Running Head: Publication Bias in the EI?

An Examination of the PreVisor™ Employment Inventory for Publication Bias

Jeffrey M. Pollack & Michael A. McDaniel

Department of Management, School of Business

Virginia Commonwealth University

P. O. Box 844000

Richmond, Virginia 23284-4000

(804) 397-0818

Fax: (804) 828-8884

pollackjm@vcu.edu

Paper presented at SIOP (April, 2008). San Francisco.

Authors' note: The authors appreciate feedback received from Ken Lahti, George Paajanen, and Tom Janz. The paper has benefited from their thoughtful feedback. Our acknowledgement of their helpful feedback should not be viewed as their endorsement of this paper.

Abstract

We examined the technical manuals of the Employment Inventory, a product of PreVisor™, for potential publication bias in the validity data reported in the test manuals. We looked for publication bias separately by three criteria: dependable job behavior, rehireability, and termination code. For the dependable job behavior criterion, results were consistent with the inference that lower magnitude validity coefficients were suppressed such that the validity data in the Employment Inventory manuals overestimate the typical validity. Alternative explanations for the results are also offered. For the rehireability and termination code criteria, there was little evidence of publication bias suggesting that the validity coefficients are not overestimated.

KEYWORDS: publication bias, Employment Inventory, trim and fill,

An Examination of the PreVisor™ Employment Inventory for Publication Bias

Users of employment tests often rely on technical data provided by test publishers to assist in evaluating the usefulness of a test. In a review of test vendor's reporting practices, McDaniel, Rothstein and Whetzel (2006) found evidence of likely publication bias in the technical manuals of two of the four test vendors reviewed. Specifically, there was evidence consistent with the inference that some test vendors were systematically suppressing lower magnitude validity coefficients to make their test products appear more valid than they actually were.

Publication bias can occur when studies are more likely to be published depending on the magnitude, direction, or statistical significance of the findings (McDaniel, Rothstein, & Whetzel, 2006). Studies that are suppressed (i.e., missing from the literature) may be absent for a number of reasons. For example, studies with statistically non-significant or marginally significant results are less likely to be submitted by authors for publication, or if submitted are less likely to be published, whereas statistically significant results are more likely to be submitted and may benefit from favorable editorial policy (Dickersin, 2005; Greenwald, 1975). Despite the importance of accounting for publication bias in reporting results, the topic has received limited attention in the personnel selection literature. Given the implications for employers as well as job applicants, this is unfortunate.

The purpose of this study is to examine the technical documentation for the Employment Inventory (Paajanen, Hansen, & McLellan, 1993) to determine if there is any evidence supporting a conclusion of publication bias. The Employment Inventory is a

well known personality test for personnel selection that commands substantial market share:

> "Since it was created in 1986, the PDI Employment Inventory has become the leader in selection measurement. More than three million employment inventories are administered each year in approximately 300 organizations worldwide. Many of the largest retailers in the United States – including supermarket chains, car rental companies, fast-food organizations, airlines, and discount stores – use PDI's Employment Inventory to help select high-quality, reliable employees." (Personnel Decisions, Inc., undated, page 9).

The test was originally developed by Personnel Decisions, Inc (PDI) (Paajanen, Hansen, & McLellan, 1993). The test has changed ownership due to mergers and acquisitions of various human resource consulting firms. After PDI, the Employment Inventory was a product of ePredix and is now a product of PreVisor™. The Employment Inventory has been marketed as an instrument designed to "identify applicants most likely to become productive and successful employees in entry-level and non-exempt positions (ePredix, 2001, pp. 5)." The test has two scales. The Performance Scale is intended to "measure personality characteristics related to employee dependability, which basically underlies the full range of productive and counterproductive behavior (ePredix, 2001, pp. 15)." The Tenure Scale is used to predict "how long a candidate would voluntarily stay on the job (ePredix, 2001, pp. 15)." In the validity documentation (ePredix, 2001; Paajanen, Hansen, & McLellan, 1993), validities are reported for a single composite of the performance and tenure scales and it is this composite scale that is referred to as the Employment Inventory in this paper.

The Employment Inventory has been the subject of a substantial number of validity studies as documented in its technical manuals. Although most of the validity studies were done by PDI for its clients, the instrument has also been used in research unconnected to the test publisher (Carless et al., 2007).

Recent analyses indicate that the effect of publication bias can be severe. The results of a publication bias analysis by Duval (2005) and Oh, Postlethwaite, Schmidt and McDaniel (2007) indicated that, without publication bias, the estimated validity of structured interviews, analyzed by McDaniel, Whetzel, Schmidt, and Maurer (1994), would likely be lower and more similar to unstructured interviews. Additionally, results of other publications bias analyses indicated that the magnitude of Black/White mean differences in job performance may be underestimated in the published literature (McDaniel, McKay, & Rothstein, 2006).

Considering the issues surrounding publication bias, test publishers and researchers should examine its impact. However, extant methods for assessing publication bias have been used rarely. The most prevalent method of assessing publication bias is the failsafe N method, commonly known as the “file drawer problem” (Rosenthal, 1979). The failsafe N method has been used with relative frequency by industrial organizational psychologists, but it has substantial deficiencies in assessing publication bias (Becker, 2005). Other analysis tools are more sensitive to publication bias (Duval, 2005; Hedges & Vevea, 2005, Sterne & Egger, 2005, Sutton & Pigott, 2005). Here we briefly review several publication bias analysis methods and then apply the appropriate ones to an exploration of potential publication bias in the Employment Inventory.

Rosenthal Failsafe N (The "File Drawer" Problem)

Due to concerns that relevant studies were not being published, and subsequently were missing from the available literature, Rosenthal (1979) introduced the term "file drawer problem." Rosenthal (1979) suggested that, rather than estimate the number of "file drawer" studies needed to "nullify" an observed effect, the number of studies required to "nullify" an effect can be calculated. This number is referred to by Cooper (1979) as the failsafe N. Rosenthal (1979) asserted that a large failsafe N could indicate that a high level of confidence may be held in the results of an analysis, and although the effect size may be overestimated due to missing studies, it would likely not be zero.

The failsafe N method is flawed in two main ways. First, the assumption is made that all suppressed studies have an effect size of zero, as opposed to considering that the effect size may, in fact, be negative (Becker, 1994; 2005). Second, the failsafe N method focuses on statistical significance rather than practical significance (Becker, 1994; 2005). Becker (2005) recommended that the failsafe N method not be used due to these flaws. Following this recommendation, we do not report failsafe N analyses.

Begg and Mazumdar Rank Correlation Test

When publication bias exists, large studies tend to be included in analyses regardless of their effect size, but small studies are more likely to be included only when reporting a large effect (e.g., validity coefficient). Generally, this leads to an inverse correlation between study size and effect size. Begg and Mazumdar (1994) argued that this correlation can serve as a test for publication bias. Specifically, they suggest that one compute the rank order correlation (Kendall's tau b) between the treatment effect and the standard error (the standard error is primarily a function of sample size and is highly and negatively correlated with sample size). However, this approach has some limitations.

