Paper presented at the annual conference of the Academy of Management Convention, Honolulu. August 2005

Do Employment Test Publisher Manuals Provide Accurate Information On Test Validity?

Michael A. McDaniel
Virginia Commonwealth University
Hannah R. Rothstein
Baruch College
Deborah L. Whetzel
Work Skills First, Inc.

Abstract

This paper has two goals. First, we discuss the issue of publication bias and explain why it presents a problem for organizational and human resources research. After reviewing the traditional failsafe N, or file drawer analysis, we introduce a more sophisticated method of publication bias analysis (Trim and Fill) that has been developed in the medical literature but which is largely unfamiliar to human resources, organizational behavior, and management researchers .We demonstrate the Trim and Fill approach to conducting publication bias analyses by applying it to validity information reported in three test publishers' technical manuals. In doing so, we assess the likelihood that criterion-related validity information provided by test publishers may overestimate test validity. Our application of the Trim and Fill publication bias method to twelve validity distributions (five from one publisher, three from a second publisher, and four from a third publisher) found evidence of either no bias or minimal bias for most of the distributions. However, in the case of one publisher, the level of bias in two of three cases was serious enough to substantially change validity estimates.

Publication bias is the term used to refer to the fact that not all completed studies on a topic make their way into the published literature, and that these studies are likely to be systematically different from those that do appear in the literature. When the results of publicly available research differ from the results of **all** the research that has been done in an area, readers and reviewers are in danger of being misled. Publication bias can have powerful consequences, particularly when a dangerous or ineffective practice is viewed as effective or safe because of selective publication of results. For example, Merck pharmaceutical recently recalled Vioxx, its popular arthritis drug. While Merck maintained that it recalled Vioxx as soon as the data indicated a high prevalence of cardiovascular problems among those who took this drug for more than 18 months, media reports claimed that Merck hid adverse event evidence for years (Wall Street Journal, November 1, 2004). Similarly, the attorney general of New York State, filed a 2004 lawsuit against GlaxoSmithKline, in which it was claimed that they concealed data about the lack of efficacy, and about the increased likelihood of suicide associated with the use of Paxil by children and teenagers (NY vs. GlaxoSmithKline).

In the human resources and organizational behavior literature, the consequences of publication bias are not likely to be fatal; nevertheless we can think of several areas where they can be quite serious. One of the most important is the area of employment test validities. Specifically, there are important negative consequences for employers, and job applicants, if an invalid test is falsely viewed as valid, or if a test with low validity is viewed as a having high validity. Unfortunately, we know next to nothing about the likelihood that publication bias is affecting the published information on employment test validities; in fact we know very little about the extent to which publication bias threatens the validity of most research literatures in human resources or organizational behavior. While there is a burgeoning literature examining publication bias in healthcare, (Dickersin, 2005; Halpern & Berlin, 2005) and while the topic has received increased attention among some areas of psychological research (Bogg & Roberts, 2004; Cooper, 2003; Mezulis, Abramson, Hyde & Hankin, 2004; Moyer, Rounds & Hannum, 2004; Robbins, Lauver, Le, Davis, Langley & Carlstrom, 2004; Smith, McCullough & Poll,2003; Vitaliano, Zhang & Scanlan, 2003), the potential problems posed by publication bias have generally been ignored in the personnel, human resources, organizational behavior and management literatures. To the extent that publication bias has been addressed in our literature, the outdated failsafe N method, also known as the “file drawer problem” method is used. This method has been shown to be a poor procedure for assessing publication bias (Becker, 2005) and more accurate and powerful approaches have been offered (Duval, 2005; Hedges & Vevea, 2005; Sterne & Egger, 2005; Sutton & Pigott, 2005).

The current paper has a two-fold purpose. First, we describe two methods for the assessment of publication bias. After reviewing and critiquing the familiar failsafe N, or file drawer analysis method, we describe a newly introduced procedure to assess publication bias. This procedure, called Trim and Fill (Duval

& Tweedie, 2000a, 2000b), has been used widely in the healthcare literature, but is largely unfamiliar to, and therefore unused by, human resources, organizational behavior, and management researchers. Second, we illustrate the operation of the Trim and Fill procedure by using it to explore the possibility that criterion-related validity information in test publisher manuals is subject to publication bias. Specifically, we apply Trim and Fill to sets of employment test validities drawn from the manuals of three test publishers.

Publication Bias and Procedures for Its Detection

Publication bias is classically defined as the tendency to publish studies depending on the magnitude, direction or statistical significance of their results. There are several possible causes for the suppression of negative, low, or statistically non-significant research results; most of them are clearly relevant to the domain of test validities. We outline these here. There is considerable evidence that one cause of bias is that researchers are not likely to submit negative results for publication (Dickersin, 2005). We suggest that researchers also are not likely to share weak (poor) validity results with the test publisher. For example, a researcher who has found that a test in use by his or her company has poor validity may withhold these data due to concerns about employment litigation or embarrassment to the organization. There also is some evidence that editorial policy, at least in some journals, favors the publication of significant results (Dickersin, 2005; Greenwald, 1975). Similarly, test manual publishers might favor the reporting of positive, statistically significant validities. This could be because they view these as more important or more interesting, or because they are concerned that insignificant, low or negative validities present the test in a poor light.

Rosenthal Failsafe N (the file drawer problem)

Rosenthal (1979) introduced what he called the “file drawer problem.” His concern was that some non-significant studies may be missing from an analysis (i.e., hidden in a file drawer) and that these studies, if included, would nullify the observed effect. By “nullify,” he meant to reduce the cumulated effect to a level where it was 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 indeed cause for concern. However, if this number is large, we can be confident that the mean observed correlation, while possibly inflated by the exclusion of some studies, is nevertheless not zero. The failsafe N method has been used to assess the likelihood of publication bias in several human resources and organizational behavior meta-analyses including Jenkins, Mitra, Gupta and Shaw (1998), Mitra, Jenkins and Gupta (1992) and Rhoades and Eisenberger (2002).

This approach is limited in two important ways (Becker, 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 shown 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. That is, it may allow one to assert that the mean correlation is not zero, but does not address what the unbiased correlation is and whether it remains of useful size after the missing studies have been included.

Consider an employer choosing between two tests A and B 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 what the validity of tests A and B is likely to be in the absence of the 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 in no way helps to determine what the validity of the tests are in the absence of publication bias.

In acknowledgment of the fact that statistical significance levels are not typically as of much concern as are effect sizes, Orwin (1983) extended the idea of the failsafe N to effects 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 has also been used in I/O and OB meta-analyses, (a good example is 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 help us estimate what we really want to know, namely the likely magnitude of the population effect, taking into account the studies that may exist, but that are missing from our analysis. We concur with Becker (2005) who maintains that the failsafe N should no longer be used to assess publication bias.

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 graph (Light & Pillemer, 1984). A funnel graph, or 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. This means that, on average, 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. In summary, funnel graphs display a collection of correlation coefficients using the sample size and the magnitude of the correlation to determine the location of the study on the graph. The plotted studies tend to form into an inverted funnel shape with large sample studies clustering tightly around the center at the top of the funnel and small sample studies being dispersed widely at the bottom of the funnel.

The concept of precision is of relevance to funnel graphs. 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 graph (see Sterne and Egger, 2005 for a discussion of why precision is a better choice than sample size for the Y axis).

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 1.0 or -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 the publication bias methods we will describe examine the symmetry of a funnel plot of the correlations, these plots of correlations need to be based on correlations transformed to Fisher *z* rather than on a plot of the correlations themselves. Figure 2a shows a symmetrical funnel graph plotting correlations, expressed as Fisher *z* as a function of precision.

Using the example of employment test validities, if a test publisher's manual reported all the criterion-related validity studies that had been conducted, we expect that the studies in the funnel plot would 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 correlations are censored we expect an asymmetric funnel plot, with a relatively high number of small studies falling toward the right (representing a large validity coefficient) 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. The Trim and Fill method of publication bias assessment focuses on detecting asymmetry in the distribution of effect sizes.

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. As discussed above, the key idea behind the funnel plot is that in the absence of bias, the plot would be symmetric about the summary effect (e.g., 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.

Trim and Fill uses an iterative procedure to remove the most extreme small studies from the positive side of the funnel plot, re-computing the effect size at each iteration, until 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. This “fill” has no impact on the point estimate but serves as a correction to the variance (Duval & Tweedie, 2000a, 2000b).

We believe that 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. Following the application of Trim and Fill to a given distribution of effects (again, in our case, test validities) we can assess the degree of divergence between the original mean effect and the adjusted mean effect. It could be that we fill find that the adjusted effect is basically similar to the original effect, or we could find that the size of effect size has changed but the core finding (e.g., that the test does or does not have a useful level of validity) remains unchanged. In some cases, however, the adjusted result may call the original findings into question. We suggest that the potential impact of bias could be regarded as “minimal” when the unadjusted and adjusted effect distributions yield essentially similar estimates of the effect size, that its impact be called “moderate” when the size of the validity coefficients change substantially but the key finding (e.g., that a test is or is not useful) remains in force, and that the impact of potential bias be labeled “severe” when the basic conclusion of the analysis (e.g., that the test is operationally useful) is called into question.

In the remainder of this paper, we apply Trim and Fill analysis to twelve distributions of criterion-related validity data drawn from the manuals of three test publishers and demonstrate how results can be affected if validities presented in test publisher manuals have been selected based on direction, significance, or magnitude.

Method

Data source. We obtained test publisher technical manuals on several tests from three test publishers. The source of the data and the decision rules used for each data set are described.

We obtained data from the Personal Characteristics Inventory (Wonderlic, Inc., 2002) which is a measure of the Big 5 personality traits (Digman, 1989). Table 20 of the manual (p. 5-45) presents multiple validity coefficients from each of the Big 5 scales: extraversion, agreeableness, conscientiousness, (emotional) stability, and openness. The validity data for each scale were analyzed separately. We restricted our analyses to the criteria listed as supervisory ratings and excluded other criteria such as sales, new accounts opened, and voluntary turnover. We did this because different criterion types (supervisory ratings, sales, etc.) often yield different mean validities, that is, the criterion type is often a moderator. Since the Trim and Fill method assumes that the correlations are homogeneous (a moderator-free distribution), we limited the analyses to supervisory ratings so as not to contaminate the analyses with any potential criterion-type moderator. Also, supervisor ratings are the most frequently used criterion in validation studies. With this decision rule, the data set contains only one validity coefficient per sample.

We also obtained validity data on the three occupational scales of the Hogan Personality Inventory (Hogan & Hogan, 1995). This test reports scores on a seven factor implementation of the Big 5. It achieves seven factors by splitting extraversion into ambition and sociability. It also splits openness into intellectance and school success. The technical manual does not provide criterion-related validity coefficients for the seven factor scales; however, it does provide criterion-related validity coefficients for three occupational scales: Service Orientation, Stress Tolerance, and Reliability. These occupational scales are derived from item clusters of the seven factor scales. The criterion-related validity data of the occupational scales are listed in Table 6.2 of their technical manual (p. 66-67).

