict our analyses to studies of jobs that were at a high level of complexity to control for this moderator.

Although researchers can isolate distributions that are apparently free from moderators, there is always some chance that a moderator unknown to the researcher may be distorting the trim and fill results.

Limitations of the failsafe N

We believe that our analyses provide a useful case study concerning why the Rosenthal failsafe N should not be used in publication bias analysis. In cases such as Vendors B and D, in which only significant positive correlations tended to be reported, the trim and fill analysis found evidence of this bias; however, the failsafe N analysis did not. Thus, even when there is moderate to severe publication bias, the failsafe N method can yield incorrect conclusions. Thus, we concur with Becker (2005) that the failsafe N method should not be used.

Non-statistical methods for detecting publication bias

There are three non-statistical methods that could be used to make inferences about publication bias. First, when sample sizes are relatively small and the mean correlations are not far from zero, one should expect some of the correlations to be in directions counter to that expected. So, for example, our research literature suggests positive correlations between measures of certain constructs and job performance. Thus, in Vendor’s A data, we expect to see positive predictor-criterion relations for tests measuring those constructs. Because we do not see complete positive relationships, this gives us greater confidence that Vendor A reported all available validity data. Second, when sample sizes are relatively small and effect sizes are not large, one should expect some correlations to not be statistically significant. Thus, Vendors A and C reported at least one non-significant validity thus giving us greater confidence that the test vendor reported all available data. Third, one can examine correlations by data source if data source is a reasonable correlate of publication bias. McKay and McDaniel (2006) noted that journals reported lower mean racial differences in job performance than unpublished technical reports. They speculated that there might be publication bias in this literature. This finding prompted McDaniel, McKay and Rothstein (2006) to conduct a trim and fill publication bias analysis on the data that indicated that larger mean racial differences are suppressed in the journal literature while the unpublished technical report data appears unbiased.

Conclusion

This paper has presented a review of two statistical methods for detecting of publication bias. We suggest that the use of the failsafe N procedure be discontinued and replaced by procedures that are more informative. We also recommend the use of the trim and fill procedure and find it appropriate for the data analyzed in this study. As a case study for the trim and fill methodology, we reviewed the validity coefficients provided in the technical manuals of four test vendors. For two of the vendors, the distributions of scale validities were found to be symmetrical and consistent with the finding of no publication bias. At least one distribution from each of the other two test vendors showed asymmetry, consistent with the finding of publication bias. This finding is consistent with their stated reporting practices of typically providing only statistically significant validity coefficients in their manuals. We encourage all test publishers to present unbiased validity data, and we urge industrial and organizational psychologists to routinely incorporate trim and fill or other appropriate methods of publication bias assessment into their meta-analyses.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Becker, B.J. (1994). Combining significance levels. In H. Cooper and L. Hedges (Eds). *The Handbook of Research Synthesis,* NY: Russell Sage , 215-230.

Becker, B.J. (2005). The Failsafe N or File-drawer Number. In H. Rothstein, A.J. Sutton, & M. Borenstein (Eds). *Publication Bias in Meta Analysis: Prevention, Assessment and Adjustments*. Wiley. 111-126.

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology, 78*, 387-409.

Borenstein, M., Hedges, L., Higgins, J. & Rothstein, H. (2005). *Comprehensive meta-analysis. Version 2*. Englewood, NJ: Biostat.

Brewer, E.W., & Shapard, L. (2004). Employee burnout: A meta-analysis of the relationship between age or years of experience. *Human Resource Development Review, 3*, 102-123.

Camara, W. J. & Schneider, D. L. (1994). Integrity tests: Facts and unresolved issues. *American Psychologist*, 49, 112-119.

Cooper, H.M. (2003). Editorial. *Psychological Bulletin, 129*, 3-9.

Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology*, 37, 131-146.

Del Vecchio, T. & O’Leary (2004). Effectiveness of anger treatments for specific anger problems: A meta-analytic review. *Clinical Psychology Review, 24*, 15-34.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. Rothstein, A.J. Sutton, & M. Borenstein (Eds). *Publication Bias in Meta Analysis: Prevention, Assessment and Adjustments*. Wiley. 11-34.

Digman, J.M. (1989). Five robust trait dimensions: Development, stability, and utility. *Journal of Personality, 57*, 195-214.

Duval, S. (2005). The “Trim and Fill” method. In H. Rothstein, A.J. Sutton, & M. Borenstein (Eds). *Publication Bias in Meta Analysis: Prevention, Assessment and Adjustments*. Wiley. 127-144.