Though a significant correlation does suggest that bias exists, the method does not provide an estimate of the effect size (e.g., the validity coefficient) if the suppressed studies were available and included in the analysis. Also, a non-significant correlation may be due to low statistical power, and cannot be taken as evidence that bias is absent. The relative low power of this method is related to the number of studies (*k*) being the sample size for the analysis and because it is a nonparametric test.

Egger's Test of the Intercept

Egger, Smith, Schneider, and Minder (1997) offered a regression-based approach to examine the relationship between precision (a statistic highly correlated with sample size) and publication bias. In this method, the publication bias is estimated by a test of the intercept. Although the method has somewhat greater power than the Begg and Mazumdar approach because the Egger test is parametric, a non-significant correlation may also be due to low statistical power. As with the Begg and Mazumdar approach, the low power of the Egger test is related to the numbers of studies (*k*) being the sample size for the analysis. Due to the low power, the test cannot be taken as evidence that bias is absent. Also, the method does not provide an estimate of the effect size (e.g., the validity coefficient) if the suppressed studies were available and included in the analysis.

Trim and Fill

Duval and Tweedie (Duval, 2005; Duval & Tweedie, 2000a; 2000b) introduced a method called trim and fill to detect publication bias. An understanding of the method requires familiarity with a funnel plot (Light & Pillemer, 1984). Figure 1a graphs, in open circles, the correlations from a set of studies (reproduced from McDaniel et al., 2006). The X axis reflects the magnitude of the correlations and the Y axis reflects the sample sizes. In absence of publication bias, we would expect a normal distribution whereas

asymmetry is considered an indicator of bias (Duval & Tweedie, 2000a; 2000b). When sampling error is the only source of variance in the distribution, correlations from large samples will cluster towards the center line of the funnel and correlations from smaller samples may underestimate or overestimate the population correlation. Thus, the correlations form a roughly symmetrical funnel distribution. Trim and fill analysis methods were developed in the medical literature where studies are typically weighted by precision (the inverse of the standard error) rather than by sample size, which is more commonly used in personnel selection meta-analyses. As a result, all trim and fill software weights studies by precision rather than sample size. For correlation coefficients, precision is substantially correlated with sample size. Thus, one can think of the Y axis as sample size. Trim and fill is based on examining asymmetry so the effect size being examined needs to have a symmetrical sampling error distribution. The sampling error distribution of correlation coefficients is asymmetrical (except when the population correlation is zero). To conduct a trim and fill analysis with correlation coefficients, the correlations are transformed into Fisher z's because that statistic has a symmetrical sampling distribution (McDaniel et al., 2006).

An example of a symmetrical funnel plot correlation, plotted as Fisher z as a function of precision, is shown in Figure 1a (reproduced from McDaniel et al., 2006). If the variance in correlations across studies is solely a function of random sampling error the funnel plot would be symmetrical.

Figure 1b (reproduced from McDaniel et al., 2006) illustrates an asymmetric funnel plot consistent with a suppression effect. Note that low validity small samples studies are not present in the funnel plot. Larger samples with lower validities are less likely to be suppressed because they are more likely to reach statistical significance.

Assumptions of trim and fill

There are two key assumptions underlying the trim and fill method (McDaniel et al., 2006). First, trim and fill assumes that there may be existing correlations not known to the reviewer. Trim and fill analyses incorporate an algorithm to estimate the magnitude of these missing studies. Second, trim and fill assumes that sampling error is the sole source of variance in a sample and thus, to the extent possible, the analyses should be applied to validity distributions that are homogenous (e.g., free of moderators). Thus, for example, if sex moderated the relationship between two variables, one would subset the data by the sex of the sample and conduct trim and fill analyses separately for the male samples and for the female samples.

In the case of validity data where low magnitude correlations may be suppressed, the trim and fill method searches for missing studies on the left side of the funnel. First, correlations on the right side of the funnel without close counterparts on the left side of the funnel are removed using an iterative process such that the resulting distribution is reasonably symmetrical. This "trimming" produces a truncated but symmetrical distribution, the mean of which is assumed to be the mean of a bias free distribution. The algorithm then restores the correlations trimmed from the right side of the distribution and imputes a mirror image correlation on the left side of the funnel such that the resulting distribution is symmetrical. An estimate of the mean correlation and variance of the correlation distribution is then calculated on the new "filled" funnel (Duval & Tweedie, 2000a, 2000b; McDaniel et al., 2006). An example of this "filled" funnel is shown in Figure 1c (reproduced from McDaniel et al., 2006).

The main benefit of the trim and fill method is that an adjusted effect size estimate for the funnel plot asymmetry is provided. These mean values based on the

observed and imputed correlations should not necessarily be viewed as the unbiased estimate of the validity. However, the mean of the "filled" distribution can be compared to the mean original distribution to see how much they differ. There are three ways in which the trim and fill method can be used to assess publication bias. First, one can visually inspect the original validity distribution to draw inferences about publication bias. Second, one can count the number of imputed studies and use this information to inform conclusions about publication bias. Third, a difference between the means of the observed and the trim and fill distribution can be viewed as evidence of publication bias.

Method

Test and data source.

The Employment Inventory Performance/Tenure (EI/PT) consists of a set of 25 predictor constructs organized into 10 categories which are: undependability, socialization, attitudes, problems with authority relationships, excitement seeking, work motivation, social influence, unstable upbringing, drug/alcohol use, and unmet needs (ePredix, 2001, pp. 11-14). Outcomes related to performance were categorized by the criteria of dependable job behavior, rehireability, and termination code. Dependable job behavior and rehrieability are supervisor ratings. Termination code characterized employees who were classified as either 1) satisfactory performers who stayed on the job for at least three months and would be rehired, or who left before three months but would also be rehired, 2) marginal performers who had been laid off or fired but may be rehired, and those who had quit but would not be rehired, or 3) problem performers who had been laid off or fired and who would not be rehired (ePredix, 2001, pp. 17).

The three criteria that were examined were those for which ten or more validity studies were available. These data came from Table 7 in the technical manual of the Employment Inventory Performance/Tenure (EI/PT) published by PDI (Paajanen, Hansen, & McLellan, 1993). At some later time, ePredix obtained rights to the Employment Inventory and we obtained a (2001) technical manual from ePredix. Both the Personnel Decisions, Inc. technical manual from 1993 and the ePredix Incorporated technical manual from 2001 provide the exact same validity data for the Employment Inventory. Thus, any publication bias found in the Employment Inventory validity distribution analyzed in this paper can best be attributed to data suppression that occurred at PDI in 1993 or earlier. PreVisor™ (personal communication to Michael A. McDaniel, on August 6, 2007) stated that the PreVisor™ Employment Inventory technical manual contains no additional validity data than those that appear in the ePredix manual. Because the authors have the ePredix manual, we know that we have all the data in the PreVisor™ manual. We were unable to obtain a copy of the PreVisor™ manual, because unlike most test publishers (e.g., Hogan Assessment Systems, Psychological Services, Inc., SHL, and Wonderlic), PreVisor™ refuses to release or sell its technical manuals without the recipient signing a confidentiality agreement, the terms of which would preclude the current authors from publishing this paper.

