Journal of Applied Psychology
2012, Vol. 97, No. 3, 499–530

0021-9010/12/$12.00 DOI: 10.1037/a0021196

The Criterion-Related Validity of Integrity Tests: An Updated Meta-Analysis

Chad H. Van Iddekinge
Florida State University

Philip L. Roth and Patrick H. Raymark
Clemson University

Heather N. Odle-Dusseau
Gettysburg College

Integrity tests have become a prominent predictor within the selection literature over the past few decades. However, some researchers have expressed concerns about the criterion-related validity evidence for such tests because of a perceived lack of methodological rigor within this literature, as well as a heavy reliance on unpublished data from test publishers. In response to these concerns, we meta-analyzed 104 studies (representing 134 independent samples), which were authored by a similar proportion of test publishers and non-publishers, whose conduct was consistent with professional standards for test validation, and whose results were relevant to the validity of integrity-specific scales for predicting individual work behavior. Overall mean observed validity estimates and validity estimates corrected for unreliability in the criterion (respectively) were .12 and .15 for job performance, .13 and .16 for training performance, .26 and .32 for counterproductive work behavior, and .07 and .09 for turnover. Although data on restriction of range were sparse, illustrative corrections for indirect range restriction did increase validities slightly (e.g., from .15 to .18 for job performance). Several variables appeared to moderate relations between integrity tests and the criteria. For example, corrected validities for job performance criteria were larger when based on studies authored by integrity test publishers (.27) than when based on studies from non-publishers (.12). In addition, corrected validities for counterproductive work behavior criteria were larger when based on self-reports (.42) than when based on other-reports (.11) or employee records (.15).

Keywords: integrity, honesty, personnel selection, test validity, counterproductive work behavior

In recent years, integrity tests have become a prominent predictor within the selection literature. Use of such tests is thought to offer several advantages for selection, including criterion-related validity for predicting a variety of criteria (Ones, Viswesvaran, & Schmidt, 1993) and small subgroup differences (Ones & Viswesvaran, 1998). Researchers also have estimated that across a range of selection procedures, integrity tests may provide the largest amount of incremental validity beyond cognitive ability tests (Schmidt & Hunter, 1998). Furthermore, relative to some types of selection procedures (e.g., structured interviews, work sample tests), integrity tests tend to be cost effective and easy to administer and score.

Several meta-analyses and quantitative-oriented reviews have provided the foundation for the generally favorable view of the criterion-related validity of integrity tests (e.g., J. Hogan & Hogan, 1989; Inwald, Hurwitz, & Kaufman, 1991; Kpo, 1984; McDaniel & Jones, 1988; Ones et al., 1993). Ones et al. (1993) conducted the most thorough and comprehensive review of the literature. Their meta-analysis revealed correlations (corrected for predictor range restriction and criterion unreliability) of .34 and .47 between integrity tests and measures of job performance and counterproductive work behavior (CWB), respectively. These researchers also found support for several moderators of integrity test validity. For instance, validity estimates for job performance criteria were somewhat larger in applicant samples than in incumbent samples.

Chad H. Van Iddekinge, College of Business, Florida State University; Philip L. Roth, Department of Management, Clemson University; Patrick H. Raymark, Department of Psychology, Clemson University; Heather N. Odle-Dusseau, Department of Management, Gettysburg College.

An earlier version of this article was presented at the 70th Annual Meeting of the Academy of Management, Montreal, Quebec, Canada, August 2010. We gratefully acknowledge the many researchers and test publishers who provided unpublished primary studies for possible inclusion in this meta-analysis. This study would not have been possible without their assistance. We are particularly grateful to Linda Goldinger (Creative Learning, Atlanta, Georgia), Matt Lemming and Jeff Foster (Hogan Assessment Systems, Tulsa, Oklahoma), and Kathy Tuzinski and Mike Fetzer (PreVisor, Minneapolis, Minnesota) for helping us locate some of the older unpublished work in this area, and to Saul Fine (Midot, Israel) and Bernd Marcus (University of Hagen, Germany) for providing unpublished data on some newer integrity tests. Finally, we thank Huy Le for his guidance concerning several technical issues, and Mike McDaniel for his helpful comments on an earlier version of the article.

Correspondence concerning this article should be addressed to Chad H. Van Iddekinge, College of Business, Florida State University, 821 Academic Way, P.O. Box 3061110, Tallahassee, FL 32306-1110. E-mail: cvanidde@fsu.edu

Several variables also appeared to moderate relations between integrity tests and CWB criteria, such that validity estimates were larger for overt tests, incumbent samples, concurrent designs, self-reported deviance, theft-related criteria, and high-complexity jobs. The work of Ones et al. is highly impressive in both scope and sophistication.

Despite these positive results, some researchers have been concerned that the majority of validity evidence for integrity tests comes from unpublished studies conducted by the firms who develop and market the tests (e.g., Camara & Schneider, 1994, 1995; Dalton & Metzger, 1993; Karren & Zacharias, 2007; Lilienfeld, 1993; McDaniel, Rothstein, & Whetzel, 2006; Morgeson et al., 2007; Sackett & Wanek, 1996). For example, conclusions from several reviews of particular integrity tests (e.g., J. Hogan & Hogan, 1989; Inwald et al., 1991), or of the broader integrity literature (e.g., Sackett, Burris, & Callahan, 1989), have been based primarily or solely on test-publisher-sponsored research. The same holds for meta-analytic investigations of integrity test criterion-related validity. For instance, only 10% of the studies in Ones et al.'s (1993) meta-analysis were published in professional journals (p. 696), and all the studies cumulated by McDaniel and Jones (1988) were authored by test publishers. This situation has led to two main concerns.

First, questions have been raised about the methodological quality of some of this unpublished test publisher research. For instance, during the 1980s, when there was great interest in the integrity test industry to publish its work, very few studies submitted for publication at leading journals were accepted because of their poor quality (Morgeson et al., 2007). Various methodological issues have been noted about these studies (e.g., Lilienfeld, 1993; McDaniel & Jones, 1988; Sackett et al., 1989), including an overreliance on self-report criterion measures, selective reporting of statistically significant results, and potentially problematic sampling techniques (e.g., use of "extreme groups"). Such issues have prompted some researchers to note that "gathering all of these low quality unpublished studies and conducting a meta-analysis does not erase their limitations. We have simply summarized a lot of low quality studies" (Morgeson et al., 2007, p. 707).

The second concern is that test publishers have a vested interest in the validity of their tests. As Michael Campion noted, "my concern is not the 'file drawer' problem (i.e., studies that are written but never published). I believe that non-supportive results were never even documented" (Morgeson et al., 2007, p. 707). Karren and Zacharias (2007) reached a similar conclusion in their review of the integrity literature, stating that "since it is in the self-interest of the test publishers not to provide negative evidence against their own tests, it is likely that the reported coefficients are an overestimate of the tests' validity" (p. 223).

Concerns over test-publisher-authored research in the integrity test literature resemble concerns over research conducted by for-profit organizations in the medical research literature. The main concern in this literature has been conflicts of interest that may occur when for-profit organizations (e.g., drug companies) conduct studies to test the efficacy of the drugs, treatments, or surgical techniques they produce. Several recent meta-analyses have addressed whether for-profit and non-profit studies produce different results (e.g., Bekelman, Li, & Gross, 2003; Bhandari et al., 2004; Kjaergard & Als-Nielsen, 2002; Ridker & Torres, 2006; Wahlbeck & Adams, 1999). The findings of this work consistently suggest that studies funded or conducted by for-profit organizations tend to report more favorable results than do studies funded or conducted by non-profit organizations (e.g., government agencies). Research of this type also may provide insights regarding validity evidence reported by researchers with and without vested interests in integrity tests.

Present Study

The aim of the current study was to reconsider the criterion-related validity of integrity tests, which we did in three main ways. First, questions have been raised about the lack of methodological rigor within the integrity test literature. This is of particular concern because several of the noted methods issues are likely to result in inflated estimates of validity. These include design features, such as contrasted groups and extreme groups, and data analysis features, such as stepwise multiple regression and the reporting of statistically significant results only. We address these issues by carefully reviewing each primary study and then meta-analyzing only studies whose design, conduct, and analyses are consistent with professional standards for test validation (e.g., Society for Industrial and Organizational Psychology [SIOP], 2003). This approach is in line with calls for meta-analysts to devote greater thought to the primary studies included in their research (e.g., Berry, Sackett, & Landers, 2007; Bobko & Stone-Romero, 1998).

Second, the results of prior meta-analyses primarily are based on test-publisher research, and there are unanswered questions concerning potential conflicts of interest and the comparability of publisher and non-publisher research results (Sackett & Wanek, 1996). However, such concerns largely are based on anecdotal evidence rather than on empirical data. We address this issue by examining whether author affiliation (i.e., test publishers vs. non-publishers) moderates the validity of integrity tests.

Finally, almost 20 years have passed since Ones et al.'s (1993) comprehensive meta-analysis. We do not attempt to replicate this or other previous reviews, but rather to examine the validity evidence for integrity tests from a different perspective. For example, whereas prior reviews have incorporated primary studies that used a wide variety of samples, designs, and variables, our results are based on studies that met a somewhat more focused set of inclusion criteria (which we describe in the Method section). Further, in addition to job performance and CWB, we investigate relations between integrity tests and two criteria that to our knowledge have not yet been cumulated individually: training performance and turnover. We also investigate the potential role of several previously unexplored moderators, including author affiliation (i.e., test publishers vs. non-publishers), type of job performance (i.e., task vs. contextual performance), and type of turnover (i.e., voluntary vs. involuntary). Finally, we incorporate results of integrity test research that has been conducted since the early 1990s.

We believe the results of the present research have important implications for research and practice. From a practice perspective, practitioners may use meta-analytic findings to guide their decisions about which selection procedures—among the wide variety of procedures that exist—to use or to recommend to managers and clients. Accurate meta-analytic evidence may be particularly important for practitioners who are unable to conduct local validation

studies (e.g., due to limited resources, small sample jobs, or lack of good criterion measures) and, thus, may rely more heavily on cumulative research to identify, and help justify the use of, selection procedures than practitioners who do not have such constraints. For instance, if meta-analytic evidence suggests a selection procedure has lower criterion-related validity than actually is the case, then practitioners may neglect a procedure that could be effective and, in turn, end up with a less optimal selection system (Schmidt, Hunter, McKenzie, & Muldrow, 1979). On the other hand, if meta-analytic evidence suggests a selection procedure has higher criterion-related validity than actually is the case, this could lead practitioners to incorporate the procedure into their selection systems. This, in turn, could diminish the organization's ability to identify high-potential employees and possibly jeopardize the defensibility of decisions made on the basis of the selection process.

Professional organizations devoted to personnel selection and human resources management also use meta-analytic findings as a basis for the assessment and selection information they provide their membership and the general public. For example, materials from organizations such as SIOP and the U.S. Office of Personnel Management (OPM) describe various selection procedures with respect to factors such as validity, subgroup differences, applicant reactions, and cost. Both SIOP and OPM indicate criterion-related validity as a key benefit of integrity tests. For instance, OPM's Personnel Assessment and Selection Resource Center website states that "integrity tests have been shown to be valid predictors of overall job performance as well as many counterproductive behaviors . . . The use of integrity tests in combination with cognitive ability can substantially enhance the prediction of overall job performance" (http://apps.opm.gov/ADT).

Meta-analytic evidence also can play an important role in legal cases involving employee selection and promotion. For instance, in addition to the use of meta-analyses to identify and defend the use of the selection procedures, expert witnesses may rely heavily on meta-analytic findings when testifying about what is known from the scientific literature concerning a particular selection procedure.

Lastly, a clear understanding of integrity test validity has implications for selection research. For one, results of meta-analyses can influence the direction of future primary studies in a particular area. As McDaniel et al. (2006, p. 947) noted, "meta-analytic studies have a substantial impact as judged by citation rates, and researchers and practitioners often rely on meta-analytic results as the final word on research questions"; meta-analysis may "suppress new research in an area if there is a perception that the meta-analysis has largely settled all the research questions." Meta-analysis also can highlight issues that remain unresolved and thereby influence the agenda for future research.

Second, meta-analytic values frequently are used as input for other studies. For example, criterion-related validity estimates from integrity meta-analyses (e.g., Ones et al., 1993) have been used in meta-analytic correlation matrices to estimate incremental validity beyond cognitive ability tests (e.g., Schmidt & Hunter, 1998) and in simulation studies to examine the predicted performance or adverse impact associated with different selection procedures (e.g., Finch, Edwards, & Wallace, 2009). Thus, the validity of inferences drawn from the results of such studies hinges, in part, on the accuracy of meta-analytic values that serve as input for analysis.

In sum, results of the present meta-analysis address questions and concerns about integrity tests that have been debated for years, but until now have not been systematically investigated. This study also incorporates the results of almost 20 years of additional integrity test data that have not been cumulated. We believe the end result is a better understanding of integrity test validity, which is vital to both practitioners and researchers involved in personnel selection. Before we describe the method of our study, we discuss the basis for the potential moderator variables we examine.

Potential Moderators of Integrity Test Validity

Type of Integrity Test

The first potential moderator we examine is type of integrity test. Integrity tests can be either *overt* or *personality-based* (Sackett et al., 1989). Overt or "clear-purpose" tests ask respondents directly about integrity-related attitudes and past dishonest behaviors. Conversely, personality-based or "disguised-purpose" tests are designed to measure a broader range of constructs thought to be precursors of dishonesty, including social conformity, impulse control, risk-taking, and trouble with authority (Wanek, Sackett, & Ones, 2003).

Two theoretical perspectives provide a basis for expecting test type to moderate relations between test scores and CWB criteria. According to the theory of planned action (Ajzen, 1991; Ajzen & Fishbein, 2005), the most immediate precursor of behavior is one's intentions to engage in the behavior. This theory also specifies three main determinants of intentions: attitudes toward the behavior, subjective norms regarding the behavior, and perceived control over engaging in the behavior. The second perspective is the theory of behavioral consistency (Wernimont & Campbell, 1968), which is based on the premise that past behavior is a good predictor of future behavior. More specifically, the more a predictor measure samples behaviors that are reflected in the criterion measure, the stronger the relationship between the two measures should be.

Most overt integrity tests focus on measuring attitudes, intentions, and past behaviors related to dishonesty. For example, such tests ask respondents to indicate their views about dishonesty, such as their acceptance of common rationalizations for dishonest behavior (i.e., attitudes), their perceptions regarding the ease of behaviors such as theft (i.e., perceived control), and their beliefs about the prevalence of dishonesty (i.e., subjective norms) and how wrongdoers should be punished (Wanek et al., 2003). Further, many overt tests ask respondents to report past dishonest behaviors, such as overcharging customers and stealing cash or merchandise (i.e., behavior consistency). Thus, on the basis of the theories of planned action and behavioral consistency, people who have more positive attitudes about dishonesty, who believe that most people are somewhat dishonest, and who have engaged in dishonest behaviors in the past, should be more likely to behave dishonestly in the future.

In contrast, personality-based integrity tests primarily focus on personality-related traits, such as social conformity and risk-taking. Although potentially relevant to CWB, such traits are more distal to actual behavior than are the attitudes, intentions, and behaviors on which overt tests tend to focus. This leads to our first hypothesis:

> *Hypothesis 1:* There will be a stronger relationship between overt integrity tests and CWB than between personality-based integrity tests and CWB.

We also investigate whether test type moderates relations between integrity tests and measures of job performance.[1] Scores on overt tests may relate to performance because supervisors and peers consider CWB (which such tests were designed to predict) when forming an overall evaluation of an employee's performance (Rotundo & Sackett, 2002). Scores on personality-based tests may relate to performance because some of the traits these tests measure are relevant to performance in certain types of jobs. For example, some personality-based tests assess elements of conscientiousness, such as rule abidance, orderliness, and achievement orientation (Wanek et al., 2003). However, we are not aware of a compelling theoretical basis to predict that either type of test will be strongly related to job performance (particularly to task-related performance), or to predict that one test will be a better predictor of performance than will the other. Thus, we explore test type as a potential moderator of validity with respect to performance criteria.

> *Research Question 1:* Does type of integrity test (overt vs. personality-based) moderate relations between test scores and job performance?

Study Design and Sample

The next two potential moderators we examine are study design (i.e., predictive vs. concurrent) and study sample (i.e., applicants vs. incumbents), which typically are concomitant within the selection literature (Van Iddekinge & Ployhart, 2008). We expect to find higher validity estimates in concurrent designs than in predictive designs because in concurrent studies, respondents complete an integrity test and a self-report CWB measure at the same time. As a result, relations between scores on the two measures are susceptible to common method factors, such as transient mood state and measurement context effects (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). In contrast, predictive designs are less susceptible to such influences, because completion of the integrity test and CWB measure are separated by time, context, and so forth.

Another reason why we expect to find larger validity estimates in concurrent designs concerns the potential for the predictor and criterion in these studies to assess the same behavioral events. For example, many integrity tests (particularly overt tests but also some personality-based tests) ask respondents to report dishonest or counterproductive behaviors they have displayed recently at work. In a concurrent design, participants are then asked to complete a self-report measure of work-related CWB. Thus, the two measures may ask the respondent about the same types of behaviors but using different questions. In fact, some have suggested that correlations between overt integrity tests and self-reported CWB are more like alternate form or test–retest reliability estimates than like criterion-related validity estimates (e.g., Morgeson et al., 2007; Sackett & Wanek, 1996).

This same logic also may apply to other criteria used to validate integrity tests, such as employee records of CWB and ratings of job performance. If an integrity test asks respondents to report CWB they recently demonstrated, and then test scores are related to employee records that reflect the same instances of this CWB (e.g., of theft, absenteeism, insubordination), then relations between test scores and employee records may be stronger than if the two measures were separated in time (and thus assessed different instances of behavior). Similarly, supervisors may be asked to evaluate employees' performance over the past 6 months or a year, and although these ratings may focus primarily on productive behaviors, they may (explicitly or implicitly) capture counterproductive behaviors as well. This, in turn, may result in stronger relations between integrity test scores and performance ratings than if test scores reflected employees' pre-hire attitudes and behaviors.

> *Hypothesis 2:* Criterion-related validity estimates for integrity tests will be larger in concurrent designs than in predictive designs.

We also expect to find higher validity estimates in incumbent samples than in applicant samples. Although the debate continues concerning the prevalence and effects of applicant response distortion on personality-oriented selection procedures (e.g., Morgeson et al., 2007; Ones, Dilchert, Viswesvaran, & Judge, 2007; Tett & Christiansen, 2007), meta-analytic research suggests that integrity tests, particularly overt tests, are susceptible to faking and coaching (e.g., Alliger & Dwight, 2000). Thus, to the extent that faking is more prevalent among applicants than among incumbents, lower criterion-related validities may be found in applicant samples.

Finally, a finding of stronger validity evidence for concurrent designs and incumbent samples would be consistent with the results of primary and meta-analytic studies that have examined the moderating effects of validation design or sample on other selection procedures, including personality tests (e.g., Hough, 1998), biodata inventories (e.g., Harold, McFarland, & Weekley, 2006), situational judgment tests (e.g., McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001), and employment interviews (e.g., Huffcutt, Conway, Roth, & Klehe, 2004).

> *Hypothesis 3:* Criterion-related validity estimates for integrity tests will be larger in incumbent samples than in applicant samples.