Duval S.J. and Tweedie R.L. (2000a). A non-parametric "Trim and Fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89-98.

Duval, S. J, & Tweedie, R.L. (2000b). Trim and fill: A simple funnel plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 276-284.

Field, A.P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods,* 10, 444-467.

Frei, R.L., & McDaniel, M.A. (1998). Validity of customer service measures in personnel selection: A review of criterion and construct evidence. *Human Performance, 11*, 1-27.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.

Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hedges, L. & Vevea, J. (2005). The selection model approach to publication bias. In H. Rothstein, A.J. Sutton, & M. Borenstein (Eds). *Publication Bias in Meta Analysis: Prevention, Assessment and Adjustments*. Wiley.

Hunter, J.E., & Hunter,R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72-98.

Hurtz, G.M. & Donovan, J.J. (2000). Personality and job performance: The big five revisited. *Journal of Applied Psychology, 85*, 869-879.

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

Jenkins, G.D., Mitra, A., Gupta, N., Shaw, J.D. (1998). Are financial incentives related to performance? A meta-analytic review of empirical research. *Journal of Applied Psychology, 83*, 777-787.

Light, R. J. & Pillemer, D. B. (1984). *Summing up*. Boston, MA: Harvard University Press.

Ma, X. & Kishor, N. (1997). Attitude toward self, social factors, and achievement in mathematics: A meta-analytic review. *Educational Psychology Review, 9,* 89-120.

McDaniel, M.A., Hurtz, G.M., & Donovan, J.J. (2006). *An evaluation of publication bias in Big 5 validity data.* Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology. Dallas, TX

McDaniel, M.A., McKay, P. & Rothstein, H. (2006, May). *Publication bias and racial effects on job performance: The elephant in the room.* Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology. Dallas, TX

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*, 599-616.

McKay, P. & McDaniel, M.A. (2006). A re-examination of Black-White mean differences in work performance: More data, more moderators. *Journal of Applied Psychology*.

McNatt, D. B. (2000). Ancient Pygmalion joins contemporary management: A meta-analysis of the result. *Journal of Applied Psychology, 85*, 314-322.

Mitra, A.; Jenkins, G. D. & Gupta, N. (1992). A meta-analytic review of the relationship between absence and turnover. *Journal of Applied Psychology, 77*, 879-889.

Mullen, B., Muellerleile, P. & Bryant, B. (2001). Cumulative meta-analysis: A consideration of indicators of sufficiency and stability. *Personality and Social Psychology Bulletin*, *27,* 1450-146.

Ones, D.L., Viswesvaran, C. & Schmidt, F.L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*, 679-703.

Ones, D.L., Viswesvaran, C. & Schmidt, F.L. (1995). Integrity tests: Overlooked facts, resolved issues, and remaining questions. *American Psychologist, 50*, 456-457.

Ones, D.L., Viswesvaran, C. & Schmidt, F.L. (2003). Personality and absenteeism: a meta-analysis of integrity tests. Personality and industrial, work and organizational applications. *European Journal of Personality, 17(Suppl1)*, S19-S38.

Orwin, R.F. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics, 8*, 157-159.

Parker, C.P., Baltes, B.B., Young, S.A., Huff, .W., Altmann, R.A., Lacost, H.A., & Roberts, J.E. (2003). Relationships between psychological climate perceptions and work outcomes: a meta-analytic review. *Journal of Organizational Behavior, 24*, 389-416.

Rhoades, L. & Eisenberger, R. (2002). Perceived organizational support: A review of the literature. *Journal of Applied Psychology, 87*, 698–714.

Rivkin, D., & McDaniel, M. A. (1990). *The measurement and validation of occupational aptitude requirements*. In A. Lancaster (Chair), The enhancement of the Department of Defense Student Testing Program. Symposium presented at the 98th Annual conference of the American Psychological Association, Boston, MA.

Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin, 86*, 638-641.

Roth, P.L., Bevier, C.A., Bobko, P., Switzer, III, F.S., Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*, 297-330.

Rothstein, H., Sutton, A.J., & Borenstein, M. (Eds) (2005). *Publication Bias in Meta Analysis: Prevention, Assessment and Adjustments*. Wiley.

Salgado, J.F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., Rolland, J.P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology, 88*, 1068-1081.

Society of Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures. Fourth edition*. Bowling Green, OH: Author.

Sterne, J.A.C. & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. Rothstein, A.J. Sutton, & M. Borenstein (Eds). *Publication Bias in Meta Analysis: Prevention, Assessment and Adjustments*. Wiley. 99-110.