Analysis

We conducted a meta-analysis of the observed validity coefficients obtained from the technical manuals of the Employment Inventory Performance/Tenure (EI/PT) for three criteria: dependable job behavior, rehireability, and termination code. Seventy validity coefficients were available for the dependable job behavior criterion. For the rehireability criterion, thirteen validity coefficients were available. Forty-one validity

coefficients were available for the termination code criterion. Because the type of criterion measure might moderate the validity of the Employment Inventory, we conducted publication bias analyses separately be criteria type. We used the Comprehensive Meta-Analysis (CMA) software (Borenstein, Hedges, Higgins, & Rothstein, 2005) for the meta-analyses and for the publication bias analyses.

Results

Results are first presented for the dependable job performance criteria. Then results are presented for the rehire criteria. Finally, we present the results for the termination code criteria.

Publication bias analyses for the dependable job performance criterion

Begg and Mazumdar Rank Correlation Test. The Begg and Mazumdar tests indicated a correlation of .23 ($p < .01$) between the validity correlations and their standard errors. Expressed in terms of sample size, there is a negative relationship between sample size and the validity coefficients. With respect to a funnel plot, this would indicate that the smaller magnitude, smaller samples sizes are largely absent from the distribution. These results are consistent with an inference that the smaller magnitude, smaller sample studies have been suppressed from the Employment Inventory technical manual such that the data that are reported are a likely overestimate of the validity of the criterion.

Egger's Test of the Intercept. The Egger test yielded an intercept of .99 that was statistically significant from zero ($p < .01$). With respect to a funnel plot, this would indicate that the smaller magnitude, smaller samples sizes are largely absent from the distribution. As with the Begg and Mazumdar results, these results are consistent with an

inference of study suppression in the Employment Inventory technical manual such that the data that are reported are a likely overestimate of the validity of the criterion.

Trim and Fill Results. If the current meta-analysis incorporated all studies we would expect the funnel plot to be symmetrical and the variability in the distribution would be a function of simple random sampling error. However, if it does not incorporate all studies, then we can expect asymmetry. Figure 2 represents the observed data points and Figure 3 illustrates the observed and imputed data points. A visual inspection of the funnel plot in Figure 2 indicates that the funnel plot is substantially asymmetrical. As with the Begg and Mazumdar and Egger analyses, these results are consistent with an inference that the smaller magnitude, smaller sample studies have been suppressed in the Employment Inventory technical manual such that the data reported likely overestimate the validity of the criterion. The trim and fill analysis indicated that 22 correlations would need to be imputed to make the funnel plot symmetrical. If this value is an accurate estimate of the number of studies that were potentially suppressed, then approximately one-fourth of the validity coefficients (22 / (22 + 70)) might been excluded from the Employment Inventory technical manuals. The mean validity (and 95% confidence interval) of the 70 observed correlations is 0.25 (0.23, 0.27). Using trim and fill distribution (the 70 observed correlations and the 22 imputed correlations is 0.23 (0.21, 0.25), a difference of .02. Although this difference is not large, it supports the inference that the smaller magnitude, smaller sample studies have been suppressed in the Employment Inventory technical manual such that the data that are reported are a likely overestimate of the validity of the dependability criterion.

Publication bias analyses for the rehireability criteria

Begg and Mazumdar Rank Correlation Test. The Begg and Mazumdar tests indicated a correlation of .19 (p = .180) between the validity correlations and their standard errors. This result does not support an inference of publication bias.

Egger's Test of the Intercept. The Egger test yielded an intercept of .50 that was not statistically significant (p = .60). As with the Begg and Mazumdar results, these results are consistent with an inference that the smaller magnitude, smaller sample studies have not been suppressed in the Employment Inventory technical manual for the rehireability criterion.

Trim and Fill Results. Figure 4 represents the observed data points and Figure 5 illustrates the observed and imputed data points. A visual inspection of the funnel plot of observed data indicates that the funnel plot is slightly asymmetrical. The trim and fill analysis (see Figure 5) indicated that two correlations would need to be imputed to make the funnel plot symmetrical. If this value is an accurate estimate of the number of studies that were potentially suppressed, then approximately ten percent of the validity coefficients (2 / (2 + 13)) might been excluded from the Employment Inventory technical manuals in the case of the rehireability criterion. The mean validity (and 95% confidence interval) of the 13 observed correlations is 0.26 (0.23, 0.28). Using trim and fill distribution (the 13 observed correlations and the two imputed correlations is 0.25 (0.23, 0.27), a difference of .01. This difference is not large. Based on the Begg and Mazumdar test, the Egger test, and the trim and fill analysis, there is no evidence of publication bias in the Employment Inventory technical manual for the validity of the test in the prediction of the rehireability criterion.

Publication bias analyses for the Termination Code

Begg and Mazumdar Rank Correlation Test. The Begg and Mazumdar tests

indicated a correlation of .17 (p = .07) between the validity correlations and their standard errors. Had this correlation been statistically significant, these results would be consistent with an inference that the smaller magnitude, smaller sample studies have been suppressed from the Employment Inventory termination code criterion data. The results of the Begg and Mazumdar test suggest that this is not the case.

Egger's Test of the Intercept. The Egger test yielded an intercept of .05 that was statistically not significant (p = .45). As with the Begg and Mazumdar results, these results are consistent with an inference that the smaller magnitude, smaller sample studies have not been suppressed in the Employment Inventory technical manual regarding the termination code criterion.

Trim and Fill Results. Figure 6 represents the observed data points and Figure 7 illustrates the observed and imputed data points. A visual inspection of the funnel plot indicates that the funnel plot is somewhat asymmetrical. The trim and fill analysis indicated that 2 correlations would need to be imputed to make the funnel plot symmetrical. If this value is an accurate estimate of the number of studies that were potentially suppressed, then approximately five percent of the validity coefficients (2 / (2 + 41)) might been excluded from the Employment Inventory technical manuals regarding termination code criterion. The mean validity (and 95% confidence interval) of the 41 observed correlations is 0.25 (0.25, 0.26). Using trim and fill distribution (the 41 observed correlations and the 2 imputed correlations is 0.25 (0.25, 0.26), a zero difference. Based on the Begg and Mazumdar test, the Egger test, and the trim and fill analysis, there is little evidence of publication bias in the Employment Inventory technical manual for the validity of the test in the prediction of the termination code criterion.