The HPI occupational scale validity studies use a variety of criteria including supervisory ratings, times counseled for aberrant behavior, number of absences above allowable and annual commission. As with the data from the Personal Characteristics Inventory, we used data based solely on supervisory 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 table often listed more than one validity coefficient for each sample. For example, citing data from Muchinsky (1987), the data records four validity coefficients for 50 office managers. We adopted a decision rule of choosing the supervisory rating of “overall performance” for inclusion in the analysis and excluded the other coefficients from each sample. When a validity coefficient was not reported for an overall performance measure, we averaged across the available validity coefficients. For example, citing Muchinsky (1987), two validity coefficients were reported for 102 customer service representatives: one was a rating of quality and the other was a rating of quantity. Since neither of these criteria was described as overall performance, we used the average of the two coefficients as the validity coefficient for the sample. Consistent with the Personal Characteristics Inventory data set, each sample provided only one validity coefficient for the data analysis and all criteria were supervisory ratings.

In addition, we obtained data from the Sales Solution Technical Manual (Klein & McLellan, 2001), a publication of the test publisher ePredix. This manual summarizes criterion-related validity data for several ePredix employment tests. Validity data on the EI-Customer Service test are summarized on page 12 of the manual. Validity data were reported for four criteria: behavior composite, trait composite, rehireability, and overall job performance. 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 example, when a sample of hourly employees in a US fast food restaurant reported sample sizes of 427 to 442 for the four criteria, we used the midpoint of the range (435) as the sample size.

The ePredix technical manual also reports validity data for the Customer Service and Clerical Potential Index. The validity data for this test are reported on page 19. Results are presented for four samples, labeled Group I through Group IV. The data from these four groups also were reported by race, sex, and job classification. Whereas that data would be duplicative with the data from the four groups, we did not use it in the analysis. For the four groups, the criterion was a supervisor rating collected from two raters. Validities were reported separately by rater and then averaged. Data from Group I using one of the rater’s ratings as the criterion was used to key the instrument. Since the validity data for Group I 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 Groups II through IV, we used the validity coefficient for the criterion that combined the ratings of raters one and two.

The ePredix manual also reports validity data on two leadership tests. The Leadership Inventory (LI) is a personality-based measure. The Leadership Inventory Plus (LI+) contains the Leadership Inventory plus three sets of cognitive ability items. The validity data for these two instruments are based on the same samples. The validity for the LI scale is reported on page 29 of the technical manual and the validity for the L1+ scale is reported on page 30. 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, in order to make them comparable to those from the other data sets.

Meta-analysis procedure. We conducted a meta-analysis of the observed validity coefficients and a meta-analysis of observed plus imputed correlations generated by the Trim and Fill procedure. Most meta-analyses of employment test validity data use the psychometric meta-analysis method (Hunter & Schmidt, 2004). However, statistical software for publication bias has not been developed for psychometric meta-analysis as it has been 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 and Rothstein, 2005) which follows procedures associated with Hedges and Olkin. We note that this procedure is not much different from a “bare bones” psychometric meta-analysis (i.e., correlations are not corrected for statistical artifacts such as measurement error and range restriction) where observed validity coefficients are analyzed. Also similar to psychometric meta-analysis, validity coefficients are weighted. While psychometric meta-analyses weight coefficients by sample size, the CMA software weights the data by the correlation’s precision (1 divided by the standard error). For correlation coefficients, precision and sample size are highly correlated so that weighting under the two schemes is very similar. In a departure from psychometric meta-analysis, we transformed the correlation coefficients to the Fisher z metric during computation and re-expressed them as correlation coefficients for presentation of results. As mentioned above, for the typical range of correlations in the data sets to be examined, the correlations and their Fisher z counterparts have very similar values. For this paper, we report the precision-weighted mean of the correlation distribution and the confidence interval of the mean.

Results

Tables 1 through 12 present the results of the publication bias analyses for the 12 scales. Each table is divided into two sections labeled A and B.

Section A of each table presents a summary of the meta-analytic results to orient the reader to the data set. Column one displays a description of the sample. The next five columns present the observed validity coefficient, the lower and upper values of the 95% confidence interval, the *Z* value associated with a statistical test of the correlation (is not to be confused with the Fisher *z* transformation of the correlation), and the significance level associated with the *Z* test that the correlation does not equal zero. The right section of the table presents a forest plot of the data. In the forest plot, each correlation coefficient (*r*) is presented as a box which indicates the point estimate, and a line which graphically illustrates the width of the confidence interval. The last row of section A shows the sample-size-weighted mean observed correlation and its associated statistics.

Sections B presents the Trim and Fill results. Consistent with the Trim and Fill procedure, the plot uses Fisher *z* rather then the observed correlations. The Fisher *z*s are plotted on the X axis and the precision (inverse of the standard error) of the correlation is presented on the Y axis. For correlations, the precision value is highly correlated with the sample size. Thus, the large sample, high precision studies are at the top of the plot and the low sample, low precision studies are at the bottom of the plot. The original data are displayed as clear circles. If any correlations were imputed by the Trim and Fill procedure, they are displayed as dark circles.

Table 1 presents the results for the PCI Extraversion scale. The 14 correlations yielded a mean correlation of .034 with a confidence interval from -.006 to .073. The Trim and Fill analysis suggested that five additional studies would be needed to make the distribution symmetrical. However, the missing studies do not appear to have much of an effect on the estimate of the population correlation. The Trim and Fill adjusted mean correlation is -0.006, a difference of .040 from the observed mean of .034. In the case of PCI extraversion, there may be some publication bias operating, but it appears to have only a minimal impact on validity, which is extremely low in any case.

The results for the PCI Agreeableness scale are presented in Table 2. The 14 correlations yielded a mean correlation of .074 with a confidence interval from .034 to .113. The Trim and Fill analysis suggested that no additional studies were required to achieve symmetry, and there does not appear to be any evidence that publication bias is present in the distribution of reported PCI agreeableness validities.

Table 3 presents the results for the PCI Conscientiousness scale. The 14 correlations yielded a mean correlation of .233 with a confidence interval from .195 to .270. The Trim and Fill analysis suggested that four additional studies were needed to make the distribution symmetric. The observed mean correlation was .233 and the Trim and Fill adjusted mean correlation is .219, a difference of .014. Thus, imputing the potentially missing studies has only a minimal effect on the estimate of the population correlation. In both cases the validity of the test is moderate (in the low .20s), and the difference between the two seems unlikely to alter a decision about whether to use the test.

The results for the PCI Stability scale are shown in Table 4. The 14 correlations yielded a mean correlation of .091 with a confidence interval from .052 to .130. The Trim and Fill analysis suggested that four additional studies were needed to bring the distribution into symmetry. The Trim and Fill corrected mean correlation is .064, indicating a difference of .027 from the observed mean of .091. Therefore, imputing the “missing studies” has only a minimal effect on the estimate of the population correlation. Although the Trim and Fill analysis suggests that some studies may be missing, and that these studies would reduce the estimated mean validity from .09 to .06, the population estimate of validity in both cases is quite low, and is unlikely to alter a decision about whether to use the test.

Table 5 presents the results for the PCI Openness scale. The 14 correlations yielded a mean correlation of .055 with a confidence interval from .016 to .095. The Trim and Fill analysis found no evidence of publication bias. We also note that the validity is of small magnitude.

The results for the HPI Service Orientation scale are presented in Table 6. The 12 correlations yielded a mean correlation of .286 with a confidence interval from .243 to .328. The Trim and Fill analysis suggested six studies were missing. The mean correlation after the Trim and Fill imputed studies is .184. In this case, the observed and the mean of the distribution containing the imputed studies differ by .102. An observed mean validity of .286 is likely to have different implications for test use than a validity of .186, and, using the terminology we introduced earlier in the paper, we would suggest that there is a moderate to severe amount of publication bias operating in this case. The reader may recall that we suggested use of the term “moderate bias” when the size of the validity coefficients change substantially but the key finding (e.g., that it is or is not useful) remains in force, and recommended calling the impact of bias “severe” when the basic conclusion of the analysis (e.g., that the test is operationally useful) is called into question.

Table 7 presents the results for the HPI Reliability scale. The 11 correlations yielded a mean correlation of .295 with a confidence interval from .247 to .342. The Trim and Fill analysis suggested that four studies were missing, and that the mean correlation including the studies imputed by Trim and Fill is .234. The observed and the Trim and Fill means differ by .061. An observed mean validity of .295 may have different implications for test use than a validity of .234; therefore using the terminology we introduced earlier in the paper, we would suggest that there is a moderate amount of publication bias operating in this case.

The results for the HPI Stress Tolerance scale are shown in Table 8. The nine correlations yielded a mean correlation of .408 with a confidence interval from .337 to .475. The Trim and Fill analysis found no evidence of publication bias. We also note that the validity is of high magnitude (.408).

Table 9 presents the results for the ePredix EI-Customer Service scale. The eight correlations yielded a mean correlation of .218 with a confidence interval from .184 to .252. The Trim and Fill analysis suggested two studies were needed to bring the distribution into symmetry. The observed mean correlation was .218 and the Trim and Fill adjusted mean correlation is .211. Thus, the observed and the Trim and Fill means are very similar and offer no evidence of publication bias.

The results for the ePredix EI-Customer Service and Clerical Potential Index is presented in Table 10. The four correlations yielded a mean correlation of .304 with a confidence interval from .280 to .327. The Trim and Fill analysis suggested that no studies were missing. Based on this analysis, there appears to be no evidence of publication bias in the EI-Customer Service and Clerical Potential Index.

Table 11 presents the results for the ePredix Leadership Inventory scale. The seven correlations yielded a mean correlation of .243 with a confidence interval from .182 to .303. The Trim and Fill analysis suggested that two studies were needed to bring the distribution into symmetry. The observed mean correlation was .243 and the Trim and Fill mean correlation including two imputed studies is .211. Thus, the observed and the Trim and Fill means differ by .032. This finding suggests that publication bias may be operating, but that its impact on validity is minimal. Given the relationship between this scale and the following one, we will reserve judgment on this scale until we review the publication bias evidence for the Leadership Inventory Plus. The correlations for the Leadership Inventory in Table 11 are from the same samples as the correlations for the ePredix Leadership Inventory Plus scale (Table 12). The difference between the two scales is the inclusion of cognitive ability items in the Leadership Inventory Plus. Whereas the coefficients are drawn from the same samples, conclusions for publication bias for one test should be consistent with conclusions for the other test.