Sutton, A. J. & Pigott, T.D. (2005) Bias in meta-analysis induced by incompletely reported studies. In H. Rothstein, A.J. Sutton, & M. Borenstein (Eds). *Publication Bias in Meta Analysis: Prevention, Assessment and Adjustments*. Wiley. 223-240.

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

Vevea, J.L., Clements, N.C., Hedges, L.V. (1993). Assessing the effects of selection bias on validity data for the General Aptitude Test Battery. *Journal of Applied Psychology, 78*, 981-987.

Whetzel, D.L. (2006, May). *Publication bias the validity of customer service measures.* Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology. Dallas, TX

Table 1. Results from meta-analysis, trim and fill analysis, and failsafe N

	Meta-analysis Results			**Trim and Fill Results**			**Difference between Meta-analysis and Trim and Fill**	**Rosenthal Results**	
Vendor/Scale	# of studies	Mean r	95% CI	# studies missing	Mean r	95% CI	Mean r difference	Failsafe N	5K + 10[1]
A1	14	.06	.01-.13	5	.00	-.08-.07	.06	11	80
A2	14	.06	.03-.11	0	.06	.03-.11	.00	21	80
A3	14	.23	.20-.27	4	.22	.18-.25	.01	468	80
A4	14	.10	.05-.15	4	.07	.01-.12	.03	71	80
A5	14	.06	.02-.10	0	.06	.02-.10	.00	14	80
B1	12	.33	.23-.42	6	.20	.09-.31	.13	436	70
B2	11	.36	.23-.48	4	.24	.09-.38	.12	373	65
B3 (with outlier)	9	.38	.21-.53	0	.38	.21-.53	.00	193	55
B3 (without outlier)	8	.29	.20-.38	2	.26	.18-.34	.03	574	50
C1	8	.22	.18-.25	2	.21	.18-.24	.01	255	50
C2	4	.31	.27-.36	0	.31	.27-.36	.00	574	30
C3	7	.24	.18-.30	2	.21	.15-.28	.03	98	45
C4	7	.28	.22-.34	0	.28	.22-.34	.00	126	45
D1	10	.20	.14-.26	2	.18	.11-.24	.02	138	60
D2	8	.24	.16-.32	4	.16	.07-.25	.08	113	50
D3	11	.26	.19-.32	3	.24	.18-.29	.02	162	65
D4	12	.26	.20-.31	0	.26	.20-.31	.00	201	70
D5	6	.23	.15-.30	0	.23	.15-.30	.00	40	40

[1] 5K + 10 is a decision rule for interpreting a failsafe N. A failsafe N equal to or greater than 5 times the number of studies in the original meta-analysis, plus 10 studies (5K + 10) would indicate that the meta-analytic results were robust to the threat of publication bias (Rosenthal, 1979; Mullen, Muellerleile & Bryant, 2001)