Discussion

The results of our analyses are consistent with the inference that lower magnitude validity coefficients were suppressed in the Employment Inventory technical manuals such that the validity data reported overestimates the typical validity of the PreVisor™ Employment Inventory on the criterion of dependable job behavior. The five decision rules on which we make base this conclusion are as follows. First, the Begg and Mazumdar analyses suggested data suppression in that there is a relationship between the standard errors and the validity coefficients. Second, the Egger test suggested data suppression in that the intercept was significantly different from zero. Third, a visual inspection of the funnel plot showed substantial asymmetry suggesting data suppression. Fourth, the trim and fill analysis indicated that 22 correlations would need to be imputed to bring the funnel plot into symmetry. Based on other applications with this method, we argue that 22 imputed studies represent a large number of missing studies. Fifth, the mean of the trim and fill distribution was smaller by .02 than the observed distribution. Although this difference is small, the direction of the differences is consistent with the inference that low magnitude validity coefficients have been suppressed from the technical manuals of the Employment Inventory. In contrast, the analyses suggest no suppression of small magnitude validity coefficients for the criterion of rehireability and termination code.

So what conclusions can be drawn about the validity of the PreVisor™ Employment Inventory? For rehireability and termination code, the validities in the Employment Inventory manuals appear to be an unbiased sample of observed validity coefficients. However, for the criterion of dependable job behavior, the results of the analyses may cause many to conclude that the validity is lower than that offered in the

Employment Inventory technical manuals. It would be prudent for PreVisor™, the current publisher of the Employment Inventory, to attempt to determine the extent to which data suppression may have occurred in the selection of the studies for the 1993 Personnel Decisions, Inc. manual. If the passage of time or the turnover of staff precludes locating the original validation studies, PreVisor™ may wish to conduct new validity studies to insure that the validity data that it reports on its product is not biased due to data suppression.

Alternative explanations

We have concluded that our results are consistent with an inference that low magnitude validities results were suppressed. There are other possible inferences that could be drawn from these results. In guiding our discussion of alternative explanations, we sought feedback from those who have a stake in the accuracy of the Employment Inventory validity data. Specifically, we sent an earlier draft of our paper to PDI, the initial owner of the Employment Inventory, PreVisor™ the current owner of the Employment Inventory, and two of the three authors of the 1993 Employment Inventory test manual. One of the three authors could not be located. Our discussion of the alternative explanations is in part based on two sets of feedback that were received. Other alternative explanations have been developed by the authors of this paper.

One of the authors of the PDI employment inventory wrote us and stated that "there was no reporting bias of validity results in the test manual. All coefficients from all studies were included in the manual tables, and many more studies were conducted since its publication." The author stated that he did not know why the validity coefficients showed asymmetry. Based on this statement by one of the technical manual authors, we speculate that if the validity studies were not conducted by the authors of the technical

manual, the organizations who did conduct the validity studies may have only shared results with PDI if the results were positive. Thus, it is possible that the authors of the technical manual were not suppressing data known to them. Those with knowledge about the source of the validity studies in the technical manuals have not shared with us information relevant to this argument. We also note that the methods for detecting evidence of publication bias that were used in this article did not exist at the time the 1993 technical manual was written and thus the manual authors could not have used these methods in assessing potential publication bias. To the extent that publication bias exists in these data, we do not know its origin and believe that arguments suggesting that the technical manual authors are the source of the publication bias are speculative and should not be made in the absence of evidence.

A representative of PreVisor™ also provided some feedback. In addition to other feedback, the representative wrote: "It is important to note that this response and acknowledgement of our review of the manuscript is in no way an endorsement by PreVisor of the research design, the obtained results, or the conclusions presented in the manuscript. In particular, the Discussion section presents conclusions about publication bias that we do not believe are adequately supported." We provide this quote so that PreVisor™'s position would not be misunderstood. We believe that a review of alternative explanations for the findings enhances one's understanding of potential publication bias. Below, we offer five alternative explanations for the asymmetry in the validity data.

The first alternative explanation to the inference that the data were suppressed is known as a small study effect (Sterne, Becker, & Egger, 2005). Small studies may be conducted with greater care than larger studies. For example, in small studies it may be

easier to provide extensive training to raters of job performance and such training might enhance validity. Under this reasoning, the asymmetry might be due to the better studies having smaller sizes. This argument would be a more compelling explanation for our results if someone with access to the primary validity studies summarized in the Employment Inventory technical manuals could offer evidence that this did occur.

A second alternative explanation may be that that small sample studies might be subject to outlier analysis resulting in case deletion. Thus, studies may not have been suppressed but small sample studies might have higher validities because cases in those studies that were outliers to the general trend of the data were deleted. Outlier data deletion may be more common with small sample studies because the effects of the data deletion may be more pronounced with small samples than with larger samples. For example, dropping five outliers from a small sample would likely result in a greater enhancement of the validity coefficient than dropping five outliers from a large sample. Opinions vary as to the appropriateness of outlier removal. Some might find it reasonable to drop cases for a supervisor when the supervisor did not appear to provide criterion data with care. Others might find this practice to be a mechanism of publication bias such that small magnitude validities are falsely reported as larger magnitude validities. Those who conducted the primary studies that contributed data to the technical manual might shed light on this speculative alternative explanation.

A third alternative explanation is that when the researchers are constrained to collecting data from only a small sample, they might select the sample such that it includes only those employees who had exceptionally high or exceptionally low criterion scores. This would create range enhancement in the data set and overestimate the validity relative to the population validity. The Employment Inventory manual did not

address this issue and we do not know how credible this alternative explanation might be. Although this would be an alternative explanation to an inference of data suppression, many would consider this practice to be a mechanism for publication bias. Those who conducted the primary studies that contributed data to the technical manual might shed light on this speculative alternative explanation.

A fourth alternative explanation may be that the results are randomly asymmetrical and that no data suppression has occurred. Given the number of validity studies and the degree of asymmetry, the authors do not find this explanation to be compelling.

A fifth alternative explanation is that there is an unknown moderator operating in the data set that co-varies with sample size. The symmetry-based publication bias statistics used in this paper assume that sampling error is the only source of variance in the distribution (that is, there are no moderators present). If small studies were associated with a moderator condition that yielded higher magnitude validity, and the moderator was randomly distributed in the larger sample studies, one might obtain an asymmetric validity distribution. For example, in medical studies, the initial trials of a drug treatment may be based on a small sample of severely ill individuals. If the drug was effective, it may show a larger effect size for these severely ill individuals than in comparatively less ill subjects found in later larger sample studies. The Employment Inventory technical manual did not suggest that a moderator was operating in these data and this paper's authors have no idea of what such a moderator might be or how it might be correlated with sample size. Given this, we do not know how to judge the credibility of this alternative hypothesis.