Performance Construct

In recent years, researchers have devoted increased attention to understanding the criteria used to validate selection procedures. One important trend in this area concerns the identification and testing of multidimensional models of job performance (Campbell, McCloy, Oppler, & Sager, 1993). One model that has received support partitions the performance domain into three broad dimensions: task performance, contextual or citizenship performance, and counterproductive performance or CWB (e.g., Rotundo & Sackett, 2002).[2] *Task performance* involves behaviors that are a formal part of one's job and that contribute directly to the products or services an organization provides. *Contextual performance* involves behaviors that sup-

[1] As we discuss later, CWB can be considered an aspect of job performance (e.g., Rotundo & Sackett, 2002). However, we use *job performance* to refer to "productive" performance behaviors (i.e., task and contextual behaviors) and *CWB* to refer to counterproductive behaviors.

[2] Some models also include adaptive performance, which concerns the proficiency with which individuals alter their behavior to meet the demands of the work environment (Pulakos, Arad, Donovan, & Plamondon, 2000). However, relations between integrity tests and adaptive performance have not been widely examined, and thus we do not consider this performance construct here.

port the organizational, social, and psychological context in which task behaviors are performed. Examples of citizenship behaviors include volunteering to complete tasks not formally part of one's job, persisting with extra effort and enthusiasm, helping and cooperating with coworkers, following company rules and procedures, and supporting and defending the organization (Borman & Motowidlo, 1993). Finally, *counterproductive performance* (i.e., CWB) reflects voluntary actions that violate organizational norms and threaten the well-being of the organization and/or its members (Robinson & Bennett, 1995; Sackett & Devore, 2001). Researchers have identified various types of CWB, including theft, property destruction, unsafe behavior, poor attendance, and intentional poor performance.

We expect both overt and personality-based tests will relate more strongly to CWB than to productive work behaviors that reflect task or contextual performance. Integrity tests primarily are designed to predict CWB, and as we noted, some integrity tests and CWB measures even include the same or highly similar items concerning past or current CWB. We also note that researchers have tended to measure CWB using self-reports, whereas productive work behaviors often are measured using supervisor or peer ratings. Thus, common method variance also may contribute to stronger relations between integrity tests and CWB than between integrity tests and productive work behaviors.

> *Hypothesis 4:* Criterion-related validity estimates for integrity tests will be larger for CWB than for productive work behaviors that reflect task and contextual performance.

We also explore whether integrity tests relate differently to task performance versus contextual performance. A common belief among researchers is that ability-related constructs (e.g., cognitive ability) tend to be better predictors of task performance, whereas personality-related constructs (e.g., conscientiousness) tend to be better predictors of contextual performance (e.g., Hattrup, O'Connell, & Wingate, 1998; LePine & Van Dyne, 2001; Van Scotter & Motowidlo, 1996). If true, then integrity tests—which are thought to capture personality traits, such as conscientiousness, emotional stability, and agreeableness (Ones & Viswesvaran, 2001)—may demonstrate stronger relations with contextual performance than with task performance.

However, some studies have found that personality constructs do not demonstrate notably stronger relationships with contextual behaviors than with task behaviors (e.g., Allworth & Hesketh, 1999; Hurtz & Donovan, 2000; Johnson, 2001). One possible contributing factor to this finding is that measures of task and contextual performance tend to be highly correlated (e.g., Hoffman, Blair, Meriac, & Woehr, 2007), which may make it difficult to detect differential relations between predictors and these two types of performance. Thus, although a theoretical rationale exists to expect that integrity tests will relate more strongly to contextual performance than to task performance, we might not necessarily find strong empirical support for this proposition.

> *Research Question 2:* Does job performance construct (task performance vs. contextual performance) moderate the criterion-related validity of integrity tests?

Breadth and Source of CWB Criteria

Researchers have used various types of CWB measures to validate integrity tests. One factor that differentiates CWB measures is the "breadth" of their content. Some measures are broad in scope and assess multiple types of CWB, such as theft, withdrawal, substance abuse, and violence. Other measures are narrower and assess only one type of CWB, such as theft. Ones et al. (1993) addressed this issue by comparing validity evidence for integrity tests for criteria that assessed theft only to validity evidence for broader criteria that reflected violence, withdrawal, and other CWB. Results revealed a somewhat larger corrected mean validity for theft-related criteria (.52) than for broader CWB criteria (.45).

We also examine whether criterion breadth moderates relations between integrity tests and CWB. Specifically, we compare validities for criteria that reflect multiple types of CWB to validities for criteria that reflect only one type of CWB, namely, substance abuse, theft, or withdrawal, which are among the most commonly measured CWB dimensions. Integrity tests are considered relatively broad, non-cognitive predictors (Ones & Viswesvaran, 1996). Most overt tests assess multiple aspects of integrity, including thoughts and temptations to behave dishonestly, admissions of past dishonesty, norms about dishonest behavior, beliefs about how dishonest individuals should be punished, and assessments of one's own honesty. Personality-based tests also tend to be quite broad and tap into constructs such as conscientiousness, social conformity, risk-taking, impulsivity, and trouble with authority. Because integrity tests are broad in scope, we expect they will relate more strongly to criteria that also are broad in scope than to narrower criteria.

> *Hypothesis 5:* Criterion-related validity estimates for integrity tests will be larger for broad CWB criteria than for more narrow CWB criteria.

We also investigate whether the "source" of CWB information moderates integrity test validity. Researchers have measured CWB using self-reports from applicants or employees, supervisor or peer ratings, and employee records of disruptive behavior. Integrity tests and self-report measures are subject to several of the same method factors (Podsakoff et al., 2003). For instance, because the same individuals complete both measures, relations between integrity and CWB may be influenced by common rater effects, such as social desirability, consistency motif, and mood state (e.g., negatively affectivity). Further, because studies often have had participants complete an integrity test and a self-report CWB measure on the same occasion, relations between the two also are subject to transient mood and measurement context effects. Such factors are likely to result in larger validities when the criterion data are provided by a common rater (i.e., self-reports) than when they are provided by a different rater (e.g., a supervisor) or a different source of information (e.g., company records).

> *Hypothesis 6:* Criterion-related validity estimates for integrity tests will be larger for self-reported CWB than for external measures of CWB.

Type of Turnover

Theory and research on antecedents of employee turnover have tended to focus on job attitudes, such as job satisfaction and organizational commitment (e.g., Griffeth, Hom, & Gaertner, 2000), which are difficult or impossible to assess during the selection process, as most applicants have not yet been exposed to the job or organization. Recently, however, the use of selection procedures to reduce turnover has received attention. Importantly, studies have found that personality variables, such as conscientiousness and emotional stability, may be useful for predicting turnover (e.g., Barrick & Mount, 1996; Barrick & Zimmerman, 2009).

We attempt to add to this recent stream of research by investigating relations between integrity tests and turnover. On the one hand, integrity tests may capture variance in the types of CWB that lead to involuntary turnover. They also may capture variance in personality traits related to voluntary turnover (see below). On the other hand, myriad reasons may cause an employee to leave an organization, and thus any single predictor is unlikely to account for a large portion of the variance in turnover. Further, turnover typically is difficult to predict because of low base rates, imprecise coding of reasons for turnover, and the dichotomous nature of this criterion. Thus, if a relationship between integrity tests and turnover exists, we expect it will be a modest one.

We also examine whether the type of turnover moderates integrity–turnover relations. Involuntary turnover typically results from substandard job performance or CWB, such as absenteeism, theft, or substance abuse. Thus, to the extent integrity tests are related to job performance or CWB, such tests also may predict involuntary turnover.

Because integrity tests are thought to tap into personality traits like conscientiousness and emotional stability, such tests also may predict voluntary turnover. For example, it has been suggested that conscientious individuals are more likely to believe they have a moral obligation to remain with an organization, which affects their commitment to the organization and, in turn, retention decisions (Maertz & Griffeth, 2004). Further, employees low on emotional stability are more likely to experience negative states of mind or mood, which can lead to conflict with coworkers and lack of socialization, which can lead to stress and ultimately influence decisions to leave the organization (Barrick & Zimmerman, 2009).

However, if a relationship exists between integrity tests and voluntary turnover, it would seem to be less direct, and thus more modest, than the relationship between integrity tests and involuntary turnover. In addition, voluntary turnover often is due to factors other than employees' personality and behavior, including poor work conditions, availability of better jobs, and work-life issues, such as relocation due to a spouse's job change (Griffeth et al., 2000). This leads to our next hypothesis:

> *Hypothesis 7:* Criterion-related validity estimates for integrity tests will be larger for involuntary turnover than for voluntary turnover.

Author Affiliation

As discussed, a prevalent concern about integrity tests is that test-publisher research may provide an overly optimistic view of criterion-related validity. For one, questions have been raised about methodological approaches used by some test publishers that tend to overestimate validity (e.g., extreme group designs, reporting significant results only). Second, because publishers have a vested interest in the success of their tests, questions have been raised about the possible suppression of studies that may reveal less optimistic results. The documentation of publication bias in data from some selection test publishers has served to further increase awareness of this issue (McDaniel et al., 2006). Despite this, we are not aware of any empirical evidence that supports or refutes the claim that test publishers report more positive validity evidence for integrity tests than do non-publishers.

As we noted earlier, studies in the medical literature have used meta-analysis to assess the comparability of results from for-profit versus non-profit research and have found that for-profit research tends to report more favorable results than does non-profit research (e.g., Bhandari et al., 2004; Kjaergard & Als-Nielsen, 2002; Wahlbeck & Adams, 1999). Although we are not aware of any analogous studies within the selection literature, Russell et al. (1994) examined the influence of investigator characteristics on reported criterion-related validity estimates for various personnel selection procedures. Studies whose first author was employed in private industry were associated with somewhat higher mean validities ($r = .29$) than studies whose first author was an academician ($r = .24$). Furthermore, studies conducted to address some organizational need tended to yield higher validities than studies conducted for research purposes. For example, studies conducted for legal compliance were associated with higher mean validities ($r = .33$) than studies conducted for theory testing and development ($r = .22$).

We adopted a similar approach to try to understand the potential influence of author affiliation on the reported validity evidence for integrity tests. Specifically, we estimate criterion-related validity for three separate categories of integrity studies: (a) studies authored by test publishers only, (b) studies authored by non-publishers only, and (c) studies authored by publishers and non-publishers. Additionally, among non-publisher studies, we compare validity evidence from studies authored by researchers who developed the integrity test to validity evidence from studies whose authors did not develop the test.

> *Research Question 3:* Does author affiliation (test publisher vs. non-publisher) moderate the criterion-related validity of integrity tests?

Publication Status

Finally, we examine whether published and unpublished studies on integrity tests report similar or different levels of validity evidence. Publication bias can occur when studies are more likely to be published depending on the magnitude, direction, or statistical significance of the results (Dickerson, 2005; McDaniel et al., 2006). Indeed, a common assumption is that studies that find large or statistically significant results are overrepresented in the published literature because journals have limited space and consider such results more interesting than small or non-significant results. Researchers also may contribute to this phenomenon by submitting studies with significant findings while putting studies with non-significant findings in the "file drawer" (R. Rothstein, 1979).

The only study we know of to examine publication status and integrity test validity is Ones et al. (1993), who found an observed correlation of –.02 between publication status and validity (i.e., published studies tended to report slightly larger validity estimates). We add to their analyses by reporting separate validity estimates for published and unpublished studies for each criterion in our analyses. Conventional wisdom would suggest that published integrity test studies will yield larger validity estimates. However, because a large portion of test-publisher research is unpublished, and given concerns that test-publisher studies may provide an overly optimistic view of integrity test validity, there might not be a strong association between publication status and validity in this literature.

Research Question 4: Does publication status (published vs. unpublished) moderate the criterion-related validity of integrity tests?

Method

Literature Search

We started by searching for published articles on integrity test criterion-related validity, beginning with the articles included in Ones et al.'s (1993) comprehensive meta-analysis. We then searched available electronic databases (e.g., PsycINFO, ABI/INFORM Global, ERIC) for additional studies. We searched for words such as "integrity," "honesty," and "dishonesty" appearing in study titles and abstracts. We also performed separate searches on the names of each of the approximately 30 integrity tests we identified through our research as well as the names of known integrity test developers and researchers. We then reviewed the reference sections of all of the obtained articles to identify additional publications.

Our search process for unpublished studies was much more involved. We started by attempting to obtain copies of all the unpublished papers, technical reports, and test manuals cited by Ones et al. (1993). If we could not locate or obtain a response from the original authors of a given study, we contacted other authors who have cited the study in their work to see whether they had a copy of the paper. For example, we contacted authors of several qualitative reviews and books related to integrity testing to obtain copies of the primary studies they reviewed (e.g., Goldberg, Grenier, Guion, Sechrest, & Wing, 1991; O'Bannon, Goldinger, & Appleby, 1989; U.S. Congress, Office of Technology Assessment, 1990).

We also attempted to contact all the publishers whose integrity tests were cited in Ones et al. (1993). This was challenging because many of these publishers are no longer in existence, several publishers have changed names, and some tests are now published by different companies. In addition, we identified several newer integrity tests during this process, and we attempted to contact the publishers of these tests.

We encountered a range of responses from our attempts to contact the approximately 30 test publishers we identified. Several publishers did not respond, and a few others responded but declined to participate (e.g., because of concerns about how the studies would be used). One major test publisher declined to participate after several months of discussions. Furthermore, this publisher advised us (under threat of legal recourse) that we could not use unpublished studies on their tests we obtained from other researchers, such as those who authored the qualitative reviews noted earlier (see Footnote 10). Of the publishers who responded to our inquiries and expressed interest in helping us, almost all required us to submit a formal research proposal and/or to sign a non-disclosure agreement. In the end, only two publishers provided us with more than just a few studies for potential inclusion in the meta-analysis. Our overall experience appears to be similar to some of the experiences described by past researchers who have attempted to obtain unpublished studies from integrity test publishers (e.g., Camara & Schneider, 1995; Lilienfeld, 1993; Martelli, 1988; Snyman, 1990; Woolley & Hakstian, 1993).

Finally, we took several steps to obtain additional unpublished studies. In addition to requesting newer validity studies from each of the publishers we contacted, we searched the Dissertation Abstracts database for unpublished doctoral dissertations and master's theses. We also searched electronic and hard copies of programs from the annual conventions of the Academy of Management, American Psychological Association, and SIOP. Last, we contacted numerous researchers who have published in the integrity test area for any "file drawer" or in-progress studies.

Overall, we located 324 studies that appeared relevant to the criterion-related validity of integrity tests. Of these studies, 153 were included in Ones et al.'s (1993) meta-analysis, and 171 were not. Most of the studies not included in Ones et al. were completed subsequent to their meta-analysis.

Inclusion Criteria

Our interest was to identify primary studies whose results were relevant to the criterion-related validity of integrity-specific scales for predicting individual work behavior and whose conduct was consistent with professional standards for test validation. With this in mind, we set up several criteria to foster a careful review of the available studies.

The first criterion for inclusion in the meta-analysis concerned study design. We only included studies that collected both predictor and criterion data on individual participants. We excluded studies that compared integrity test scores between a group of known deviants (e.g., prisoners) and a group of "normal" individuals (e.g., job applicants). In addition to lacking criterion data, this "contrasted group" approach can overestimate validity, because it is easier to differentiate between criminals and non-criminals than it is to differentiate among non-criminals (Coyne & Bartram, 2002; Murphy, 1995).

We also excluded studies that examined relations at the unit-level of analysis, such as how use of an integrity test for selection correlated with theft or inventory shrinkage for entire stores, rather than for individual employees. First, relations among aggregate data are not necessarily the same as relations among the underlying individual data (E. L. Thorndike, 1939). Therefore, inclusion of unit-level integrity studies could have distorted validity estimates for individual-level criteria (McDaniel & Jones, 1986). Second, most unit-level studies have used some form of time-series design, whereby changes in an outcome (e.g., theft) were measured before and after implementation of an integrity testing program. Although the results of such research are interesting and potentially valuable,

they do not provide validity estimates for individual-level outcomes.

The second inclusion criterion concerned the type of integrity test examined. We only included studies that examined one or more integrity-specific scales contained within an integrity test whose content appeared consistent with overt or personality-based integrity tests. We excluded studies that only reported scores from scales of related, but different, constructs contained within an integrity test. For instance, although some overt tests include items that assess views and behaviors concerning substance abuse, we did not include scales that focus on substance abuse only. As an example, certain versions of the Personnel Selection Inventory (Vangent, 2007) include both the Honesty Scale, which is regarded in the literature as an integrity test, and the Drug Avoidance Scale, which focuses more specifically on the use and sale of illegal drugs and is not regarded as an integrity test per se. Likewise, we excluded scales that focus solely on attitudes and behaviors about workplace safety and customer service. We also excluded studies in which an integrity scale was included in a composite of several scales measured within the same instrument, and no integrity scale-specific validity estimates were reported. In sum, we focused on integrity-specific scales only.

In addition, we excluded studies that examined organizational surveys designed to assess the integrity of existing employees, such as the Employee Attitude Inventory (London House, 1982). Such measures were designed to assess integrity-related attitudes and behaviors among an organization's current employees (e.g., employee's perceptions about the prevalence of dishonest behavior in their workplace) and are not intended for preemployment selection (Jones, 1985). Finally, we excluded a few studies that used other methods to measure integrity, such as interviews (e.g., Gerstein, Brooke, & Johnson, 1989) and situational judgment tests (e.g., Becker, 2005). Although we encourage exploration of such methods, our primary interest was to estimate the criterion-related validity of traditional integrity tests.

The third inclusion criterion concerned the type of criterion. To be included, studies had to relate integrity test scores to scores on one or more *work-related* criteria, including measures of job performance, training performance, CWB, or turnover. We excluded studies that used measures of non-work deviance as criteria, such as academic cheating, traffic violations, and shoplifting. In addition, we excluded studies in which students participated in lab studies in which their (non-work related) integrity-related attitudes or behaviors were measured in response to an experimental manipulation (e.g., whether students kept or returned an overpayment for participation; Cunningham, Wong, & Barbee, 1994). Although these types of studies and outcomes are important, we focused on validity evidence for criteria that were more directly relevant to workplace behavior. Due to longstanding concerns regarding the validity and reliability of polygraphs (e.g., Lykken, 1981; Sackett & Decker, 1979; Saxe, Dougherty, & Cross, 1985; U.S. Congress, Office of Technology Assessment, 1983), we also excluded studies that used polygrapher ratings as the criterion to validate an integrity test.

Furthermore, we excluded studies in which the criterion reflected different types of, or reasons for, turnover. For example, several studies from a particular test publisher used purportedly interval-scaled turnover measures, such as 1 = *voluntary turnover-would rehire*, 2 = *reduction in force-may rehire*, 3 = *probationary-may not rehire*, 4 = *involuntary turnover-minor offense*, and 5 = *involuntary turnover-major offense*. In addition to questions about treating these as equal intervals, such scales appear to confound turnover with performance (e.g., *voluntary turnover would rehire* vs. *reduction in force may rehire*). Finally, although we coded studies that used employee tenure as a criterion, we did not include results from these studies in our analysis of turnover because tenure and turnover are related, but different, criteria (Williams, 1990).

The fourth criterion for inclusion concerned the reporting of validity results. Each study had to describe an original study and to provide sufficient details concerning the research design, data analysis, and results. For example, we did not include secondary reports of studies from qualitative reviews (e.g., Sackett et al., 1989) and meta-analytic studies of specific integrity measures (e.g., McDaniel & Jones, 1988), as such studies tend to provide only basic results for the primary studies analyzed, such as sample size, study design (e.g., predictive vs. concurrent), and validity coefficients. Instead, as noted above, we attempted to obtain the primary studies cited in these secondary sources to judge whether the conduct and results of the original study met all the criteria described herein.