Table 12 presents the results for the ePredix Leadership Inventory Plus scale. The analysis yielded a mean correlation of .281 with a confidence interval from .220 to .339. The Trim and Fill analysis suggested that no studies were missing. Thus, there is no evidence of publication bias for the Leadership Inventory Plus scale. Given that the asymmetry in the Leadership Inventory (Table 11) was not large and the same samples did not yield asymmetry in the Leadership Inventory Plus scale, we suggest that the small amount of publication bias in the Leadership Inventory is not meaningful.

Discussion

Our results indicate that while no or minimal bias is operating for any of the five scales of the PCI or the four scales presented in ePredix’s Sales Solution Technical Manual, there was evidence of moderate bias for one of the three occupational composites of the HPI, and moderate to severe bias for a second HPI composite. In trying to locate sources of the possible bias in these data, we believe it is informative to examine the reporting practices of the three test publishers for clues about the reasons for the asymmetry in the distributions of HPI scale validities.

One point of comparison is the consistency across scales in the samples providing validity data. To our knowledge, the five PCI scales are scored from the same instrument. That is, one takes the PCI and scores can be generated on all five scales of the PCI. Likewise, to our knowledge, the three HPI occupational scales are also scored from the same instrument. That is, one takes the HPI and scores can be generated on all three of the occupational scales. Thus, if a study provides validity data for one scale of the PCI, it should be able to provide validity data for the remaining four scales of the PCI. Likewise, for the HPI, if a study provides validity data for one of the HPI occupational scales, it should be able to provide validity data for the remaining two HPI occupational scales. Of the four ePredix tests examined, two of the four tests, the Leadership Inventory and the Leadership Plus Inventory, were scored from the same instrument. Specifically the Leadership Inventory scale is a subset of the Leadership Plus Inventory. Therefore, validity data should be available for both scales.

An inspection of the PCI validity data shows that if a sample contributes a validity coefficient to one PCI scale, it provides a validity coefficient for every other PCI scale. This was also true for the ePredix Leadership Inventory and the Leadership Plus Inventory. However, the HPI validity data look different. Consider a study by Cage (1989) that is cited in the HPI manual. Cage used the HPI with a sample of 20 nannies. Validity data are reported for the Service Orientation scale and the Stress Tolerance scale but not the Reliability scale. This raises the possibility that the validity coefficient for the Reliability scale was calculated but not reported. We cannot resolve this issue because we could not obtain the Cage study, an unpublished technical report from 16 years ago. We could not locate the author and were unable to obtain the study from Hogan Assessment Systems.

Next, consider a study by Muchinsky (1987) reported in the HPI technical manual. Muchinsky reported validity data for a sample of 102 customer service representatives. While the HPI technical manual reports validity data for this sample for all three occupational scales, the criterion measures vary somewhat across the scales. For the Service Orientation scale, validity coefficients are reported for supervisory ratings of quality and quantity. For the Stress Tolerance scale, validity coefficients are reported for supervisory ratings of quality, quantity, teamwork, and overall performance. For the Reliability scale, validity coefficients are reported for the supervisory rating scales of quality, teamwork, and overall performance. Thus, there appear to be at least four supervisory rating scales: quality, quantity, teamwork, and overall performance. For the Service Orientation scale, it is possible that the validity coefficients for teamwork and overall performance exist but were not included in the technical manual. For the Reliability scale, the validity coefficient for quantity may exist but was not included in the technical manual. In other words, outcomes may be selectively reported. Selective outcome reporting bias has already been shown to be problematic in healthcare and psychological research literatures (Chan, Hrobjartsson Haahr, Gøtzsche & Altman, 2004, Hutton, & Williamson, 2000, Orwin & Cordray,1985). Unfortunately, we could not ascertain if this was the case here; Dr. Muchinsky no longer had a copy of this 18 year old study to provide to the authors and we could not obtain the study from Hogan Assessment Systems.

The Stress Tolerance scale showed no evidence of publication bias based on asymmetry in the validity distribution. However, given that some validity coefficients appear to be omitted from the Hogan technical manual, we have some concerns about the potential publication bias in the Stress Tolerance scale.

We next compared and contrasted the magnitude, direction and statistical significance of the validity data reported in teach of the tables from the manuals of the three test publishers. An inspection of the criterion-related validity data in the PCI technical manual shows that both statistically significant and statistically non-significant validity correlations are reported. Of the 100 criterion validity coefficients reported (page 5-45), 40 are statistically significant and 60 are not. An inspection of the ePredix data shows that one of eight validity coefficients for the EI-Customer Service test is not statistically significant. All four of the EI-Customer Service and Clerical Potential Index validity coefficients are significant, as are all seven of the coefficients for the Leadership Inventory and the Leadership Plus inventory. An inspection of the criterion-related validity data in the HPI technical manual for the three occupational scales shows that all 95 validity coefficients are statistically significant. (Note that the PCI listing of 100 coefficients exceeds the 70 (14 for each of 5 scales) used in our analysis because we included only supervisory rating criteria. Similarly, the HPI listing of 95 coefficients exceeds the 35 (9, 12, and 11 respectively for the three scales) used in our analysis because we only used supervisory rating criteria and only allowed one coefficient per sample per scale.) In sum, all validity coefficients reported in the HPI manual are statistically significant, while this is not the case for the Wonderlic PCI and the ePredix Sales Solution manual. We viewed this as additional evidence of possible selective reporting in the HPI manual.

We also inspected the test publisher manuals to identify how many validity coefficients have a sign in the opposite direction of that expected for the scale. For example, a negative correlation between a measure of conscientiousness and a supervisory rating of job performance would be in the opposite direction than expected for a valid test. For the PCI, we identified conscientiousness and stability as the two dimensions where directional hypotheses are warranted. Specifically, we know of no body of validity data showing that the undependable and/or the neurotic have good job performance on average. Of the 40 coefficients for these two scales, three of the 40 (7.5%) are in the opposite direction to the one expected. That is, 7.5% of the validity coefficients reported for the two scales suggest that, functionally, the scales do not perform as expected. For the four distributions examined for the ePredix products, all scales permit directional hypotheses. None of the 26 correlations in the ePredix manual are in a counter-intuitive direction. For the HPI, all three scales permit directional hypotheses. Of the 95 coefficients, none of the 95 are in the direction opposite to what was expected.

Given that our analyses revealed that that two of the three HPI occupational scales show the type of asymmetry that would occur if the test manual reported only statistically significant correlations in a direction supportive of the test's validity, we decided to seek direct verification from the test publisher. To clarify whether selective reporting might be responsible for the asymmetric funnels, the senior author contacted Hogan Assessment Systems regarding their reporting practices. Staff at Hogan Assessment Systems confirmed that their manuals report only statistically significant correlations (personal communication to Michael McDaniel from Nicole Bourdeau, January, 10, 2005.) The test publisher also stated “The table in question from the HPI manual was designed to illustrate what types of specific criteria are best predicted by the specific occupational scales being described, and was not intended to be a comprehensive list of validity coefficients” (personal communication to Michael McDaniel from Scott Davies, January, 10, 2005).

The standards described in 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 that is 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. Indeed, in the introduction, we noted how practices of researchers independent of the test publisher could contribute to publication bias. This is even more likely to be the case when consultants collect validity data and such data are not shared with the publisher because of clients’ proprietary rights. That being said, test publishers also have a role in preventing publication bias by publicly reporting all known validity data. We also encourage test publishers to conduct their own publication bias analyses and report the results in their test manuals. This will permit users of the test manuals to make informed decisions concerning the validity data presented.

We note that clear reporting of test validity data is an exception and not common practice, and that we are not singling out the Hogan tests for special attention; rather we are using them as an illustrative example. Many test publishers have no validity evidence. Other test publishers provide some narrative summaries of past validity studies but sample sizes and validity coefficients are often not provided. Some test publishers provide copies of primary validity studies but do not have a technical manual that summarizes the data. As researchers in personnel selection, we find it disheartening that so few test publishers offer any validity data. Although the HPI technical manual shows some evidence of data censoring, the test publisher does provide a substantial amount of validity data and they are responsive to inquiries about its reporting practices.

Limitations of the research

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. Real world data do not always behave neatly. 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.

Given these considerations, the finding that there is asymmetry in some test validity distributions for the HPI, should not be viewed as a definitive demonstration that all asymmetry is due to publication bias. However, given what we know about the reporting practices involved in the current case, it is clear that the data in the HPI technical manual have been selectively chosen for publication. We are hopeful that the reporting practices in the HPI technical manual that cause concern are issues that can be addressed in the next revision of the test manual. We also hasten to point out that, even if our results are interpreted to suggest that the validity data in the HPI technical manual are upwardly biased, our analyses also show that the HPI occupational scales have useful levels of validity even after the validity has been adjusted for potential bias. However, users of tests rely on test manual validity data to choose among tests from various vendors and thus it is important that validity data in test manuals be an unbiased description of validity.

Conclusion

This paper has presented a review of two methods for the detection of publication bias. We argue that the use of the failsafe N procedure be discontinued and supplanted by better procedures. We also recommend the use of the Trim and Fill procedure and found it appropriate for the data analyzed in this study. Of the 12 test validity distributions analyzed, the five distributions from the PCI showed little to no evidence of publication bias. The same was true of the four test validity distributions in the ePredix sales solution test manual. Thus, test users can have increased confidence in the validity data presented for the Wonderlic PCI and the ePredix Sales Solution products. In contrast, two of the three occupational scales of the HPI showed evidence of asymmetry consistent with the operation of publication bias and the third scale shared questionable reporting practices with the two scales showing evidence consistent with a conclusion of publication bias. We encourage all test publishers to present complete validity data in order to fulfill their roles as scientists as well as practitioners, and organizational and human resources researchers also should 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. (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.

Bogg, T. & Roberts, B.W. (2004). Conscientiousness and health-related behaviors: a meta-analysis of the leading behavioral contributors to mortality. *Psychological Bulletin. 130*, 887-919.

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

Cage, J. (1989). *Development and validation of a nanny selection inventory*. Tulsa: St. John Hospital.

Chan A-W, Hrobjartsson A, Haahr M.T, Gøtzsche P.C and Altman D.G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of the American Medical Association*, 291, 2457–2465.

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.

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.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.

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.

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

Halpern, S.D. & Berlin, J.A. (2005). Beyond conventional publication bias: other determinants of data suppression. In H. Rothstein, A.J. Sutton, & M. Borenstein (Eds). *Publication Bias in Meta Analysis: Prevention, Assessment and Adjustments*. Wiley.

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.

Hogan, R. & Hogan, J. (1995). *Hogan Personality Inventory Manual. Second Edition*. Tulsa, OK: Hogan Assessment Systems.

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

Hutton, J.L. & Williamson, P.R. (2000) Bias in meta-analysis due to outcome variable selection within studies. *Applied Statistics,* 49, 359-370.

Klein, S.R. & McLellan, R.A. (2001). *Sales solution technical manual.* Minneapolis, MN: ePredix.