In summary, there are other possible explanations for the asymmetry than the inference that small sample validity studies were suppressed from the Employment Inventory technical manual. Some of the alternative explanations are more credible than others. Some could gain or lose credibility through an examination of the primary studies whose validity statistics were summarized in the Employment Inventory technical manuals. Those who hold the studies are encouraged to conduct such an examination.

Responses to criticism of this paper

Because this paper raises questions about a well-known and respected employment test, the authors distributed drafts of this paper to various researchers and solicited feedback. Here, we respond to two criticisms of the paper. One reviewer of this paper has argued that the difference between the mean observed distribution and the trim and fill distribution of .02 is sufficiently small that the best conclusion is there is no evidence of publication bias. If this was the sole bit of evidence relevant to publication bias, we would agree with this reviewer. However, the reviewer ignores the four other lines of evidence, summarized in the first paragraph of the discussion, that are consistent with an inference of publication bias. Related to this criticism is the argument that, if .02 is the best estimate of the amount of publication bias in the distribution ,any publication bias that occurred had little substantive impact. Duval (2005), the co-developer and primary methodologist of the trim and fill method, argued that the mean of the trim and fill adjusted distribution "should be used primarily as a form of sensitivity analysis, to assess the potential impact of missing studies on the meta-analysis, rather than as a means of adjusting results.." (p. 131). Thus, although it tempting to argue that the mean validity of trim and fill adjusted distribution is the operational validity of the test in the absence of bias, this conclusion is speculative. If the operational validity of the Employment

Inventory was only .02 less than that offered in the technical manuals, we would concur that any operating publication bias had little effect. It is the authors' position that we do not know the operational validity of the Employment Inventory if publication bias were operating.

Additional implications of publication bias

Questions about the accurate reporting of the Employment Inventory raises five additional concerns. First, validity studies on the Employment Inventory were one source of data in the Ones et al. (1993) meta-analysis of integrity tests. Another test classified as a personality based integrity measure used in the Ones et al. analyses was evaluated in McDaniel et al. (2006; see vendor B results) who reported that the results were consistent with an inference of publication bias. Thus, publication bias analyses of two of the integrity tests contributing data to the Ones et al. results (1993) have offered evidence consistent with the inference of data suppression of small validity coefficients such that technical manuals may overestimate the validity of these tests. Thus, it would be prudent for the Ones et al. (1993) authors to examine their data for publication bias.

Second, we note that instances of data suppression are inconsistent with the *Principles for the Validation and Use of Selection Procedures* (Society of Industrial and Organizational Psychology, Inc., 2003) and the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999) which require reporting of all validity data. Because PreVisor™ is one of the primary employment test publishers and an employer of about twenty SIOP members, it would be prudent if these SIOP member employees examined PreVisor™'s practices to ensure that they are consistent with the ethical principles of their profession.

Third, we believe the findings of this study and that of McDaniel et al. (2006) present an ethical dilemma for SIOP's officers and the executive committee. Test vendors provide substantial money to SIOP in the form of booth reservations and program advertisements. If a test vendor is violating the SIOP Principles, is it ethical for SIOP to accept advertising money from the test vendor? On November 2, 2006, twelve SIOP Fellows wrote SIOP's President expressing concern over this matter, and as of April 1, 2008, they have not received a response from SIOP's president and/or executive committee.

Fourth, for test vendors, the accurate and comprehensive reporting of all validity data is essential. Misrepresentation of data to current and potential customers is misleading. For example, Merck has been accused of data suppression for its drug Vioxx (Curfman et al., 2005) and is facing a variety of legal challenges. Likewise, if one test publisher buys the product of another, the purchasing buyer would likely expect the product information provided by the seller to be accurate.

Fifth, publication bias impedes the growth of scientific knowledge and raises serious questions about whether our journals provide accurate summaries of scientific knowledge. As an example, prior to Duval (2005) and Oh et al. (2007), there was little doubt among personnel psychologists and practitioners that structured interviews were more valid than unstructured interviews. This institutional certainty was based, in part, on meta-analyses such as McDaniel et al., (1994). However, Duval (2005) and Oh et al. (2007) found evidence for publication bias in the distribution of effect sizes for structured interviews suggesting that their validity was likely much closer to that of unstructured interviews. McDaniel, McKay and Rothstein (2006) reported findings suggesting that journals studies underestimate the magnitude of race differences in job performance. To

maintain the scientific value of our literatures, we join McDaniel et al. (2006) in calling for publication bias analyses to be routinely included in any meta-analysis.

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 Cooper H, Hedges L (Eds.), *The handbook of research synthesis* (pp. 215-230). New York: Russell Sage.

Becker, B. J. (2005). The failsafe N or file-drawer number. In Rothstein, H. R., Sutton, A. J., Borenstein, M. (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments* (pp. 111-126). Chichester, UK: Wiley.

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

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

Carless, S. A., Fewings-Hall, S., Hall, M., Hay, M., Hemsworth, P. H., & Coleman, G. J. (2007). Selecting unskilled and semi-skilled blue-collar workers: The criterion-related validity of the PDI-employment inventory. *International Journal of Selection and Assessment, 15*, 335-340.

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.

Curfman, G. D., Morrissey, S., Drazen, J. M. (2005, December). Editorial expression of Concern: Bombardier et al., “Comparison of Upper Gastrointestinal Toxicity of

Rofecoxib and Naproxen in Patients with Rheumatoid Arthritis" N Engl J Med 2000;343:1520-8. New England Journal of Medicine, 353 (26), 2813-2814.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In Rothstein, H. R., Sutton, A. J., Borenstein M. (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments* (pp. 11-34). Chichester, UK: Wiley.

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

Duval, S. J., 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.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315,* 629-634.

Epredix. (2001). *PDI Employment Inventory: Technical Manual.* Minneapolis, MN: Author.

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

Hedges, L., Vevea, J. (2005). The selection model approach to publication bias. In Rothstein, H., Sutton, A. J, Borenstein, M. (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments*. Chichester, UK: Wiley.

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

McDaniel, M. A., McKay, P., Rothstein, H. R. (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., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case study of four test vendors. *Personnel Psychology, 59*, 927-953.

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.

Oh, I., Postlethwaite, B. E., Schmidt, F. L. & McDaniel, M. A. (2007). *Do structured and unstructured interviews have near equal validity?* Paper presented at the 22nd Annual Conference of the Society for Industrial and Organizational Psychology. New York.

Ones, D. S, 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.

Paajanen, G. E., Hansen, T. L. & McLellan, R. A. (1993). *PDI Employment Inventory and PDI Customer Service Inventory manual*, Minneapolis: Personnel Decisions, Inc.

Personal communication from a PreVisor™ consultant to Michael A. McDaniel, on August 6, 2007.