We also had to exclude many studies for which the study particulars (e.g., sampling procedures, data analysis, validity results) were not fully described, or for which the description of these elements was so unclear that we could not be reasonably confident about the resulting validity estimates.[3] This situation appears to be consistent with previous quantitative and qualitative reviews of the integrity test literature. For example, Ones et al. (1993, p. 696) noted that the test publisher technical reports included in their meta-analysis were "sketchy, often omitting important information," and O'Bannon et al. (1989, p. 70) noted that

[3] Space limitations prevent us from describing each of the studies that we excluded because of unclear reporting. However, we briefly describe three studies as examples. One study reported the results of a validation study of an integrity test across two samples. However, some validity results were reported for only one of the samples, some for both samples separately, and some for both samples combined, without explanation as to why or how the samples were (were not) combined. Further, although a particular subscale of the integrity test is designed to predict length of service of the job, only the correlation between the other subscale and turnover was reported, and the sample size on which the correlation was based did not match any of the other sample sizes reported in the article. In another study, 1,657 job applicants were hired during the study period, but the integrity test was administered to only 367 of these applicants (even though recruiters were instructed to give the test to all applicants). Then, 90-day performance evaluations could be located for only 146 of these individuals, yet the authors did not indicate the number of applicants who actually were hired. Further, the validity of integrity test scores for a subset of these individuals was only −.05, which led the researchers to remove these data points from the final sample, which comprised only 71 employees. In a third study, the authors had store managers provide job performance and turnover information for 131 subordinates, who completed an integrity test as job applicants. However, the authors indicated that only 100 of these individuals actually had been hired. Then, only 44 of these employees were included in the validation sample. Finally, no information was provided concerning the job performance criteria used to validate the test, and the resultant validity coefficients were described as both "correlations" and "weights assigned by the discriminant function."

many reports were "ambiguous, incomplete, or not detailed enough to be properly evaluated" (also see Sackett & Harris, 1984). However, before excluding such studies, we first made several attempts to contact the authors to clarify our questions or to obtain the necessary information to estimate validity. Finally, we excluded several studies that were referenced in previous integrity test criterion-related validity meta-analyses, but that did not appear to report any validity results.

The fifth criterion concerned the exclusion of studies for specific methodological issues we encountered in this literature. First, we excluded studies that reported statistically significant validity results only, as exclusion of non-significant results can lead to overestimates of validity (McDaniel & Jones, 1988). Second, and possibly related to the above, we excluded studies that collected data on multiple integrity test scales and/or multiple criterion measures, but reported validity estimates for only certain test scales or criteria and did not explain why this was done. Third, we excluded studies for which variance on the predictor, criterion, or both was artificially increased. For instance, we excluded studies that oversampled low performing employees. We also excluded studies that used an extreme-groups approach that, for example, collected job performance data on a range of employees but then used data from the top and bottom 25% of performers only to validate an integrity test. Such studies were excluded because they can produce higher correlations than if participants were selected at random (e.g., Sackett & Harris, 1984) as well as reduce the representativeness of the sample (i.e., because cases between the extremes are omitted; Butts & Ng, 2009).

Finally, we only included results based on independent samples. To help ensure this, we used Wood's (2008) method to identify (and exclude) studies in which a sample appeared to overlap with a sample from another article authored by the same researchers. When possible, we also tried to confirm apparent instances of sample overlap with the study authors.

Of the 324 studies we found, 104 (32.1%) met all the criteria. These 104 studies comprised 42 published studies and 62 unpublished studies, and a total of 134 independent samples. Table 1 shows the number and percentage of studies we had to exclude according to each inclusion criterion. Although studies were excluded for a range of reasons (and many studies could have been excluded for multiple reasons), the three most prevalent were (a) lack of details concerning the research design, data analysis, and/or results; (b) use of polygrapher ratings as validation criteria; and (c) use of contrasted group designs that compared integrity test scores between a group of known deviants (e.g., prisoners) and a group of "normal" individuals (e.g., job applicants).

Table 1
Number and Percentage of Excluded Studies by Inclusion Criterion

Outcome/inclusion criterion	*k*	%
Total studies reviewed	324	
Studies that passed all inclusion criteria	104	32.1
Studies that did not pass one or more inclusion criteria	220	67.9
1. Study design criterion		
Contrasted group design	24	9.6
Unit-level analysis	8	3.2
Time-series design	17	6.8
2. Integrity test criterion		
Predictor was not integrity-specific	14	5.6
Composite included both integrity and non-integrity scales	6	2.4
Integrity survey for existing employees	16	6.4
Alternative type of integrity measure	6	2.4
3. Validation criteria criterion		
Criteria reflected non-job related behaviors	17	6.8
Laboratory experiment	9	3.6
Polygraph as criterion	34	13.7
Criterion reflected different types/reasons for turnover	6	2.4
4. Reporting of validity results criterion		
Lack of sufficient details regarding study particulars	34	13.7
Unclear methods and/or results	8	3.2
No apparent criterion-related validity results	13	5.2
5. Methodological issues criterion		
Reported significant results only	15	6.0
Reported results for only certain predictors or criteria	4	1.6
Extreme groups/range enhancement	12	4.8
6. Independent sample criterion		
Sample overlapped with a sample from another study	6	2.4

Note. The *k*s and associated percentages for the inclusion criteria reflect the percentage of excluded studies ($k = 220$) that were excluded because of each criterion. Because many excluded studies failed to meet multiple criteria, the total *k* exceeds 220.

Coding of Primary Studies

Two of the authors coded all the studies. Both were professors with more than 10 years of research experience. We coded whether the integrity test was overt or personality-based, whether the sample comprised applicants or incumbents (including students who were employed or recently employed), and whether the integrity test and criterion data were collected using a concurrent or predictive design. We also coded whether the criterion measured task performance, contextual performance, CWB, or some combination thereof. We used definitions from the work of Borman and colleagues (e.g., Borman & Motowidlo, 1993; Coleman & Borman, 2000) to differentiate task performance from contextual performance. We categorized measures as task or contextual only when the primary authors specifically indicated such, or when we could be reasonably confident that a measure reflected primarily task (contextual) performance according to the dimension descriptions provided. Further, we categorized CWB according to the dimensions of counterproductivity and workplace deviance identified by Gruys and Sackett (e.g., Gruys & Sackett, 2003; Sackett, 2002) and Bennett and Robinson (Bennett & Robinson, 2000; Robinson & Bennett, 1995). We also coded the source of the measures, namely, self-report, other-report (i.e., supervisors or peers), or employee records.

Finally, with respect to author affiliation, the studies in our data set were authored by test publishers, non-test publishers, and a combination of publishers and non-publishers. We also noted two types of non-publishers: researchers who developed the integrity test they studied and researchers who did not develop the test. Thus, we categorized the authors of each study as follows: (a) test publishers, (b) non-publishers who developed the test, (c) non-publishers who did not develop the test, and (d) test publishers and non-publishers.

Before analyzing the data, we estimated interrater agreement on the coding of key study variables, including the validity coefficients, sample sizes, and proposed moderators. The percentage of judgments on which the two authors agreed ranged from 95% to 100% across the coded variables, with a mean 97% agreement. Instances of disagreement were resolved after discussion. In the Appendix, we provide the main codes and input values for 74 of the 104 primary studies included in the meta-analysis. Information for the remaining studies was withheld to protect the confidentiality of client-specific technical reports from certain test publishers.

Analyses

Observed validities. We implemented Hunter and Schmidt's (2004) psychometric approach to meta-analysis. We began by identifying (and/or computing) the observed validity coefficient(s) within each primary study. Most primary studies reported zero-order correlations. Instead of correlations, several studies reported means and standard deviations or frequency counts in 2×2 tables (e.g., did vs. did not receive a particular score on an integrity test, and stayed on vs. left the job). Thus, we first converted such statistics to correlation coefficients.

Studies reported a variety of validity coefficients depending on the nature and number of integrity tests and criteria examined. If studies reported a validity coefficient based on an overall integrity test score and an overall criterion score, we used that coefficient in our analyses. If studies reported validities only for subscales of a test (e.g., the Performance and Tenure scales of the PDI Employment Inventory; ePredix, 2001) and/or facets of a criterion measure (e.g., two dimensions of task performance), we estimated a unit-weighted composite validity coefficient using the predictor and criterion correlations (Schmidt & Le, 2004). If the primary authors did not report the correlations needed to estimate a composite validity, we tried to obtain correlations from the authors. In instances for which we could not obtain the necessary information to estimate a composite validity, we computed the mean validity across the predictors and/or criteria for the given study.[4] Because our interest was to estimate the validity of using a single integrity test for selection, we did not estimate a composite validity for studies that examined the validity of multiple integrity tests, as such estimates would not be comparable with those from single-test studies. In these cases, we computed the mean validity across integrity tests.

Furthermore, some test publishers used multiple items or subscales from the same integrity test to predict criteria. For instance, the Inwald Personality Inventory (Institute for Personality and Ability Testing, 2006) includes 26 separate scales, and many of the studies based on this measure used discriminant function analysis or multiple regression analysis to estimate criterion-related validity. The coefficients from such analyses reflect the combined validity of all 26 scales, which are optimally weighted to predict the outcome of interest. Because the weights assigned to each scale often are chosen on the basis of the research sample, and given the large number of predictors considered, there is a high likelihood of capitalization on chance in such situations.

Although we were somewhat hesitant to include validity estimates from these studies with the more common bivariate validity estimates that the vast majority of other studies in our dataset reported, this is the approach the developers of these integrity tests have tended to use. Because we wanted our results to reflect how integrity tests have been (are being) used, we cautiously included this small set of studies in the meta-analysis. However, we first attempted to obtain correlations among the items or subscales so that we could calculate a unit-weighted composite validity estimate. For studies for which we could not obtain predictor intercorrelations, we adjusted the reported validity estimate for shrinkage using the positive-part Pratt population *R* formula described by Shieh (2008). The resulting values estimate what validity would be if the coefficients were calculated based on the full population. We used the unit-weighted composite validities or shrinkage-adjusted validities, rather than the original validities, in our analyses.[5] The one exception is that for the author affiliation moderator analyses, we report results using validity estimates that the test publishers originally reported (which we label "reported validity" in the tables) and using validity estimates that we computed (which we label "computed validity" in the tables). That is, for the computed validity estimates, we replaced the reported validities with the corresponding composite or shrunken validities.

Corrected validities. We also report validity estimates corrected for measurement error in the criterion to estimate the operational validity of integrity tests. Supervisor or peer ratings represent the main way integrity test researchers measured job performance. A few studies also used ratings to measure CWB. Only three studies reported an estimate of interrater reliability. The mean estimate across these studies was .72. All three studies used two raters, so the estimates reflect the reliability of a performance measure based on mean ratings of two raters. Using the Spearman–Brown formula, the mean single-rater reliability was .56. This value is highly similar to interrater estimates in the mid .50s to low .60s found in other meta-analyses involving job performance (e.g., H. R. Rothstein, 1990; Viswesvaran, Schmidt, & Ones, 2005).

Our approach was to use reliability estimates from the studies within the meta-analysis whenever possible. Thus, for studies that reported an interrater estimate, we used the actual estimates in our analyses. For studies that did not report such an estimate, but did report the number of raters, we took the mean single-rater reliability estimate (.56) and used the Spearman–Brown formula to estimate reliability based on the number of individuals whose ratings contributed to the performance or CWB criterion. For studies that did not report the number of raters, we assumed a single rater and inputted the mean reliability estimate.

None of the studies in our data set reported reliability estimates for training performance, and so we had to use estimates from other research. A meta-analysis by McNatt (2000) reported 11 internal consistency reliability estimates for training exams (see Table 1, p. 317), and we calculated the mean reliability estimate to be .81. Similarly, a meta-analysis by Taylor, Russ-Eft, and Chan (2005) reported mean internal consistency estimates of .76 and .85

[4] See the Limitations and Directions for Future Research section for a discussion of the possible implications of having to use mean validities instead of composite validities in such cases.

[5] We had to use shrinkage-adjusted *R*s (instead of zero-order or unit-weighted composite correlations) for five job performance samples, one training performance sample, three CWB samples, and four turnover samples.

for training tests that measured declarative knowledge and procedural knowledge and skills, respectively (mean $\alpha = .81$). Thus, we used a reliability estimate of .81 for studies that used training exam scores or grade point average as a criterion. We could not find reliability estimates for instructor ratings of training performance. We therefore used the mean interrater reliability of .56 from the job performance criteria studies.

Self-reports are the primary way researchers measured CWB. Twenty-five studies reported an internal consistency reliability estimate for a global measure of CWB (mean $\alpha = .72$). For studies that did not report a reliability estimate, we took the available reliability estimates and the number of items on which each estimate was based and used the Spearman–Brown formula to estimate the mean reliability for a single-item measure. We then used this estimate (i.e., .21) to calculate a reliability estimate for each study based on the number of items in the criterion for that study. For a few studies that did not report a reliability estimate or the number of items within the criterion measure, we used the mean reliability estimate.

We also were able to cumulate validity evidence for three specific types of CWB: on-the-job substance abuse, theft, and withdrawal. Seven studies reported reliability estimates (alpha) for self-reported substance abuse (mean $\alpha = .50$), 12 studies reported a reliability estimate for theft (mean $\alpha = .51$), and four studies reported a reliability estimate for withdrawal (mean $\alpha = .80$). As with the global CWB measures, we used these reliability estimates and the number of items on which each estimate was based to estimate the mean reliability for a single-item measure. We then used these estimates to calculate a reliability estimate for each study based on the number of items in the criterion for that study. We again used the mean reliability estimate for studies that did not report such an estimate or the number of items within the criterion measure.

Studies also used employee records to measure job performance and CWB. No studies reported reliability estimates for such measures, so we had to draw upon results from other research. A few studies used measures of employee productivity, such as sales, error rate, and accidents. A meta-analysis by Hunter, Schmidt, and Judiesh (1990) estimated the reliability of various productivity measures. They reported a mean reliability estimate of .83 for a 1-month period. Four of the five studies in our database that used a productivity criterion reported the period of measurement. For these studies, we took the reliability estimate from Hunter et al. and used the Spearman–Brown formula to derive a reliability estimate for each study. The mean reliability estimate was .99. We used this mean reliability estimate for one study that did not report the measurement period.

Other studies used records of employee absenteeism. A meta-analysis by Ones, Viswesvaran, and Schmidt (2003; using a subset of data from their original 1993 meta-analysis) determined the test–retest reliability of absence records. They identified 79 studies from the general absenteeism literature that reported test–retest information and the period for which the absence records were kept. Using the Spearman–Brown formula, the mean test–retest estimate for a 1-month period was .17. Eight of the 10 studies in our data set that used absence records as a criterion also reported the length of measurement period. For these studies, we took the .17 estimate from Ones et al. (2003) and used the Spearman–Brown formula to derive a reliability estimate for each study. The mean estimate across the eight studies was .72. For the two studies that did not report measurement period, we used this mean reliability estimate.

In addition, a few studies used records of detected theft. Unfortunately, we could not find any data concerning the reliability of such records. For these studies, we used the reliability estimate for self-reported theft ($\alpha = .51$).

There were two criteria we did not correct for measurement error. First, a few studies used number of disciplinary actions as a criterion, but we could not find any information concerning the reliability of this type of measure. We view records of disciplinary actions as similar to records of turnover (discussed below) in that employees either were disciplined or they were not. Although some instances of discipline may fail to be recorded, it seems like this might be rare. Thus, we did not make any corrections for number of disciplinary actions.

Second, to be consistent with prior research (e.g., Griffeth et al., 2000; Harrison, Newman, & Roth, 2006; Zimmerman, 2008), we did not correct relations between integrity tests and turnover for unreliability. However, as other researchers have done (e.g., Griffeth et al., 2000; Zimmerman, 2008), we adjusted all integrity–turnover correlations to reflect a 50–50 split between "stayers" and "leavers." This correction estimates what the maximum correlation would be if there was an "optimal" turnover base rate of .50. It also controls the potential influence of turnover base rate across studies, which, if not corrected, could falsely indicate the existence of moderator effects (Zimmerman, 2008). The mean turnover base rate across primary studies was .31. One primary study did not report the turnover base rate, and thus we used the mean value for this study.

We also made corrections when integrity test or criterion scores were artificially dichotomized (Schmidt & Hunter, 1998). Some test publishers transformed test scores into a dichotomous variable that reflected whether participants achieved a particular cutoff score, and other publishers dichotomized scores on the criteria (e.g., three or fewer absences vs. more than three absences). This practice can alter relations between the dichotomized variable and other variables depending on where in the score distribution the cut is made. For example, a cutoff on the integrity test could be selected that maximizes validity in the current sample, or a cutoff could be set a priori based on previous research with the integrity test or criterion measures (Sackett & Harris, 1984). Although publishers rarely described the rationale for use of a particular cutoff score, we decided to correct for dichotomization in such cases. For the one study that did not report information needed to correct for dichotomization in the criteria, we corrected the validity estimates to reflect a 50/50 distribution of criterion scores.

Finally, we attempted to determine the likelihood and nature of range restriction within each primary study (Berry et al., 2007).[6] Only seven studies in our sample (comprising 10 independent samples) reported the statistics necessary to estimate range restriction. Given the general lack of range restriction information in the primary studies, we chose not to correct for this artifact in our main

[6] We identified five categories of studies with respect to range restriction. The largest category (61% of our sample) comprised studies that used incumbent samples for which the authors provided limited or no information concerning how employees originally were selected. Given this, the presence and nature of range restriction in these studies could not be determined. The second largest category (22% of our sample) comprised

analyses. However, we present some illustrative analyses later in the Discussion section.

Results

We first report validity evidence for job performance criteria (and also training performance) and then for CWB criteria. We report these two sets of results separately, given that productive and counterproductive behaviors typically are considered separately in the literature (e.g., Ones et al., 1993). Furthermore, as noted, researchers have tended to use self-reports to measure CWB and supervisor or peer ratings to measure task and contextual behaviors. Thus, analyzing the results separately provides a clearer picture concerning how integrity tests relate to different criteria. We conclude by presenting validity evidence with respect to turnover.

Meta-Analysis Results for Job Performance Criteria

Overall validity evidence. Table 2 displays the meta-analytic validity estimates for integrity tests and measures of job performance. Across 74 independent samples, the overall, sample-size weighted mean observed validity was .13, and the mean validity corrected for unreliability in the criterion was .18. The 90% confidence interval (CI) for the corrected validity was .15–.20.

While reviewing studies from the publisher of a particular personality-based test, we noticed that most studies used a standard, publisher-developed criterion measure. Although the publisher referred to this as a measure of job performance, the measure also assessed CWB, including suspected theft, withdrawal behaviors (e.g., absenteeism), drug abuse, and antagonistic behaviors. Given this, we analyzed studies on this particular integrity test and compared the validity estimates based on criterion measures that appeared to reflect both productive and counterproductive behaviors ($k = 24$; $n = 3{,}127$) to validity estimates based on criterion measures that appeared to reflect productive behaviors only ($k = 18$; $n = 6{,}223$). The mean validities for these two groups of studies were .26 and .16, respectively. This finding is consistent with Hypothesis 4, which predicted that integrity tests would be more strongly related to CWB than to productive work behaviors.