Figure 1. Funnel plot

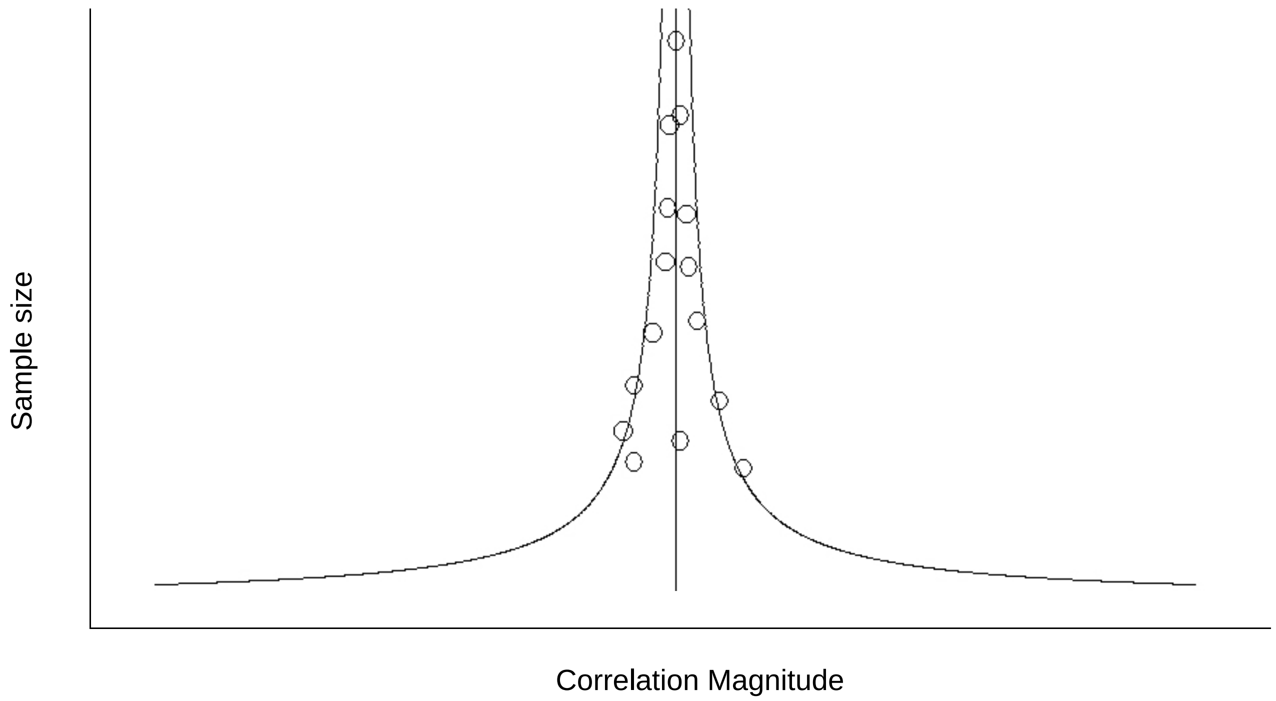


Figure 2. Illustrative symmetrical and