Personnel Decisions International (undated). *Building successful organizations*. Minneapolis: Author. Available at http://www.personneldecisions.com.sg/press/pdf/brochure_corporate.pdf

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

Society of Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures (4th ed.).* Bowling Green, OH: Author.

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In Rothstein, H., Sutton, A. J., Borenstein, M. (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments* (pp. 99-110). Chichester, UK: Wiley.

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H.R. Rothstein, A.J. Sutton, and M. Borenstein (Eds), In Rothstein, H., Sutton, A. J., Borenstein, M. (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments.* (pp. 75-98). Chichester, UK: Wiley.

Sutton, A. J., & Pigott, T. D. (2005). Bias in meta-analysis induced by incompletely reported studies. In Rothstein H, Sutton AJ, Borenstein M (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments* (pp. 223-240). Chichester, UK: Wiley.

Figure 1: Symmetrical and Asymmetrical Funnel Plots: (a) Symmetrical, (b) Asymmetrical, and (c) Asymmetrical funnel plot with imputed studies. Reproduced with permission from McDaniel et al. (2006).

Figure 1a. Symmetrical funnel plot

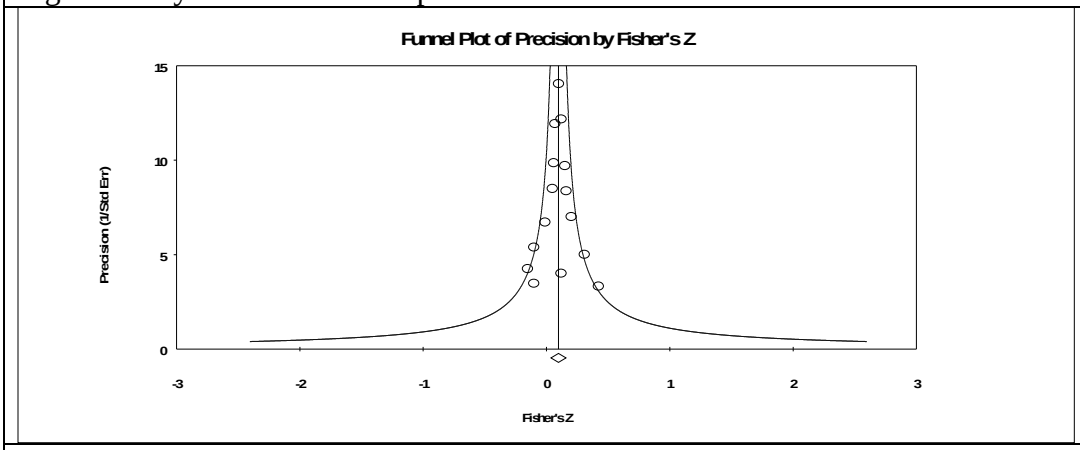

Figure 1b. Non-Symmetrical funnel plot

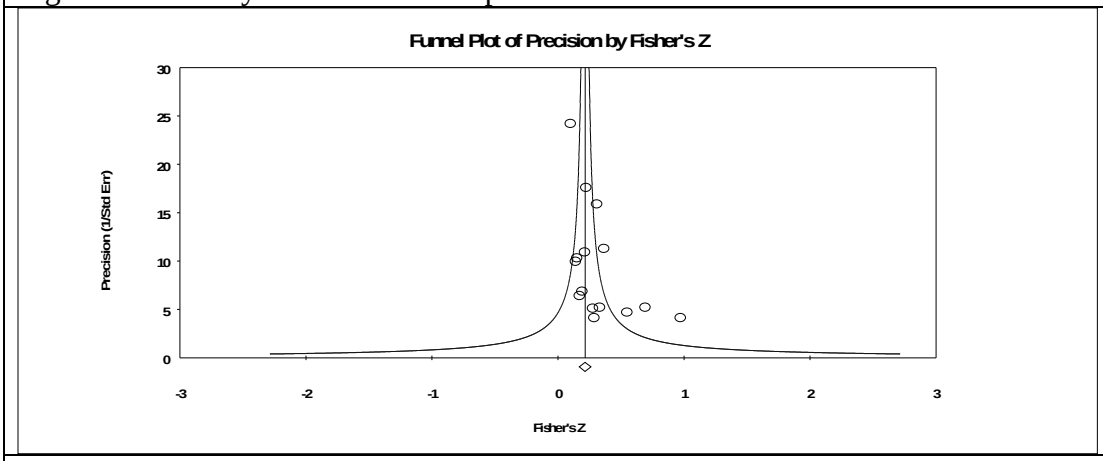

Figure 1c. Non-Symmetrical funnel plot with imputed studies

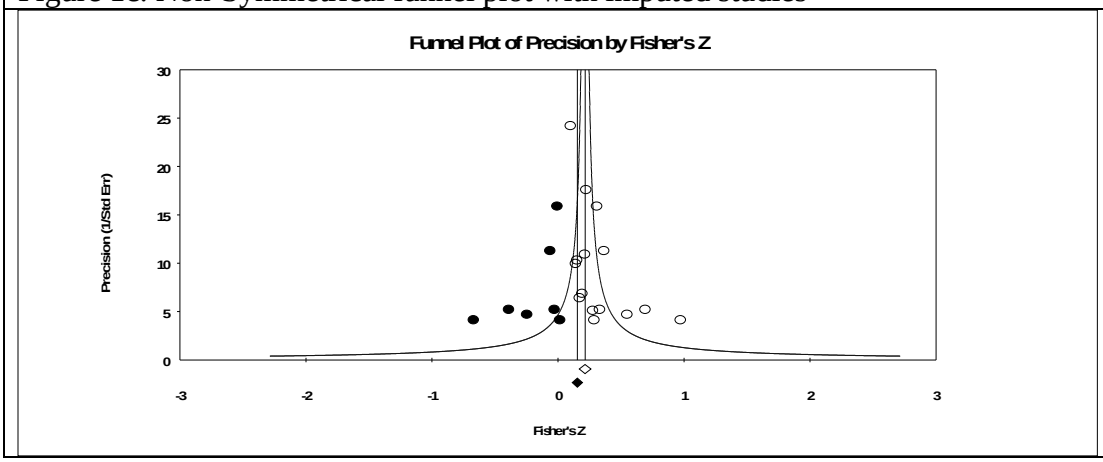

Figure 2. Observed validity data for the dependable job behavior criterion.