We then reran the overall analysis excluding estimates based on criteria that included CWB. Some of these studies also reported a validity estimate based on a subset of ratings that reflected task or overall performance. For these studies, we replaced the validity estimates based on criteria that included CWB with validity estimates based on task or overall performance. The resulting mean observed and corrected validities were .12 and .15, respectively ($k = 63$, $n = 11{,}995$). These values may provide the best overall estimates of the relationship between integrity tests and measures of productive performance.

It has been suggested that studies that use predictive designs with applicant samples provide the best estimates of operational validity for integrity tests (e.g., Ones et al., 1993; Sackett & Wanek, 1996). Therefore, we also cumulated validity evidence for these studies separately and report the results in Table 3. There were 24 predictive-applicant studies in the data set, and the mean observed and corrected validities from these studies were .11 and .15, respectively. Considering the author affiliation moderator results discussed below, we also present validity estimates based on studies conducted by non-publishers. There were eight such studies for job performance, all of which were authored by non-publishers who did not develop the integrity test they examined. The mean observed and corrected validity estimates from these studies were .03 and .04, respectively. Excluding an influential case decreased both of the observed and corrected validity to $-.01$.[7]

Moderator analyses. Statistical artifacts accounted for 55.4% of the variance in corrected validity estimates for job performance, which indicates the possible existence of moderator variables. Beginning with our job performance-related hypotheses, Hypothesis 2 predicted larger validities for concurrent designs than for predictive designs, and Hypothesis 3 predicted larger validities for incumbent samples than for applicant samples. Although the results were consistent with Hypothesis 2, the difference in corrected validities between concurrent and predictive designs was small (.19 vs. .17). Hypothesis 3 received somewhat greater support in that corrected validities were somewhat larger for incumbent samples (.20) than for applicant samples (.15). However, the CIs around the corrected validities for these two types of samples overlapped slightly.

We also posed several research questions with respect to job performance criteria. Research Question 1 pertained to the type of integrity test. Validity estimates were somewhat larger for personality-based tests than for overt tests (.18 vs. .14), and excluding an influential case decreased the corrected validity for overt tests to .11. Research Question 2 focused on task versus contextual performance criteria. Results revealed slightly larger corrected validity estimates for task than for contextual performance (.16 vs. .14). The highly similar validity estimates for these two types of performance is consistent with other research (e.g., Hurtz & Donovan, 2000) that suggests that measures of personality-related constructs do not tend to demonstrate notably stronger relationships with contextual behaviors than with task behaviors.

studies whose samples were subject to some form of direct range restriction because the integrity test originally was used (typically along with one or more other selection procedures) to select participants. However, because few, if any, of these studies used an integrity test as the sole basis for selection, these actually represent instances of indirect rather than direct range restriction (Hunter et al., 2006). The third range restriction category (9% of our sample) comprised studies in which the authors stated that the job incumbent participants were not selected on the basis of the integrity test. Therefore, the validity estimates from these studies are subject to indirect range restriction because of a possible correlation between integrity test scores and the procedure(s) on which incumbents originally were selected. The fourth category (4%) comprised studies that used an applicant sample, but the authors did not indicate how the applicants were selected, including whether the integrity test was used in the process. Finally, there were studies (4%) in which all applicants completed both an integrity test and the criterion (i.e., a self-report measure of prior CWB); hence, the resulting validity estimates are not subject to range restriction.

[7] To identify potential influential studies, we used a modified version of the sample adjusted meta-analytic deviancy (SAMD) statistic (Beal, Corey, & Dunlap, 2002; Huffcutt & Arthur, 1995) available in Meta-Analysis Mark XIII, a Microsoft Excel-based program developed by Piers Steel. If exclusion of a study changed the original corrected validity estimate by 20% or more, we report the results with and without the influential study (Cortina, 2003).

Table 2
Meta-Analytic Estimates of Integrity Test Criterion-Related Validity for Job Performance and Training Performance

Criterion/analysis	*k*	*N*	*r*	ρ	SD_ρ	% VE	90% CI	80% CV
Job performance								
Overall	74	13,706	.13	.18	.08	55.4	.15, .20	.07, .28
Without criteria that included CWB	63	11,955	.12	.15	.07	61.1	.13, .18	.06, .25
Type of integrity test[a]								
Overt	18	2,213	.10	.14	.13	46.4	.07, .21	−.03, .30
Without influential case[b]	17	1,891	.08	.11	.11	56.0	.04, .17	−.04, .25
Personality-based	60	12,017	.14	.18	.07	64.1	.16, .21	.09, .27
Study design[c]								
Concurrent	38	4,586	.16	.19	.09	60.4	.16, .24	.07, .32
Predictive	32	8,608	.12	.17	.09	49.9	.13, .20	.06, .27
Study sample[d]								
Incumbents	47	6,191	.16	.20	.08	65.8	.17, .23	.10, .31
Applicants	24	7,104	.11	.15	.09	44.9	.11, .19	.04, .26
Performance construct								
Task performance	13	1,464	.13	.16	.00	100.0	.12, .21	.16, .16
Contextual performance	8	799	.11	.14	.00	100.0	.07, .22	.14, .14
Author affiliation								
Test publishers								
Computed validity	45	5,946	.17	.21	.09	60.3	.18, .25	.09, .33
Reported validity	45	5,946	.22	.27	.13	43.3	.23, .31	.10, .44
Non-publishers								
Overall	25	3,247	.09	.12	.10	56.1	.07, .17	−.01, .25
Developed integrity test	7	798	.15	.20	.09	66.8	.10, .29	.08, .31
Did not develop integrity test	18	2,449	.07	.10	.09	58.4	.04, .15	−.02, .22
Publishers and non-publishers	4	4,513	.12	.17	.00	100.0	.15, .18	.17, .17
Publication status								
Published studies	25	3,533	.12	.15	.10	53.4	.10, .20	.02, .28
Unpublished studies	49	10,173	.14	.18	.08	57.8	.16, .21	.09, .28
Type of criterion[e]								
Ratings of performance	73	13,517	.13	.18	.09	53.6	.15, .20	.06, .29
Productivity records	6	799	.15	.15	.06	54.7	.07, .22	.05, .25
Training performance								
Overall	8	1,530	.13	.16	.09	40.2	.08, .23	.05, .28
Grades	5	824	.20	.23	.03	91.7	.16, .29	.19, .26
Instructor ratings	3	706	.05	.06	.07	61.3	−.05, .17	−.03, .15

Note. *k* = number of validity coefficients (*k*s for some moderator categories are larger or smaller than the overall *k* due to unique design features of particular studies that comprise these categories); *r* = sample-size weighted mean observed validity estimate; ρ = validity estimate corrected for measurement error in the criterion only; SD_ρ = standard deviation of ρ; % VE = percentage of variance in ρ accounted for by sampling error and measurement error in the criterion; 90% CI = lower and upper bounds of the 90% confidence interval for ρ; 80% CV = lower and upper bounds of the 80% credibility value for ρ; CWB = counterproductive work behavior.

[a] Three studies (comprising four independent samples) reported separate validity estimates for both overt and personality-based tests. Thus, the total *k* for this moderator analysis is larger than the *k* for the overall analysis. [b] See Footnote 7 regarding identification of influential cases. [c] Results of two studies are based on a combination of concurrent and predictive designs, and two studies did not clearly specify the type of design used. These four studies were excluded from this moderator analysis. [d] Results of three studies are based on both incumbents and applicants and thus were excluded from this moderator analysis. [e] Five studies reported separate validity estimates for both performance ratings and a productivity measure. Thus, the total *k* for this moderator analysis is larger than the *k* for the overall analysis.

Research Question 3 pertained to the possible influence of author affiliation. Corrected validity estimates were larger for studies authored by test publishers (.21) than for studies authored by non-publishers (.12). However, this test-publisher estimate is based on some validities for which we computed a unit-weighted composite validity or adjusted the reported validity for shrinkage (i.e., for studies that used multiple items or subscales of an integrity test as predictors). When we used the validity estimates the test publishers originally reported, the difference between corrected validities from publishers and non-publishers was .27 versus .12. Moreover, corrected validities from studies conducted by non-publishers who developed the integrity test they examined (.20) were larger than validity estimates from non-publishers who did not develop the test (.10), although the two sets of CIs overlapped to some extent. Finally, the mean corrected validity estimate from studies authored by both test publishers and non-publishers was .17. Overall, validity evidence reported by test publishers and test developers tended to be somewhat more optimistic than validity evidence reported by non-publishers who did not develop the integrity test.

Research Question 4 explored potential validity differences between published and unpublished studies. Interestingly, published studies were associated with slightly smaller corrected validity estimates (.15) than were unpublished studies (.18). Finally, we also separated validity estimates by type of criterion measure and found slightly larger mean validities for performance ratings than for productivity measures (.18 vs. .15)

We then conducted a weighted least squares (WLS) multiple regression analysis to examine relations among the moderator variables

Table 3
Meta-Analytic Estimates of Integrity Test Criterion-Related Validity From Studies Using Predictive Designs, Applicant Samples, and Non-Self-Report Criterion Measures

Criterion/analysis	*k*	*N*	*r*	ρ	SD_ρ	% VE	90% CI	80% CV
Job performance								
Overall	24	7,104	.11	.15	.09	44.9	.11, .19	.04, .26
Non-publishers only	8	928	.03	.04	.11	54.8	−.06, .14	−.10, .19
Without influential case[a]	7	735	−.01	−.01	.08	70.5	−.10, .09	−.12, .10
Training performance								
Overall	5	962	.05	.07	.00	100.0	.01, .12	.07, .07
Non-publishers only	4	782	.06	.08	.00	100.0	.02, .14	.08, .08
CWB								
Overall	10	5,056	.09	.11	.02	76.0	.08, .14	.08, .14
Non-publishers only	2	340	.13	.13	.09	100.0	.05, .22	.12, .14
Theft[a]	3	1,481	.03	.04	.03	76.1	−.02, .11	.00, .09
Withdrawal behaviors[b]	5	5,873	.12	.15	.00	42.2	.11, .19	.09, .20
Turnover								
Overall	13	22,647	.06	.09	.05	20.1	.06, .11	.03, .15
Without influential case	12	4,652	.11	.16	.06	38.4	.12, .20	.08, .24
Non-publishers only	5	2,407	.08	.15	.07	28.5	.09, .21	.06, .24

Note. *k* = number of validity coefficients; *r* = sample-size weighted mean observed validity estimate; ρ = validity estimate corrected for measurement error in the criterion only; SD_ρ = standard deviation of ρ; % VE = percentage of variance in ρ accounted for by sampling error and measurement error in the criterion; 90% CI = lower and upper bounds of the 90% confidence interval for ρ; 80% CV = lower and upper bounds of the 80% credibility value for ρ; CWB = counterproductive work behavior.
[a] No studies in this category were conducted by non-publishers. [b] See Footnote 7 regarding identification of influential cases.

and validity estimates (see Steel & Kammeyer-Mueller, 2002). The observed validity coefficients from the 74 primary studies within this category served as the dependent variable. The moderators served as the independent variables, which we binary-coded (i.e., 0 vs. 1) to represent the two levels of each moderator (see the note to Table 4 for details regarding these codes). We also included the year the study was published as an additional (continuous) predictor. We did not include the task versus contextual performance moderator, as this distinction was examined in only a subset of the studies. Finally, we weighted each study by the inverse of the sampling error variance, such that studies with less sampling error received greater weight than studies with more sampling error (Hedges & Olkin, 1985; Steel & Kammeyer-Mueller, 2002).

Table 4 displays the moderator intercorrelations and WLS regression results. Study design, which was not a sizeable or statistically significant predictor of validity within the regression model, correlated .78 with study sample and appeared to produce multicollinearity effects when included in the model. Thus, to more clearly interpret the effects of the other moderators, we excluded this variable from the final model. In addition, one test-publisher-authored study (n = 87, r = .66) emerged as in influential case (Cook's D = 0.61 vs. a mean D of 0.02 across the remaining primary studies), and we chose to exclude this study from the final model. However, the results with and without the study sample moderator and the one influential case were not drastically different than the results reported herein.

As a group, the moderators accounted for 38% of the variance in observed validities (R = .62). Four moderators emerged as sizeable (and statistically significant) individual predictors within the regression model. Study sample was related to validity (β = .29), such that incumbent samples were associated with larger validities than applicant samples. Type of criterion was related to validity (β = .25), such that ratings of performance were associated with larger validities than objective measures. Author affiliation was related to validity (β = .38), such that validities were larger in studies authored by test publishers than by non-publishers. Finally, year of publication was related to validity (β = −.31), such that older studies tended to report larger validities than did more recent studies.

Meta-Analysis Results for Training Performance Criteria

There were eight independent samples for the relationship between integrity tests and performance during training (see Table 2). The overall mean observed validity was .13, and the mean validity corrected for unreliability in the criterion was .16 (90% CI [.08, .23]). We also separated the validity estimates by type of criterion and found that corrected validities were larger for training grades (.23) than for instructor ratings (.06). In addition, we estimated validity based on predictive studies with job applicants (see Table 3). The observed and corrected validities based on the results of these five studies were .05 and .07, respectively. The corresponding validities for the four studies authored by non-publishers (none of whom developed the integrity test) were .06 and .08. Of course, all the training performance results need to be interpreted with caution given the small number of samples on which they are based.[8]

Meta-Analysis Results for CWB Criteria

Overall validity evidence. Table 5 presents validity evidence for integrity tests and CWB. Across 65 independent samples, the mean observed validity estimate was .26, and the mean validity

[8] Given the small number of samples available for training performance, we did not examine additional potential moderators of validity for this criterion.

Table 4
Results of Weighted Least Squares Regression of Integrity Test–Job Performance Validity Estimates on Coded Moderators

Variable	1	2	3	4	5	6	7	8
Correlations among moderators and validity estimates								
1. Validity	—							
2. Type of integrity test	.08	—						
3. Study design	.10	.22*	—					
4. Study sample	.28**	.19*	.78**	—				
5. Type of criterion	.02	.10	−.34*	−.26*	—			
6. Author affiliation	.43**	−.01	−.11	−.04	−.22*	—		
7. Publication status	.01	.43*	.21*	.28**	.15	−.35**	—	
8. Year of publication	−.40**	.27*	−.01	−.11	.30**	−.52**	.37**	—

Variable	*B*	*SE*	90% CI	β	*t*
Multiple regression analysis results[a]					
Type of integrity test	.01	.04	−.06, .08	.03	0.24
Study sample	.07	.03	.02, .12	.29	2.57*
Type of criterion	.13	.06	.03, .23	.25	2.28*
Author affiliation	.12	.04	.05, .19	.38	3.04**
Publication status	.03	.04	−.04, .10	.12	0.96
Year of publication	−.01	.00	−.01, .00	−.31	−2.34*
$F(6, 61) = 6.21^{**}$, $R = .62$, $R^2 = .38$					

Note. $N = 73$ independent samples (excludes one influential sample). Both correlation and regression analyses are based on primary study results weighted by the inverse of the sampling error variance. Validity = observed validity coefficient between integrity test scores and job performance; *B* = unstandardized regression coefficient; *SE* = standard error of *B*; 90% CI = lower and upper bounds of the 90% confidence interval for *B*; β = standardized regression coefficient. Type of integrity test was coded 0 for personality and 1 for overt. Study design was coded 0 for predictive and 1 for concurrent. Study sample was coded 0 for applicants and 1 for incumbents. Type of criterion was coded 0 for objective performance measures and 1 for subjective performance measures (i.e., ratings). Author affiliation was coded 0 for non-publishers and 1 for test publishers. Publication status was coded 0 for unpublished and 1 for published.
[a] Study design was excluded from the final regression analysis because of collinearity with other predictors.
* $p < .05$. ** $p < .01$.

estimate corrected for unreliability in the criterion was .32 (90% CI [.27, .35]). These results provide additional support for Hypothesis 4, which predicted that integrity tests would be more strongly related to CWB than to productive work behaviors. Indeed, the validity estimates for CWB criteria were approximately two times larger than the overall observed and corrected validity estimates for performance criteria that did not include CWB (.12 and .15). However, as we describe below, the source of CWB criteria (i.e., self-reports vs. other-reports and employee records) had a large effect on validity. When we excluded self-reported CWB criteria, the validity evidence for job performance and CWB was more comparable. In fact, for ratings criteria, the mean corrected validity for job performance was slightly larger (.15) than the mean corrected validity for CWB (.11), although the CWB validity estimate is based on only seven independent samples (see Table 5). The mean corrected validity for employee records (.15) was the same for both job performance and CWB.

As before, we highlight evidence that may provide the best indication of validity for the operational use of integrity tests. For CWB criteria, this includes studies that used a predictive design, an applicant sample, and a non-self-report criterion (Sackett & Wanek, 1996). The mean observed and corrected validity estimates from the 10 such studies in our data set were .09 and .11, respectively (see Table 3).[9] Only two studies in this category were conducted by non-publishers (one who developed the test and one who did not), and both the observed and corrected validities from these studies were .13.

Moderator analyses. Statistical artifacts explained only 11% of the variance in CWB validities, and several factors appeared to moderate validity (see Table 5). In support of Hypotheses 1, 2, and 3 (respectively), corrected validities were larger for (a) overt tests than for personality-based tests (.38 vs. .27), (b) concurrent designs than for predictive designs (.40 vs. .13), and (c) incumbent samples than for applicant samples (.45 vs. .22).

Hypothesis 5 predicted that validity estimates would be larger for broad CWB criteria than for more narrow CWB criteria. To test this hypothesis, we compared validities for measures that assessed multiple types of CWB to validities for measures of three specific CWB for which there were a sufficient number of studies: substance abuse, theft, and withdrawal. Measures of this latter CWB assessed tardiness, leaving work early, taking long or unauthorized work breaks, or being absent from work altogether, which is consistent with how the withdrawal construct has been operationalized in the literature (e.g., Hulin, 1991; Schmitt, Cortina, Ingerick, & Wiechmann, 2003). Because these types of CWB vary in how they have been measured (e.g., all substance abuse studies used self-reports), we used only self-report criteria for this analysis to control the potential influence of criterion source.

[9] Table 3 also presents these results for the narrower criteria of theft and withdrawal, which we discuss below with respect to Hypothesis 5. No results are presented for substance abuse because all of these studies used self-report criterion measures.