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.

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

Mezulis, A. H. Abramson, L. Y.; Hyde, J. S.; Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin, 130*, 711-747.

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.

Moyer, C. A., Rounds, J., & Hannum, J. W. (2004). A meta-analysis of massage therapy research. *Psychological Bulletin,130*, 3-18.

Muchinsky, P. (1987). *Validation documentation for the development of personnel selection batteries for telecommunications service jobs.* Ames: Iowa State University.

NY vs. GlaxoSmithKline. Filing to the Supreme Court of the State of New York. June 2, 2004.

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

Orwin, R. G. & Cordray, D. S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. *Psychological Bulletin*, *97*, 134-147.

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

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? a meta-analysis. *Psychological Bulletin, 130*, 261-288.

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

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

Smith, T. B., McCullough, M. E. & Poll, J. (2003). Religiousness and depression: Evidence for a main effect and the moderating influence of stressful life events. *Psychological Bulletin, 129*, 614-636.

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.

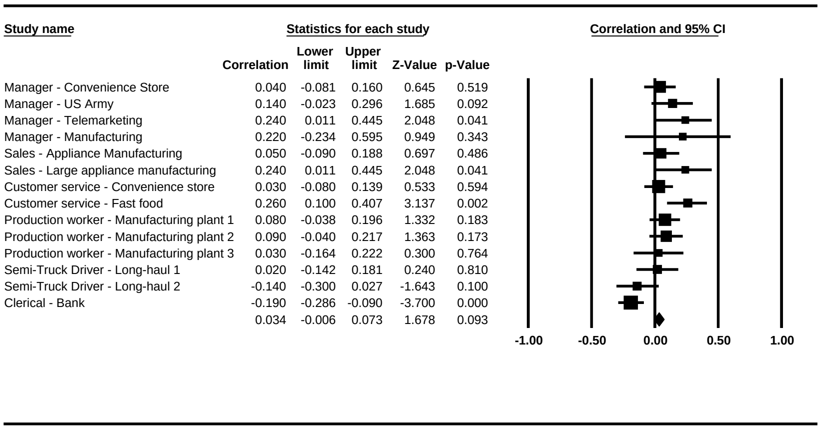
Vitaliano, P. P., Zhang, J. & Scanlan, J. M. (2003). Is caregiving hazardous to one's physical health? a meta-analysis. *Psychological Bulletin, 129*, 946-972.

Wall Street Journal, November 1, 2004. *E-Mails Suggest Merck Knew Vioxx's Dangers at Early Stage.*

Wonderlic, Inc. (2002). *Personal Characteristics Inventory*. Libertyville, IL: Author.

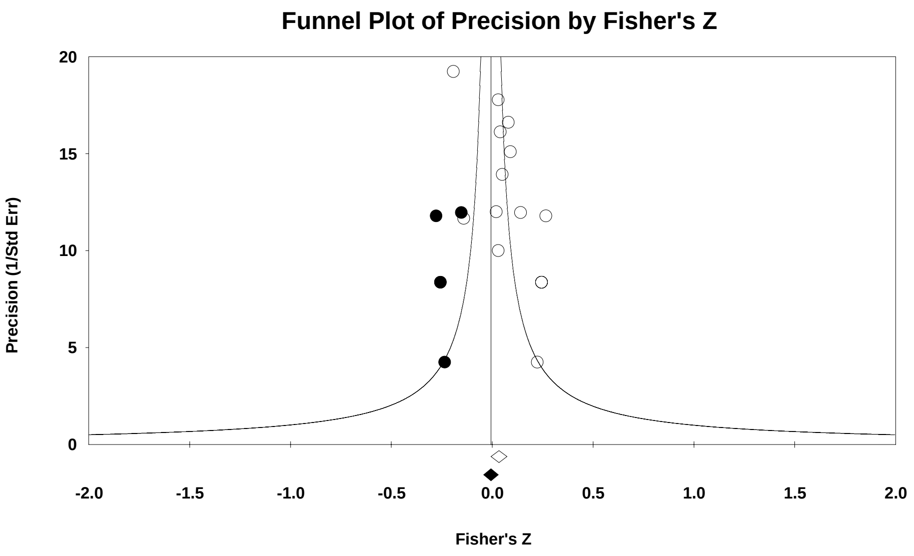
Table 1. Publication Bias Results for PCI Extraversion

A. Meta-Analysis Results



Study name	Statistics for each study				
	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Manager - Convenience Store	0.040	-0.081	0.160	0.645	0.519
Manager - US Army	0.140	-0.023	0.296	1.685	0.092
Manager - Telemarketing	0.240	0.011	0.445	2.048	0.041
Manager - Manufacturing	0.220	-0.234	0.595	0.949	0.343
Sales - Appliance Manufacturing	0.050	-0.090	0.188	0.697	0.486
Sales - Large appliance manufacturing	0.240	0.011	0.445	2.048	0.041
Customer service - Convenience store	0.030	-0.080	0.139	0.533	0.594
Customer service - Fast food	0.260	0.100	0.407	3.137	0.002
Production worker - Manufacturing plant 1	0.080	-0.038	0.196	1.332	0.183
Production worker - Manufacturing plant 2	0.090	-0.040	0.217	1.363	0.173
Production worker - Manufacturing plant 3	0.030	-0.164	0.222	0.300	0.764
Semi-Truck Driver - Long-haul 1	0.020	-0.142	0.181	0.240	0.810
Semi-Truck Driver - Long-haul 2	-0.140	-0.300	0.027	-1.643	0.100
Clerical - Bank	-0.190	-0.286	-0.090	-3.700	0.000
	0.034	-0.006	0.073	1.678	0.093

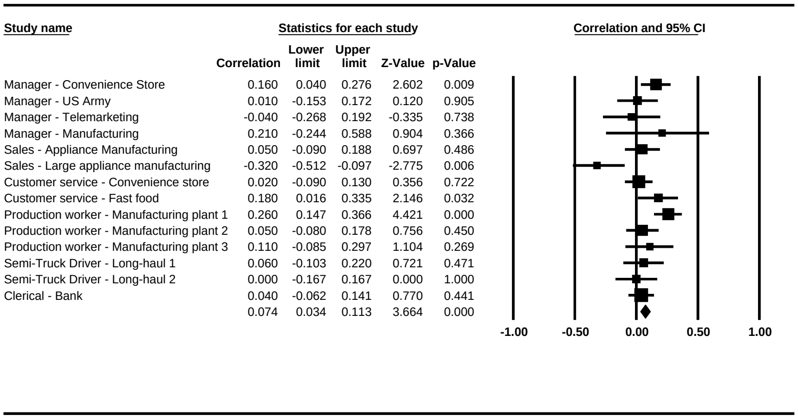
B. Duval and Tweedie's Trim and Fill



The clear circles represent the observed data. The dark circles represent imputed studies. The trim and fill procedure suggests that five studies are missing. The point estimate (i.e., the mean correlation) and 95% confidence interval for the combined studies is 0.034 (-0.006, 0.073). Using trim and fill, the imputed point estimate is -0.006 (-0.042, 0.030).

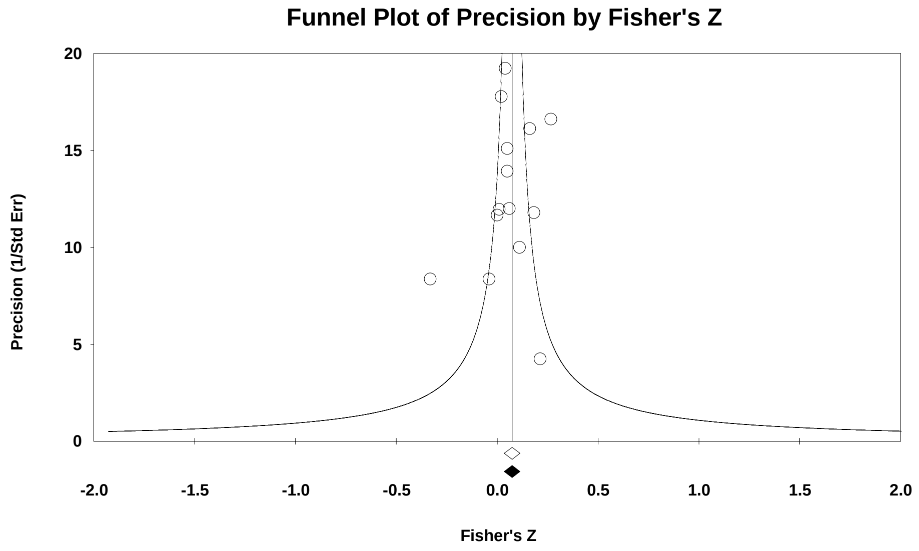
Table 2. Publication Bias Results for PCI Agreeableness

A. Meta-Analysis Results



Study name	Statistics for each study				
	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Manager - Convenience Store	0.160	0.040	0.276	2.602	0.009
Manager - US Army	0.010	-0.153	0.172	0.120	0.905
Manager - Telemarketing	-0.040	-0.268	0.192	-0.335	0.738
Manager - Manufacturing	0.210	-0.244	0.588	0.904	0.366
Sales - Appliance Manufacturing	0.050	-0.090	0.188	0.697	0.486
Sales - Large appliance manufacturing	-0.320	-0.512	-0.097	-2.775	0.006
Customer service - Convenience store	0.020	-0.090	0.130	0.356	0.722
Customer service - Fast food	0.180	0.016	0.335	2.146	0.032
Production worker - Manufacturing plant 1	0.260	0.147	0.366	4.421	0.000
Production worker - Manufacturing plant 2	0.050	-0.080	0.178	0.756	0.450
Production worker - Manufacturing plant 3	0.110	-0.085	0.297	1.104	0.269
Semi-Truck Driver - Long-haul 1	0.060	-0.103	0.220	0.721	0.471
Semi-Truck Driver - Long-haul 2	0.000	-0.167	0.167	0.000	1.000
Clerical - Bank	0.040	-0.062	0.141	0.770	0.441
	0.074	0.034	0.113	3.664	0.000

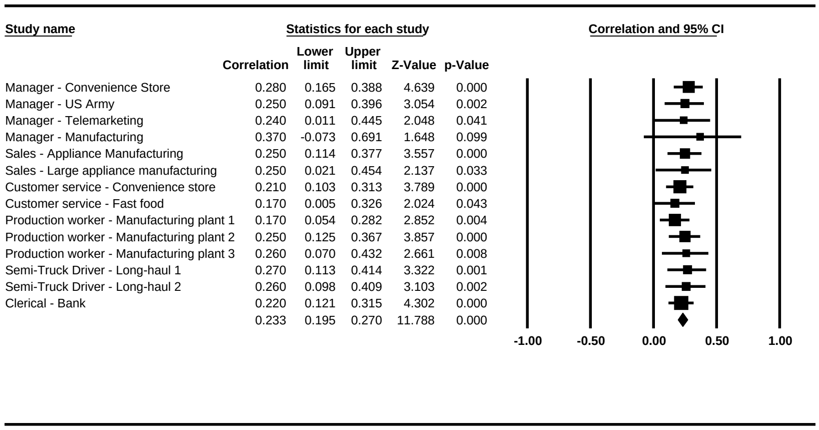
B. Duval and Tweedie's Trim and Fill



The clear circles represent the observed data. The dark circles represent imputed studies, but in this case there are none because no data were imputed. The trim and fill procedure suggests that zero studies are missing. The point estimate and 95% confidence interval for the combined studies is 0.074 (0.034, 0.113). Using trim and fill, the imputed point estimate is 0.074 (0.034, 0.113).