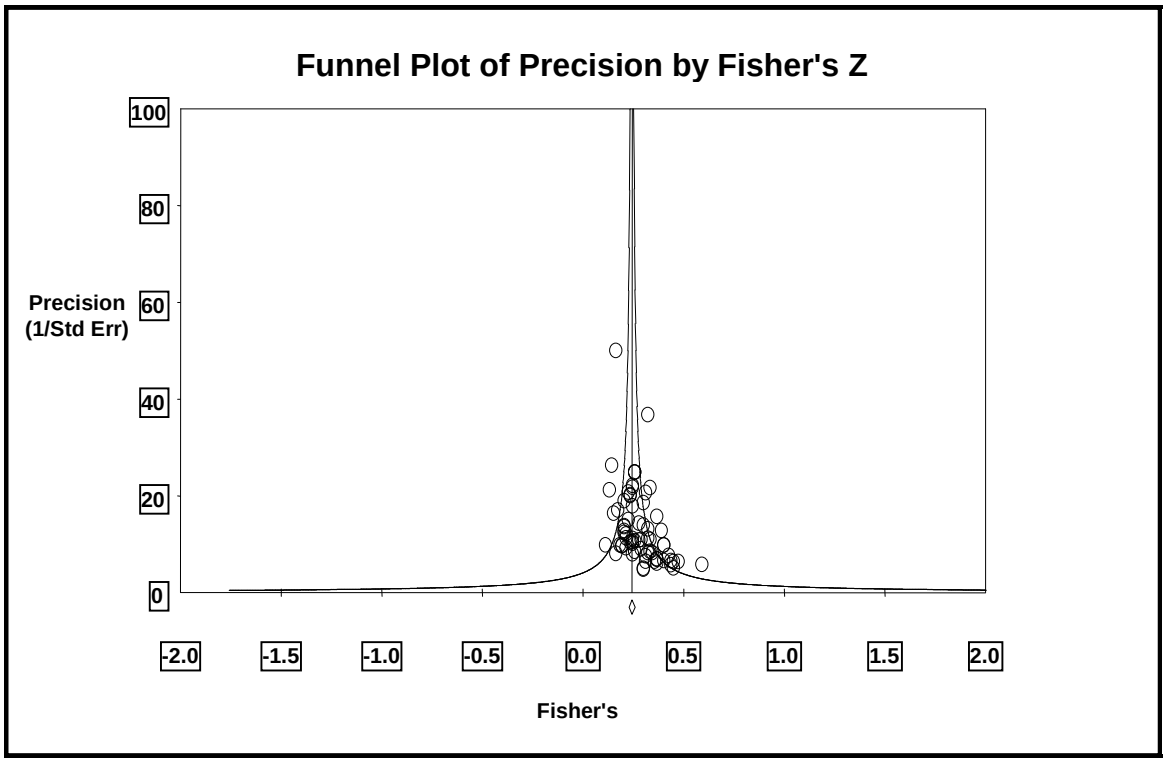

Figure 3. Observed and imputed validity data for the dependable job behavior criterion.

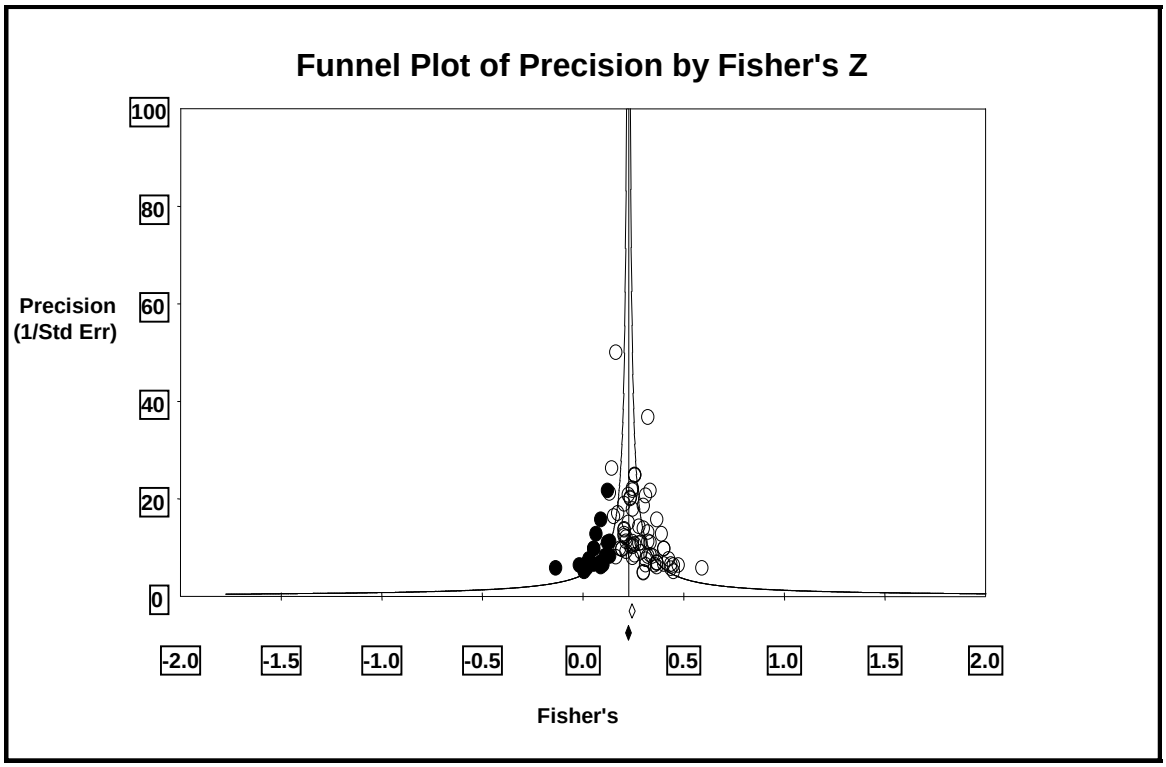

Figure 4. Observed validity data for the rehireability criterion.