asymmetrical funnel plots

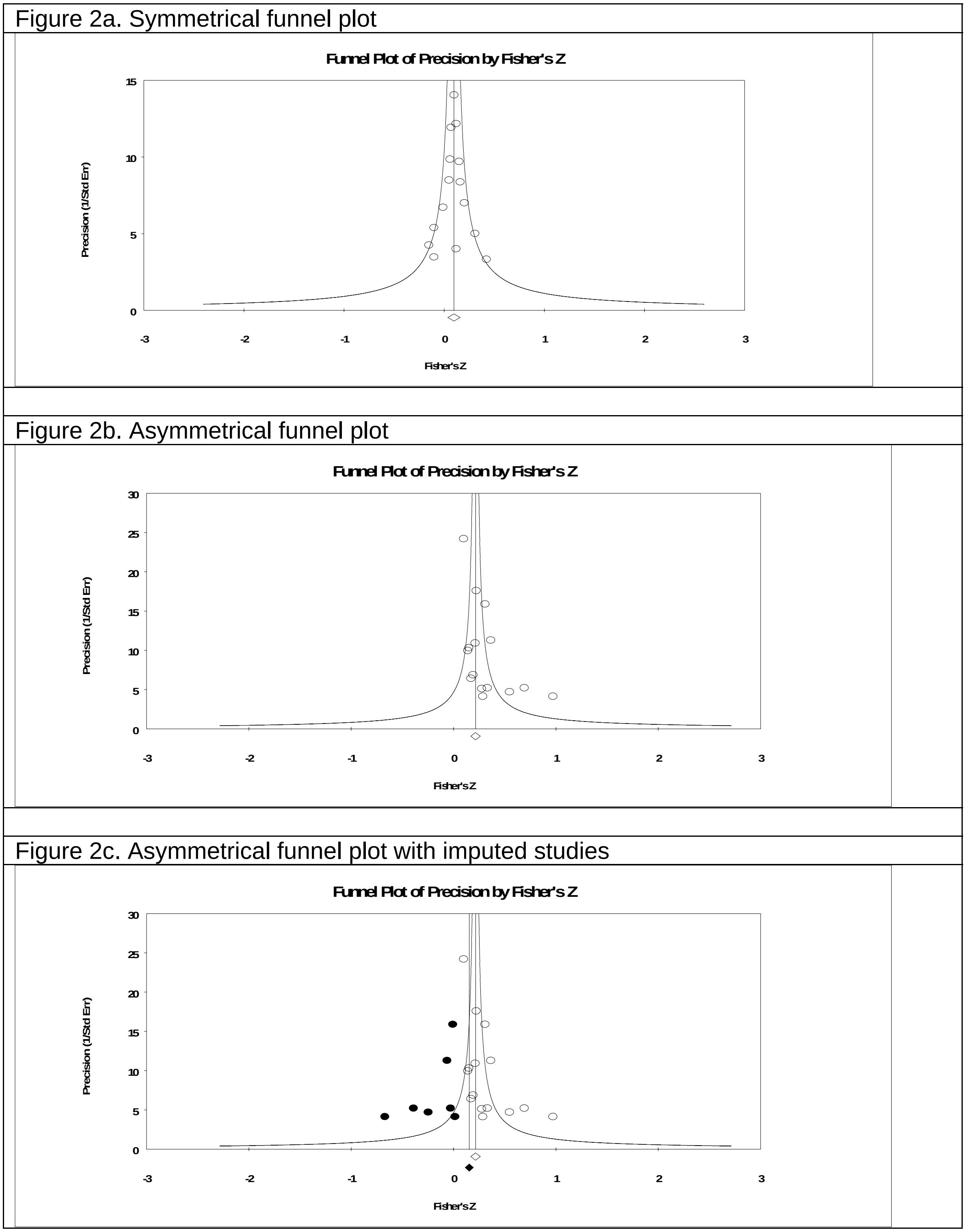
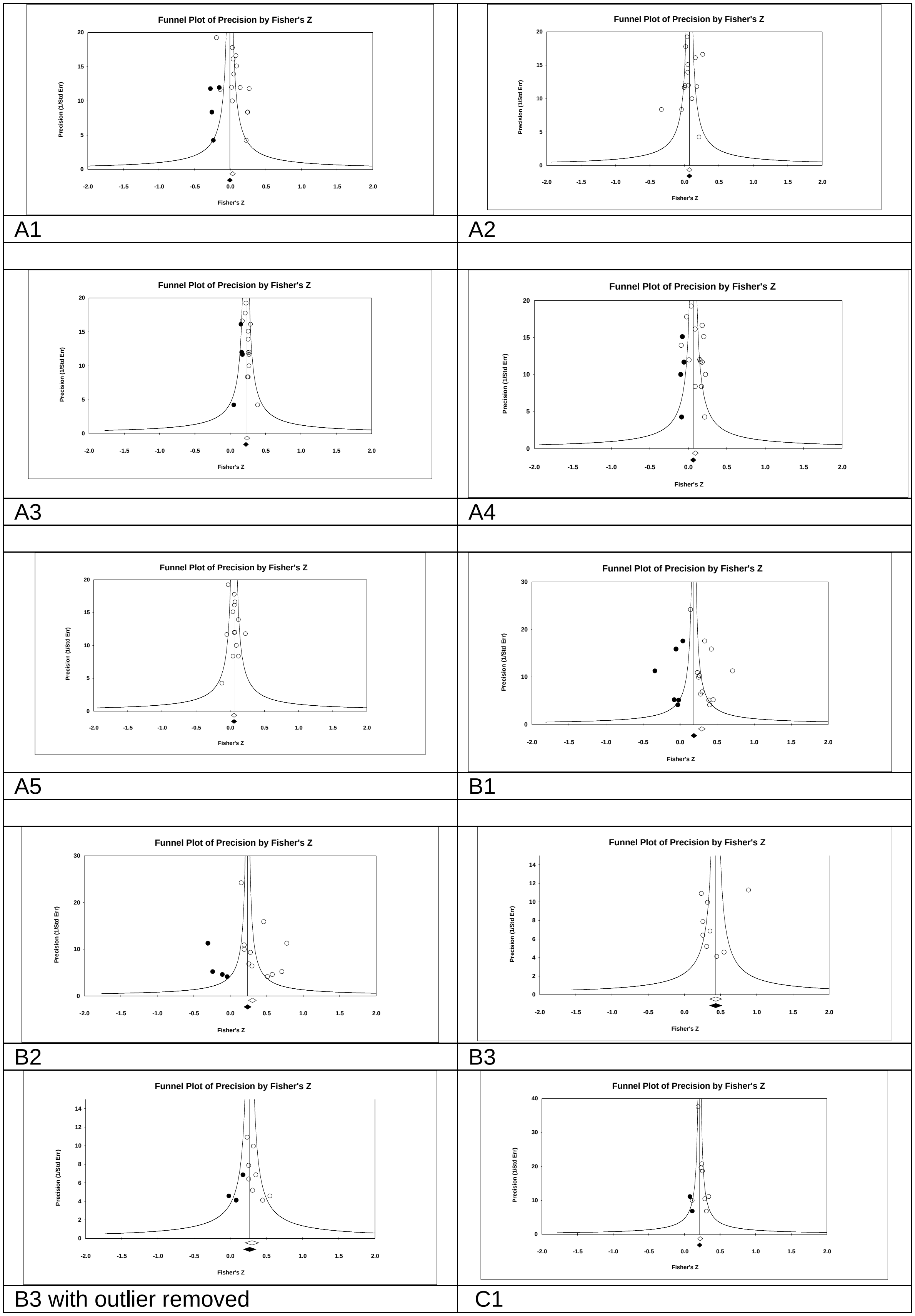