Table 5
Meta-Analytic Estimates of Integrity Test Criterion-Related Validity for CWB

Analysis	*k*	*N*	*r*	ρ	SD_ρ	% VE	90% CI	80% CV
Overall	65	19,449	.26	.32	.18	11.0	.27, .35	.08, .55
Type of integrity test[a]								
Overt	43	11,751	.30	.38	.16	14.5	.33, .42	.18, .58
Personality-based	32	9,364	.23	.27	.19	9.6	.20, .32	.03, .51
Study design[b]								
Concurrent	49	13,457	.32	.40	.15	13.7	.35, .43	.20, .59
Predictive	13	5,481	.11	.13	.09	31.7	.08, .18	.02, .24
Study sample[c]								
Incumbents	45	7,047	.38	.45	.18	14.2	.39, .50	.21, .68
Applicants	16	10,802	.18	.22	.12	14.7	.18, .28	.07, .37
Breadth of criterion[d]								
Broad criterion measures	34	11,222	.35	.43	.13	13.7	.38, .47	.27, .59
Narrower criterion measures								
Substance abuse	14	5,106	.20	.25	.17	15.1	.24, .43	.03, .47
Without influential case[e]	12	3,106	.28	.40	.10	40.4	.44, .61	.27, .53
Theft	25	6,797	.23	.33	.11	30.5	.28, .38	.19, .47
Withdrawal	11	3,989	.23	.33	.10	22.2	.25, .39	.20, .46
Source of criterion[f]								
Self-reports	43	13,085	.33	.42	.13	13.1	.37, .45	.25, .59
Other-reports	7	3,645	.09	.11	.00	95.3	.08, .15	.11, .11
Employee records	17	3,420	.14	.15	.10	38.9	.11, .21	.02, .29
Author affiliation: Self-reports								
Test publishers[g]	26	10,010	.31	.39	.11	14.4	.34, .43	.25, .54
Non-publishers[h]								
Overall	14	2,218	.48	.55	.14	28.8	.50, .61	.37, .72
Developed integrity test	11	1,756	.49	.55	.15	26.2	.50, .63	.35, .75
Did not develop integrity test	8	1,466	.47	.54	.15	19.8	.45, .63	.35, .73
Publishers and non-publishers	3	857	.15	.28	.00	100.0	.20, .37	.28, .28
Author affiliation: Non-self-reports								
Test publishers								
Computed validity	16	3,174	.14	.15	.11	38.2	.10, .21	.01, .29
Reported validity	16	3,174	.21	.24	.11	31.8	.18, .29	.10, .38
Non-publishers								
Overall	4	656	.17	.17	.00	100.00	.12, .23	.17, .17
Developed integrity test	1	91	.25	.27				
Did not develop integrity test	3	565	.16	.16	.00	100.0	.10, .22	.16, .16
Publishers and non-publishers	4	3,235	.08	.10	.00	100.0	.07, .14	.10, .10
Publication status								
Published	37	6,554	.32	.35	.22	12.0	.30, .42	.07, .63
Unpublished	28	12,895	.23	.29	.15	10.9	.24, .34	.10, .48

Note. CWB = counterproductive work behavior; *k* = number of validity coefficients; *r* = sample-size weighted mean observed validity estimate; ρ = validity estimate corrected for measurement error in the criterion only; SD_ρ = standard deviation of ρ; % VE = percentage of variance in ρ accounted for by sampling error and measurement error in the criterion; 90% CI = lower and upper bounds of the 90% confidence interval for ρ; 80% CV = lower and upper bounds of the 80% credibility value for ρ.

[a] Ten studies reported separate validity estimates for both overt and personality-based tests. Thus, the total *k* for this moderator analysis is larger than the *k* for the overall analysis. [b] Results of three studies are based on a combination of concurrent and predictive designs and thus were excluded from this moderator analysis. [c] Results of four studies are based on both incumbents and applicants and thus were excluded from this moderator analysis. [d] We limited the criterion breadth analyses to self-report criteria. Observed and corrected validity estimates across all sources of criterion information (i.e., self-reports, other-reports, and employee records) were .27 and .33 for broad CWB criteria ($k = 46$, $N = 16{,}562$), .20 and .28 for theft ($k = 30$, $N = 8{,}608$), and .16 and .21 for withdrawal ($k = 24$, $N = 10{,}764$) (the values for substance abuse are the same as the tabled values because all studies used self-report criteria). [e] See Footnote 7 regarding identification of influential cases. [f] Two studies reported separate validity estimates for both self-report and other-report criteria. Thus, the total *k* for this moderator analysis is larger than the *k* for the overall analysis. [g] We did not have to compute any alternate validity estimates for test publisher studies that used self-report CWB criteria. [h] In five non-publisher samples, the researchers examined an integrity test they developed and one or more tests they did not develop, which we analyzed separately. Thus, the sum of the *k*s for the two subcategories of non-publishers is larger than the overall *k*.

The results of this analysis provide support for Hypothesis 5 in that the mean corrected validity for broad measures (.43) was larger than the corrected validities for substance abuse, theft, and withdrawal, which ranged from .25 to .33 (although excluding two influential cases increased the corrected validity for substance abuse from .25 to .40).

Hypothesis 6 predicted larger validity estimates for self-reported CWB than for external CWB measures. The results provided strong support for this hypothesis, as corrected validities were much larger when based on self-reports (.42) than when based on other-reports (.11) and employee records (.15).

Regarding author affiliation (Research Question 3), we first separated studies that used self-report criteria and non-self-report criteria, so that criterion source would not obscure potential relations between author affiliation and validity. For self-report criteria, corrected validity estimates from test publisher studies actually

were smaller than validity estimates from non-publisher studies (.39 vs. .55).[10] In contrast to the job performance results, validities from non-publishers who developed the test (.55) were comparable to validities from non-publishers who did not develop the test (.54). Further, although the mean corrected validity for studies authored by publishers and non-publishers (.28) was notably lower than the other validities, this estimate is based on only three studies and thus needs to be interpreted with caution.

For non-self-report CWB criteria, corrected validity estimates from test publisher studies (.15) were slightly smaller than validity estimates from non-publisher studies (.17). However, replacing the validities we computed with the validities publishers originally reported increased the corrected validity for test publishers from .15 to .24. Once again, corrected validities from non-publishers who developed the test (.27) were larger than validities from non-publishers who did not develop the test (.16). However, this comparison is based on just a few studies and thus should be interpreted very cautiously. Although studies authored by both test publishers and non-publishers yielded the smallest corrected validities (.10), all five of these studies used other-reported CWBs (which tend to be associated with the smallest validities), and thus these results are not readily comparable to those of the other two groups of studies. Finally, with respect to publication status (Research Question 4), corrected validities from published studies (.35) were somewhat larger than the validities from unpublished studies (.29).

As with the job performance criteria, we used WLS regression analysis to examine relations among the moderator variables and the validity estimates. The results of this analysis are shown in Table 6. As before, study design correlated highly with some of the other moderators, including .85 with criterion source, and including study design in the regression model appeared to produce multicollinearity effects. Thus, we excluded this variable from the final model.

The moderators as a group accounted for 69% of the variance in observed CWB validities ($R = .83$). Three moderators were sizeable (and statistically significant) individual predictors within the regression model. Source of criterion demonstrated the strongest relationship with validity ($\beta = .70$). Studies that used self-reported CWBs as criteria were associated with larger validity estimates than were studies that used other-ratings or information from employee records to measure counterproductivity. Consistent with the job performance regression analyses, study sample was related to validity ($\beta = .29$), such that incumbent samples were associated with larger validities than applicant samples. In addition, publication status ($\beta = .26$) was related to validity, such that published studies were associated with somewhat more positive validity evidence than were unpublished studies.

Meta-Analysis Results for Turnover

Overall validity evidence. Validity evidence for integrity tests and turnover is presented in Table 7. Across 20 independent samples, the mean observed validity was .07, and the mean validity adjusted to an "optimal" turnover base rate of 50% was .09 (90% CI [.07, .11]). However, one large-sample study ($n = 17{,}995$, $r = .05$) emerged as an influential case, and excluding it from the analysis yielded observed and adjusted validities of .11 and .15 (90% CI [.12, .18]). Five independent samples were available to estimate relations between integrity tests and tenure. The resulting mean observed validity estimate was .10. We did not adjust these validities for base rate differences because tenure was measured as continuous variable in these studies.

Table 3 displays validity evidence for turnover studies that used predictive designs with job applicants. The mean observed and adjusted validity estimates across these 13 studies were .06 and .09, respectively. Excluding the same influential case as noted above yielded validities of .11 and .16, respectively. Five of these studies were conducted by non-publishers, none of whom developed the integrity test examined. The observed and adjusted validities based on these studies were .08 and .15.

Moderator analyses. Statistical artifacts explained only 25.6% of the variance in turnover validity estimates. Hypothesis 7 predicted that validity estimates for integrity tests would be larger for involuntary turnover than for voluntary turnover. In support of this hypothesis, the base rate adjusted validity for involuntary turnover was .19, whereas the corresponding validity for voluntary turnover was .08. Consistent with the job performance and CWB criteria results, validities also were somewhat larger for incumbent samples than for applicant samples (.14 vs. .09).[11] However, these results are based on small numbers of samples, and excluding an influential study increased the corrected validity for applicant samples to .16. Further, only one study used an overt test, so we could not examine the influence of test type on validity.

Concerning author affiliation (Research Question 3), corrected validity estimates from test publisher studies (.08 and .10 for computed and reported validities, respectively) were somewhat smaller than validity estimates from non-publisher studies (.15). Nonetheless, when the influential study was excluded, the computed and reported validities for test publishers increased to .16 and .26, respectively. Lastly, with respect to publication status (Research Question 4), corrected validities from published studies were larger than validities from unpublished studies (.15 vs. .08), although the unpublished validity estimate increased to .16 when the influential study was excluded.[12]

Discussion

Integrity tests have become a prominent selection procedure over the past few decades. Use of such tests for selection often is encouraged because they are thought to predict both job performance and counterproductive work behaviors, but yield small subgroup differences. The purpose of the present study was to conduct an updated meta-analysis of the criterion-related validity of integrity tests. The key findings and their implications for research and practice are summarized below.

[10] We had nine additional CWB validity estimates from four unpublished technical reports authored by a particular test publisher. However, the publisher did not grant us permission to include these reports in our study. Inclusion of these validity estimates would have substantially increased the mean validity for the publisher authored studies.

[11] Because all turnover studies used predictive designs, we could not examine whether study design moderates integrity test–turnover relations.

[12] We did not regress the validity coefficients on the coded moderators for turnover (as we did for job performance and CWB criteria), given the smaller number of available primary studies.

Table 6
Results of Weighted Least Squares Regression of Integrity Test–CWB Validity Estimates on Coded Moderators

Variable	1	2	3	4	5	6	7	8
Correlations among moderators and validity estimates								
1. Validity	—							
2. Type of integrity test	.15	—						
3. Study design	.50**	.49**	—					
4. Study sample	.64**	−.14	.48**	—				
5. Source of criterion	.58**	.58**	.85**	.30**	—			
6. Author affiliation	−.34**	.09	−.10	−.41**	−.07	—		
7. Publication status	.43**	.06	−.07	.32**	−.01	−.54**	—	
8. Year of publication	.01	.31**	.46**	−.06	.54**	−.10	−.37**	—

Variable	*B*	*SE*	90% CI	β	*t*
Multiple regression analysis results[a]					
Type of integrity test	−.06	.04	−.13, .01	−.17	1.52
Study sample	.10	.04	.03, .17	.29	2.81**
Source of criterion	.27	.05	.19, .35	.70	5.41**
Author affiliation	−.02	.05	−.10, .06	−.04	−0.38
Publication status	.09	.04	.02, .16	.26	2.23*
Year of publication	−.00	.00	−.01, .00	−.21	−1.78
$F(6, 52) = 19.24^{**}$, $R = .83$, $R^2 = .69$					

Note. $N = 65$ independent samples. Both correlation and regression analyses are based on primary study results weighted by the inverse of the sampling error variance. CWB = counterproductive work behavior; Validity = observed validity coefficient between integrity test scores and job performance; *B* = unstandardized regression coefficient; *SE* = standard error of *B*; 90% CI = lower and upper bounds of the 90% confidence interval for *B*; β = standardized regression coefficient. Type of integrity test was coded 0 for personality-based and 1 for overt. Study design was coded 0 for predictive and 1 for concurrent. Study sample was coded 0 for applicants and 1 for incumbents. Source of criterion was coded 0 for non-self-report (i.e., ratings and employee records) and 1 for self-report. Author affiliation was coded 0 for non-publishers and 1 for test publishers. Publication status was coded 0 for unpublished and 1 for published.

[a] Study design was excluded from the final regression analysis because of collinearity with other predictors.

* $p < .05$. ** $p < .01$.

Key Findings and Implications

Current database of integrity test validity research. We located and reviewed over 300 published and unpublished studies whose results were potentially relevant to the criterion-related validity of integrity tests. A large number of these studies (about two thirds) did not meet one or more of our inclusion criteria. Many of these studies were well-conducted, but were not relevant to the validity of integrity-specific scales for predicting individual work behavior. For example, although research that has related integrity tests to general deviant behavior (e.g., academic cheating, shoplifting) is important, it is not as directly relevant to the prediction of work-specific deviance. Some of the time-series studies we reviewed also were interesting and may provide insights concerning how integrity tests can influence unit- or organizational-level outcomes. However, this type of design does not directly address the validity of such tests for individual-level outcomes, which was our primary focus.

We also reviewed many studies with potentially problematic designs and reporting, such as the use of extreme group designs and the reporting of statistically significant results only. Further, a notable portion of studies we reviewed did not provide sufficient details concerning key elements, such as how the sample was obtained and who it comprised, how the integrity test was scored, how the data were analyzed, and what the validity results represent. These findings lend support to concerns researchers have raised about some of the methodological issues within the integrity literature. It also suggests that, despite the extensive research base for integrity tests, the number of studies that provide direct and rigorous evidence regarding criterion-related validity appears to be much smaller.

Criterion-related validity evidence for job performance and training performance. For integrity tests as predictors of job performance, the overall estimated observed validity is .13, and the estimated validity corrected for unreliability in the criterion is .18. However, for studies that used criteria that focused on task, contextual, or overall job performance (and did not directly measure CWB), the observed and corrected validities are .12 and .15, respectively. Several factors appear to moderate relations between integrity tests and job performance. For example, validity estimates are larger for incumbent samples than for applicant samples, for ratings criteria than for objective criteria, and for older studies than for more recent studies (see Table 4).

Recall that we did not correct validities for predictor range restriction given that so few primary studies reported information to estimate range restriction or to determine the nature of the possible restriction (e.g., direct vs. indirect). However, to illustrate what the range restriction-corrected validity estimates may be, we took the mean job performance-specific corrected validity estimate of .15 and further corrected it for indirect range restriction using Hunter, Schmidt, and Le's (2006) correction procedure (see their Table 2, p. 603). We used a range restriction value (i.e., *u*) of .90, which we derived from 10 samples in our data set that reported

Table 7
Meta-Analytic Estimates of Integrity Test Criterion-Related Validity for Turnover and Tenure

Analysis	*k*	*N*	*r*	ρ	SD_ρ	% VE	90% CI	80% CV
Turnover								
Overall	20	24,808	.07	.09	.05	25.6	.07, .11	.03, .15
Without influential case[a]	19	6,813	.11	.15	.06	43.6	.12, .18	.08, .23
Type of turnover								
Voluntary turnover	7	17,185	.06	.08	.00	100.0	.07, .09	.08, .08
Involuntary turnover	12	8,248	.16	.19	.03	34.0	.16, .22	.12, .26
Type of integrity test								
Overt	1	140	.06	.06				
Personality-based	19	24,668	.07	.09	.05	24.9	.07, .11	.03, .15
Without influential case	18	6,673	.12	.16	.05	48.1	.13, .19	.09, .22
Study sample								
Incumbents	7	2,161	.13	.14	.00	100.0	.10, .17	.14, .14
Applicants	13	22,647	.06	.09	.05	20.1	.06, .11	.03, .15
Without influential case	12	4,652	.11	.16	.06	38.4	.12, .20	.08, .24
Author affiliation								
Test publishers								
Computed validity	13	21,857	.06	.08	.04	29.5	.06, .10	.03, .13
Without influential case	12	3,862	.13	.16	.04	67.0	.13, .19	.11, .21
Reported validity	13	21,857	.08	.10	.09	7.3	.06, .14	−.01, .21
Without influential case	12	3,862	.20	.26	.11	18.8	.20, .32	.12, .40
Non-publishers[b]	7	2,951	.09	.15	.07	33.9	.10, .20	.06, .24
Publication status								
Published	8	2,394	.11	.15	.07	37.6	.10, .20	.06, .24
Unpublished	12	22,414	.06	.08	.04	25.8	.06, .11	.03, .13
Without influential case	11	4,419	.12	.16	.04	63.9	.13, .19	.11, .20
Tenure[c]	5	875	.10		.02	90.7	.04, .15	.07, .13

Note. *k* = number of validity coefficients; *r* = sample-size weighted mean observed validity estimate; ρ = mean validity after adjusting each validity coefficient to a turnover base rate of 50% (no corrections were made for criterion unreliability); SD_ρ = standard deviation of ρ; % VE = percentage of variance in ρ accounted for by sampling error and measurement error in the criterion; 90% CI = lower and upper bounds of the 90% confidence interval for ρ; 80% CV = lower and upper bounds of the 80% credibility value for ρ.
[a] All noted influential cases represent the same study (n = 17,995; r = .05). See Footnote 7 regarding identification of such cases. [b] All studies in this category were conducted by non-publishers who did not develop the integrity test; there were no test developer-authored studies. [c] No corrections were made to the validity estimates for tenure.

restricted and unrestricted standard deviations for integrity test scores.

After correcting for indirect range restriction, the validity of .15 increased to .18. As a point of comparison, Ones et al.'s (1993) meta-analysis yielded a fully corrected validity of .34 for job performance. However, they used a mean *u* value of .81 based on 79 range restriction values. Using their *u* value would increase our validity estimate from .15 to .21. This corrected value should be interpreted very cautiously given that many of the studies from which Ones et al. derived their *u* value did not meet our inclusion criteria, and thus are not represented in the present meta-analysis.

We also highlighted studies that used predictive designs with job applicant samples, as the results of such studies are thought to provide the best estimates of operational validity. Across 24 predictive-applicant studies, the estimated corrected validity is .15. This estimate increases to .18 when we further correct for indirect range restriction. These values also are weaker than corrected validities for predictive-applicant studies reported in previous meta-analyses (e.g., .41 from Ones et al., 1993). Finally, some researchers may wish to consider validity evidence from predictive-applicant studies conducted by non-publishers. The corrected validity across eight such studies is .04 (.05 when corrected for range restriction), which decreases to −.01 when an influential case is excluded.

Regardless of the specific estimates and artifact corrections we consider, our results suggest that relations between integrity tests and measures of job performance tend to be rather weak. This suggests that integrity tests may not be as useful for selection as previously thought, particularly when predicting productive work behaviors (e.g., task or contextual performance) as a primary interest. These weaker validities also may have implications for conclusions that researchers and practitioners may draw from results of studies that have used earlier meta-analytic estimates of integrity test validity as input for analysis. For example, integrity tests may not provide the level of incremental validity beyond other selection procedures (e.g., cognitive ability tests) that previous studies have estimated. Additionally, given the present results, practitioners who would like to consider an integrity test for selection, but who are unable to conduct a local validation study (e.g., because of a small sample job), might not be able to rely heavily on meta-analytic validity evidence to help justify use of such a test.

Finally, this is the first study we know of to cumulate validity evidence for integrity tests as predictors of performance during training. The estimated observed validity for training performance is .13, and the estimated validity corrected for unreliability in the criterion is .16. When corrected for indirect range restriction, the .16 validity estimate increases to .19. Thus, the overall validity

evidence for training performance appears to be very similar to the validity evidence for job performance. Furthermore, the corrected validities for integrity tests are stronger when the criteria reflect training grades (.23) than when they reflect instructor ratings (.06). Although these results are interesting and, for example, suggest integrity tests may hold some promise for predicting training grades, we urge caution given the small number of studies ($k = 8$) that were available for these analyses.