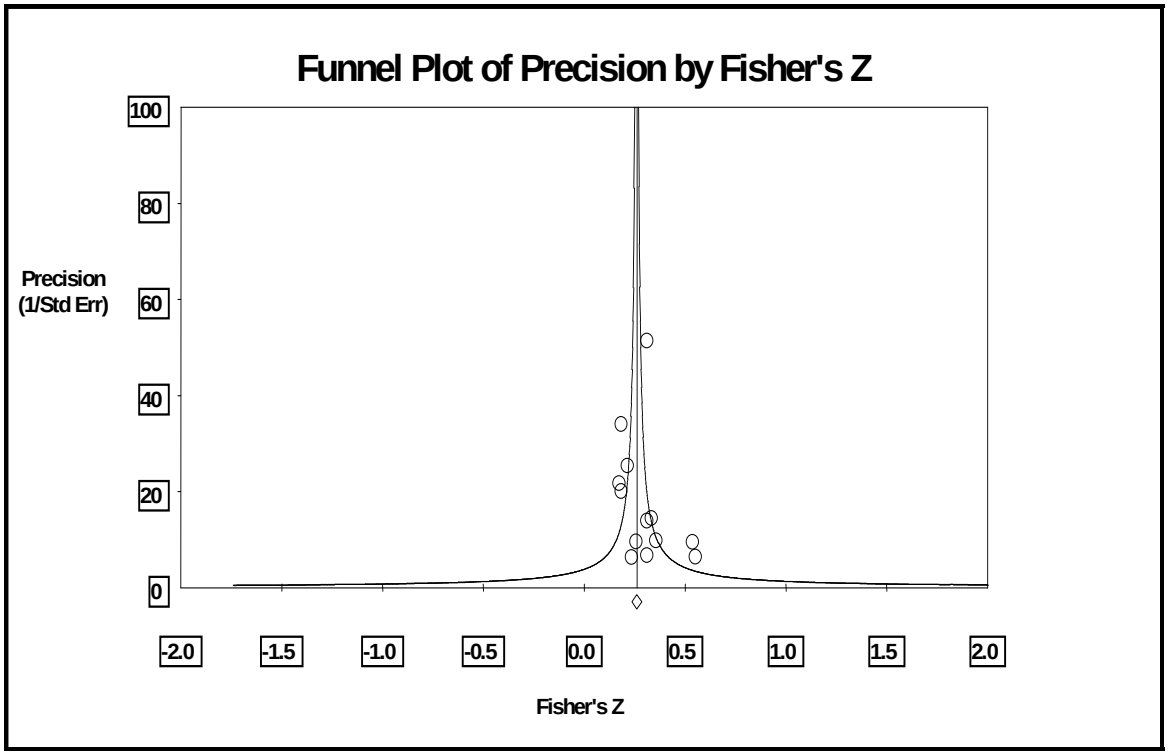

Figure 5. Observed and imputed validity data for the rehireability criterion.

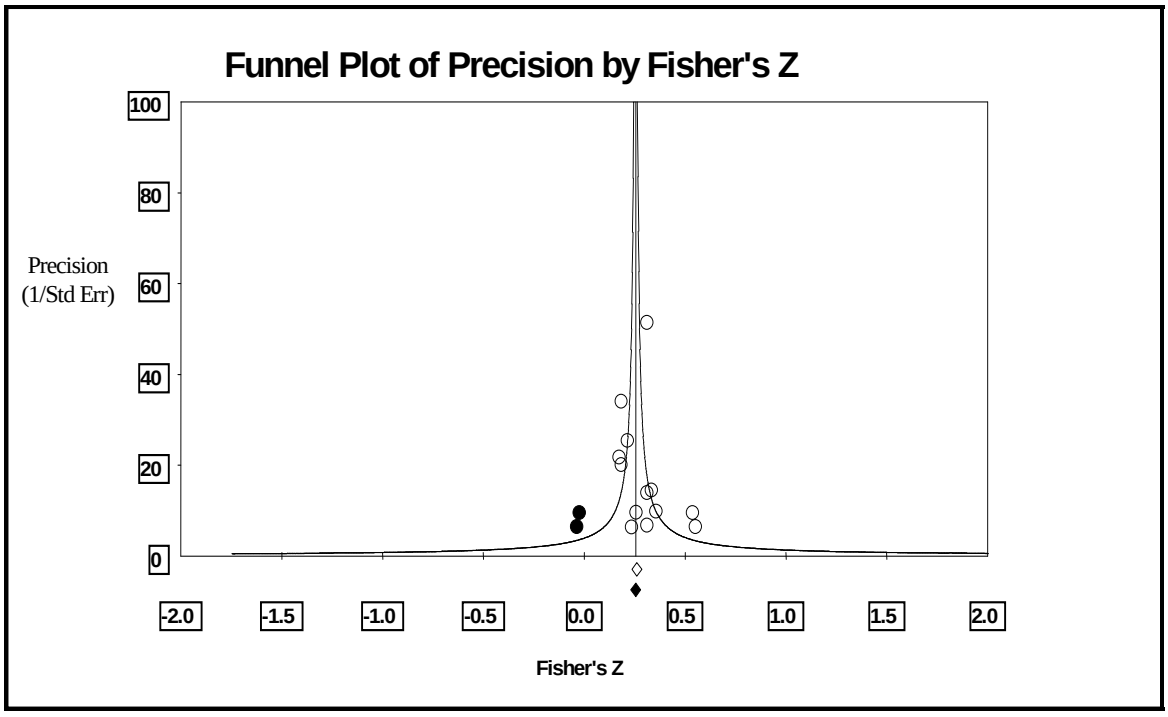

Figure 6. Observed validity data for the termination code criterion.

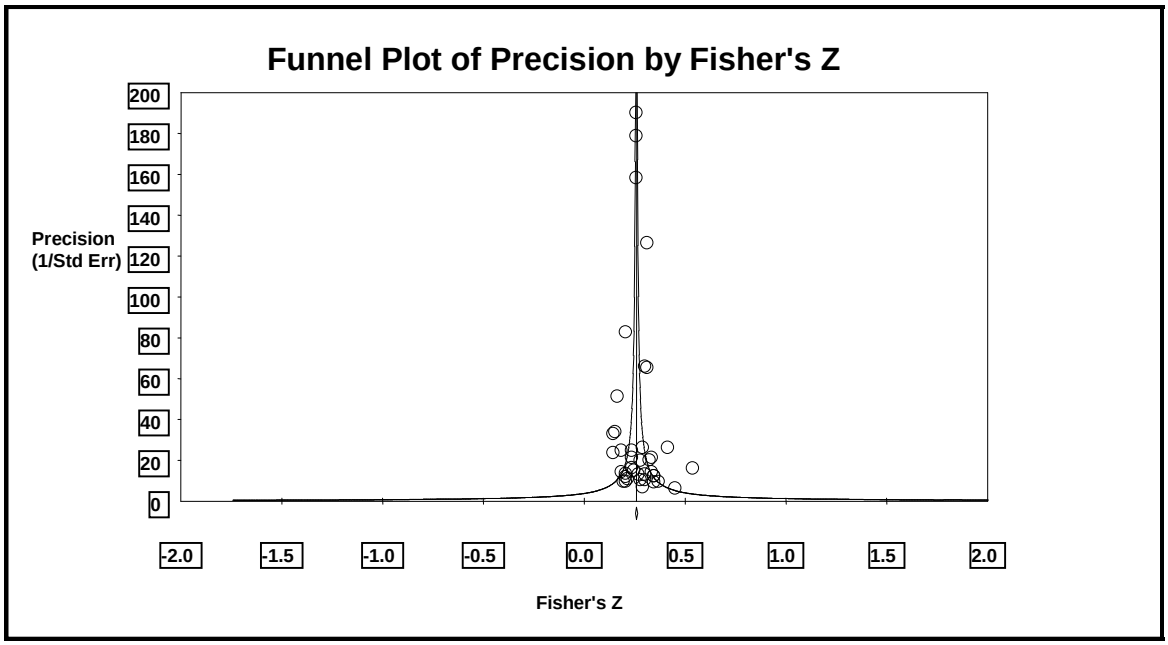

Figure 7. Observed and imputed validity data for the termination code criterion.