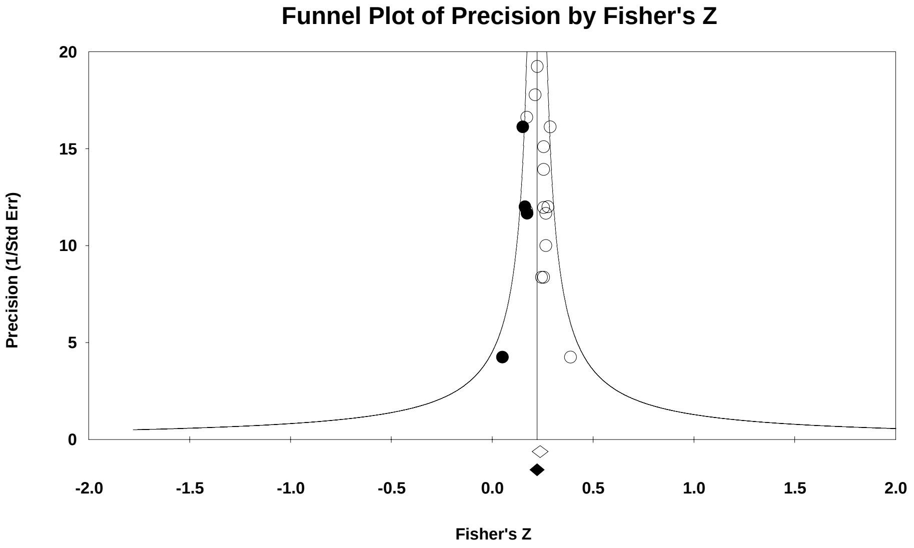
Table 3. Publication Bias Results for PCI Conscientiousness

A. Meta-Analysis Results



Study name	Statistics for each study				
	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Manager - Convenience Store	0.280	0.165	0.388	4.639	0.000
Manager - US Army	0.250	0.091	0.396	3.054	0.002
Manager - Telemarketing	0.240	0.011	0.445	2.048	0.041
Manager - Manufacturing	0.370	-0.073	0.691	1.648	0.099
Sales - Appliance Manufacturing	0.250	0.114	0.377	3.557	0.000
Sales - Large appliance manufacturing	0.250	0.021	0.454	2.137	0.033
Customer service - Convenience store	0.210	0.103	0.313	3.789	0.000
Customer service - Fast food	0.170	0.005	0.326	2.024	0.043
Production worker - Manufacturing plant 1	0.170	0.054	0.282	2.852	0.004
Production worker - Manufacturing plant 2	0.250	0.125	0.367	3.857	0.000
Production worker - Manufacturing plant 3	0.260	0.070	0.432	2.661	0.008
Semi-Truck Driver - Long-haul 1	0.270	0.113	0.414	3.322	0.001
Semi-Truck Driver - Long-haul 2	0.260	0.098	0.409	3.103	0.002
Clerical - Bank	0.220	0.121	0.315	4.302	0.000
	0.233	0.195	0.270	11.788	0.000

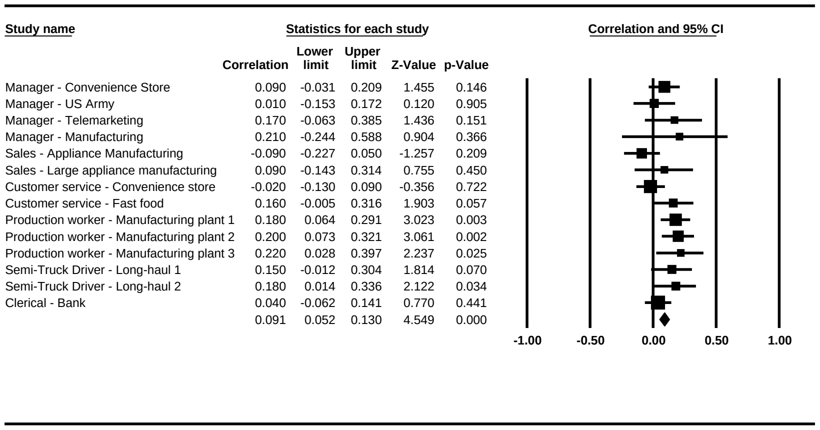
B. Duval and Tweedie’s Trim and Fill



The clear circles represent the observed data. The dark circles represent imputed studies. The trim and fill procedure suggests that four studies are missing. The point estimate and 95% confidence interval for the combined studies is 0.233 (0.195, 0.270). Using trim and fill, the imputed point estimate is 0.219 (0.185, 0.252).

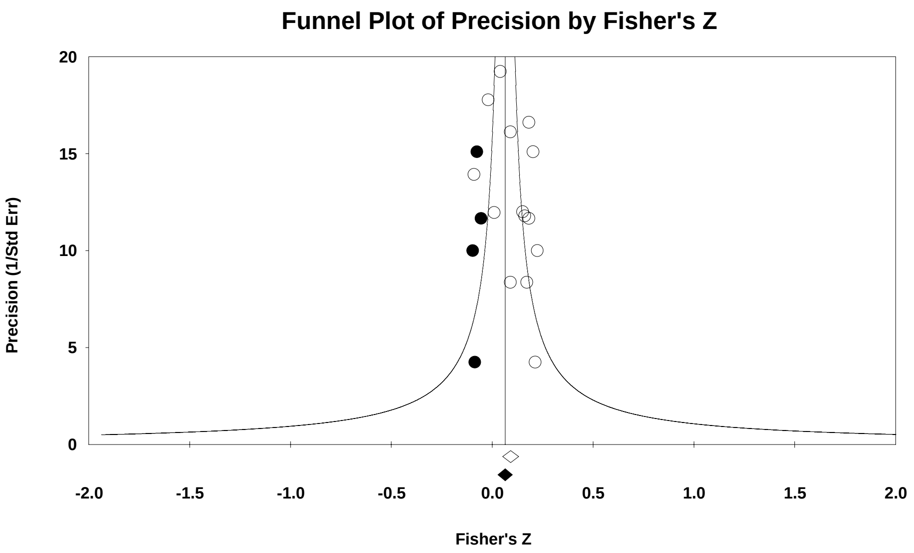
Table 4. Publication Bias Results for PCI Stability

A. Meta-Analysis Results



Study name	Statistics for each study					Correlation and 95% CI
	Correlation	Lower limit	Upper limit	Z-Value	p-Value	
Manager - Convenience Store	0.090	-0.031	0.209	1.455	0.146	
Manager - US Army	0.010	-0.153	0.172	0.120	0.905	
Manager - Telemarketing	0.170	-0.063	0.385	1.436	0.151	
Manager - Manufacturing	0.210	-0.244	0.588	0.904	0.366	
Sales - Appliance Manufacturing	-0.090	-0.227	0.050	-1.257	0.209	
Sales - Large appliance manufacturing	0.090	-0.143	0.314	0.755	0.450	
Customer service - Convenience store	-0.020	-0.130	0.090	-0.356	0.722	
Customer service - Fast food	0.160	-0.005	0.316	1.903	0.057	
Production worker - Manufacturing plant 1	0.180	0.064	0.291	3.023	0.003	
Production worker - Manufacturing plant 2	0.200	0.073	0.321	3.061	0.002	
Production worker - Manufacturing plant 3	0.220	0.028	0.397	2.237	0.025	
Semi-Truck Driver - Long-haul 1	0.150	-0.012	0.304	1.814	0.070	
Semi-Truck Driver - Long-haul 2	0.180	0.014	0.336	2.122	0.034	
Clerical - Bank	0.040	-0.062	0.141	0.770	0.441	
	0.091	0.052	0.130	4.549	0.000	

B. Duval and Tweedie’s Trim and Fill

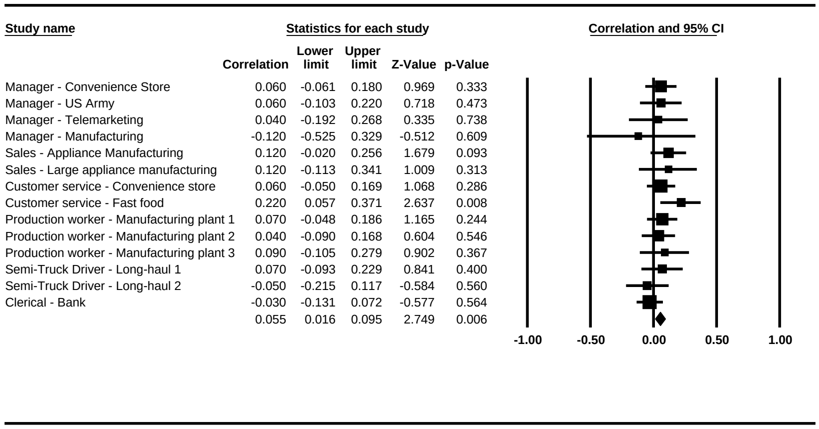


Funnel Plot of Precision by Fisher's Z

The clear circles represent the observed data. The dark circles represent imputed studies. The trim and fill procedure suggests that four studies are missing. The point estimate and 95% confidence interval for the combined studies is 0.091 (0.052, 0.130). Using trim and fill, the imputed point estimate is 0.064 (0.028, 0.101).