Criterion-related validity evidence for CWB. For integrity tests as predictors of counterproductive work behaviors, the overall estimated observed validity is .26, and the estimated validity corrected for unreliability in the criterion is .32. Further correcting the corrected CWB validity estimate for indirect range restriction yields a validity of .36. For comparison, Ones et al. (1993) reported observed and corrected validities of .33 and .47 for their overall analysis of CWB criteria. Thus, the validities we found are somewhat smaller, yet are still moderately large in magnitude and suggest a relationship between integrity tests and CWB.

Several factors appear to moderate integrity test-CWB relations. For example, validity estimates are larger for incumbent samples than for applicant samples and for published studies than for unpublished studies. However, by far the strongest moderator of validity is the source of the criterion: corrected validity estimates are notably larger when CWB is measured using self-reports (.42) than when it is measured using other-reports (.11) or employee records (.15). Thus, whether one accepts self-report measures as appropriate criteria appears to be critical for interpreting the validity evidence for integrity tests and CWB. Other method factors also may be relevant when participants fill out an integrity test and a self-report criterion measure on the same occasion and when both the test and the criterion capture similar behaviors. In this regard, it has been suggested that the most relevant validity evidence for integrity tests and CWB comes from studies that use predictive designs, applicant samples, and non-self-report criteria (although non-self-report criteria also have limitations, such as the fact that some degree of employee deviance goes undetected; Sackett et al., 1989). The estimated corrected validity from the 10 such samples in our dataset is .11, which increases to .13 when correct for indirect range restriction.

Criterion-related validity evidence for turnover. To our knowledge, the present study is the first to cumulate relations between integrity tests and turnover. The mean overall observed validity estimate for this relationship is .07 with an influential case and .11 without this case. If there is a "statistically optimal" 50–50 turnover base rate within each primary study, the corresponding validity estimates are .09 (with the influential case) and .15 (without the influential case). Further correcting the values of .09 and .15 for indirect range restriction yields corrected validity estimates of .11 and .18, respectively.

The modest number of turnover studies, as well as a large-sample influential case, made it difficult to assess potential moderators of criterion-related validity. That being said, the mean corrected validity for involuntary turnover (.19) was notably larger than the corrected validity for voluntary turnover (.08). Thus, although relations between integrity tests and turnover generally are weak, such tests may hold some promise for predicting involuntary turnover.

Test publisher versus non-publisher research. Although questions about test-publisher-sponsored research are longstanding in the integrity literature (e.g., Goldberg et al., 1991; O'Bannon et al., 1989; Sackett et al., 1989), until now, there have been no direct quantitative comparisons of test publisher and non-publisher research results. For studies in which the criteria reflected job performance, test publishers consistently reported more positive validity estimates than non-publishers. For example, when we use the validity estimates that test publishers originally reported (vs. the validity estimates that we computed based on information from those studies), the mean validity (corrected for criterion unreliability) is over two times larger than the corresponding mean validity from non-publisher studies (.27 vs. .12). These validity differences remain when we control (i.e., via the regression analyses) factors that might account for the larger test-publisher validities, such as type of integrity test and study design.

The nature and degree of validity differences between test publishers and independent non-publishers are somewhat less clear among studies that have used CWB and turnover as criteria. This may be due to the more modest sample sizes for some analyses and to confounding factors such as the source and type of criterion measure. For instance, among studies that used self-reported CWB as a criterion, corrected validity estimates from test publisher studies are smaller (.39) than validity estimates than non-publisher studies (.55). Moreover, the validity evidence from test publishers and non-publishers is similar for both non-self-report CWB criteria and turnover. However, one consistent finding is that validity estimates that test publishers originally reported are larger than validity estimates that we computed based on what the publishers reported.

We also found some evidence that validity estimates from researchers who developed the focal integrity test can be somewhat more positive than validity estimates from researchers who examined a test developed by another researcher or a test publisher. For example, the two mean corrected validities were .20 and .10, respectively, for job performance. Although these comparisons are based on small numbers of studies (particularly the CWB criterion analyses), they suggest that test publisher versus non-publisher might not be the only factor to consider; results of research from test developers versus non-developers also may be different.

Overall, these findings lend some empirical support to anecdotal claims that test-publisher research tends to provide a more optimistic view of integrity test validity, particularly for criteria that reflect job performance. Thus, one possible contributing factor to the smaller validity estimates found in the present study is that our estimates are based on a similar proportion of studies authored by test publishers versus non-publishers (approximately 60% vs. 40%),whereas results of previous meta-analyses are based primarily or solely on test publisher studies. In fact, our results may underestimate differences between test publisher and non-publisher studies, because we excluded a much larger percentage of publisher-authored studies that used designs, analysis techniques, and so forth (e.g., extreme group designs) that are likely to result in inflated estimates of criterion-related validity.

It is important to note that the more favorable validity evidence test publishers appear to report in certain cases does not necessarily indicate biased reporting (e.g., suppression of less positive results). We reviewed research from some publishers whose methods generally were sound and gave us no reason to doubt their results. Thus, we cannot rule out the possibility that test-publisher

research provides a more accurate picture of integrity test validity and that non-publisher research, for whatever reason, is overly pessimistic. Perhaps the key takeaway from our results is that the overall validity evidence from test publishers and independent researchers (and in some cases, non-publishers who do and do not develop integrity tests) is not always comparable, and that researchers and practitioners should consider this fact when drawing conclusions from this literature.

Moreover, the higher percentage of non-publisher research results is only one factor that may contribute to the smaller validity estimates we found compared with previous meta-analyses. For one, we included the results of research conducted after the earlier meta-analyses were published. This is notable because more recent integrity test studies tend to yield smaller validities than older studies (see Tables 4 and 6). We also used a somewhat more stringent set of inclusion criteria than what some previous meta-analyses may have used. For instance, we did not include validity estimates from studies that used contrasted or extreme groups designs, polygraph ratings as criteria, and that reported statistically significant results only, all of which may yield relatively larger validity estimates than the types of studies we cumulated. In addition, we reported results with and without validity estimates (e.g., *R*) adjusted for shrinkage (for validities based on multiple integrity test scales or items), which previous meta-analyses may not have done.

Thus, there may be several reasons why the present results appear to be less optimistic about the criterion-related validity of integrity tests. Although we could have included a wider range of studies to try to uncover the specific factors that led to differences between this study and previous studies, this was not a goal of our work. Rather, our goal was to cumulate studies that met criteria we think are important to best understand the validity of integrity-specific scales for predicting individual behavior at work and, in turn, the potential usefulness of such scales for personnel selection.[13]

Limitations and Directions for Future Research

We conclude by noting some limitations of the present study as well as some possible directions for future research. First, some of the moderator results we report are based on relatively small subsets of studies. Consequently, confidence and credibility intervals for the estimated mean validity are rather wide in some instances and overlap with intervals from other levels of a given moderator. The small number of available studies for some of the subanalyses highlights how little we know about the extent to which integrity tests predict certain criteria and under what conditions they may or may not predict those criteria.

Second, the scant information on statistical artifacts may have limited our ability to provide more precise estimates of true validity. For instance, only a few studies reported estimates of interrater reliability for job performance criteria, and many authors did not indicate how many raters contributed to their performance measures. Thus, the single-rater estimates we used for such studies may overestimate corrected validity if the criteria were based on data from multiple raters. Of course, there is debate about whether interrater coefficients provide appropriate estimates of measurement error in the first place (e.g., Murphy & DeShon, 2000; Schmidt, Viswesvaran, & Ones, 2000).

There also was a somewhat limited number of reliability estimates for CWB criteria. When reliabilities were reported, they almost always were alpha coefficients. Reliability estimates that account for other likely sources of error in CWB measures, including transient error (i.e., via test–retest or parallel form estimates) and rater error (i.e., via interrater estimates), appear to be very limited within this literature. We hope future researchers can employ designs that allow for more comprehensive reliability assessments of the various ways CWB can be measured.

Further, although the majority of primary studies that met our criteria used job incumbents as participants, very few authors provided details necessary to determine the nature or degree of range restriction, such as how incumbents originally were selected, the variance in test scores for both applicants and the research sample, and the extent to which the original predictors were related to incumbents' subsequent integrity test scores. We encourage integrity test researchers to provide such information to increase understanding of range restriction in this area.

Another potential limitation of our work is that because several studies did not provide predictor or criterion correlations, we sometimes had to use mean validity estimates rather than composite validity estimates. This is a possible concern because mean validities can underestimate composite validities (Hunter & Schmidt, 2004). To estimate the extent to which our use of mean validities may have underestimated validity, we identified 32 independent samples for which we were able to compute a composite validity (e.g., because the authors reported, or we were able to obtain, predictor correlations). For these studies, we calculated what the corresponding mean validity estimates would be had we used them. The composite validities (mean $r = .24$) were indeed somewhat larger than the mean validities (mean $r = .17$). We used this finding to simulate the effect of having to use the mean validities for studies that did not provide information necessary to compute composite validities. Specifically, we increased each mean validity estimate by 41% to reflect the difference we found between the composite and mean validities (i.e., .24 minus .17 = .07, and .07 divided by .17 = .41). We then reran the relevant meta-analyses using these values instead of the original values.

The overall impact of inclusion of the mean estimates was quite small. When we included the simulated composite validities, the only change was that the mean observed validity for job performance changed slightly from .12 to .13, and the mean corrected validity changed from .15 to .17 (actually, .154 to .168). There were no changes in the overall observed or corrected validity estimates for training performance, CWB, or turnover. Regardless, we encourage researchers to report correlations among all measures so that meta-analysts in the future can estimate composite validities when appropriate.

[13] In fact, it would not have been possible to fully replicate earlier meta-analyses. For one, many of the primary studies summarized in past research were conducted 30 or more years ago, and in many instances, the original reports could not be located. In addition, some test publishers were not willing to share unpublished studies that previous researchers apparently were able to obtain. Finally, even if we had been able to get all the unpublished studies summarized in the past, we were unable to obtain information concerning how previous authors coded each primary study.

Finally, many integrity tests include multiple subscales, or at least the potential to compute subscales given the heterogeneity of items. For example, the Inwald Personality Inventory comprises 26 scales, and 25 constructs underlie the two primary scales contained in the PDI Employment Inventory. However, most of the studies we reviewed reported a validity estimate for overall integrity test scores only or some multivariate statistic that reflected the combined validity of all the subscales. When scale-level validity information was reported, the information frequently would not have been suitable to include in the meta-analysis, such as beta weights for subscales that emerged as statistically significant predictors within a multiple regression analysis.

Therefore, we are unable to cumulate validity evidence for different facets of integrity. This was unfortunate, because there appears to be evidence of differential validity across subscales from the same integrity test (e.g., Carless et al., 2007; Kauder & Thomas, 2003; Van Iddekinge, Taylor, & Eidson, 2005). Clearly, much more research is needed to increase understanding about what integrity tests measure and whether and how the underlying facets relate to valued criteria.

Conclusions

The goal of this study was to provide an updated understanding of the criterion-related validity of integrity tests. Overall, the results reinforce some of the concerns that have been raised about the general quality of studies that comprise the integrity test literature and the validity evidence based upon this research. Indeed, when we estimate validity on the basis of studies whose conduct is consistent with professional standards for test validation, and whose results focus on the validity of integrity tests for predicting individual work behavior, the validity evidence appears to be somewhat less optimistic than that suggested by earlier reviews. With the notable exception of self-report CWB criteria, most of the corrected validity estimates for integrity tests are smaller than .20, and many estimates are closer to .10. Thus, although integrity tests yield small subgroup differences and low correlations with cognitive ability, the present results suggest the criterion-related validity of these tests generally is quite modest. We hope our findings may be informative to researchers and practitioners who wish to consider integrity tests for research purposes or personnel selection.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Ahart, A. M., & Sackett, P. R. (2004). A new method of examining relationships between individual difference measures and sensitive behavior criteria: Evaluating the unmatched count technique. *Organizational Research Methods, 7,* 101–114.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50,* 179–211.

Ajzen, I., & Fishbein, M. (2005). The influence of attitudes on behavior. In D. Albarracin, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 173–221). Mahwah, NJ: Erlbaum.

Alliger, G. M., & Dwight, S. A. (2000). A meta-analytic investigation of the susceptibility of integrity tests to faking and coaching. *Educational and Psychological Measurement, 60,* 59–72.

Allworth, E., & Hesketh, B. (1999). Construct-oriented biodata: Capturing change-related and contextually relevant future performance. *International Journal of Selection and Assessment, 7,* 97–111.

*Ash, P. (1985). Predicting dishonesty with the Reid Report. *Journal of the American Polygraph Association, 5,* 139–145.

*Barge, B. N., & Skilling, N. J. (1986). *An analysis of requirements, employee characteristics, and performance in Kelly assisted living assignments* (Report No. 121). Minneapolis, MN: Personnel Decisions Research Institutes.

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the validity of personality constructs. *Journal of Applied Psychology, 81,* 261–272.

Barrick, M. R., & Zimmerman, R. D. (2009). Hiring for retention and performance. *Human Resource Management, 48,* 183–206.

Beal, D. J., Corey, D. M., & Dunlap, W. P. (2002). On the bias of Huffcutt and Arthur's (1995) procedure for identifying outliers in the meta-analysis of correlations. *Journal of Applied Psychology, 87,* 583–589.

Becker, T. E. (2005). Development and validation of a situational judgment test of employee integrity. *International Journal of Selection and Assessment, 13,* 225–232.

Bekelman, J. E., Li, Y., & Gross, C. P. (2003). Scope and impact of financial conflicts of interest in biomedical research. *Journal of the American Medical Association, 289,* 454–465.

Bennett, R. J., & Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology, 85,* 349–360.

Berry, C. M., Sackett, P. R., & Landers, R. N. (2007). Revisiting the interview-cognitive ability relationships: Attending to specific range restriction mechanisms in meta-analysis. *Personnel Psychology, 60,* 837–874.

Bhandari, M., Busse, J. W., Jackowski, D., Montori, V. M., Schunemann, H., Sprague, S., . . . Devereaux, P. J. (2004). Association between industry funding and statistically significant pro-industry findings in medical and surgical randomized trials. *Canadian Medical Association Journal, 170,* 477–480.

*Billings, S. W. (2001). *Clarifications of the relationship between conscientiousness and integrity* (Unpublished doctoral dissertation). Wayne State University, Detroit, MI.

*Bing, M. N., Stewart, S. M., Davison, H. K., Green, P. D., McIntyre, M. D., & James, L. R. (2007). An integrative typology of personality assessment for aggression: Implications for predicting counterproductive workplace behavior. *Journal of Applied Psychology, 92,* 722–744.

Bobko, P., & Stone-Romero, E. F. (1998). Meta-analysis is another useful research tool but it is not a panacea. In G. R. Ferris (Ed.), *Research in personnel and human resources management* (Vol. 16, pp. 359–379). Greenwich, CT: JAI Press.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 71–98). San Francisco, CA: Jossey-Bass.

*Boye, M. W., & Wasserman, A. R. (1996). Predicting counterproductivity among drug store applicants. *Journal of Business and Psychology, 10,* 337–349.

*Bradley, P. A. (1984). *Milby Profile validity study.* Minneapolis, MN: Milby Systems.

Butts, M. M., & Ng, T. W. H. (2009). Chopped liver? OK. Chopped data? Not OK. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical myths and methodological urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 361–386). New York, NY: Routledge.

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

Camara, W. J., & Schneider, D. L. (1995). Questions of construct breadth and openness of research in integrity testing. *American Psychologist, 50,* 459–460.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A

theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35–70). San Francisco, CA: Jossey-Bass.

*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.

*Chibnall, J. T., & Detrick, P. (2003). The NEO PI-R, Inwald Personality Inventory, and MMPI-2 in the prediction of police academy performance: A case for incremental validity. *American Journal of Criminal Justice, 27,* 233–248.

Coleman, V. I., & Borman, W. C. (2000). Investigating the underlying structure of the citizenship performance domain. *Human Resource Management Review, 10,* 25–44.

Cortina, J. M. (2003). Apples and oranges (and pears, oh my!): The search for moderators in meta-analysis. *Organizational Research Methods, 6,* 415–439.

*Cotton, J. R. (1990). *A study to evaluate the use of the Phase II Profile as an indicator of employment longevity for hospitality industry employees* (Unpublished master's thesis). Johnson and Wales University, Providence, RI.

Coyne, I., & Bartram, D. (2002). Assessing the effectiveness of integrity tests: A review. *International Journal of Testing, 2,* 15–34.

Cunningham, M. R., Wong, D. T., & Barbee, A. P. (1994). Self-presentation dynamics on overt integrity tests: Experimental studies of the Reid Report. *Journal of Applied Psychology, 79,* 643–658.

Dalton, D. R., & Metzger, M. B. (1993). Integrity testing for personnel selection: An unsparing perspective. *Journal of Business Ethics, 12,* 147–156.

*Detrick, P., & Chibnall, J. T. (2002). Prediction of police officer performance with the Inwald Personality Inventory. *Journal of Police and Criminal Psychology, 17,* 9–17.

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

*Dunnette, M. D., Paullin, C., & Motowidlo, S. J. (1989). *Development of the Kelly Applicant Profile for use in screening candidates for positions with Kelly Assisted Living Services* (Report No. 170). Minneapolis, MN: Personnel Decisions Research Institutes.

ePredix. (2001). *PDI Employment Inventory technical manual.* Minneapolis, MN: Author.

Finch, D. M., Edwards, B. D., & Wallace, J. C. (2009). Multistage selection strategies: Simulating the effects on adverse impact and expected performance for various predictor combinations. *Journal of Applied Psychology, 94,* 318–340.

*Fine, S. (2009a, December). *Integrity testing and counterproductive behaviors in Israel.* Paper presented at the 2nd Academic Research Conference of Transparency International (TI) Israel, Tel Aviv, Israel.

*Fine, S. (2009b, May). *Managing personnel risk: Integrity, disengagement, and counterproductive behaviors.* Paper presented at the 14th Annual European Congress of Work and Organizational Psychology, Santiago de Compostela, Spain.

*Fine, S. (2010, July). *Cross-cultural validity of integrity testing: A tale of three banks.* Paper presented at the 27th International Congress of Applied Psychology, Melbourne, Australia.

*Fine, S., & Horowitz, I. (2008, October). *Integrity testing in personnel management.* Invited address presented at the Annual Brinks HOS Conference, Tel Aviv, Israel.

*Fine, S., Horowitz, I., Weigler, H., & Basis, L. (2010). Is good character good enough? The effects of situational variables on the relationship between integrity and counterproductive work behaviors. *Human Resource Management Review, 20,* 73–84.

*Fortmann, K., Leslie, C., & Cunningham, M. (2002). Cross-cultural comparisons of the Reid Integrity Scale in Latin America and South Africa. *International Journal of Selection and Assessment, 10,* 98–108.

*Frost, A. G., & Orban, J. A. (1990). An examination of an appropriateness index and its effect on validity coefficients. *Journal of Business and Psychology, 5,* 23–36.

*Frost, A. G., & Rafilson, F. M. (1989). Overt integrity tests versus personality-based measures of delinquency: An empirical comparison. *Journal of Business and Psychology, 3,* 269–277.

*Gardner, J. F. (1998). *The Inwald Personality Inventory (IPI) and law enforcement officer response to domestic violence* (Unpublished doctoral dissertation). University of Alabama, Tuscaloosa.

Gerstein, L. H., Brooke, C. R., & Johnson, S. D. (1989). Internal validity studies of a telephone pre-employment measure. *Journal of Employment Counseling, 26,* 77–83.