Figure 2a. Symmetrical funnel plot

Figure 2b. Asymmetrical funnel plot

Figure 2c. Asymmetrical funnel plot with imputed studies

Figure 3. Funnel plots for each test distribution



A1

A2

A3

A4

A5

B1

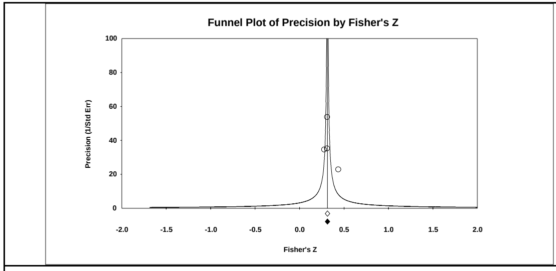
B2

B3

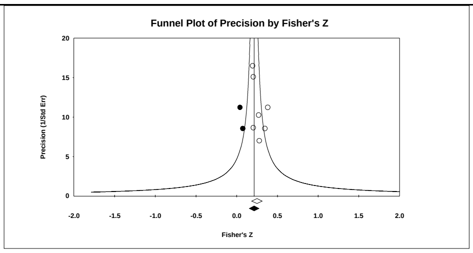
B3 with outlier removed

C1

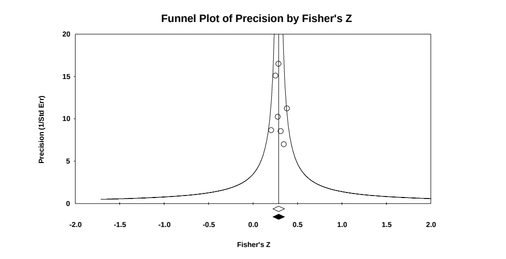
Figure 3. continued



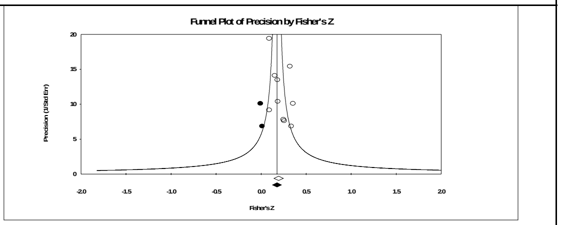
C2



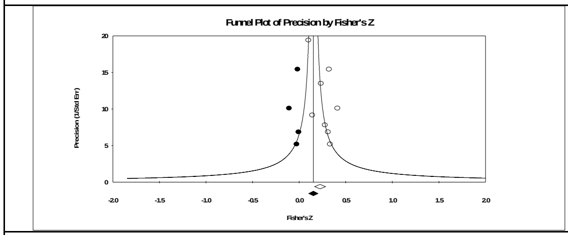
C3



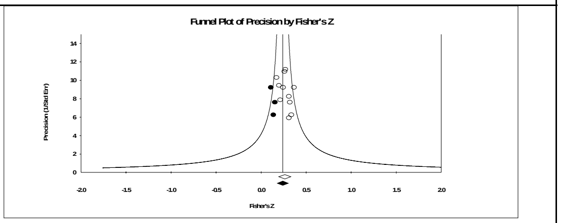
C4



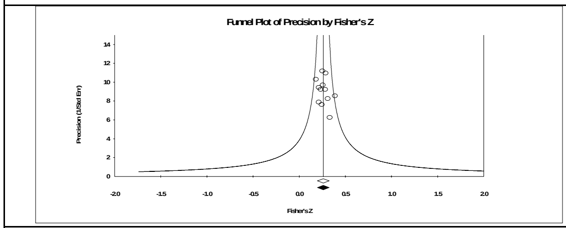
D1



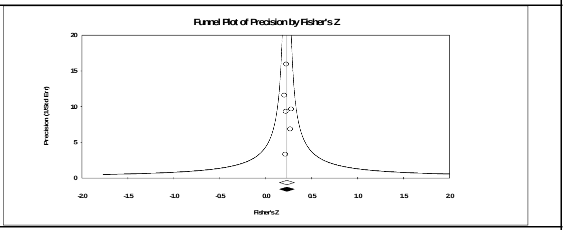
D2



D3



D4



D5