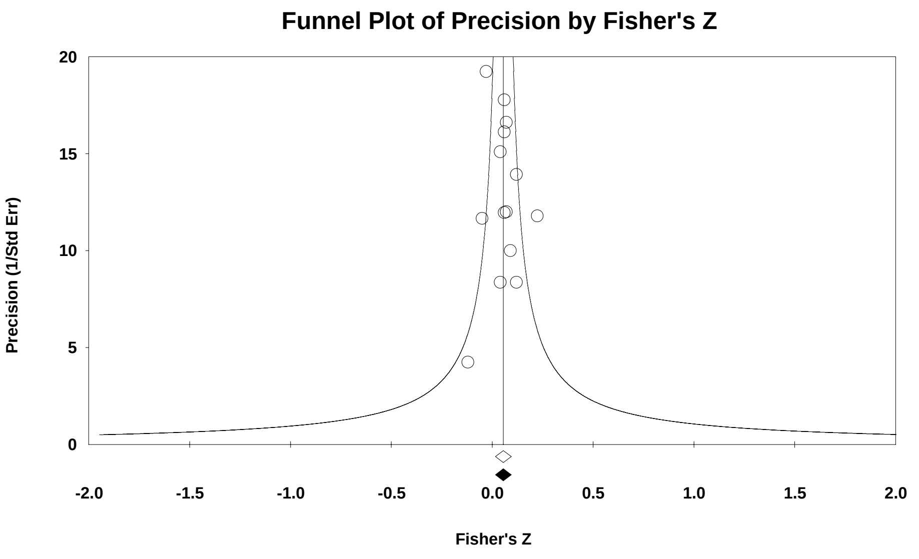
Table 5. Publication Bias Results for PCI Openness

A. Meta-Analysis Results



Study name	Statistics for each study				
	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Manager - Convenience Store	0.060	-0.061	0.180	0.969	0.333
Manager - US Army	0.060	-0.103	0.220	0.718	0.473
Manager - Telemarketing	0.040	-0.192	0.268	0.335	0.738
Manager - Manufacturing	-0.120	-0.525	0.329	-0.512	0.609
Sales - Appliance Manufacturing	0.120	-0.020	0.256	1.679	0.093
Sales - Large appliance manufacturing	0.120	-0.113	0.341	1.009	0.313
Customer service - Convenience store	0.060	-0.050	0.169	1.068	0.286
Customer service - Fast food	0.220	0.057	0.371	2.637	0.008
Production worker - Manufacturing plant 1	0.070	-0.048	0.186	1.165	0.244
Production worker - Manufacturing plant 2	0.040	-0.090	0.168	0.604	0.546
Production worker - Manufacturing plant 3	0.090	-0.105	0.279	0.902	0.367
Semi-Truck Driver - Long-haul 1	0.070	-0.093	0.229	0.841	0.400
Semi-Truck Driver - Long-haul 2	-0.050	-0.215	0.117	-0.584	0.560
Clerical - Bank	-0.030	-0.131	0.072	-0.577	0.564
	0.055	0.016	0.095	2.749	0.006

B. Duval and Tweedie’s Trim and Fill

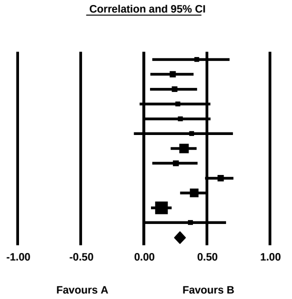


The clear circles represent the observed data. The dark circles represent imputed studies, but in this case, none were imputed. The trim and fill procedure suggests that zero studies are missing. The point estimate and 95% confidence interval for the combined studies is 0.055 (0.016, 0.095). Using trim and fill, the imputed point estimate is 0.055 (0.016, 0.095).

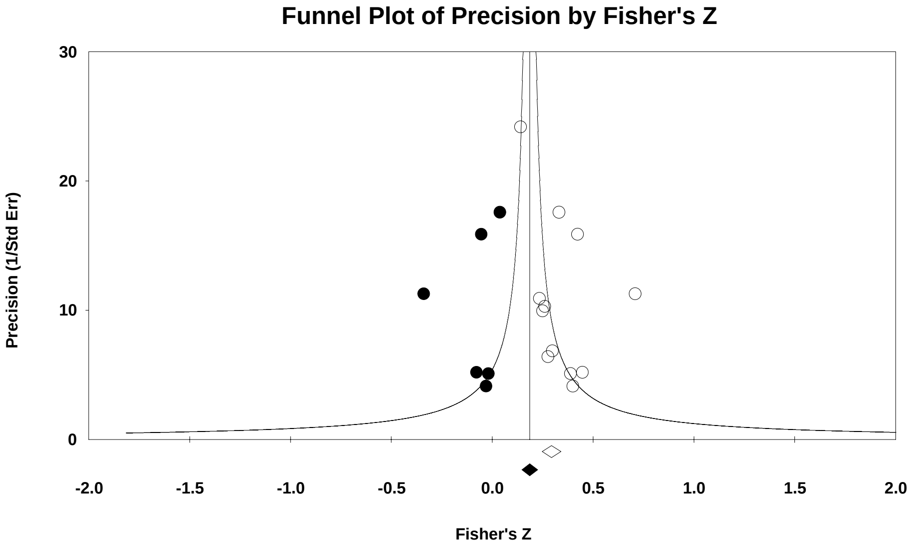
Table 6. Publication Bias Results for Hogan Service Orientation

A. Meta-Analysis Results

Study name	Statistics for each study				
	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Hogan, Hogan, & Busch (1984)	0.420	0.070	0.678	2.326	0.020
Montgomery, Butler & McPhail (1987)	0.230	0.054	0.392	2.555	0.011
Muchinsky (1987) Customer Service Representatives	0.245	0.053	0.419	2.488	0.013
Muchinsky (1987) Field service representatives	0.270	-0.029	0.525	1.773	0.076
Muchinsky (1987) Office managers	0.290	0.013	0.526	2.047	0.041
Cage (1989)	0.380	-0.075	0.704	1.649	0.099
Curphy, Gibson, Asiu, Horn, & Macomber	0.320	0.217	0.416	5.830	0.000
Muchinsky (1993)	0.255	0.070	0.423	2.685	0.007
Hayes, Roehm & Castellano (1994)	0.610	0.489	0.708	7.989	0.000
Hogan, Hogan, & Brinkmeyer (1994)	0.400	0.291	0.498	6.725	0.000
Landy, Jacobs, & Associates (1994)	0.140	0.060	0.218	3.409	0.001
Klippel (1995)	0.370	0.004	0.649	1.981	0.048
	0.286	0.243	0.328	12.326	0.000



B. Duval and Tweedie’s Trim and Fill

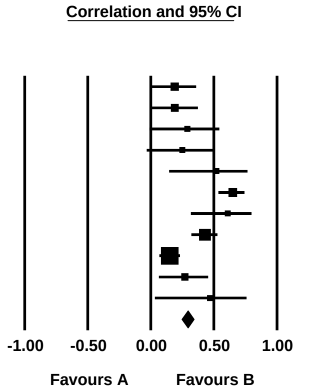


The clear circles represent the observed data. The dark circles represent imputed studies. The trim and fill procedure suggests that six studies are missing. The point estimate and 95% confidence interval for the combined studies is 0.286 (0.243, 0.328). Using trim and fill, the imputed point estimate is 0.184 (0.146, 0.221).

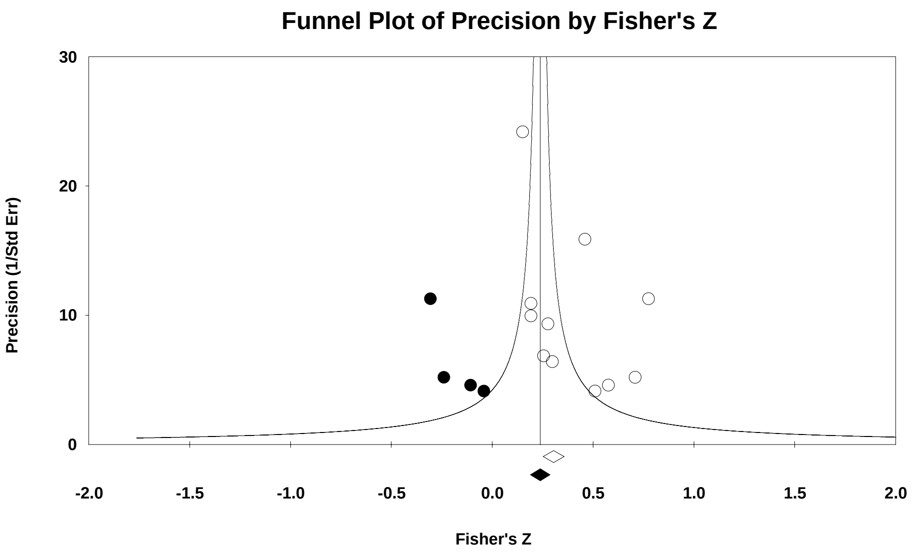
Table 7. Publication Bias Results for Hogan Reliability

A. Meta-Analysis Results

Study name	Statistics for each study		
	Correlation	Lower limit	Upper limit
Montgomery et al (1987)	0.190	0.013	0.356
Muchinsky (1987) 1	0.190	-0.005	0.371
Muchinsky (1987) 2	0.290	-0.008	0.540
Muchinsky (1987) 3	0.250	-0.030	0.494
Hogan & Gerhold (1994)	0.520	0.148	0.763
Hayes et al. (1994)	0.650	0.538	0.739
Hogan et al (1994) 1	0.610	0.320	0.795
Hogan et al (1994) 2	0.430	0.324	0.525
Landy, Jacons & Associates (1994)	0.150	0.070	0.228
Hogan & Gerhold (1995b)	0.270	0.067	0.452
Klippel (1995)	0.470	0.035	0.755
	0.295	0.247	0.342



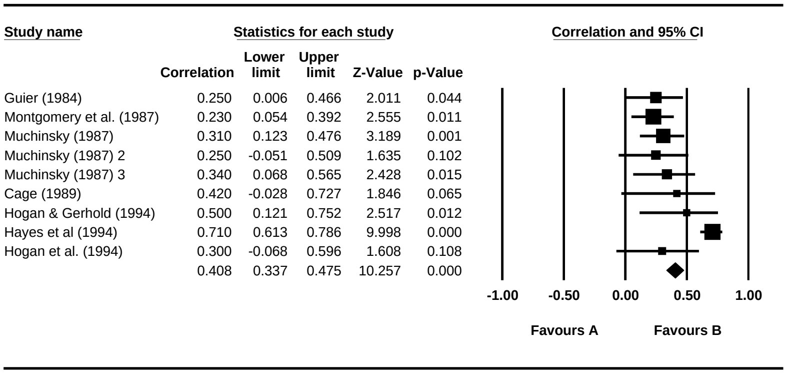
B. Duval and Tweedie’s Trim and Fill



The clear circles represent the observed data. The dark circles represent imputed studies. The trim and fill procedure suggests that four studies are missing. The point estimate and 95% confidence interval for the combined studies is 0.295 (0.247, 0.342). Using Trim and Fill, the imputed point estimate is 0.234 (0.188, 0.279).