Goldberg, L. R., Grenier, J. R., Guion, R. M., Sechrest, L. B., & Wing, H. (1991). *Questionnaires used in the prediction of trustworthiness in pre-employment selection decisions: An APA task force report.* Washington, DC: American Psychological Association.

*Gonder, M., & Gilmore, D. C. (2004). Personality profiles of police officers who successfully completed academy training. *Applied H.R.M. Research, 9,* 59–62.

*Gough, H. G. (1965). Reliability as a qualitative factor in employees' work. *Bollettino di Psicologia Applicata, 69–70,* 3–7.

*Gough, H. G., & Freddi, G. (1967). The prediction of performance in an industrial training program. *Bollettino di Psicologia Applicata, 83–84,* 93–102.

Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of Management, 26,* 463–488.

Gruys, M. L., & Sackett, P. R. (2003). Investigating the dimensionality of counterproductive work behavior. *International Journal of Selection and Assessment, 11,* 30–42.

*Haaland, D., Christiansen, N. D., & Kaufmann, G. (1999, April). Applicant distortion of personality measures in police selection: Reasons for optimism and caution. In M. L. Kelly & A. D. Mead (Chairs), *Using personality in police selection.* Symposium presented at the 14th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

*Hakstian, A. R., Farrell, S., & Tweed, R. G. (2002). The assessment of counterproductive tendencies by means of the California Psychological Inventory. *International Journal of Selection and Assessment, 10,* 58–86.

Harold, C. M., McFarland, L. A., & Weekley, J. A. (2006). The validity of verifiable and nonverifiable biodata items: An examination across applicants and incumbents. *International Journal of Selection and Assessment, 14,* 336–346.

Harrison, D. A., Newman, D. A., & Roth, P. L. (2006). How important are job attitudes? Meta-analytic comparisons of integrative behavioral outcomes and time sequences. *Academy of Management Journal, 49,* 305–325.

Hattrup, K., O'Connell, M. S., & Wingate, P. H. (1998). Prediction of multidimensional criteria: Distinguishing task and contextual performance. *Human Performance, 11,* 305–319.

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

*Hilliard, P. A. (2000). *Comparison of the predictive validity of a written test, an integrity test, a conscientiousness questionnaire, a structured behavioral interview and a personality inventory in the assessment of job applicants' background investigations, and subsequent task and contextual job performance* (Unpublished doctoral dissertation). University of Southern California, Berkeley.

Hoffman, B. J., Blair, C. A., Meriac, J. P., & Woehr, D. J. (2007). Expanding the criterion domain? A quantitative review of the OCB literature. *Journal of Applied Psychology, 92,* 555–566.

*Hogan, J., Brinkmeyer, K., & Kidwell, D. (1994). *Validity of the Hogan Personality Inventory for selecting drivers at* [client name withheld for confidentiality] (Report No. 62). Tulsa, OK: Hogan Assessment Systems.

Hogan, J., & Hogan, R. (1989). How to measure employee reliability. *Journal of Applied Psychology, 74,* 273–279.

*Hogan, J., Peterson, S., Hogan, R., & Jones, S. (1985). *Development and validation of a line haul driver selection inventory* (Report No. 11). Tulsa, OK: University of Tulsa.

*Hogan, R., & Gerhold, C. (1994). *Validity of the Hogan Personality Inventory for selecting certified nursing assistants at* [client name withheld for confidentiality] (Report No. 63). Tulsa, OK: Hogan Assessment Systems.

*Hogan, R., & Gerhold, C. (1995). *Validity of the Hogan Personality Inventory for selecting managers and assistant managers at* [client name withheld for confidentiality] (Report No. 67). Tulsa, OK: Hogan Assessment Systems.

*Hogan, R., Hogan, J., & Brinkmeyer, K. (1994). *Validity of the Hogan Personality Inventory for selecting drivers at* [client name withheld for confidentiality] (Report No. 64). Tulsa, OK: Hogan Assessment Systems.

*Hogan, R., Hogan, J., Lock, J., & Brinkmeyer, K. (1994). *Validity of the Hogan Personality Inventory for selecting managers at* [client name withheld for confidentiality] (Report No. 61). Tulsa, OK: Hogan Assessment Systems.

*Hogan, R., Jacobson, G. K., Hogan, J., & Thompson, B. (1987). *Development and validation of a service operations dispatcher selection battery* (Report No. 20). Tulsa, OK: Hogan Assessment Systems.

Hough, L. M. (1998). Personality and work: Issues and evidence. In M. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 131–159). Hillsdale, NJ: Erlbaum.

Huffcutt, A. I., & Arthur, W. A. (1995). Development of a new outlier statistic for meta-analytic data. *Journal of Applied Psychology, 80,* 327–334.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Klehe, U.-C. (2004). The impact of job complexity and study design on situational and behavior description interview validity. *International Journal of Selection and Assessment, 12,* 262–273.

Hulin, C. L. (1991). Adaptation, persistence, and commitment in organizations. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 445–505). Palo Alto, CA: Consulting Psychologist Press.

*Hunt, S. T., Hansen, T. L., & Paajanen, G. E. (1997, April). *The empirical structure and construct validity of a widely used, personality-based integrity test.* Paper presented at the 12th Annual Meeting of the Society for Industrial and Organizational Psychology, St. Louis, MO.

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

Hunter, J. E., Schmidt, F. L., & Judiesh, M. K. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology, 75,* 28–42.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications for direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91,* 594–612.

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

Institute for Personality and Ability Testing. (2006). *Inwald Personality Inventory technical manual.* Champaign, IL: Author.

*Inwald, R. E. (1988). Five-year follow-up study of departmental terminations as predicted by 16 preemployment psychological indicators. *Journal of Applied Psychology, 73,* 703–710.

*Inwald, R. E., & Brockwell, A. L. (1991). Predicting the performance of government security personnel with the IPI and MMPI. *Journal of Personality Assessment, 56,* 522–535.

Inwald, R. E., Hurwitz, H., Jr., & Kaufman, J. C. (1991). Uncertainty reduction in retail and public safety-private security screening. *Forensic Reports, 4,* 171–212.

*Inwald, R. E., & Patterson, T. (1989). *Use of psychological testing to predict performance of law enforcement trainees* (Unpublished technical report). New York, NY: Hilson Research.

*Inwald, R. E., & Shusman, E. J. (1984). Personality and performance sex differences of law enforcement officer recruits. *Journal of Police Science and Administration, 12,* 339–347.

*Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13,* 371–388.

Johnson, J. W. (2001). The relative importance of task and contextual performance dimensions to supervisor judgments of overall performance. *Journal of Applied Psychology, 86,* 984–996.

*Jones, J. W. (1979, November). *Employee deviance: Attitudinal correlates of theft and one-the-job alcohol.* Paper presented at the 5th Annual Meeting of the Society of Police and Criminal Psychology, Chicago, IL.

*Jones, J. W. (1980a). Attitudinal correlates of employees' deviance: Theft, alcohol use, and nonprescribed drug use. *Psychological Reports, 47,* 71–77.

*Jones, J. W. (1980b, October). *Correlates of police misconduct: Violence and alcohol use on the job.* Paper presented at the 6th Annual Meeting of the Society of Police and Criminal Psychology, Atlanta, GA.

*Jones, J. W. (1981). Dishonesty, burnout, and unauthorized work break extensions. *Personality and Social Psychology Bulletin, 7,* 406–409.

*Jones, J. W. (1982, August). *Psychological predictors of employee theft.* Paper presented at the 90th Annual Meeting of the American Psychological Association, Washington, DC.

Jones, J. W. (1985). *Conducting theft audits with the Employee Attitude Inventory: A review and critique* (Unpublished manuscript).

*Jones, J. W., Brasher, E. E., & Huff, J. W. (2002). Innovations in integrity-based personnel selection: Building a technology-friendly assessment. *International Journal of Selection and Assessment, 10,* 87–97.

*Jones, J. W., Joy, D. S., & Martin, S. L. (1990). A multidimensional approach for selecting child care workers. *Psychological Reports, 67,* 543–553.

*Jones, J. W., & Scruggs, D. P. (1981, April). *Psychologically profiling endorsers of nuclear crime and sabotage.* Paper presented at the 53rd Annual Meeting of the Midwestern Psychological Association, Detroit, MI.

*Jones, J. W., & Terris, W. (1983). Predicting employees' theft in home improvement centers. *Psychological Reports, 52,* 187–201.

Karren, R. J., & Zacharias, L. (2007). Integrity tests: Critical issues. *Human Resource Management Review, 17,* 221–234.

*Kauder, B. S., & Thomas, J. C. (2003). Relationship between MMPI-2 and Inwald Personality Inventory (IPI) scores and ratings of police officer probationary performance. *Applied HRM Research, 8,* 81–84.

Kjaergard, L. L., & Als-Nielsen, B. (2002). Association between competing interests and authors' conclusions: Epidemiological study of randomized clinical trials published in the BMJ. *British Medical Journal, 325,* 249–252.

Kpo, W. (1984). *Application of validity generalization to honesty testing* (Unpublished master's thesis). Illinois Institute of Technology, Chicago.

*LaFosse, W. G. (1992). *Employee theft: The relationship of shrinkage rates to job satisfaction, store security, and employee reliability* (Unpublished master's thesis). University of North Texas, Denton.

LePine, J. A., & Van Dyne, L. (2001). Voice and cooperative behavior as contrasting forms of contextual performance: Evidence of differential relationships with Big Five personality characteristics and cognitive ability. *Journal of Applied Psychology, 86,* 326–336.

*Leslie, C. S. (2006). *A comparison of self-disclosure measures between*

Internet and paper-based integrity assessments: Extending research to a job application setting (Unpublished doctoral dissertation). Capella University, Minneapolis, MN.

Lilienfeld, S. O. (1993). Do "honesty" tests really measure honesty? *Skeptical Inquirer, 18,* 32–41.

London House. (1982). *The Employee Attitude Inventory.* Park Ridge, IL: Author.

*Loudermilk, K. M. (1964). *The relationship between aptitude, personality, physical fitness, and personal data, and job performance in a combined lumber and paper mill* (Unpublished doctoral dissertation). University of Idaho, Boise.

*Luther, N. (2000). Integrity testing and job performance within high performance work teams: A short note. *Journal of Business and Psychology, 15,* 19–25.

Lykken, D. T. (1981). *A tremor in the blood: Uses and abuses of the lie detector.* New York, NY: McGraw-Hill.

Maertz, C. P., & Griffeth, R. W. (2004). Eight motivational forces and voluntary turnover: A theoretical synthesis with implications for research. *Journal of Management, 30,* 667–683.

*Malin, S. Z., Luria, J., & Morgenbesser, L. I. (1987, August). *New York state pre-employment psychological screening program: Longitudinal validation study.* Paper presented at the 95th Annual Meeting of the American Psychological Association, New York, NY.

*Marcus, B. (2006). Relationships between faking, validity, and decision criteria in personnel selection. *Psychological Science, 48,* 226–246.

*Marcus, B., Lee, K., & Ashton, M. C. (2007). Personality dimensions explaining relationships between integrity tests and counterproductive behavior: Big Five or one in addition? *Personnel Psychology, 60,* 1–34.

*Marcus, B., Schuler, H., Quell, P., & Hümpfer, G. (2002). Measuring counterproductivity: Development and initial validation of a German self-report questionnaire. *International Journal of Selection and Assessment, 10,* 18–35.

Martelli, T. A. (1988). *Preemployment screening for honesty: The construct validity, criterion-related validity, and test–retest reliability of a written integrity test* (Unpublished doctoral dissertation). Ohio University, Athens.

*Mastrangelo, P. M., & Jolton, J. A. (2001). Predicting on-the-job substance abuse with a written integrity test. *Employee Responsibilities and Rights Journal, 13,* 95–106.

McDaniel, M. A., & Jones, J. W. (1986). A meta-analysis of the validity of the Employee Attitude Inventory theft scales. *Journal of Business and Psychology, 1,* 31–50.

McDaniel, M. A., & Jones, J. W. (1988). Predicting employee theft: A quantitative review of the validity of a standardized measure of dishonesty. *Journal of Business and Psychology, 2,* 327–343.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Predicting job performance from common sense. *Journal of Applied Psychology, 86,* 730–740.

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

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

*Moretti, D. M. (1980, October). *Employee counterproductivity: Attitudinal predictors of industrial damage and waste.* Paper presented at the 6th Annual Meeting of the Society of Police and Criminal Psychology, Atlanta, GA.

*Moretti, D. M. (1986). The predictions of employee counterproductivity through attitude assessment. *Journal of Business and Psychology, 1,* 134–147.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K. R., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60,* 683–729.

Murphy, K. R. (1995). Integrity testing. In N. Brewer & C. Wilson (Eds.), *Psychology and policing* (pp. 205–228). Hillsdale, NJ: Erlbaum.

Murphy, K. R., & DeShon, R. P. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53,* 873–900.

*Nicol, A. M., & Paunonen, S. V. (2002). Overt honesty measures predicting admissions: An index of validity or reliability. *Psychological Reports, 90,* 105–115.

O'Bannon, R. M., Goldinger, L. A., & Appleby, G. S. (1989). *Honesty and integrity testing: A practical guide.* Atlanta, GA: Applied Information Resources.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60,* 995–1027.

Ones, D. S., & Viswesvaran, C. (1996). Bandwidth-fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior, 17,* 609–626.

Ones, D. S., & Viswesvaran, C. (1998). Gender, age, and race differences on overt integrity tests: Results across four large-scale job applicant data sets. *Journal of Applied Psychology, 83,* 35–42.

Ones, D. S., & Viswesvaran, C. (2001). Personality at work: Criterion-focused occupational personality scales used in personnel selection. In B. W. Roberts & R. Hogan (Eds.), *Personality psychology in the workplace* (pp. 63–92). Washington, DC: American Psychological Association.

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

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2003). Personality and absenteeism: A meta-analysis of integrity tests. *European Journal of Personality, 17,* 19–38.

*Paajanen, G. E. (1988). *The prediction of counterproductive behavior by individual and organizational variables* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis.

*Palmatier, J. J. (1996). *The Big Five factors and hostility in the MMPI and IPI: Predictors of Michigan State troopers' job performance* (Unpublished doctoral dissertation). Michigan State University, East Lansing.

*Personnel Decisions Incorporated. (1986). *Validation of the PDI Employment Inventory for store security officers* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (1987a). *Retail department store evaluation of the PDI Employment Inventory* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (1987b). *Validation of the PDI Employment Inventory for fast food restaurants* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (1988a). *Validation of the PDI Employment Inventory for amusement parks* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (1988b). *Validation of the PDI Employment Inventory for food wholesalers* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (1990a). *Preliminary findings: Validity of the PDI Employment Inventory for a beverage bottler* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (1990b). *Validity of the PDI Employment Inventory for drug stores* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (1990c). *Validity of the PDI Employment Inventory for retail and discount supermarkets* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (1991a). *Validation of the PDI Employment Inventory for car rental services* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (1991b). *Validation of the PDI Employment Inventory for a medical center* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (1992a). *Validation of the PDI Employment Inventory for a major international airline* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (1992b). *Validation of the PDI Employment Inventory for office supply stores* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (1993). *Validity of the PDI Employment Inventory and Customer Service Inventory for a major airline* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (n.d.). *Validation of the PDI Employment Inventory for bank tellers* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (n.d.). *Validation of the PDI Employment Inventory for courier service* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (n.d.). *Validation of the PDI Employment Inventory for department stores* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (n.d.). *Validation of the PDI Employment Inventory for a furniture store* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (n.d.). *Validation of the PDI Employment Inventory for gasoline station/store employees* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (n.d.). *Validity of the PDI Employment Inventory for a hard goods chain* (Unpublished technical report). Minneapolis, MN: Author.

*Personnel Decisions Incorporated. (n.d.). *Validation of the PDI Employment Inventory for supermarkets* (Unpublished technical report). Minneapolis, MN: Author.

Podsakoff, P., MacKenzie, S., Lee, J.-Y., & Podsakoff, N. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88,* 879–903.

*Powell, D. M., Poole, A. E. R., Carswell, J., & Marcus, B. (2008, April). *Predicting counterproductive workplace behavior with narrow facets of the HEXACO.* Paper presented at the 23rd Annual Meeting of the Society for Industrial and Organizational Psychology, San Francisco, CA.

*Provines, J. L. (2006). *Investigation of police officer selection procedures* (Unpublished doctoral dissertation). Wichita State University, Wichita, KS.

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology, 85,* 612–624.

*Rafilson, F. M. (1988). Development of a standardized measure to predict employee productivity. *Journal of Business and Psychology, 3,* 199–213.

*Raza, S., Metz, D., Dyer, P., Coan, T., & Hogan, J. (1986). *Development and validation of personnel selection procedures for hospital service personnel* (Report No. 13). Tulsa, OK: University of Tulsa.

Ridker, P. M., & Torres, J. (2006). Reported outcomes in major cardiovascular clinical trials funded by for-profit and not-for-profit organizations: 2000–2005. *Journal of the American Medical Association, 295,* 2270–2274.

Robinson, S. L., & Bennett, R. J. (1995). A typology of deviant workplace behaviors: A multidimensional scaling study. *Academy of Management Journal, 38,* 555–572.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymtote level with increasing opportunity to observe. *Journal of Applied Psychology, 75,* 322–327.

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

Rotundo, M., & Sackett, P. R. (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy-capturing approach. *Journal of Applied Psychology, 87,* 66–80.

Russell, C. R., Settoon, R. P., McGrath, R. N., Blanton, A. E., Kidwell, R. E., Lohrke, F. T., . . . Danforth, G. W. (1994). Investigator characteristics as moderators of personnel selection research: A meta-analysis. *Journal of Applied Psychology, 79,* 163–170.

Sackett, P. R. (2002). The structure of counterproductive work behaviors: Dimensionality and relations with facets of job performance. *International Journal of Selection and Assessment, 10,* 5–11.

Sackett, P. R., Burris, L. R., & Callahan, C. (1989). Integrity testing for personnel selection: An update. *Personnel Psychology, 42,* 491–529.

Sackett, P. R., & Decker, P. J. (1979). Detection of deception in the employment context: A review and critical analysis. *Personnel Psychology, 32,* 487–506.

Sackett, P. R., & Devore, C. J. (2001). Counterproductive work behaviors. In N. Anderson, D. S. Ones, H. K. Sinangil, & V. Viswesvaran (Eds.), *International handbook of work psychology* (Vol. 1, pp. 145–164). London, England: Sage.

Sackett, P. R., & Harris, M. M. (1984). Honesty testing for personnel selection: A review and critique. *Personnel Psychology, 37,* 221–245.

Sackett, P. R., & Wanek, J. E. (1996). New developments in the use of measures of honesty, integrity, conscientiousness, dependability, trustworthiness, and reliability for personnel selection. *Personnel Psychology, 49,* 787–829.

Saxe, L., Dougherty, D., & Cross, T. (1985). The validity of polygraph testing: Scientific analysis and public controversy. *American Psychologist, 40,* 355–366.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel selection: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology, 64,* 609–626.

Schmidt, F. L., & Le, H. (2004). *Software for the Hunter-Schmidt meta-analysis methods.* Iowa City, IA: University of Iowa, Department of Management and Organization.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53,* 901–912.

Schmitt, N., Cortina, J. M., Ingerick, M. J., & Wiechmann, D. (2003). Personnel selection and employee performance. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology* (Vol. 12, pp. 77–105). Hoboken, NJ: Wiley.