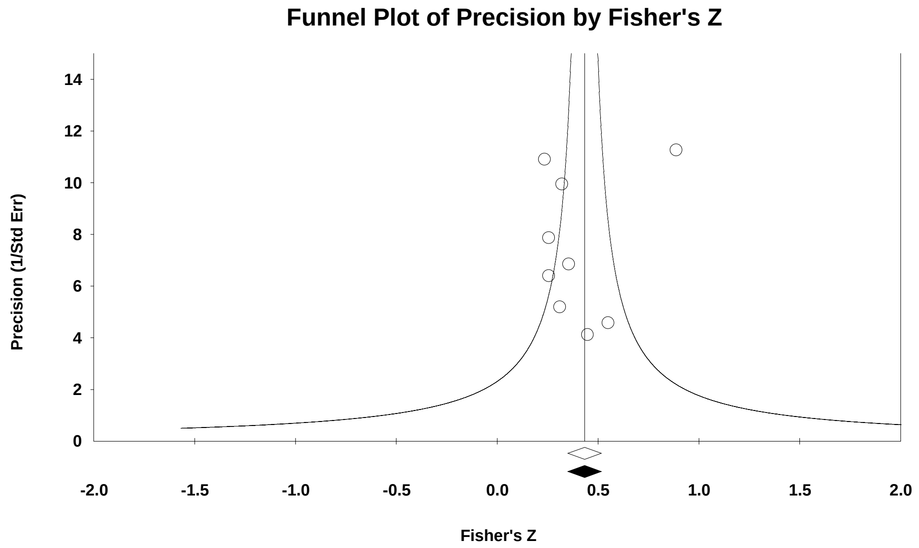
Table 8. Publication Bias Results for Hogan Stress Tolerance

A. Meta-Analysis Results



Study name	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Guier (1984)	0.250	0.006	0.466	2.011	0.044
Montgomery et al. (1987)	0.230	0.054	0.392	2.555	0.011
Muchinsky (1987)	0.310	0.123	0.476	3.189	0.001
Muchinsky (1987) 2	0.250	-0.051	0.509	1.635	0.102
Muchinsky (1987) 3	0.340	0.068	0.565	2.428	0.015
Cage (1989)	0.420	-0.028	0.727	1.846	0.065
Hogan & Gerhold (1994)	0.500	0.121	0.752	2.517	0.012
Hayes et al (1994)	0.710	0.613	0.786	9.998	0.000
Hogan et al. (1994)	0.300	-0.068	0.596	1.608	0.108
	0.408	0.337	0.475	10.257	0.000

B. Duval and Tweedie’s Trim and Fill

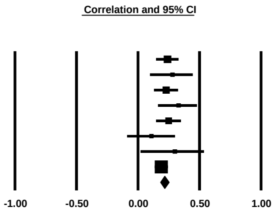


The clear circles represent the observed data. The dark circles represent imputed studies, but in this case none were imputed. The trim and fill procedure suggests that zero studies are missing. The point estimate and 95% confidence interval for the combined studies is 0.408 (0.337, 0.475). Using trim and fill, the imputed point estimate is 0.408 (0.337, 0.475).

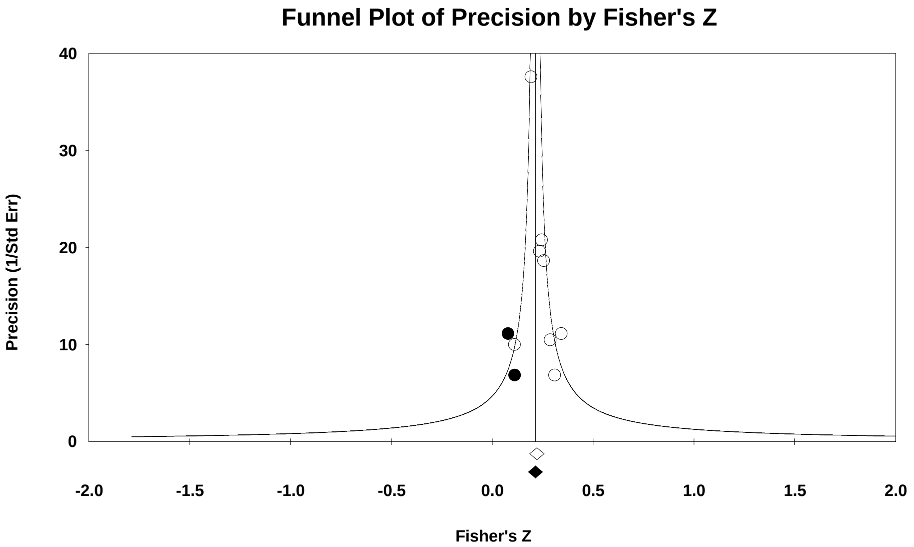
Table 9. Publication Bias Results for EI-Customer Service Scale

A. Meta-Analysis Results

Study name	Statistics for each study				
	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Quick Service Restaurant (US) Hourly Crew Members	0.240	0.149	0.327	5.088	0.000
Quick Service Restaurant (Canada) Hourly Crew Members	0.280	0.100	0.442	3.017	0.003
Regional Department Store Chain Sales Associates	0.230	0.133	0.322	4.595	0.000
Truck Rental Company Rental Representative	0.330	0.165	0.477	3.818	0.000
Convenience Store Chain Cashiers & Pump Attendants	0.250	0.149	0.346	4.765	0.000
Shoe Store Chain Store Associates	0.110	-0.085	0.297	1.104	0.269
Discount Retailer Liquidator Store Associates & Stockers	0.300	0.024	0.534	2.122	0.034
National Grocery Store Chain All Hourly Public Contact Jobs	0.190	0.139	0.240	7.232	0.000
	0.218	0.184	0.252	12.067	0.000



B. Duval and Tweedie's Trim and Fill

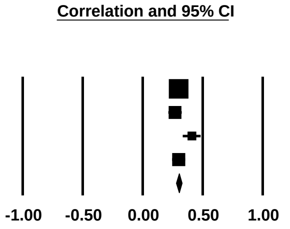


The clear circles represent the observed data. The dark circles represent imputed studies. The trim and fill procedure suggests that two studies are missing. The point estimate and 95% confidence interval for the combined studies is 0.218 (0.184, 0.252). Using Trim and Fill the imputed point estimate is 0.211 (0.177, 0.244).

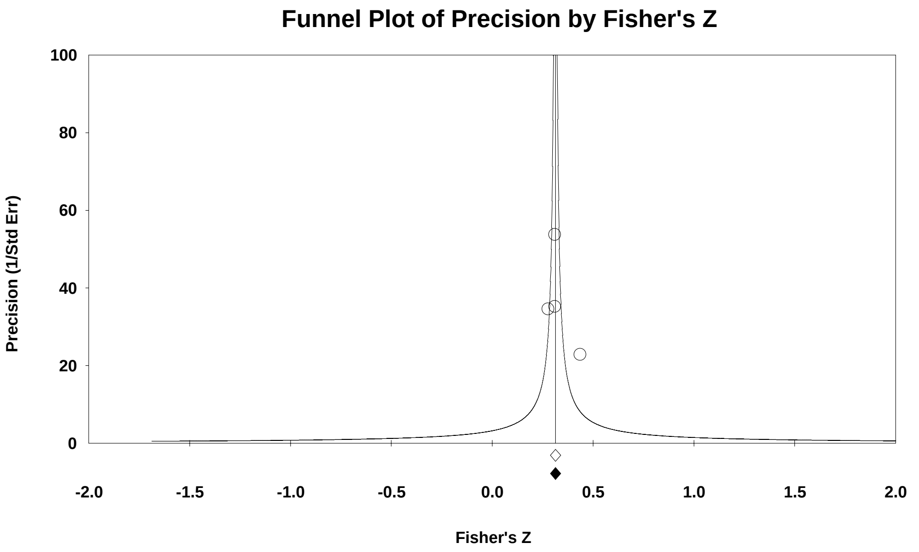
Table 10. Publication Bias Results for EI-Customer Service and Clerical Potential Index

A. Meta-Analysis Results

Study name	Statistics for each study				
	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Group 1	0.300	0.266	0.333	16.648	0.000
Group 2	0.270	0.217	0.322	9.583	0.000
Group 3	0.410	0.336	0.479	9.962	0.000
Group 4	0.300	0.249	0.350	10.912	0.000
	0.304	0.281	0.327	24.038	0.000



B. Duval and Tweedie's Trim and Fill

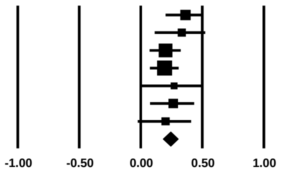


The clear circles represent the observed data. The dark circles represent imputed studies, but in this case none were imputed. The trim and fill procedure suggests that zero studies are missing. The point estimate and 95% confidence interval for the combined studies is 0.304 (0.280, 0.327). Using Trim and Fill the imputed point estimate is 0.304 (0.280, 0.327).

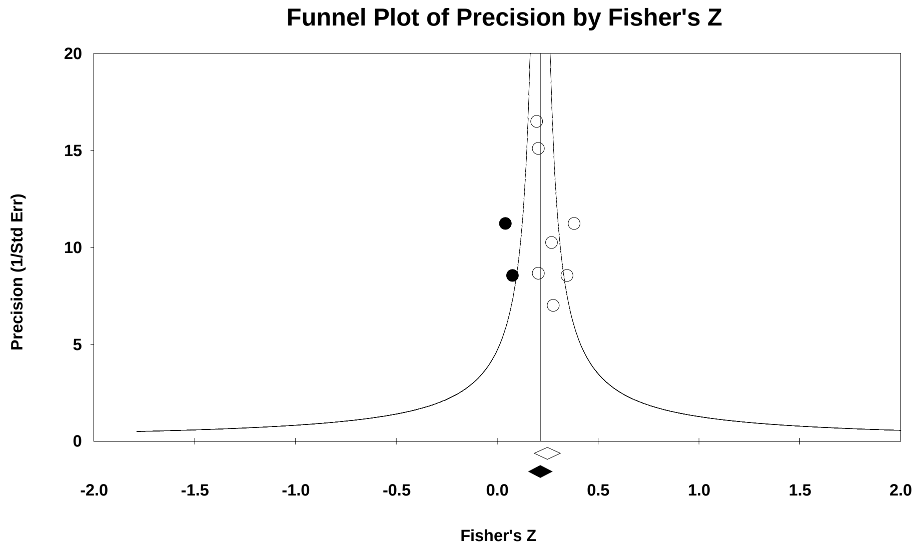
Table 11. Publication Bias Results for Leadership Inventory

A. Meta-Analysis Results

Study name	Statistics for each study					Correlation and 95% CI
	Correlation	Lower limit	Upper limit	Z-Value	p-Value	
Consumer Direct Sales Managers	0.364	0.204	0.505	4.283	0.000	
Department Store Managers	0.333	0.116	0.520	2.959	0.003	
Department Store, Department Managers	0.201	0.074	0.322	3.083	0.002	
Fast-Food Restaurant Managers	0.194	0.077	0.305	3.235	0.001	
Grocery Distribution Management	0.271	-0.002	0.507	1.946	0.052	
Banking Center Managers	0.263	0.078	0.431	2.764	0.006	
Banking Center Assistant Managers	0.201	-0.022	0.406	1.768	0.077	
	0.243	0.182	0.303	7.567	0.000	



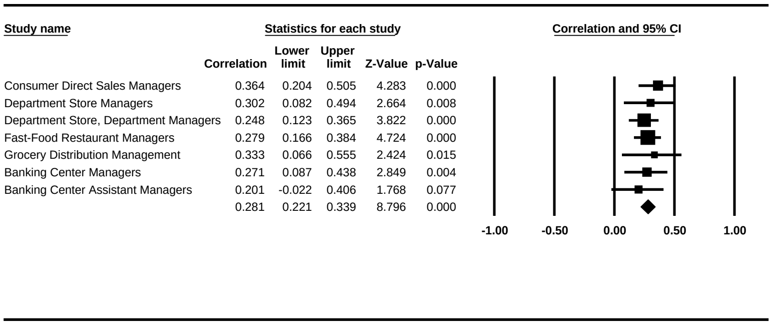
B. Duval and Tweedie's Trim and Fill



The clear circles represent the observed data. The dark circles represent imputed studies, but in this case none were imputed. The trim and fill procedure suggests that two studies are missing. The point estimate and 95% confidence interval for the combined studies is 0.243 (0.182, 0.303). Using Trim and Fill the imputed point estimate is 0.211 (0.155, 0.265).