*Schneider, B. M. (2002). *Using the Big-Five personality factors in the Minnesota Multiphasic Personality Inventory, California Psychological Inventory, and Inwald Personality Inventory to predict police performance* (Unpublished doctoral dissertation). Florida International University, Miami.

*Science Research Associates. (1987). *Personal Outlook Inventory (POI) examiner's manual.* Chicago, IL: Author.

*Sevy, B. A. (1987). *Validity of the PDI employment inventory for bus driver selection* (Unpublished technical report). Minneapolis, MN: Personnel Decisions Incorporated.

Shieh, G. (2008). Improved shrinkage estimation of squared multiple correlation coefficient and squared cross-validity coefficient. *Organizational Research Methods, 11,* 387–407.

*Shusman, E. J., & Inwald, R. E. (1991). Predictive validity of the Inwald Personality Inventory. *Criminal Justice and Behavior, 18,* 419–426.

*Shusman, E. J., Inwald, R. E., & Knatz, H. F. (1987). A cross-validation study of police recruit performance as predicted by the IPI and MMPI. *Journal of Police Science and Administration, 15,* 162–169.

*Shusman, E. J., Inwald, R. E., & Landa, B. (1984). Correction officer job performance as predicted by the IPI and the MMPI: A validation and cross-validation study. *Criminal Justice and Behavior, 11,* 309–329.

*Sigma Assessment Systems. (2009). *Employee Screening Questionnaire–2 technical manual.* Port Huron, MI: Author.

Snyman, J. H. (1990). *Fakability of pre-employment paper-and-pencil honesty tests* (Unpublished master's thesis). Radford University, Radford, VA.

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

Steel, P. D., & Kammeyer-Mueller, J. D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology, 87,* 96–111.

*Surrette, M. A., & Serafino, G. (2003). Relationship between personality and law enforcement performance. *Applied HRM Research, 8,* 89–92.

Taylor, P. J., Russ-Eft, D. F., & Chan, D. W. L. (2005). A meta-analytic review of behavior modeling training. *Journal of Applied Psychology, 90,* 692–709.

*Terris, W., & Jones, J. (1980). Attitudinal and personality correlates of theft among supermarket employees. *Journal of Security Administration, 3,* 65–78.

*Terris, W., & Jones, J. (1982). Psychological factors related to employees' theft in the convenience store industry. *Psychological Reports, 51,* 1219–1238.

Tett, R. P., & Christiansen, N. D. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007). *Personnel Psychology, 60,* 968–993.

Thorndike, E. L. (1939). On the fallacy of imputing the correlations for groups to the individuals or smaller groups composing them. *American Journal of Psychology, 52,* 122–124.

*Tolbirt, M. E. (1992). *Relationship between honesty, ability, and personality tests and job performance* (Unpublished master's thesis). California State University, Long Beach.

U.S. Congress, Office of Technology Assessment. (1983). *Scientific validity of polygraph testing: A research review and evaluation* (OTA-TM-H-15). Washington, DC: U.S. Government Printing Office.

U.S. Congress, Office of Technology Assessment. (1990). *The use of integrity tests for pre-employment screening* (OTA-SET-442). Washington, DC: U.S. Government Printing Office.

Vangent. (2007). *Personnel Selection Inventory (PSI): A compilation of scientific brief information.* Chicago, IL: Author.

*Van Hein, J. L., Kramer, J. J., & Hein, M. (2007). The validity of the Reid Report for selection of corrections staff. *Public Personnel Management, 36,* 269–280.

Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection procedures: A critical review and recommendations for practice. *Personnel Psychology, 61,* 871–925.

*Van Iddekinge, C. H., Taylor, M. A., & Eidson, C. E., Jr. (2005). Broad versus narrow facets of integrity: Predictive validity and subgroup differences. *Human Performance, 18,* 151–177.

Van Scotter, J. R., & Motowidlo, S. J. (1996). Interpersonal facilitation and job dedication as separate facets of contextual performance. *Journal of Applied Psychology, 81,* 525–531.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analysis framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90,* 108–131.

Wahlbeck, K., & Adams, C. (1999). Beyond conflict of interest: Sponsored drug trials show more favourable outcomes. *British Medical Journal, 318,* 465.

Wanek, J. E., Sackett, P. R., & Ones, D. S. (2003). Towards an understanding of integrity test similarities and differences: An item-level analysis of seven tests. *Personnel Psychology, 56,* 873–894.

*Werner, S. H., Jones, J. W., & Steffy, B. D. (1989). The relationship between intelligence, honesty, and theft admissions. *Educational and Psychological Measurement, 49,* 921–927.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52,* 372–376.

Williams, C. R. (1990). Deciding when, how, and if to correct turnover correlations. *Journal of Applied Psychology, 75,* 732–737.

Wood, J. (2008). Methodology for dealing with duplicate effects in a meta-analysis. *Organizational Research Methods, 11,* 79–95.

*Woolley, R. M., & Hakstian, A. R. (1993). A comparative study of integrity tests: The criterion-related validity of personality-based and overt measures of integrity. *International Journal of Selection and Assessment, 1,* 27–40.

Zimmerman, R. A. (2008). Understanding the impact of personality traits on individuals' turnover decisions: A meta-analytic path model. *Personnel Psychology, 61,* 309–348.

(*Appendix follows*)

Appendix

Main Codes and Input Values for the Primary Studies in the Meta-Analysis

Study	Publication status	Author affiliation	Test type	Study design	Study sample	Criterion	Criterion source	*N*	r_{yy}	*r*
Ahart and Sackett (2004)	Published	Non-publishers, developed test	Overt	Concurrent	Students with work experience	CWB	Self-reports	122	.31	.60
Ash (1985)	Published	Publisher	Overt	Predictive	Applicants	CWB	Records	140	.51	.04
						Turnover	Records	140	n/a	.06
						Tenure	Records	140	n/a	.00
Barge and Skilling (1986)	Unpublished	Non-publishers, did not develop test	Personality	Concurrent	Incumbents	Job performance	Ratings	121	.56	.17
Billings (2001)	Unpublished	Publisher	Both types	Concurrent	Incumbents	Job performance	Ratings	181	.56	.13
						CWB	Ratings	166	.56	.08
Bing et al. (2007)	Published	Non-publishers, did not develop test	Personality	Concurrent	Incumbents	Job performance	Ratings	184	.83	.13
						CWB	Ratings	184	.83	.21
Boye and Wasserman (1996)	Published	Publishers	Overt	Predictive	Applicants	CWB	Self-reports	95	.82	.31
Bradley (1984)	Unpublished	Publisher	Overt							
Sample 1				Predictive	Applicants	Job performance	Ratings	83	.56	.00
Sample 2				Unknown	Incumbents	Job performance	Ratings	101	.56	.00
						Tenure	Records	101	n/a	.00
Sample 3				Unknown	Both types	Job performance	Ratings	56	.56	.00
Carless et al. (2007)	Published	Non-publishers	Personality							
Study 1				Concurrent	Incumbents	Job performance	Ratings	61	.56	.09
Study 2				Predictive	Incumbents	Job performance	Ratings	120	.56	.17
						Turnover	Records	243	n/a	.08
Chibnall and Detrick (2003)	Published	Non-publishers, did not develop test	Personality	Predictive	Applicants	Training performance	Grades	79	.81	.20
Cotton (1990)	Unpublished	Non-publishers, did not develop test	Overt	Concurrent	Incumbents	Tenure	Records	36	n/a	.34
Detrick and Chibnall (2002)	Published	Non-publishers, did not develop test	Personality	Predictive	Applicants	Job performance	Ratings	108	.56	.10
						Turnover	Records	138	n/a	.00
Dunnette et al. (1989)	Unpublished	Non-publishers, did not develop test	Personality	Predictive	Applicants	Job performance	Ratings	202	.56	.12
Fine (2009a)	Unpublished	Publisher	Overt	Concurrent	Applicants	CWB	Self-reports	2,308	.73	.26
Fine (2009b)	Unpublished	Publisher	Overt	Concurrent	Incumbents	CWB	Self-reports			
Sample 1								298	.45	.06
Sample 2								257	.49	.21
Fine (2010)	Unpublished	Publisher	Overt	Concurrent	Applicants	CWB	Self-reports	1,005	.76	.26
Fine and Horowitz (2008)	Unpublished	Publishers	Overt	Concurrent	Both types	CWB	Self-reports	1,089	.51	.24
Fine et al. (2010)	Published	Publishers	Overt	Concurrent	Incumbents	CWB	Self-reports	429	.70	.35
Fortmann et al. (2002)	Published	Publishers and a non-publisher	Overt							

(*Appendix continues*)

Appendix (*continued*)

Study	Publication status	Author affiliation	Test type	Study design	Study sample	Criterion	Criterion source	*N*	r_{yy}	*r*
Sample 1				Predictive	Applicants	Job performance	Ratings	97	.56	.19
						CWB	Ratings	97	.56	.20
Sample 2				Both types	Both types	Job performance	Ratings	177	.56	.18
						CWB	Ratings	177	.56	.02
Sample 3				Both types	Both types	Job performance	Ratings	178	.56	.12
						CWB	Ratings	178	.56	.17
Sample 4				Both types	Both types	CWB	Self-reports	156	.42	.24
Frost and Orban (1990)	Published	Publishers	Overt	Concurrent	Students with work experience	CWB	Self-reports	155	.70	.62
Frost and Rafilson (1989)	Published	Publishers	Both types	Concurrent	Incumbents	CWB	Self-reports	105	.84	.33
Gardner (1998)	Unpublished	Non-publisher, did not develop test	Personality	Concurrent	Incumbents	Job performance	Ratings	76	.64	.15
Gonder and Gilmore (2004)	Published	Non-publishers	Personality	Predictive	Applicants	Turnover	Records	199	n/a	.03
Gough (1965)	Published	Non-publisher, developed test	Personality	Concurrent	Incumbents	Job performance	Ratings	79	.56	.38
Gough and Freddi (1967)	Published	Non-publisher, developed test	Personality	Predictive	Incumbents	Training performance	Ratings	84	.56	.27
Haaland et al. (1999)	Unpublished	Non-publishers	Personality	Predictive	Applicants	Training performance	Ratings	442	.56	.02
Hakstian et al. (2002)	Published	Non-publishers, developed test	Personality							
Sample 5 (males)				Concurrent	Incumbents	Job performance	Ratings	156	.56	.04
Sample 5 (females)				Concurrent	Incumbents	Job performance	Ratings	140	.56	.00
Sample 6				Predictive	Incumbents	Job performance	Ratings	79	.56	.23
Sample 7				Predictive	Applicants	CWB	Records	91	.83	.25
Hilliard (2000)	Unpublished	Non-publisher	Overt	Predictive	Applicants	Job performance	Ratings	12	.56	−.47
Hunt et al. (1997)	Unpublished	Publishers and a non-publisher	Personality	Predictive	Applicants	Job performance	Ratings	4,061	.56	.12
						CWB	Ratings	2,783	.56	.07
Inwald (1988)	Published	Publisher	Personality	Predictive	Incumbents	Turnover	Records	219	n/a	.24
Inwald and Brockwell (1991)	Published	Publishers	Personality	Predictive	Incumbents	Job performance	Ratings	307	.56	.15
Inwald and Patterson (1989)	Unpublished	Publishers	Personality	Predictive	Applicants	Turnover	Records	448	n/a	.05
Inwald and Shusman (1984)	Published	Publishers	Personality	Predictive	Applicants	Turnover	Records			
Sample 1 (males)								596	n/a	.16
Sample 2 (females)								143	n/a	.00
Jackson et al. (2000)	Published	Non-publishers, developed test	Personality	Concurrent	Students with work experience	CWB	Self-reports			
Study 1								84	.75	.48
Study 2								106	.76	.41
Jones (1979)	Unpublished	Publisher	Overt	Concurrent	Incumbents	CWB	Self-reports	36	.64	.57
Jones (1980a)	Published	Publisher	Overt	Concurrent	Incumbents	CWB	Self-reports	39	.65	.35
Jones (1980b)	Unpublished	Publisher	Overt	Concurrent	Incumbents	CWB	Self-reports	53	.68	.10
Jones (1981)	Published	Publisher	Overt	Concurrent	Incumbents	CWB	Self-reports	33	.57	.39
Jones (1982)	Unpublished	Publisher	Overt	Concurrent	Students with work experience	CWB	Self-reports	71	.80	.51

(*Appendix continues*)

Appendix (*continued*)

Study	Publication status	Author affiliation	Test type	Study design	Study sample	Criterion	Criterion source	*N*	r_{yy}	*r*
Jones et al. (2002)	Published	Publishers	Overt	Concurrent	Incumbents	Job performance	Ratings			
Study 4								94	.56	.35
Study 5								322	.56	.25
Jones et al. (1990)	Published	Publishers	Overt	Concurrent	Incumbents	Job performance	Ratings	49	.56	.22
Jones and Scruggs (1981)	Unpublished	Publishers	Overt	Concurrent	Students with work experience	CWB	Self-reports	45	.89	.33
Jones and Terris (1983)	Published	Publishers	Overt	Predictive	Incumbents	Job performance	Ratings	60	.56	.04
						CWB	Ratings	60	.56	.17
Kauder and Thomas (2003)	Published	Non-publishers, did not develop test	Personality	Predictive	Applicants	Job performance	Ratings	30	.56	−.03
LaFosse (1992)	Unpublished	Non-publisher	Personality	Concurrent	Incumbents	Job performance	Ratings	489	.56	.05
Leslie (2006)	Unpublished	Publisher	Overt	Concurrent	Applicants	CWB	Self-reports			
Sample 1								1,000	.39	.29
Sample 2								1,000	.40	.23
Loudermilk (1964)	Published	Non-publisher	Personality	Predictive	Applicants	Tenure	Records	348	n/a	.15
Luther (2000)	Published	Non-publisher	Personality	Concurrent	Incumbents	Job performance	Ratings	114	.56	.15
Malin et al. (1987)	Unpublished	Non-publishers, did not develop test	Personality	Predictive	Applicants	Turnover	Records	1,748	n/a	.07
Marcus (2006)	Published	Non-publisher, developed test	Both types	Concurrent	Incumbents	Job performance	Ratings	170	.56	.19
						Training performance	Grades	253	.81	.24
						CWB	Self-reports	266	.83	.54
Marcus et al. (2007)	Published	Non-publisher, developed test	Both types	Concurrent	Incumbents and students with work experience	CWB	Self-reports			
Sample 1 (Canadian)								169	.81	.49
Sample 2 (East German)								204	.72	.47
Sample 3 (West German)								294	.77	.60
Marcus et al. (2002)	Published	Non-publishers, developed test	Both types	Concurrent	Incumbents					
Sample 1						Job performance	Ratings	98	.56	.19
						CWB	Self-reports	98	.88	.40
Sample 2						Job performance	Ratings	76	.56	.15
						CWB	Self-reports	76	.88	.57
Mastrangelo and Jolton (2001)	Published	Non-publishers, did not develop test	Overt	Concurrent	Students with work experience	CWB	Self-reports	173	.82	.56
Moretti (1980)	Unpublished	Publisher	Overt	Concurrent	Incumbents	CWB	Self-reports			
Sample 1								17	.57	.40
Sample 2								14	.57	.31
Moretti (1986)	Published	Publishers and non-publishers	Overt	Concurrent	Incumbents	CWB	Self-reports	196	.86	.60
Nicol and Paunonen (2002)	Published	Non-publishers, developed test	Overt	Concurrent	Students with work experience	CWB	Self-reports	152	.67	.65
Paajanen (1988)	Unpublished	Publisher	Personality	Predictive	Applicants	Turnover	Records	17,995	n/a	.05

(*Appendix continues*)

Appendix (*continued*)

Study	Publication status	Author affiliation	Test type	Study design	Study sample	Criterion	Criterion source	*N*	r_{yy}	*r*
Palmatier (1996)	Unpublished	Non-publisher	Personality	Predictive	Incumbents	Job performance	Ratings	174	.56	.24
						Training performance	Grades	231	.81	.28
						Turnover	Records	301	n/a	.21
Powell et al. (2008)	Unpublished	Non-publishers, developed test	Personality	Concurrent	Students with work experience	CWB	Self-reports	186	.92	.23
Provines (2006)	Unpublished	Non-publisher	Personality	Predictive	Applicants	Job performance	Ratings	75	.56	−.15
						Training performance	Grades	86	.81	.15
						Turnover	Records	90	n/a	.41
Rafilson (1988)	Published	Publisher	Personality	Concurrent	Students with work experience	CWB	Self-reports	225	.92	.45
Schneider (2002)	Unpublished	Non-publisher	Personality	Predictive	Applicants	Job performance	Ratings	193	.56	.17
						Training performance	Grades	175	.81	.07
						Turnover	Records	232	n/a	.10
Science Research Associates (1987)	Unpublished	Publisher	Personality	Predictive	Incumbents	CWB	Records	270	.51	.39
Shusman and Inwald (1991)	Published	Publishers	Personality	Predictive	Applicants	CWB	Records	386	1.00	.08
Shusman et al. (1987)	Published	Publishers	Personality	Predictive	Applicants	Job performance	Ratings	421	.56	.00
						Training performance	Ratings	181	.56	.00
						CWB	Records	421	1.00	.20
Shusman et al. (1984)	Published	Publishers	Personality	Predictive	Applicants	CWB	Records	665	1.00	.06
						Turnover	Records	716	n/a	.11
Sigma Assessment Systems (2009)	Unpublished	Publisher	Personality	Concurrent	Incumbents	CWB	Self-reports			
Study 1								943	.88	.51
Study 2								217	.88	.58
Surette and Serafino (2003)	Published	Non-publishers, did not develop test	Personality	Predictive	Incumbents	Job performance	Ratings	30	.56	−.09
Terris and Jones (1980)	Published	Publishers	Overt	Concurrent	Incumbents	CWB	Self-reports			
Sample 1								27	.45	.66
Sample 2								15	.45	.81
Terris and Jones (1982)	Published	Publishers	Overt	Predictive	Applicants	CWB	Ratings	61	.51	.28
Tolbirt (1992)	Unpublished	Non-publisher	Overt	Predictive	Applicants	Job performance	Ratings	77	.56	−.12
Van Hein et al. (2007)	Published	Non-publishers	Overt	Predictive	Applicants	Job performance	Ratings	231	.56	−.05
						CWB	Records	249	1.00	.09

(*Appendix continues*)

Appendix (*continued*)

Study	Publication status	Author affiliation	Test type	Study design	Study sample	Criterion	Criterion source	*N*	r_{yy}	*r*
Van Iddekinge et al. (2005)	Published	Non-publishers, did not develop test	Overt	Concurrent	Incumbents	Job performance	Ratings	152	.56	−.10
Werner et al. (1989)	Published	Publishers	Overt	Concurrent	Applicants	CWB	Self-reports	338	.75	.36
Woolley and Hakstian (1993)	Published	Non-publishers	Overt and personality	Concurrent	Students with work experience	CWB	Self-reports			
Sample 1 (males)								131	.73	.42
Sample 2 (females)								158	.66	.32

Note. The values generally reflect the overall results from each study; values used in some of the secondary analyses (e.g., of more specific predictors and/or criteria) are not shown. r_{yy} = criterion reliability estimate; *r* = observed validity coefficient; CWB = counterproductive work behavior; n/a = we did not correct that type of criterion measure for unreliability.

Received July 31, 2008
Revision received June 23, 2010
Accepted July 13, 2010 ■