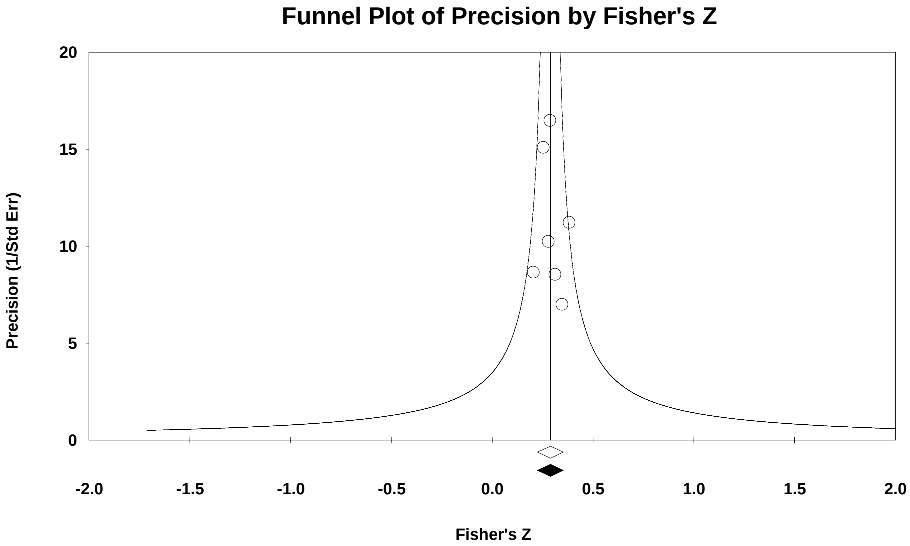
Table 12. Publication Bias Results for Leadership Inventory Plus

A. Meta-Analysis Results



Study name	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Consumer Direct Sales Managers	0.364	0.204	0.505	4.283	0.000
Department Store Managers	0.302	0.082	0.494	2.664	0.008
Department Store, Department Managers	0.248	0.123	0.365	3.822	0.000
Fast-Food Restaurant Managers	0.279	0.166	0.384	4.724	0.000
Grocery Distribution Management	0.333	0.066	0.555	2.424	0.015
Banking Center Managers	0.271	0.087	0.438	2.849	0.004
Banking Center Assistant Managers	0.201	-0.022	0.406	1.768	0.077
	0.281	0.221	0.339	8.796	0.000

B. Duval and Tweedie's Trim and Fill



The clear circles represent the observed data. The dark circles represent imputed studies, but in this case none were imputed. The trim and fill procedure suggests that zero studies are missing. The point estimate and 95% confidence interval for the combined studies is 0.281 (0.220, 0.339). Using Trim and Fill the imputed point estimate is 0.281 (0.220, 0.339).

Figure 1. Funnel plot

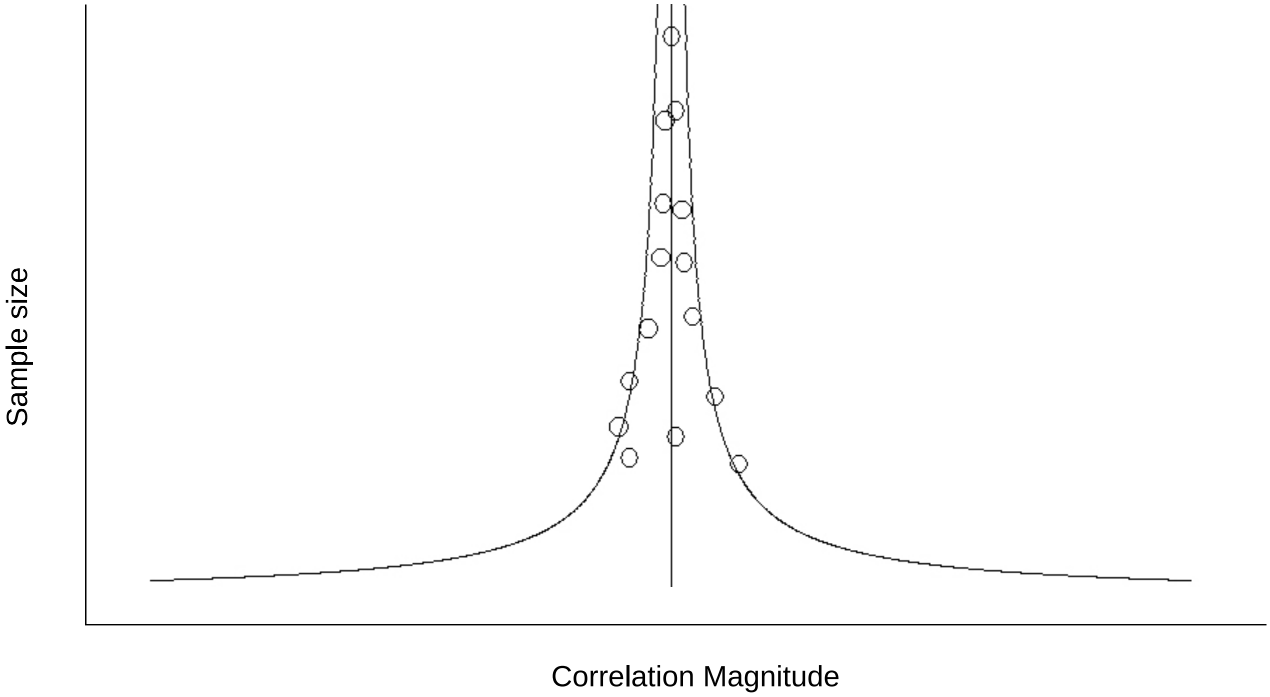


Figure 2. Illustrative symmetrical and non-symmetrical funnel plots

Figure 2a. Symmetrical funnel plot

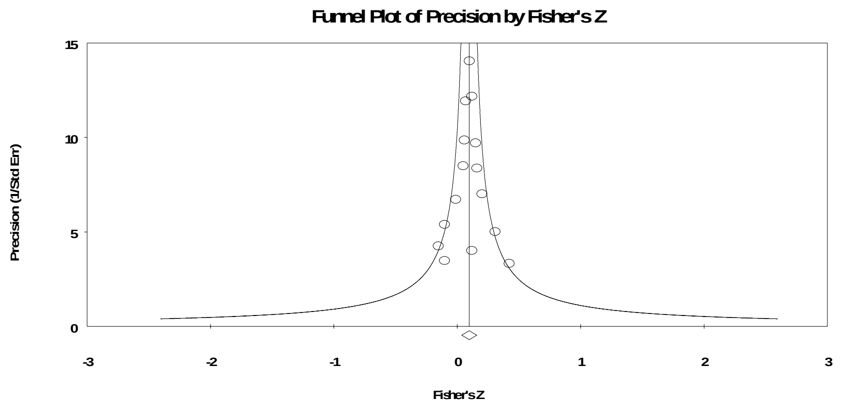


Figure 2b. Non-Symmetrical funnel plot

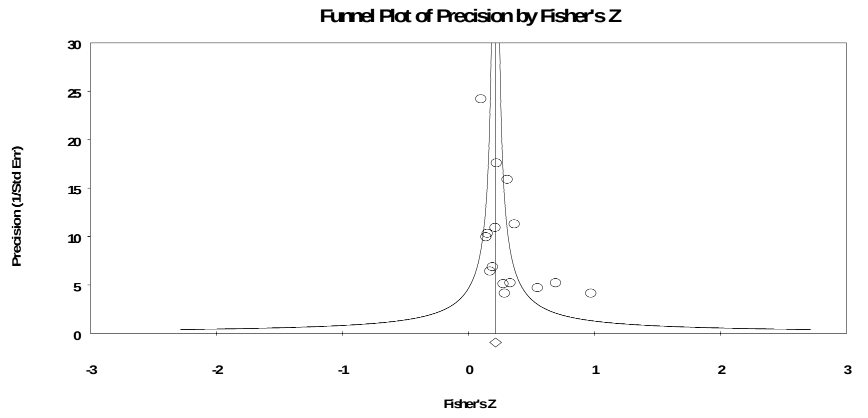
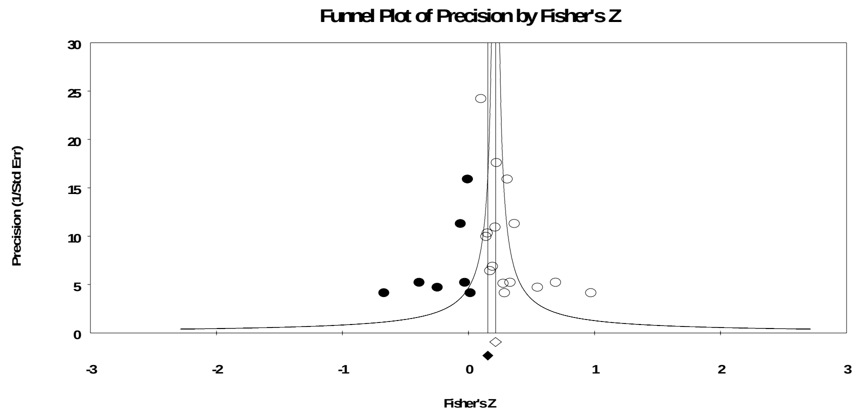


Figure 2c. Non-Symmetrical funnel plot with imputed studies


Funnel Plot of Precision by Fisher's Z
Precision (1/Std Err)
Fisher's Z